Duality-based Residual Estimation for Fully Offline Value-based Reinforcement Learning

Kohei Miyaguchi
LY Corporation Research, Tokyo, Japan

Abstract

Value-based reinforcement learning (RL) efficiently handles high-dimensional state spaces, but existing methods lack a principled method for hyperparameter tuning without online interaction, limiting use in safety-critical and data-scarce domains. We propose the **Duality-based Residual Estimator (DRE)**, a simple offline validation metric for value-based offline RL. DRE is compatible with standard value-based Off-Policy Evaluation (OPE) and enables automatic hyperparameter selection, which is formalized through an adaptive extension of the Probably Approximately Correct (PAC) guarantee for Q-function selection. Our results address a key theoretical bottleneck toward *fully offline* value-based RL, which enables deployment without extensive online tuning.

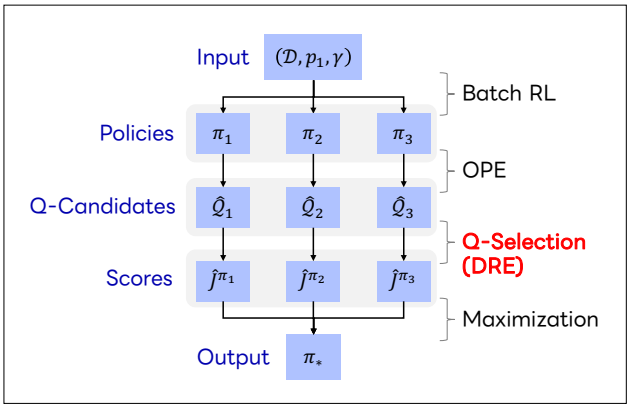

Figure 1: Pipeline of Fully Offline Value-based RL

1 INTRODUCTION

Reinforcement learning (RL) is a widely used machine learning approach for decision-making problems, applicable to robotics (Tang et al., 2025), finance (Pippas et al., 2025), recommender systems (Chen et al., 2023), healthcare (Yu et al., 2019) and energy control (Vázquez-Canteli and Nagy, 2019), among others. In particular, the value-based RL methods (Mnih et al., 2015; Schulman et al., 2017; Haarnoja et al., 2018) are widely used especially when state representations are rich and high-dimensional, since they can avoid tricky estimation of the environment dynamics.

Offline RL (or batch RL) is a variant of RL where online interaction with the environment is not allowed (Levine et al., 2020; Prudêncio et al., 2023), focusing more on safety during training and the maximal utilization of finite samples. The value-based approach has continued to be used extensively in this setting and has achieved the state-of-the-art performance (Kumar et al., 2020; Fujimoto et al., 2021; Kostrikov et al., 2022).

However, existing offline value-based RL methods are not *fully* offline because they require online access for hyperparameter tuning. The root cause of this issue is the lack of a practical method for eliminating hyperparameters from value-based off-policy evaluation (OPE) (Uehara et al., 2022), resulting in multiple scores for single policy contradicting one another. In particular, current value-based approaches for hyperparameter selection of OPE fail since they either introduce new hyperparameters or are incompatible with standard value-based OPE.

To address this challenge, we propose Duality-based Residual Estimator (DRE), an offline validation metric for Q-function estimators. Notably,

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

1. it is compatible with standard value-based OPE estimators that try to minimize L^2-Bellman residual $\|q - \mathcal{B}^\pi q\|_2$, and

2. its essential hyperparameters are automatically selected by an adaptive extension of Probably Approximately Correct (PAC) guarantee.

In particular, the L^2-Bellman residual is provably minimized with Fitted Q-Evaluation (FQE) (Ernst et al., 2005; Munos and Szepesvári, 2008; Le et al., 2019), Minimax Q-Learning (MQL) and its extensions (Uehara et al., 2021), given their hyperparameters are selected appropriately. In other words, DRE makes these methods consistent with less sensitivity to hyperparameters, resolving a key bottleneck to fully offline value-based RL (Fig. 1).

An important technical advancement behind DRE is the use of a clipped dual-norm representation of the Bellman residual (formally presented in Section A.2),

$$\|q - \mathcal{B}^\pi q\|_2 = \sup_{f:\mathcal{S}\times\mathcal{A}\to\mathbb{R}} \frac{\langle q - \mathcal{B}^\pi q, f\rangle}{\max\{1, \|f\|_2\}}. \qquad (1)$$

The clipping of the denominator stabilizes DRE, making it suitable as a criterion for Q-selection.

The rest of this paper is organized as follows. First, related work is reviewed in Section 2, and definitions and assumptions are introduced in Section 3. We then formalize the goals of this study based on a new PAC framework in Section 4. Next, we present the proposed method in Section 5 and show that it achieves the said goals in Section 6. Finally, we conclude this paper in Section 7.

2 PREVIOUS WORK

OPE has been studied in various settings such as contextual bandits, Markov decision processes (MDPs) and their extensions (Voloshin et al., 2021; Uehara et al., 2022). Furthermore, existing methods of OPE in the MDP settings are roughly categorized into model-free methods (including value-based methods) (Ernst et al., 2005; Munos and Szepesvári, 2008; Nachum et al., 2019; Uehara et al., 2020; Yang et al., 2020; Nachum and Dai, 2020) and model-based methods (Yin and Wang, 2020; Yu et al., 2020; Uehara and Sun, 2021; Wang et al., 2022).

The distinction between model-free and model-based OPE is made with their intermediate targets of estimation, which also affect how the hyperparameters of OPE are tuned offline. For instance, model-based methods estimate the transition dynamics, which can be validated with held-out datasets based on natural loss functions such as the likelihood (Uehara and Sun, 2021), total variation and Wasserstein distances (Yu et al., 2020).

The tuning of model-free (including value-based) OPE is more delicate, mainly because of the notorious double-sampling problem (Zhu and Ying, 2020; Duan et al., 2021). Residual Minimization (Baird et al., 1995) avoids double sampling by substituting the Bellman residual with an upper bound, which is biased. Modified Bellman Residual Minimization (Antos et al., 2008; Farahmand et al., 2016) de-biases the upper bound by solving an additional estimation problem, which however introduces new hyperparameters. Similarly, kernel-based loss functions (Liu et al., 2018; Feng et al., 2019; Uehara et al., 2020) can measure the goodness-of-fit of model-free estimates consistently, but the choice of kernel is open and significantly affects the loss values. Miyaguchi (2022) proposed Regret Minimization, a hyperparameter selection method for FQE. It is free from hyperparameters, but instead requires the true Bellman operator $\mathcal{B}^\pi$ to be realizable in a candidate set, which makes it effectively model-based from a theoretical perspective.

Zhang and Jiang (2021) proposed BVFT-PE, one of the closest related methods, which performs Q-selection while introducing only a single, heuristically tunable hyperparameter. However, it is incompatible with standard value-based OPE since the consistency guarantee of BVFT-PE depends on the upstream estimators minimizing the L^∞-error $\|q - q^\pi\|_\infty$, which is generally hard to achieve without strong assumptions. Liu et al. (2025) recently proposed LSTD-Tournament, a relative of BVFT-PE without additional hyperparameters, yet still incompatible with standard upstream estimators due to its dependence on the highly specialized form of consistency guarantee. The comparison between our method and these previous methods are summarized in Table 1.

Finally, there have been also estimator-agnostic methods for hyperparameter selection. However, majority of them have been developed for bandits (Su et al., 2020; Udagawa et al., 2023; Cief et al., 2025) or rely on nontrivial assumptions such as the consistency of bootstrap estimation Nie et al. (2024).

3 PRELIMINARIES

We introduce the necessary definitions (Section 3.1) and the basic assumptions we make throughout this paper (Section 3.2).

Table 1: Comparison with Previous Results on Q-Selection

Method	Compatible Error	Hyperparameter	Self Tuning
BVFT-PE (Zhang and Jiang, 2021)	$\|q - q^\pi\|_\infty$	Yes	Heuristic
LSTD-Tournament (Liu et al., 2025)	—	No	—
DRE (ours)	$\|q - \mathcal{B}^\pi q\|_2$	Yes	Adaptively PAC

3.1 Definitions

Basic Notation. $M(\mathcal{X})$ denotes the set of finite signed measures on $\mathcal{X}$. $\Delta(\mathcal{X}) \subset M(\mathcal{X})$ denotes the set of probability measures on $\mathcal{X}$. $\|f\|_\infty$ denotes the uniform (supremum) norm of functions f.

Markov Decision Process (MDP). In RL, we model the environment of interest with a time-invariant Markov decision process $\mathcal{M} := (\mathcal{S}, \mathcal{A}, p_1, P_+, P_r, \gamma)$, where $p_1 \in \Delta(\mathcal{S})$ is the initial state distribution, $P_+ : \mathcal{S}\times\mathcal{A} \to \Delta(\mathcal{S})$ is the transition kernel, $P_r : \mathcal{S} \times \mathcal{A} \to \Delta([-1, 1])$ is the reward kernel, and $\gamma \in [0, 1)$ is the discount factor. The combination of MDP $\mathcal{M}$ and policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ determines a stochastic process (trajectory) $\xi = \{(s_t, a_t, r_t)\}_{t\geq 1} \sim p_\xi^\pi$ such that $s_1 \sim p_1$, $a_t \sim \pi(s_t)$, $r_t \sim P_r(s_t, a_t)$ and $s_{t+1} \sim P_+(s_t, a_t)$ for $t \geq 1$. The value of policy π in $\mathcal{M}$ is given by the expected cumulative discounted reward,

$$J^\pi := \mathbb{E}_{\xi\sim p_\xi^\pi}\left[\sum_{t=1}^\infty \gamma^{t-1} r_t\right]. \tag{2}$$

The policy Q-function $q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is defined as the policy value conditioned on the initial state and action,

$$q^\pi(s, a) := \mathbb{E}_{\xi\sim p_\xi^\pi}\left[\sum_{t=1}^\infty \gamma^{t-1} r_t \middle| s_1 = s, a_1 = a\right]. \tag{3}$$

The marginal occupancy distribution $\mu^\pi \in M(\mathcal{S} \times \mathcal{A})$ is defined as the discounted sum of visitation probabilities

$$\mu^\pi(E) := \mathbb{E}_{\xi\sim p_\xi^\pi}\left[\sum_{t=1}^\infty \gamma^{t-1} \mathbf{1}\{(s_t, a_t) \in E\}\right]. \tag{4}$$

Offline Dataset. In offline RL, it is not allowed to access the transition and reward dynamics P_+, P_r directly. Instead, we are given an offline dataset $\mathcal{D} = \{(s_i, a_i, r_i, s_i')\}_{i=1}^n \sim p_\mathrm{b}^n$ be an i.i.d. sample of size n drawn from $p_\mathrm{b}(s, a, r, s') = \mu_\mathrm{b}(s, a)P_r(r|s, a)P_+(s'|s, a)$, where $\mu_\mathrm{b} \in \Delta(\mathcal{S} \times \mathcal{A})$ is fixed, unknown and referred to as the behavior occupancy distribution. The L^p-norm with respect to μ_b may be written as

$$\|f\|_p := \{\mathbb{E}_{(s,a)\sim\mu_\mathrm{b}}[f(s, a)^p]\}^{1/p}$$

for $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and $1 \leq p < \infty$. Moreover, the inner product of $f, g \in L^2(\mu_\mathrm{b})$ may be written as

$$\langle f, g\rangle := \mathbb{E}_{(s,a)\sim\mu_\mathrm{b}}[f(s, a)\, g(s, a)].$$

The marginal importance weight $w^\pi \in L^1(\mu_\mathrm{b})$ is given by the density ratio

$$w^\pi(s, a) := \frac{\mathrm{d}\mu^\pi}{\mathrm{d}\mu_\mathrm{b}}(s, a).$$

Off-Policy Evaluation (OPE). OPE is the problem of estimating J^π for given π based on $\mathcal{D}$, a foundation for tuning hyperparameters of policies offline. A typical signature of OPE estimators is given by

$$(\mathcal{D}, \pi, p_1, \gamma, \lambda) \mapsto \widehat{J}_\lambda,$$

where λ denotes hyperparameters and $\widehat{J}_\lambda$ denotes the corresponding estimates of J^π. In particular, value-based OPE estimators are given by the composition of Q-function estimators

$$(\mathcal{D}, \pi, \gamma, \lambda) \mapsto \widehat{q}_\lambda$$

with an expectation functional

$$\mathscr{J}_Q(q) := \mathbb{E}_{s\sim p_1, a\sim\pi(s)}[q(s, a)],$$

where $\widehat{q}_\lambda : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ is the estimate of q^π corresponding to λ.

Bellman Residual. For fixed policies π, the (L^2-) Bellman residual of Q-function estimate q is defined by

$$R^\pi(q) := \|q - \mathcal{B}^\pi q\|_2, \tag{5}$$

where $\mathcal{B}^\pi$ is the Bellman backup operator satisfying $(\mathcal{B}^\pi q)(s, a) = \mathbb{E}_{r\sim P_r(s,a), s'\sim P_+(s,a), a'\sim\pi(s')}[r + \gamma q(s', a')]$. It is known that $R^\pi(q) = 0 \Leftrightarrow q = q^\pi$ and hence $R^\pi(q)$ is commonly used as a surrogate objective for learning q^π. As shown below, the efficiency of OPE via the Bellman residual minimization is characterized by the L^2-norm of w^π (see Section C.1 for the proof).

Proposition 3.1. *For all $u \geq 0$, we have*

$$\sup_{\substack{q:\mathcal{S}\times\mathcal{A}\to\mathbb{R} \\ R^\pi(q)\leq u}} |\mathscr{J}_Q(q) - J^\pi| = \begin{cases} u\|w^\pi\|_2 & (\mu^\pi \ll \mu_\mathrm{b}) \\ \infty & (\textit{otherwise}) \end{cases}.$$

3.2 Assumptions

According to Proposition 3.1, we introduce the minimal assumptions under which standard value-based OPE provably works.

Assumption 3.1. *$\mu^\pi \ll \mu_\mathrm{b}$ and $\|w^\pi\|_2 < \infty$.*

Note that this is a weak variant of concentrability condition (Munos and Szepesvári, 2008), quantifying the discrepancy between μ_b and μ^π.

We also assume the boundedness of Q-function estimates, reflecting the boundedness of the true Q-function $\|q^\pi\|_\infty \leq 1/(1-\gamma)$.[1]

Assumption 3.2. *$\sup_\lambda \|\widehat{q}_\lambda\|_\infty \leq C_Q$, where the supremum is taken over all possible λ.*

4 PROBLEM FORMULATION

Most value-based OPE methods require careful hyperparameter tuning to estimate J^π accurately, which seems to undermine their original purpose as validation metrics. To address this challenge, we introduce the problem of *Q-selection*,

$$\mathscr{S}_Q : (\widehat{\mathcal{Q}}, \mathcal{D}', \pi, \gamma) \mapsto \widehat{q}_* \in \widehat{\mathcal{Q}}, \tag{6}$$

where $\widehat{\mathcal{Q}} := \{\widehat{q}_\lambda : \lambda \in \Lambda\}$ is the set of Q-function estimates indexed with hyperparameter λ and $\mathcal{D}' \sim p_\mathrm{b}^m$ is a held-out dataset of size m. The selected Q-function is then used to estimate J^π,

$$\widehat{J}_* = \mathscr{J}_Q(\widehat{q}_*),$$

filling the gap towards *fully* offline value-based RL (Fig. 1). Below, we formally discuss two requirements for Q-selection, namely, *upstream compatibility* and *self-hyperparameter selection*.

Upstream Compatibility. Since the Bellman residual $R^\pi(q)$ is the standard objective of Q-function estimation, at least one $q_* \in \widehat{\mathcal{Q}}$ is expected to attain $R^\pi(q_*) \approx 0$. Thus, the first requirement for $\mathscr{S}_Q$ is compatibility with such upstream estimates, i.e., to select q_* and enjoy the error guarantee

$$|\widehat{J}_* - J^\pi| \leq \|w^\pi\|_2 \min_{q\in\widehat{\mathcal{Q}}} R^\pi(q), \tag{7}$$

according to Proposition 3.1. Formalizing this idea, we introduce a new variant of the PAC-learning framework (Shalev-Shwartz and Ben-David, 2014) specialized for Q-selection.

Definition 4.1 (PAC Q-selection)**.** *Q-selectors $\mathscr{S}_Q$ are said to be $(\kappa, \epsilon, \delta)$-probably approximately correct (PAC) if*

$$\mathbb{P}\left\{|\widehat{J}_* - J^\pi| \leq \|w^\pi\|_2 \left(\kappa \min_{\lambda\in\Lambda} R^\pi(\widehat{q}_\lambda) + \epsilon\right)\right\} \geq 1-\delta, \tag{8}$$

for $\kappa \geq 1$, $\epsilon \geq 0$ and $\delta \in (0,1)$.

A unique feature of PAC Q-selection is that the threshold for being "correct" is relative to Eq. (7). Consequently, the error $|\widehat{J}_* - J^\pi|$ is bounded by the coefficient $\|w^\pi\|_2$ and the minimum residual $\min_{q\in\widehat{\mathcal{Q}}} R^\pi(q)$ as well as the PAC constants $(\kappa, \epsilon, \delta)$. Specifically, κ represents the suboptimality in the error conversion rate from residual minimization to OPE, while ϵ represents the error that persists even if the true Q-function q^π is contained in the input $\widehat{\mathcal{Q}}$.

A particularly interesting quantity in PAC Q-selection is the minimum achievable ϵ conditioned on κ and δ, denoted by

$$\epsilon^*_{\mathcal{I},m,\kappa,\delta}(\mathscr{S}_Q) := \inf\{\epsilon \geq 0 : \mathscr{S}_Q \text{ is } (\kappa,\epsilon,\delta)\text{-PAC}\}.$$

Here, $\mathcal{I} := (\mathcal{M}, \mu_\mathrm{b}, \pi, \widehat{\mathcal{Q}})$ and m in the subscript signify the dependence on the problem instance. Our first research question is thus given as follows.

Question 4.1. *Under what conditions (in addition to Assumptions 3.1 and 3.2) can we guarantee that $\epsilon^*_{\mathcal{I},m,\kappa,\delta}(\mathscr{S}_Q)$ is small?*

Self-Hyperparameter Selection. One can expect meaningful PAC error guarantees only under sufficiently strong assumptions. Formally, one may obtain a PAC error bound of the form

$$\epsilon^*_{\mathcal{I},m,\kappa,\delta}(\mathscr{S}_Q^\mathcal{F}) \leq \bar{\epsilon}_{\mathcal{I},m,\kappa,\delta}(\mathcal{F}) \tag{9}$$

where $\mathcal{F}$ is the model corresponding to the assumptions and $\mathscr{S}_Q^\mathcal{F}$ is the associated selector. The error bound $\bar{\epsilon}_{\mathcal{I},m,\kappa,\delta}(\mathcal{F})$ may decay rapidly (e.g., in $O(m^{-1/2})$) when $\mathcal{F}$ is sufficiently specific and $\mathcal{I}$ is accurately realizable with $\mathcal{F}$. However, due to this specificity, fast error bounds would be inherently sensitive to the choice of $\mathcal{F}$.

To mitigate this hyperparameter sensitivity, one may further try to obtain an adaptive PAC error bound

$$\epsilon^*_{\mathcal{I},m,\kappa,\delta}(\mathscr{S}_Q^\Xi) \leq \min_{\xi\in\Xi}\left\{\bar{\epsilon}_{\mathcal{I},m,\kappa,\delta}(\mathcal{F}_\xi) + c_A(\mathcal{F}_\xi)\sqrt{\frac{\ln|\Xi|}{m}}\right\}, \tag{10}$$

[1] Otherwise, we can force Assumption 3.2 by clipping with a sufficiently large constant.

where $\xi \in \Xi$ denotes the indices of candidate models $\mathcal{F}_\xi$ and $c_A(\mathcal{F}_\xi)$ is a model-dependent coefficient. Note that Eq. (10) mirrors typical guarantees obtained by standard hyperparameter selection procedures in supervised learning, e.g., minimizing validation losses.

We regard $\mathscr{S}_Q^\Xi$ as capable of self-hyperparameter selection if Eq. (10) holds without additional assumptions. However, unlike supervised learning, it has been unknown whether this type of error bound is feasible with Q-selection. Thus, our second research question is given as follows.

Question 4.2. *Can we construct an adaptive selector $\mathscr{S}_Q^\Xi$ satisfying Eq. (10)?*

Note that we do not attempt to eliminate all the hyperparameters here.[2] Instead, we aim to auto-select those that require additional assumptions to achieve a fast rate, while ignoring the others.

5 METHOD

We propose a simple Q-selector minimizing an empirical proxy of the Bellman residual $R^\pi(q)$, denoted by $\widehat{R}^\pi(q)$,

$$\widehat{q}_* \in \operatorname*{arg\,min}_{q \in \widehat{\mathcal{Q}}} \widehat{R}^\pi(q). \tag{11}$$

The design of the proxy is inspired by the clipped dual-norm representation (CDR, Eq. (1)). According to CDR, $R^\pi(q)$ is evaluated by maximizing its objective function,

$$R^{\pi,q}(f) := \frac{\langle q - \mathcal{B}^\pi q, f\rangle}{\max\{1, \|f\|_2\}}. \tag{12}$$

Since $R^{\pi,q}(f)$ is easily estimated with offline datasets, we can employ supervised learning algorithms to approximately maximize it, resulting in proxies of $R^\pi(q)$. However, supervised learning algorithms often come with hyperparameters, including function approximators, regularizers, and optimizers.

To avoid tuning these hyperparameters online, we employ a two-step algorithm with even data splitting $\mathcal{D}' = \mathcal{D}'_1 \cup \mathcal{D}'_2$ of size $m_j = |\mathcal{D}'_j| = m/2$, $j \in \{1,2\}$.[3] First, based on $\mathcal{D}'_1$, we compute multiple estimators $\{\widehat{f}_\xi : \xi \in \Xi\}$ maximizing $R^{\pi,q}(f)$, where $\xi \in \Xi$ denotes *any* hyperparameter configuration involved in the estimation procedure. Second, based on $\mathcal{D}'_2$, we aggregate these estimators and compute the proxy

$$\widehat{R}^\pi(q) := \max_{\xi \in \Xi} \widehat{R}^{\pi,q}_{\mathrm{LCB}}(\widehat{f}_\xi; \delta'), \tag{13}$$

[2]This would be impossible: in principle, even the smallest algorithmic design choices could be regarded as hyperparameters.

[3]For simplicity, we assume m is even.

Algorithm 1 Duality-based Residual Estimator

```
procedure DRE(q, π, γ, D'; Ξ, δ')
    (D'_1, D'_2) ← even_split(D')
    for ξ in Ξ do
        f̂_ξ ← maximize(R^{π,q}; D'_1, ξ)
        R̂_ξ ← LCB(R^{π,q}(f̂_ξ); D'_2, δ')
    return max_{ξ∈Ξ} R̂_ξ
```

where $\widehat{R}^{\pi,q}_{\mathrm{LCB}}(\cdot;\delta')$ is a (pointwise) lower confidence bound (LCB) on the objective function $R^{\pi,q}(\cdot)$ with a confidence level $\delta' \in (0,1)$.

We refer to this procedure of estimating $R^\pi(q)$ as Duality-based Residual Estimator (DRE). The summary of DRE is given in Algorithm 1. There are two explicit hyperparameters, Ξ and δ', in DRE. However, as we see in Section 6, DRE is adaptive to Ξ in the sense of Eq. (10) while δ' requires no tuning to achieve fast rates.

Below, we present exemplary implementations of the first and second steps in Section 5.1 and Section 5.2, respectively.

5.1 Implementation of First Step

We start with the simplest implementation of $\widehat{f}_\xi$ in Section 5.1.1, and then relax it in Section 5.1.2 for computational efficiency.

5.1.1 Exact ERM Estimators

A natural way to maximize $R^{\pi,q}(f)$ is by applying empirical risk minimization (ERM) with a negative sign. Let $\mathcal{F}_\xi$ denote function approximators indexed with ξ. Exact ERM estimators associated with $\mathcal{F}_\xi$ are given by

$$\widehat{f}^{(\mathrm{exact})}_\xi := \operatorname*{arg\,max}_{f \in \mathcal{F}_\xi} \widehat{R}^{\pi,q}_1(f) \tag{14}$$

with a naive empirical objective function

$$\widehat{R}^{\pi,q}_j(f) := \frac{\widehat{\mathbb{E}}_{\mathcal{D}'_j}\left[e^\pi_q(z) f(s,a)\right]}{\sqrt{\max\{1, \widehat{\mathbb{E}}_{\mathcal{D}'_j}[f^2(s,a)]\}}}, \quad j \in \{1,2\}, \tag{15}$$

where $\widehat{\mathbb{E}}_{\mathcal{D}'_j}[\cdot]$ denotes the empirical expectation with respect to $\mathcal{D}'_j$, and $e^\pi_q(z) := q(s,a) - r - \gamma \mathbb{E}_{a' \sim \pi(s')}[q(s',a')]$ denotes the temporal-difference error of $z = (s,a,r,s')$ with respect to q and π.

5.1.2 Relaxed ERM Estimators

A major drawback of the exact ERM estimators is that it can be computationally expensive due to the hard-

clipped empirical expectation in the denominator. To address this issue, we propose relaxed ERM estimators

$$\widehat{f}_\xi^\psi := \arg\max_{f \in \mathcal{F}_\xi} \widehat{R}_1^{\pi,q,\psi}(f) \tag{16}$$

with a surrogate objective

$$\widehat{R}_j^{\pi,q,\psi}(f) := \frac{\widehat{\mathbb{E}}_{\mathcal{D}'_j}[e_q^\pi(z)\, f(s,a)]}{\psi\left(\left\{\widehat{\mathbb{E}}_{\mathcal{D}'_j}[f^2(s,a)]\right\}^{1/2}\right)}, \quad j \in \{1,2\}, \tag{17}$$

where $\psi : [0,\infty) \to [1,\infty)$ is a relaxation of the hard clipping function $\psi_\infty(\alpha) := \max\{1,\alpha\}$ such that i) $\psi \geq \psi_\infty$, ii) ψ is 1-Lipschitz, and iii) $\kappa_\alpha(\psi) := \psi(\alpha)/\alpha$ is non-increasing.

In particular, taking ψ as $\psi_2(u) := \sqrt{1+u^2}$, the relaxed estimators admit analytic solutions (shown below) and stochastic gradient-based algorithms (see Section A.3).

Analytic Solutions for Linear Models. Let $\mathcal{F}_\phi$ denote the linear models given by

$$\mathcal{F}_\phi := \{(s,a) \mapsto \langle \theta, \phi(s,a)\rangle_{\mathcal{H}} : \|\theta\|_{\mathcal{H}} \leq 1\}, \tag{18}$$

where $\phi : \mathcal{S} \times \mathcal{A} \to \mathcal{H}$ maps state-action pairs into a Hilbert feature space $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}}, \|\cdot\|_{\mathcal{H}})$.[4] The following proposition gives analytic solutions for the corresponding relaxed ERM estimators (see Section D.1 for the proof).

Proposition 5.1. *Let $\mathcal{F}_\xi = \mathcal{F}_\phi$. Then, we have*

$$\widehat{f}_\xi^{\psi_2} = \left\langle \frac{(I+\Sigma)^{-1} y}{\|(I+\Sigma)^{-1} y\|_{\mathcal{H}}}, \phi \right\rangle_{\mathcal{H}}, \tag{19}$$

where $y := \widehat{\mathbb{E}}_{\mathcal{D}'_1}\left[e_q^\pi(z)\,\phi(s,a)\right]$ and $\Sigma := \widehat{\mathbb{E}}_{\mathcal{D}'_1}[\phi(s,a) \otimes \phi(s,a)]$.

5.2 Implementation of Second Step

Similarly to the first step, we begin with a simplest implementation of $\widehat{R}^{\pi,q}_{\mathrm{LCB}}(\cdot;\delta')$ in Section 5.2.1, and then extend it to a general form in Section 5.2.2 for practical applications.

5.2.1 Hoeffding-type LCB

An instance of LCB is given by simple correction of the naive empirical objective (Eq. (15))

$$\widehat{R}^{\pi,q}_{\text{LCB-H}}(f;\delta') := \widehat{R}_2^{\pi,q}(f) - \sigma_{\mathrm{H}}(f;\delta'), \tag{20}$$

where $\sigma_{\mathrm{H}}(f;\delta')$ is a Hoeffding-type uncertainty quantifier,

$$\sigma_{\mathrm{H}}(f;\delta') := \widetilde{C}_Q \|f\|_\infty \sqrt{\frac{\ln(4/\delta')}{m_2}}, \tag{21}$$

with $\widetilde{C}_Q := 10(1+2C_Q)$. The following proposition guarantees that it does bound $R^{\pi,q}(f)$ from below with high probability (see Lemma E.6 for a generalized result with proof).

Proposition 5.2. *For all $q \in [-C_Q, C_Q]^{\mathcal{S}\times\mathcal{A}}$ and $f \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, we have*

$$\mathbb{P}\left\{R^{\pi,q}(f) \geq \widehat{R}^{\pi,q}_{\text{LCB-H}}(f;\delta')\right\} \geq 1-\delta'. \tag{22}$$

5.2.2 Generalized LCB

In general, one may employ the following best-of-both-world bound,

$$\widehat{R}^{\pi,q}_{\mathrm{LCB}}(f;\delta') = \max\left\{\widehat{R}^{\pi,q}_{\text{LCB-H}}(f;\tfrac{\delta'}{2}), \widehat{R}^{\pi,q}_{\text{LCB-X}}(f;\tfrac{\delta'}{2})\right\}, \tag{23}$$

where $\widehat{R}^{\pi,q}_{\text{LCB-X}}(f)$ is any LCB satisfying

$$\mathbb{P}\left\{[R^{\pi,q}(f)]_+ \geq \widetilde{R}^{\pi,q}_{\text{LCB-X}}(f;\delta')\right\} \geq 1-\delta'. \tag{24}$$

Such an alternative LCB may be obtained by applying the empirical Bernstein inequality (Maurer and Pontil, 2009) on both the numerator and the denominator of the naive empirical objective,

$$\widehat{R}^{\pi,q}_{\text{LCB-EB}}(f) := \frac{\{\widehat{\mathbb{E}}_{\mathcal{D}'_2}\left[e_q^\pi(z) f(s,a)\right] - \widehat{\sigma}_{\mathrm{EB}}(e_q^\pi f;\delta')\}_+}{\sqrt{\max\{1, \widehat{\mathbb{E}}_{\mathcal{D}'_2}[f^2(s,a)] + \widehat{\sigma}_{\mathrm{EB}}(f^2;\delta')\}}}, \tag{25}$$

where $\widehat{\sigma}_{\mathrm{EB}}(g)$ is an empirical Bernstein-type uncertainty quantifier

$$\widehat{\sigma}_{\mathrm{EB}}(g;\delta') := \sqrt{\frac{2\widehat{\mathbb{V}}_{\mathcal{D}'_2}[g(z)]\ln(4/\delta')}{m_2-1}} + \frac{14\|g\|_\infty \ln(4/\delta')}{3(m_2-1)},$$

and $\widehat{\mathbb{V}}_{\mathcal{D}'_2}[g(z)] := \widehat{\mathbb{E}}_{\mathcal{D}'_2}[g(z) - \widehat{\mathbb{E}}_{\mathcal{D}'_2}[g(z)]]^2$ denotes the biased empirical variance of $g(z)$.

Notably, $\widehat{R}^{\pi,q}_{\text{LCB-EB}}(f;\delta')$ can be significantly tighter than $\widehat{R}^{\pi,q}_{\text{LCB-H}}(f)$ since it is independent of the potentially large constant $\widetilde{C}_Q$ and exploits variance information through sample-dependent correction terms $\widehat{\sigma}_{\mathrm{EB}}(\cdot;\delta')$.

6 RESULTS

In this section, we address Questions 4.1 and 4.2 by analyzing the DRE-based Q-selectors.

[4] For example, let $(\mathcal{H}, k)$ be a reproducing kernel Hilbert space and its kernel on $\mathcal{S}\times\mathcal{A}$. Then, $\mathcal{F}_\phi$ is the centered ball of $\mathcal{H}$ with radius ρ if $\phi(s,a) = \rho k(\cdot,[s,a])$.

Let $\mathscr{S}_Q^{\psi,\Xi,\delta'}$ denote the DRE-based selectors, where the first step is given by the relaxed ERM estimators in Eq. (16) with general ψ and the second step is given by the generalized LCBs in Eq. (23). Since its behavior is largely affected by the underlying models $\mathcal{F} \in \{\mathcal{F}_\xi : \xi \in \Xi\}$, we introduce a couple of key characteristics to describe their influence on $\mathscr{S}_Q^{\psi,\Xi,\delta'}$.

Definition 6.1 (approximation error)**.** *Let $\alpha > 0$. An α-approximation error of $\mathcal{F}$ is defined by*

$$\mathscr{E}_\alpha(\mathcal{F}) := \min\left\{\mathscr{E}_\alpha^{\mathrm{R}}(\mathcal{F}), \mathscr{E}_\alpha^{\mathrm{C}}(\mathcal{F})\right\}, \tag{26}$$

with

$$\mathscr{E}_\alpha^{\mathrm{R}}(\mathcal{F}) := \inf_{f\in\mathrm{star}(\mathcal{F})} \|f - \alpha\bar{w}^\pi\|_2, \tag{27}$$

$$\mathscr{E}_\alpha^{\mathrm{C}}(\mathcal{F}) := \sup_{q\in\widehat{\mathcal{Q}}} \inf_{f\in\mathrm{star}(\mathcal{F})} \rho^\pi(q)\,\|f - \alpha\bar{\Delta}^\pi q\|_2, \tag{28}$$

where $\bar{w}^\pi := w^\pi/\|w^\pi\|_2$, $\bar{\Delta}^\pi q := (q-\mathcal{B}^\pi q)/\|q-\mathcal{B}^\pi q\|_2$, $\rho^\pi(q) := R^\pi(q)/(1+2C_Q)$ is the standardized bellman residual, and $\mathrm{star}(\mathcal{F}) := \{cf : f \in \mathcal{F},\, c \in [0,1]\}$ is the star hull of $\mathcal{F}$.

Remark 6.1. In general, the approximation error measures the expressiveness of $\mathcal{F}$, while α scales the difficulty of approximation problem (larger is harder). Inside the minimum, the first component $\mathscr{E}_\alpha^{\mathrm{R}}(\mathcal{F})$ measures the $\alpha\bar{w}^\pi$-realizability of $\mathcal{F}$, while the second one $\mathscr{E}_\alpha^{\mathrm{C}}(\mathcal{F})$ measures the completeness of $\mathcal{F}$ with respect to $\alpha\bar{\Delta}^\pi\widehat{\mathcal{Q}}$. As a result, $\mathscr{E}_\alpha(\mathcal{F})$ becomes small if at least one of these two is small.

Definition 6.2 (metric entropy)**.** *An m-th metric entropy of $\mathcal{F}$ is defined by*

$$\mathscr{H}_m(\mathcal{F}) := \ln \mathcal{N}(m^{-1/2}\|\mathcal{F}\|_\infty; \mathcal{F}, \|\cdot\|_\infty) \tag{29}$$

where $\|\mathcal{F}\|_\infty := \sup_{f\in\mathcal{F}}\|f\|_\infty$ and $\mathcal{N}(\varepsilon; \mathcal{F}, \|\cdot\|_\infty)$ is the ε-covering number of $\mathcal{F}$ with respect to $\|\cdot\|_\infty$.

Remark 6.2. Special values of the metric entropy are known for a number of models. For instance, we have $\mathscr{H}_m(\mathcal{F}) \le \ln|\mathcal{F}|$ for finite sets, $\mathscr{H}_m(\mathcal{F}) = O(\frac{p}{2}\ln m)$ for typical bounded p-dimensional sets, and $\mathcal{H}_m(\mathcal{F}) = O(d\ln\ln m)$ for the Gaussian reproducing kernel Hilbert spaces (RKHSs) on bounded d-dimensional Euclidean subsets (Steinwart and Fischer, 2021).

Now, we are ready to present our main results. A proof sketch is deferred to Section 6.2.

Theorem 6.1. *$\mathscr{S}_Q^{\psi,\Xi,\delta'}$ is (κ,ϵ,δ)-PAC with*

$$\kappa = \kappa_\alpha(\psi), \tag{30}$$

$$\epsilon = C_0 C_Q \kappa_\alpha(\psi) \min_{\xi\in\Xi} \widetilde{\epsilon}_{\alpha,m,\delta'}(\xi), \tag{31}$$

$$\delta = |\Lambda|(1+|\Xi|)\delta', \tag{32}$$

for $\alpha > 0$ and $\delta' \in (0,1)$, where $C_0 < \infty$ is a universal constant and

$$\widetilde{\epsilon}_{\alpha,m,\delta'}(\xi) = \mathscr{E}_\alpha(\mathcal{F}_\xi) + \|\mathcal{F}_\xi\|_\infty \sqrt{\frac{\mathscr{H}_m(\mathcal{F}_\xi) + \ln\frac{8}{\delta'}}{m}}.$$

Below, we provide corollaries of Theorem 6.1 to answer both research questions individually.

6.1 Answers to Research Questions

Question 4.1 (repeated). *Under what conditions (in addition to Assumptions 3.1 and 3.2) can we guarantee that $\epsilon^*_{\mathcal{I},m,\kappa,\delta}(\mathscr{S}_Q)$ is small?*

This is addressed by Corollary 6.1, a simplification of Theorem 6.1 for a single-model input $\Xi = \{\xi\}$.

Corollary 6.1. *Let $\Xi = \{\xi\}$ and fix any $\epsilon > 0$. Then, $\mathscr{S}_Q^{\psi,\{\xi\},\delta'}$ is $(\kappa_1(\psi),\epsilon,2|\Lambda|\delta')$-PAC if*

$$m \ge \left(\frac{C_0 C_Q \kappa_1(\psi)\|\mathcal{F}_\xi\|_\infty}{\epsilon}\right)^2 \left\{\mathscr{H}_m(\mathcal{F}_\xi) + \ln\frac{8}{\delta'}\right\} \tag{33}$$

and either of the following holds:

1. *Realizability)* $\bar{w}^\pi \in \mathrm{star}(\mathcal{F}_\xi)$, (34a)
2. *Completeness)* $\bar{\Delta}^\pi\widehat{\mathcal{Q}} \subset \mathrm{star}(\mathcal{F}_\xi)$. (34b)

Proof. It immediately follows from Theorem 6.1 with $\alpha = 1$ since Eq. (34) implies $\mathscr{E}_1(\mathcal{F}_\xi) = 0$. □

In words, the additive errors of typical DRE-based selectors are guaranteed to be less than $\epsilon > 0$ if i) the validation sample is as large as $m = \Omega(\epsilon^{-2}\|\mathcal{F}_\xi\|_\infty^2 \mathscr{H}_m(\mathcal{F}_\xi))$ and ii) the underlying model $\mathcal{F}_\xi$ is rich enough to satisfy a doubly robust (DR) condition (Eq. (34)). While the first condition on m is typical one in the PAC-learning literature, the second condition on $\mathcal{F}_\xi$ is interesting since it combines two typical conditions for OPE, realizability and completeness (Chen and Jiang, 2019), in a robust way. In particular, it benefits from the complementary nature of these conditions: the target of approximation in Eq. (34a) is singular and input-independent but potentially unbounded, whereas the one in Eq. (34b) is plural and input-dependent but bounded.

Question 4.2 (repeated). *Can we construct an adaptive selector $\mathscr{S}_Q^\Xi$ satisfying Eq. (10)?*

The answer is yes with $\mathscr{S}_Q^\Xi \leftarrow \mathscr{S}_Q^{\psi,\Xi,\delta'}$, as directly implied by Theorem 6.1. More specifically, it is satisfied with $\delta' = \frac{\delta}{2|\Lambda|}$ and taking ψ and $\alpha > 0$ such that $\kappa_\alpha(\psi) \le \kappa$, resulting in $\bar{\epsilon}_{\mathcal{I},m,\kappa,\delta}(\mathcal{F}_\xi) = C_0 C_Q \kappa \widetilde{\epsilon}_{\alpha,m,\frac{\delta}{2|\Lambda|}}(\xi)$ and $c_A(\mathcal{F}_\xi) = C_0 C_Q \kappa \|\mathcal{F}_\xi\|_\infty$.

In addition, we present a slightly simplified bound that showcases the reduced hyperparameter dependence of $\mathscr{S}_Q^{\psi,\Xi,\delta'}$ achieving fast rates.

Corollary 6.2. *Take ψ and $\alpha > 0$ satisfying $\kappa_\alpha(\psi) \leq \kappa$ and let $\delta' = \frac{\delta}{|\Lambda|(1+|\Xi|)}$. Then, we have*

$$\epsilon^*_{\mathcal{I},m,\kappa,\delta}(\mathscr{S}_Q^{\psi,\Xi,\delta'}) \leq C_0 C_Q \kappa \sqrt{\frac{\mathscr{C}^*_{\alpha,m,\delta}(\Xi)}{m}}, \qquad (35)$$

where

$$\mathscr{C}^*_{\alpha,m,\delta}(\Xi) := \min_{\substack{\xi\in\Xi \\ \mathscr{E}_\alpha(\mathcal{F}_\xi)=0}} \|\mathcal{F}_\xi\|_\infty^2 \left\{ \mathscr{H}_m(\mathcal{F}_\xi) + \ln \tfrac{8|\Lambda|(1+|\Xi|)}{\delta} \right\}$$

with the convention $\min \emptyset = \infty$.

Intuitively, $\mathscr{C}^*_{\alpha,m,\delta'}(\Xi)$ measures the complexity of the simplest correct model within Ξ, and its boundedness and slow growth are sufficient conditions for $\mathscr{S}_Q^{\psi,\Xi,\delta'}$ to achieve a fast rate.

Additional Observation. We remark that the choice of the generalized clipping function ψ does not affect the decay rate of ϵ, only changing its multiplicative constant (both in Theorem 6.1 and Corollary 6.2). Thus, the tuning of ψ is beyond the scope of our automatic hyperparameter selection mechanism.

6.2 Proof Sketch of Theorem 6.1

A key property of proxy functions $\widehat{R}^\pi(q)$ is to satisfy "sandwich" inequalities,

$$\frac{|\mathscr{J}_Q(q) - J^\pi|}{\kappa' \|w^\pi\|_2} - \epsilon' \leq \widehat{R}^\pi(q) \leq R^\pi(q), \qquad (36)$$

with $\kappa' \geq 1$ and $\epsilon' \geq 0$. In fact, these inequalities directly translate to PAC selection, as shown below (see Section E.1 for the proof).

Lemma 6.1. *The Q-selector given by Eq.* (11) *is $(\kappa', \kappa'\epsilon', \delta)$-PAC if Eq.* (36) *holds simultaneously for all $q \in \widehat{\mathcal{Q}}$ with probability at least $1-\delta$.*

Therefore, the proof is completed by showing both sides of Eq. (36) in high probability separately with $\kappa' = \kappa_\alpha(\psi)$ and $\epsilon' = \widetilde{\epsilon}_{\alpha,m,\delta}(\Xi)$ for each $q \in \widehat{\mathcal{Q}}$ and combine them with the union bound.

Right inequality. The inequality on the right in Eq. (36) is relatively straightforward,

$$\begin{aligned} \widehat{R}^\pi(q) &= \max_{\xi\in\Xi} \widehat{R}^{\pi,q}_{\mathrm{LCB}}(\widehat{f}_\xi) \\ &\leq_{1-|\Xi|\delta'} [R^{\pi,q}(\widehat{f}_\xi)]_+ && \because) \textit{ Eq. (23)} \\ &\leq R^\pi(q). && \because) \textit{ Eq. (1)} \end{aligned}$$

Here, $a \leq_p b$ denotes that $a \leq b$ holds with probability $p \in [0,1]$.

Left inequality. The inequality on the left in Eq. (36) is established as follows. First, fix any $\xi_* \in \Xi$ that attains the minimum of Eq. (31) and observe that $\widehat{R}^{\pi,q}_{\mathrm{LCB}}(\widehat{f}_*) \leq \widehat{R}^\pi(q)$ by Eq. (13). Thus, it is sufficient to show that, with probability $1-\delta$,

$$\frac{|\mathscr{J}_Q(q) - J^\pi|}{\kappa_\alpha(\psi)\|w^\pi\|_2} \leq \widehat{R}^{\pi,q}_{\mathrm{LCB}}(\widehat{f}_{\xi_*}) + \widetilde{\epsilon}_{\alpha,m,\delta'}(\xi_*), \qquad (37)$$

which can be handled by the theory of ERM. The complete proof is given in Section E.2.

7 CONCLUSION

In this paper, we propose the Duality-based Residual Estimator (DRE), which fills the gap towards *fully offline value-based RL*. In doing so, we also formalize the theoretical framework of *probably approximately correct (PAC) Q-selection*, providing a foundation for future studies on Q-selection. Interesting directions of future work include selectors that achieve $\kappa = 1$ and computational efficiency simultaneously, lower bounds on ϵ, and PAC Q-selection under severer discrepancy in occupancy distributions.

References

Antos, A., Szepesvári, C., and Munos, R. (2008). Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1):89–129.

Baird, L. et al. (1995). Residual algorithms: Reinforcement learning with function approximation. In *Proceedings of the twelfth international conference on machine learning*, pages 30–37.

Chen, J. and Jiang, N. (2019). Information-theoretic considerations in batch reinforcement learning. In *International conference on machine learning*, pages 1042–1051. PMLR.

Chen, X., Yao, L., McAuley, J., Zhou, G., and Wang, X. (2023). Deep reinforcement learning in recommender systems: A survey and new perspectives. *Knowledge-Based Systems*, 264:110335.

Cief, M., Kveton, B., and Kompan, M. (2025). Cross-validated off-policy evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 16073–16081.

Duan, Y., Jin, C., and Li, Z. (2021). Risk bounds and rademacher complexity in batch reinforcement learning. In *International Conference on Machine Learning*, pages 2892–2902. PMLR.

Ernst, D., Geurts, P., and Wehenkel, L. (2005). Tree-based batch mode reinforcement learning. In *Journal of Machine Learning Research*, volume 6, pages 503–556.

Farahmand, A.-m., Ghavamzadeh, M., Szepesvári, C., and Mannor, S. (2016). Regularized policy iteration with nonparametric function spaces. *Journal of Machine Learning Research*, 17(139):1–66.

Feng, Y., Li, L., and Liu, Q. (2019). A kernel loss for solving the bellman equation. *Advances in Neural Information Processing Systems*, 32.

Fujimoto, S., Meger, D., and Precup, D. (2021). A minimalist approach to offline reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 20132–20145.

Haarnoja, T., Zhou, A., Hartikainen, K., Tucker, G., Ha, S., Tan, J., Kumar, V., Zhu, H., Gupta, A., Abbeel, P., and Levine, S. (2018). Soft actor-critic algorithms and applications. In *International Conference on Machine Learning*, pages 1861–1870. PMLR.

Kostrikov, I., Nair, A., and Levine, S. (2022). Offline reinforcement learning with implicit q-learning. In *International Conference on Learning Representations (ICLR)*.

Kumar, A., Zhou, A., Tucker, G., and Levine, S. (2020). Conservative q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 1179–1191.

Le, H., Voloshin, C., and Yue, Y. (2019). Batch policy learning under constraints. In *International Conference on Machine Learning (ICML)*, pages 3703–3712. PMLR.

Levine, S., Kumar, A., Tucker, G., and Fu, J. (2020). Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*.

Liu, P., Zhao, L., Agarwal, S., Liu, J., Huang, A., Amortila, P., and Jiang, N. (2025). Model selection for off-policy evaluation: New algorithms and experimental protocol. *arXiv preprint arXiv:2502.08021*.

Liu, Q., Li, L., Tang, Z., and Zhou, D. (2018). Breaking the curse of horizon: Infinite-horizon off-policy estimation. *Advances in neural information processing systems*, 31.

Maurer, A. and Pontil, M. (2009). Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*.

Miyaguchi, K. (2022). Almost hyperparameter-free hyperparameter selection framework for offline policy evaluation. In *AAAI Conference on Artificial Intelligence*.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.

Munos, R. and Szepesvári, C. (2008). Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(5).

Nachum, O., Chow, Y., Dai, B., and Li, L. (2019). Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. *Advances in neural information processing systems*, 32.

Nachum, O. and Dai, B. (2020). Reinforcement learning via Fenchel-Rockafellar duality. *arXiv preprint arXiv:2001.01866*.

Nie, A., Chandak, Y., Yuan, C. J., Badrinath, A., Flet-Berliac, Y., and Brunskill, E. (2024). Opera: Automatic offline policy evaluation with re-weighted aggregates of multiple estimators. *Advances in Neural Information Processing Systems*, 37:103652–103680.

Pippas, N., Ludvig, E. A., and Turkay, C. (2025). The evolution of reinforcement learning in quantitative finance: A survey. *ACM Computing Surveys*, 57(11):1–51.

Prudêncio, R. F., Maximo, M. R. O. A., and Colombini, E. L. (2023). A survey on offline reinforcement learning: Taxonomy, review, and open problems. *IEEE Transactions on Neural Networks and Learning Systems*.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Steinwart, I. and Fischer, S. (2021). A closer look at covering number bounds for Gaussian kernels. *Journal of Complexity*, 62:101513.

Su, Y., Srinath, P., and Krishnamurthy, A. (2020). Adaptive estimator selection for off-policy evaluation. In *International Conference on Machine Learning*, pages 9196–9205. PMLR.

Tang, C., Abbatematteo, B., Hu, J., Chandra, R., Martín-Martín, R., and Stone, P. (2025). Deep reinforcement learning for robotics: A survey of real-world successes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 28694–28698.

Udagawa, T., Kiyohara, H., Narita, Y., Saito, Y., and Tateno, K. (2023). Policy-adaptive estimator selection for off-policy evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 10025–10033.

Uehara, M., Huang, J., and Jiang, N. (2020). Minimax weight and q-function learning for off-policy evaluation. In *International Conference on Machine Learning*, pages 9659–9668. PMLR.

Uehara, M., Imaizumi, M., Jiang, N., Kallus, N., Sun, W., and Xie, T. (2021). Finite sample analysis of minimax offline reinforcement learning: Completeness, fast rates and first-order efficiency. *arXiv preprint arXiv:2102.02981*.

Uehara, M., Shi, C., and Kallus, N. (2022). A review of off-policy evaluation in reinforcement learning. *arXiv preprint arXiv:2212.06355*.

Uehara, M. and Sun, W. (2021). Pessimistic model-based offline reinforcement learning under partial coverage. *arXiv preprint arXiv:2107.06226*.

Vázquez-Canteli, J. R. and Nagy, Z. (2019). Reinforcement learning for demand response: A review of algorithms and modeling techniques. *Applied energy*, 235:1072–1089.

Voloshin, C., Le, H. M., Jiang, N., and Yue, Y. (2021). Empirical study of off-policy policy evaluation for reinforcement learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Wang, H., Sakhadeo, A., White, A. M., Bell, J. M., Liu, V., Zhao, X., Liu, P., Kozuno, T., Fyshe, A., and White, M. (2022). No more pesky hyperparameters: Offline hyperparameter tuning for RL. *Transactions on Machine Learning Research*.

Yang, M., Nachum, O., Dai, B., Li, L., and Schuurmans, D. (2020). Off-policy evaluation via the regularized lagrangian. *Advances in Neural Information Processing Systems*, 33:6551–6561.

Yin, M. and Wang, Y.-X. (2020). Asymptotically efficient off-policy evaluation for tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3948–3958. PMLR.

Yu, C., Liu, J., and Nemati, S. (2019). Reinforcement learning in healthcare: a survey. *arXiv preprint arXiv:1908.08796*.

Yu, T., Thomas, G., Yu, L., Ermon, S., Zou, J. Y., Levine, S., Finn, C., and Ma, T. (2020). MOPO: Model-based offline policy optimization. *Advances in Neural Information Processing Systems*, 33:14129–14142.

Zhang, S. and Jiang, N. (2021). Towards hyperparameter-free policy selection for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 34:12864–12875.

Zhu, Y. and Ying, L. (2020). Borrowing from the future: An attempt to address double sampling. In *Mathematical and scientific machine learning*, pages 246–268. PMLR.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes: all of them.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Yes: sample complexity.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. Not Applicable.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes: Assumptions 3.1 and 3.2.
   (b) Complete proofs of all theoretical results. Yes.
   (c) Clear explanations of any assumptions. Yes.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Not Applicable.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Not Applicable.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Not Applicable.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). Not Applicable.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. Not Applicable.
   (b) The license information of the assets, if applicable. Not Applicable.

(c) New assets either in the supplemental material or as a URL, if applicable. Not Applicable.

(d) Information about consent from data providers/curators. Not Applicable.

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable.

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. Not Applicable.

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable.

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable.

Contents

1 INTRODUCTION **1**

2 PREVIOUS WORK **2**

3 PRELIMINARIES **2**

3.1 Definitions . . . 3

3.2 Assumptions . . . 4

4 PROBLEM FORMULATION **4**

5 METHOD **5**

5.1 Implementation of First Step . . . 5

5.1.1 Exact ERM Estimators . . . 5

5.1.2 Relaxed ERM Estimators . . . 5

5.2 Implementation of Second Step . . . 6

5.2.1 Hoeffding-type LCB . . . 6

5.2.2 Generalized LCB . . . 6

6 RESULTS **6**

6.1 Answers to Research Questions . . . 7

6.2 Proof Sketch of Theorem 6.1 . . . 8

7 CONCLUSION **8**

A ADDITIONAL RESULTS AND DISCUSSIONS **13**

A.1 Table of Notations . . . 13

A.2 Formal Statement of Eq. (1) . . . 13

A.3 Stochastic Gradient-based Algorithms for Relaxed ERM with ψ_2. . . . 14

A.4 Special Values of Approximation Ratio $\kappa_\alpha(\psi)$ . . . 14

A.5 Additional Comparison with Modified Bellman Residual Minimization (MBRM) . . . 14

B PRELIMINARY EXPERIMENTAL RESULTS **15**

B.1 Experimental Procedure . . . 15

B.2 Results . . . 15

C PROOF OF Section 3 **15**

C.1 Proof of Proposition 3.1 . . . 15

D PROOF OF Section 5 **17**

D.1 Proof of Proposition 5.1 . . . 17

D.2 Proof of Proposition A.2 17

E PROOFS FOR Section 6 **18**

E.1 Proof of Lemma 6.1 18

E.2 Proof of Theorem 6.1 18

E.2.1 Proof of Lemma E.1 18

E.2.2 Proof of Lemma E.2 19

E.2.3 Proof of Lemma E.3 20

E.3 Additional Lemmas 20

E.3.1 Fractional Sensitivity Bounds 20

E.3.2 Concentration of Empirical Norms 21

E.3.3 Concentration of Empirical Objective 21

A ADDITIONAL RESULTS AND DISCUSSIONS

A.1 Table of Notations

Notation	Description
$\mathcal{S}$, $\mathcal{A}$	state and action spaces
$\mathcal{D}$	dataset of transitions
π	target policy
J^π	expected return of π
μ_b	behavior distribution
$\langle\cdot,\cdot\rangle$	inner product in $L^2(\mu_\mathrm{b})$
μ^π	state-action distribution induced by π
w^π	density ratio of μ^π with respect to μ_b
q^π	action-value function of π
C_Q	boundedness constant of q^π
$\mathcal{B}^\pi$	Bellman operator of π
$R^\pi(q)$	Bellman residual of q with respect to π
$\mathscr{J}_Q(q)$	Q-based OPE estimator
$\mathscr{S}_Q$	Q-selector
$\epsilon^*_{\mathcal{I},m,\kappa,\delta}(\mathscr{S}_Q)$	instance-dependent worst additive error of $\mathscr{S}_Q$
$\mathcal{F}_\xi$	model of estimator, or function class of dual variables
$c_A(\mathcal{F}_\xi)$	cost of adaptivity of $\mathcal{F}_\xi$
ψ	clipping function
$\mathscr{E}_\alpha(\mathcal{F})$	approximation error of $\mathcal{F}$
$\mathscr{H}_m(\mathcal{F})$	complexity measure of $\mathcal{F}$
$\kappa_\alpha(\psi)$	approximation ratio of ψ

Table 2: Table of notations.

A.2 Formal Statement of Eq. (1)

Below is the formal statement and the proof of the clipped dual-norm representation (CDR) given in Eq. (1).

Proposition A.1 (Clipped Dual-norm Representation)**.** *For any bounded $q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and any policy π, we have*

$$\|q - \mathcal{B}^\pi q\|_2 = \sup_{f:\mathcal{S}\times\mathcal{A}\to\mathbb{R}} \frac{\langle q - \mathcal{B}^\pi q, f\rangle}{\max\{1, \|f\|_2\}}. \tag{38}$$

Proof. Let $\mathcal{F} := \{f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}\}$ be the set of all measurable functions. Then,

$$\begin{aligned}\sup_{f \in \mathcal{F}} \frac{\langle q - \mathcal{B}^\pi q, f\rangle}{\max\{1, \|f\|_2\}} &= \sup_{f:\|f\|_2 \le 1} \langle q - \mathcal{B}^\pi q, f\rangle \\ &= \|q - \mathcal{B}^\pi q\|_2,\end{aligned}$$

where the last equation is due to the Cauchy–Schwarz inequality. This concludes the proof. □

A.3 Stochastic Gradient-based Algorithms for Relaxed ERM with ψ_2.

Consider an augmented form of the surrogate,

$$\begin{aligned}\widehat{R}_1^{\pi,q,\psi_2}(f,u) := u\widehat{\mathbb{E}}_{\mathcal{D}_1'}\left[e_q^\pi(z)\, f(s,a)\right] \\ - \frac{u^2}{2}\widehat{\mathbb{E}}_{\mathcal{D}_1'}[1 + f^2(s,a)],\end{aligned} \tag{39}$$

which is equivalent to $\widehat{R}_1^{\pi,q,\psi_2}(f)$ in the following sense (see Section D.2 for the proof).

Proposition A.2. *For all $q, f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, we have*

$$\max_{u \ge 0} \widehat{R}_1^{\pi,q,\psi_2}(f,u) = \frac{1}{2}\{\widehat{R}_1^{\pi,q,\psi_2}(f)\}_+^2, \tag{40}$$

where $\{u\}_+ := \max\{0, u\}$.

Because of this equivalence, maximizing the augmented surrogate is equivalent to maximizing the unaugmented one. In particular, the augmented form admits cheap stochastic gradient oracles,

$$\begin{aligned}\widehat{\nabla}_f &:= u\{e_q^\pi(z) - uf(s,a)\}\,\nabla f(s,a), \\ \widehat{\nabla}_u &:= e_q^\pi(z) f(s,a) - u\{1 + f^2(s,a)\}\end{aligned}$$

with $z = (s, a, r, s')$ randomly sampled from $\mathcal{D}_1'$, making it possible to approximate $\widehat{f}_\xi^{\psi_2}$ with stochastic gradient-based algorithms.

A.4 Special Values of Approximation Ratio $\kappa_\alpha(\psi)$

Hard clipping function ψ_∞. With $\psi = \psi_\infty$, we have the *optimal* approximation ratio $\kappa_\alpha(\psi) = 1$ with $\alpha \ge 1$. However, as mentioned in Section 5.1.2, the corresponding ERM estimator may be computationally expensive.

Soft clipping function ψ_2. With $\psi = \psi_2$, the approximation ratio is well bounded at $\kappa_\alpha(\psi) = \sqrt{2}$ with $\alpha = 1$. Despite being not optimal, it is sufficient for most applications and, importantly, the corresponding ERM estimator is relatively computationally inexpensive, as discussed in Section 5.1.2. If one cares about the optimality of the approximation ratio, one can increase α and obtain $\kappa_\alpha(\psi_2) \le 1 + \frac{1}{2}\alpha^{-2}$, while that, however, makes the approximation error $\mathscr{E}_\alpha(\mathcal{F})$ more difficult to minimize.

A.5 Additional Comparison with Modified Bellman Residual Minimization (MBRM)

MBRM (see Section 2 for references) is an estimator of the *squared* L^2-Bellman residual based on its regression-based correction of the biased empirical squared Bellman residual,

$$\{R^\pi(q)\}^2 = \underbrace{\mathbb{E}[e_q^\pi(z)]^2}_{\text{biased estimate}} - \underbrace{\inf_{f:\mathcal{S}\times\mathcal{A}\to\mathbb{R}} \mathbb{E}[e_q^\pi(z) - f(s,a)]^2}_{\text{correction term}}, \tag{41}$$

which simplifies to

$$\{R^\pi(q)\}^2 = \sup_{f:\mathcal{S}\times\mathcal{A}\to\mathbb{R}} \left\{2\mathbb{E}[e_q^\pi(z)\, f(s,a)] - \mathbb{E}[f^2(s,a)]\right\}. \tag{42}$$

The last equation resembles our clipped dual-norm representation (CDR), or more specifically its relaxed and augmented form given in Section A.3. A key difference of this regression-based representation and CDR lies in their estimands: the former is used to estimate the *squared* Bellman residual, rather than the residual itself. This leads to a crucial difference in their costs of adaptivity (Eq. (10)): a naive adaptation of MBRM would suffer from suboptimal costs of $c'_A(\mathcal{F}_\xi)\{\ln|\Xi|/m\}^{1/4}$, while ours enjoy $c_A(\mathcal{F}_\xi)\{\ln|\Xi|/m\}^{1/2}$, which is comparable to standard hyperparameter-tuning procedures in supervised learning.

B PRELIMINARY EXPERIMENTAL RESULTS

B.1 Experimental Procedure

We have tested the proposed method for 3 different evaluation policies and 6 different sample size, and compared the resulting OPE error $\left|\widehat{J}_* - J^\pi\right|$ with the best choice $\min_{\lambda\in\Lambda}\left|\widehat{J}_\lambda - J^\pi\right|$ as well as the worst choice $\max_{\lambda\in\Lambda}\left|\widehat{J}_\lambda - J^\pi\right|$.

The overall experimental procedure is summarized as follows:

1. Prepare three ϵ-greedy policies π_ϵ with $\epsilon \in \{0.01, 0.1, 1\}$, where π_ϵ takes random actions with probability ϵ and the optimal action with probability $1-\epsilon$.
2. Generate offline datasets $\mathcal{D}_{\mathrm{all}}$ by running π_ϵ for each ϵ up to 1000 episodes and mix the resulting 3000 episodes, then, for each $K \in \{5, 15, 50, 150, 500, 1500\}$, randomly subsample $2K$ unique episodes from $\mathcal{D}_{\mathrm{all}}$, and
   - pick K episodes as the dataset for OPE $\mathcal{D}$, and
   - let the remaining K episodes be the dataset for Q-selection $\mathcal{D}'$.
3. **OPE)** Setting each $\pi = \pi_\epsilon$ and $\gamma = 0.99$, perform value-based OPE (the minimax indirect learning (MIL) proposed by Uehara et al., 2021) with tabular function classes and different regularization parameters, resulting in 16 distinct estimates of q^π (i.e., $|\Lambda| = |\widehat{\mathcal{Q}}| = 16$).
4. **Q-Selection)** Perform Algorithm 1 with each $q \in \widehat{\mathcal{Q}}$, where
   - the first step is implemented with the linear model (Eq. (18)) with $\phi = \xi\phi_{\mathrm{onehot}}$ with radius parameters $\xi \in \Xi := \{1, 10, 100, 1000\}$, and
   - the second step is implemented with the empirical Bernstein LCB (Eq. (25)) with $\delta' = 0.05/(1+|\Xi|)/|\Lambda|$ according to Theorem 6.1.
5. Compare the resulting value estimate $\widehat{J}_*$ with the ground truth J^{π_ϵ}.

B.2 Results

We report the results in Tables 3 to 5. Here, `dre_error`, `min_error`, and `max_error` represent the errors of the proposed method, the best choice, and the worst choice, while `regret` and `normalized_regret` are computed by `dre_error - min_error` and `regret / (max_error - min_error)`.

For all ϵ, we have observed that all of `dre_error`, `regret` and `normalized_regret` converges to near zero, demonstrating that the proposed method can pick near-optimal Q-function estimates with sufficiently large sample size as our theory predicted.

C PROOF OF Section 3

C.1 Proof of Proposition 3.1

Proposition 3.1 is a corollary of the following lemma.

Table 3: Result of $\epsilon = 0.01$

K	dre_error	min_error	max_error	regret	normalized_regret
5	0.496	0.033	0.496	0.463	1
15	0.496	0.022	0.496	0.474	1
50	0.496	0.077	0.496	0.419	1
150	0.140	0.064	0.496	0.076	0.175
500	0.171	0.019	0.496	0.152	0.318
1500	0.032	0.032	0.496	0	0

Table 4: Result of $\epsilon = 0.1$

K	dre_error	min_error	max_error	regret	normalized_regret
5	0.327	0.277	0.661	0.050	0.131
15	0.327	0.048	0.327	0.279	1
50	0.327	0.047	0.327	0.280	1
150	0.015	0.015	0.327	0	0
500	0.025	0.010	0.327	0.015	0.047
1500	0.005	0.005	0.327	0	0

Lemma C.1. *We have*

$$\mathscr{J}_Q(q) - J^\pi = \mathbb{E}_{(s,a)\sim\mu^\pi}[q(s,a) - \mathcal{B}^\pi q(s,a)].$$

Consequently, if $\mu^\pi \ll \mu_{\mathrm{b}}$,

$$\mathscr{J}_Q(q) - J^\pi = \langle q - \mathcal{B}^\pi q, w^\pi \rangle. \tag{43}$$

Proof. Let $J^q := \mathscr{J}_Q(q)$ for simplicity. Let $\mathcal{P}^\pi : \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ and $\mathcal{P}'^\pi : \Delta(\mathcal{S}\times\mathcal{A}) \to \Delta(\mathcal{S}\times\mathcal{A})$ be the transition and the adjoint transition operators such that $(\mathcal{P}^\pi f)(s,a) = \mathbb{E}_{s'\sim P_+(s,a),a'\sim\pi(s')}[f(s',a')]$ and $(\mathcal{P}'^\pi\nu)(s',a') = \mathbb{E}_{(s,a)\sim\nu}[P_+(s'|s,a)\pi(a'|s)]$. Also let $\bar{r}(s,a) := \mathbb{E}_{r\sim P_r(s,a)}[r]$ be the expected reward function. Observe that $p_1^\pi = \mu^\pi - \gamma\mathcal{P}'^\pi\mu^\pi$, where $p_1^\pi \in \Delta(\mathcal{S}\times\mathcal{A})$ is the initial state-action distribution such that $p_1^\pi(s,a) = p_1(s)\pi(a|s)$. Thus, we have

$$\begin{aligned}
J^q &= \mathbb{E}_{(s,a)\sim p_1^\pi}[q(s,a)] \\
&= \mathbb{E}_{(s,a)\sim\mu^\pi}[q(s,a)] - \gamma\mathbb{E}_{(s,a)\sim\mathcal{P}'^\pi\mu^\pi}[q(s,a)] \\
&= \mathbb{E}_{(s,a)\sim\mu^\pi}[q(s,a) - \gamma\mathcal{P}^\pi q(s',a')] \\
&= \mathbb{E}_{(s,a)\sim\mu^\pi}[q(s,a) - \mathcal{B}^\pi q(s,a) + r(s,a)] \\
&= \mathbb{E}_{(s,a)\sim\mu^\pi}[q(s,a) - \mathcal{B}^\pi q(s,a)] + J^\pi.
\end{aligned}$$

This concludes the proof. □

Proof of Proposition 3.1. The first case $\mu^\pi \ll \mu_{\mathrm{b}}$ is proved by the Cauchy–Schwarz inequality on Eq. (43). Specifically, the supremum is attained with the sequence $\{q_k\}_k$ satisfying $q - \mathcal{B}^\pi q = \frac{u}{\|w_k\|_2} w_k$, or equivalently $q_k = q^\pi + \frac{u}{\|w_k\|_2}(I - \gamma\mathcal{P}^\pi)^{-1}w_k$, where $\{w_k\}_k$ is a sequence of uniformly bounded functions converging to w^π monotonically from below.

The second case $\mu^\pi \not\ll \mu_{\mathrm{b}}$ is proved by the Lebesgue decomposition $\mu^\pi = \mu^\pi_{\ll\mathrm{b}} + \mu^\pi_{\perp\mathrm{b}}$, where $\mu^\pi_{\ll\mathrm{b}}$ and $\mu^\pi_{\perp\mathrm{b}} \neq 0$ are the absolutely continuous and singular components with respect to μ_{b}. Since there exists a set $E \subset \mathcal{S}\times\mathcal{A}$ such that $\mu^\pi_{\ll\mathrm{b}}(E) = 0$ and $\mu^\pi_{\perp\mathrm{b}}(E^c) = 0$, we can attain the supremum (infinity) with $q_k = q^\pi + \frac{u}{\|w_0\|_2}(I - \gamma\mathcal{P}^\pi)^{-1}(w_0 + k1_E)$, $k \to \infty$, where w_0 is a uniformly bounded function on $\mathcal{S}\times\mathcal{A}$ and 1_E is the indicator function of E. □

Table 5: Result of $\epsilon = 1$

K	dre_error	min_error	max_error	regret	normalized_regret
5	0.017	0.017	0.699	0	0
15	0.017	0.004	0.342	0.013	0.039
50	0.018	0.013	0.177	0.005	0.032
150	0.012	0.011	0.238	0.001	0.005
500	0.007	0.004	0.192	0.003	0.014
1500	0.008	0.004	0.211	0.004	0.018

D PROOF OF Section 5

D.1 Proof of Proposition 5.1

Proof. Let $f_\theta = \langle \theta, \phi \rangle_{\mathcal{H}}$, $\|\theta\|_{\mathcal{H}} \leq 1$. Plugging it into Eq. (17), we have

$$\widehat{R}_1^{\pi,q,\psi_2}(f_\theta) = \frac{\langle y, \theta \rangle_{\mathcal{H}}}{\sqrt{1 + \langle \theta, \Sigma\theta \rangle_{\mathcal{H}}}}.$$

Without loss of generality, we assume $y \neq 0$ (otherwise all feasible solutions are maxima). Observe that the objective is maximized only if θ is on the unit sphere. Thus, restricting the solution space accordingly, we further have

$$\widehat{R}_1^{\pi,q,\psi_2}(f_\theta) = \frac{\langle y, \theta \rangle_{\mathcal{H}}}{\sqrt{\|\theta\|_2^2 + \langle \theta, \Sigma\theta \rangle_{\mathcal{H}}}}.$$

Now, due to its invariance with respect to the scale perturbation on θ, the RHS no longer requires the constraint of $\|\theta\|_2 = 1$ to preserve the maxima of interest. We examine the stationary condition of the unconstrained optimization problem,

$$\begin{aligned} 0 &= \nabla_\theta \widehat{\{R_1^{\pi,q,\psi_2}(f_\theta)\}}^2 \\ &= \frac{2\langle \theta, (I+\Sigma)\theta \rangle_{\mathcal{H}} \langle y, \theta \rangle_{\mathcal{H}} y - 2\langle y, \theta \rangle_{\mathcal{H}}^2 (I+\Sigma)\theta}{\langle \theta, (I+\Sigma)\theta \rangle_{\mathcal{H}}^2}, \end{aligned}$$

which holds only if $\langle y, \theta \rangle_{\mathcal{H}} = 0$ or $\theta \propto (I+\Sigma)^{-1} y$. Since any θ satisfying $\langle y, \theta \rangle_{\mathcal{H}} = 0$ is invalid as a maxima, the maximum should be attained with $\theta \propto (I+\Sigma)^{-1} y$. Remembering that the maxima lie in the unit sphere, we arrive at the desired result. □

D.2 Proof of Proposition A.2

Proof. Observe that $\max_{u \in \mathbb{R}} \widehat{R}_1^{\pi,q,\psi_2}(f, u)$ is attained with

$$u^* = \frac{\left\{ \widehat{\mathbb{E}}_{\mathcal{D}_1'} \left[e_q^\pi(s,a,r,s')\, f(s,a) \right] \right\}_+}{\widehat{\mathbb{E}}_{\mathcal{D}_1'}[1 + f^2(s,a)]}.$$

Substituting it to Eq. (39), we have the desired equality as

$$\begin{aligned} \max_{u \geq 0} \widehat{R}_1^{\pi,q,\psi_2}(f, u) &= \frac{1}{2} \frac{\left\{ \widehat{\mathbb{E}}_{\mathcal{D}_1'} \left[e_q^\pi(s,a,r,s')\, f(s,a) \right] \right\}_+^2}{\widehat{\mathbb{E}}_{\mathcal{D}_1'}[1 + f^2(s,a)]} \\ &= \frac{1}{2} \{ \widehat{R}^{\pi,q,\psi_2}(f) \}_+^2. \end{aligned}$$

□

E PROOFS FOR Section 6

E.1 Proof of Lemma 6.1

Proof. Let $q_* \in \arg\min_{q\in\widehat{\mathcal{Q}}} R^\pi(q)$. Then, the sandwich property implies that, with probability $1-\delta$,

$$\begin{aligned}\frac{|\widehat{J}_* - J^\pi|}{\kappa'\|w^\pi\|_2} &\le \widehat{R}^\pi(\widehat{q}_*) + \epsilon' \\ &\le \widehat{R}^\pi(q_*) + \epsilon' \\ &\le R^\pi(q_*) + \epsilon' \\ &= \min_{q\in\widehat{\mathcal{Q}}} R^\pi(q) + \epsilon',\end{aligned}$$

where the first inequality is due to the left sandwich inequality, the second one is due to Eq. (11) and the third one is due to the right sandwich inequality. This is nothing but the $(\kappa', \epsilon', \delta)$-PAC guarantee. □

E.2 Proof of Theorem 6.1

According to Section 6.2, it suffices to show Eq. (37), which is established by Lemmas E.1 to E.3 below.

Let $R^{\pi,q,\psi}(f)$ be the oracle counterpart of the generalized objective Eq. (17),

$$R^{\pi,q,\psi}(f) := \frac{\langle q - \mathcal{B}^\pi q\, f\rangle}{\psi(\|f\|_2)}, \tag{44}$$

and denote its supremum over $f \in \mathcal{F}_{\xi_*}$ by

$$R^{\pi,q,\psi}_* := \sup_{f\in\mathcal{F}_{\xi_*}} R^{\pi,q,\psi}(f). \tag{45}$$

Also let $C'_Q := 1 + 2C'_Q$.

Lemma E.1. *We have*

$$\frac{|\mathscr{J}_Q(q) - J^\pi|}{\kappa_\alpha(\psi)\|w^\pi\|_2} \le R^{\pi,q,\psi}_* + 6C'_Q\mathscr{E}_\alpha(\mathcal{F}_{\xi_*}). \tag{46}$$

Lemma E.2. *With probability $1-\delta'$, we have*

$$R^{\pi,q,\psi}_* \le R^{\pi,q}(\widehat{f}_{\xi_*}) + 38C'_Q\|\mathcal{F}_{\xi_*}\|_\infty\sqrt{\frac{\mathscr{H}_m(\mathcal{F}_{\xi_*}) + \ln(8/\delta')}{m_1}}. \tag{47}$$

Lemma E.3. *With probability $1-\delta'$, we have*

$$R^{\pi,q}(\widehat{f}_{\xi_*}) \le \widehat{R}^{\pi,q}_{\mathrm{LCB}}(\widehat{f}_{\xi_*}) + 20C'_Q\|\mathcal{F}_{\xi_*}\|_\infty\sqrt{\frac{\ln(8/\delta')}{m_2}}. \tag{48}$$

The proofs of these lemmas are given in Sections E.2.1 to E.2.3.

E.2.1 Proof of Lemma E.1

Proof. Let $\Delta^\pi q := q - \mathcal{B}^\pi q$, $\bar{\Delta}^\pi q := (\Delta^\pi q)/\|\Delta^\pi q\|_2$ and $\rho^\pi(q) := \|\Delta^\pi q\|_2/C'_Q$. Let $\mathcal{C}_\alpha(q, w^\pi) := \{\alpha\bar{\Delta}^\pi q,\, \alpha\bar{w}^\pi\}$. Fix any $\varepsilon > \mathscr{E}_\alpha(\mathcal{F})$. Then, since $0 \le \rho^\pi(q) \le 1$, we can pick $f \in \mathcal{F}_{\xi_*}$, $c \in (0,1]$, $w \in \mathcal{C}_\alpha(q, w^\pi)$ such that

$\rho^\pi(q)\|cf - w\|_2 \le \varepsilon$. Thus, we have

$$\begin{aligned}
\frac{|\mathscr{J}_Q(q) - J^\pi|}{\kappa_\alpha(\psi)\|w^\pi\|_2} &= \frac{\langle \Delta^\pi q, w^\pi\rangle}{\kappa_\alpha(\psi)\|w^\pi\|_2} && \because) \textit{ Lemma C.1}\\
&= \frac{\langle \Delta^\pi q, \bar{w}^\pi\rangle}{\kappa_\alpha(\psi)} \\
&= \frac{\langle \Delta^\pi q, \alpha\bar{w}^\pi\rangle}{\psi(\alpha)} && \because)\ \kappa_\alpha(\psi) = \psi(\alpha)/\alpha \\
&\le \frac{\langle \Delta^\pi q, w\rangle}{\psi(\alpha)} && \because)\ w \in \mathcal{C}_\alpha(q, w^\pi) \\
&= \frac{\langle \Delta^\pi q, w\rangle}{\psi(\|w\|_2)} \\
&\le \frac{\langle \Delta^\pi q, cf\rangle}{\psi(\|cf\|_2)} + \frac{6\|\Delta^\pi q\|_2\|cf - w\|_2}{\psi(\|w\|_2)} && \because) \textit{ Corollary E.1}\\
&\le \frac{\langle \Delta^\pi q, cf\rangle}{\psi(\|cf\|_2)} + 6C'_Q\varepsilon && \because)\ \psi(\cdot) \ge 1 \\
&\le \frac{\langle \Delta^\pi q, f\rangle}{\psi(\|f\|_2)} + 6C'_Q\varepsilon && \because)\ \psi(u)/u \text{ is non-increasing} \\
&\le R^{\pi,q,\psi}_* + 6C'_Q\varepsilon. && \because)\ f \in \mathcal{F}_{\xi_*}
\end{aligned}$$

where Corollary E.1 is applied with $\delta_a = |\langle \Delta^\pi q, cf - w\rangle| \le \|\Delta^\pi q\|_2\|cf - w\|_2$, $\delta_b = |\|cf\|_2 - \|w\|_2| \le \|cf - w\|_2$ and $\widetilde{c} \le \sup_f R^{\pi,q}(f) = R^\pi(q) = \|\Delta^\pi q\|_2$. Since ε can be arbitrarily close to $\varepsilon_\alpha(\mathcal{F})$, we obtain the desired inequality. □

E.2.2 Proof of Lemma E.2

Proof. Let $\ddot{\mathcal{F}}$ be a minimum ϵ-covering of $\mathcal{F}_{\xi_*}$. For $f \in \mathcal{F}_{\xi_*}$, let $\ddot{f} \in \ddot{\mathcal{F}}$ is the closest element from f. Then, we have

$$\begin{aligned}
&|\widehat{R}^{\pi,q,\psi}_1(f) - R^{\pi,q,\psi}(f)| \\
&\le |\widehat{R}^{\pi,q,\psi}_1(f) - \widehat{R}^{\pi,q,\psi}_1(\ddot{f})| + |\widehat{R}^{\pi,q,\psi}_1(\ddot{f}) - R^{\pi,q,\psi}(\ddot{f})| \\
&\qquad + |R^{\pi,q,\psi}(\ddot{f}) - R^{\pi,q,\psi}(f)| \\
&\le 10C'_Q\|\mathcal{F}_{\xi_*}\|_\infty\sqrt{\frac{\ln(4|\ddot{\mathcal{F}}|/\delta)}{m_1}} + 6C'_Q\frac{\|\mathcal{F}_{\xi_*}\|_\infty}{\sqrt{m}} \\
&\le 19C'_Q\|\mathcal{F}_{\xi_*}\|_\infty\sqrt{\frac{\mathscr{H}_m(\mathcal{F}_{\xi_*}) + \ln(4/\delta)}{m}} \\
&=: \sigma_*
\end{aligned}$$

simultaneously for all $f \in \mathcal{F}_{\xi_*}$ with probability $1 - \delta$, where the second inequality follows from evaluating the second term with the union bound of Lemma E.6 over $\ddot{f} \in \ddot{\mathcal{F}}$ and evaluating the first and third terms by Lemma E.4 with $\|f - \ddot{f}\|_\infty \le \|\mathcal{F}_{\xi_*}\|_\infty/\sqrt{m}$. Thus, letting $f_* = \arg\max_{f\in\mathcal{F}_{\xi_*}} R^{\pi,q,\psi}(f)$, we have

$$\begin{aligned}
R^{\pi,q,\psi}_* &= R^{\pi,q,\psi}(f_*) && (49)\\
&\le \widehat{R}^{\pi,q,\psi}_1(f_*) + \sigma_* && (50)\\
&\le \widehat{R}^{\pi,q,\psi}_1(\widehat{f}_{\xi_*}) + \sigma_* && (51)\\
&\le R^{\pi,q,\psi}(\widehat{f}_{\xi_*}) + 2\sigma_* && (52)
\end{aligned}$$

with the same probability. The proof is completed by bounding the RHS with $R^{\pi,q,\psi}(\cdot) \le R^{\pi,q}(\cdot)$ and letting $\delta \leftarrow \delta'/2$. □

E.2.3 Proof of Lemma E.3

Proof. Recall that $\widehat{f}_{\xi_*}$ is independent of $\mathcal{D}'_2$. Thus, by Lemma E.6 with $f = \widehat{f}_{\xi_*}$, we have

$$R^{\pi,q}(\widehat{f}_{\xi_*}) \leq \widehat{R}_2^{\pi,q}(\widehat{f}_{\xi_*}) + 10C'_Q\|\mathcal{F}_{\xi_*}\|_\infty\sqrt{\frac{\ln(4/\delta)}{m_2}} \tag{53}$$

with probability $1-\delta$. Moreover, Eq. (23) implies

$$\widehat{R}_2^{\pi,q}(\widehat{f}_{\xi_*}) = \widehat{R}^{\pi,q}_{\text{LCB-H}}(\widehat{f}_{\xi_*}) + \sigma_{\text{H}}(\widehat{f}_{\xi_*}) \tag{54}$$

$$\leq \widehat{R}^{\pi,q}_{\text{LCB}}(\widehat{f}_{\xi_*}) + \sigma_{\text{H}}(\widehat{f}_{\xi_*}). \tag{55}$$

The proof is completed by chaining these two inequalities and letting $\delta \leftarrow \delta'/2$. □

E.3 Additional Lemmas

E.3.1 Fractional Sensitivity Bounds

Lemma E.4. *For all $a, a' \in \mathbb{R}$ and $b, b' > 0$, we have*

$$\left|\frac{a'}{b'} - \frac{a}{b}\right| \leq \frac{3}{b}(\delta_a + c\delta_b), \tag{56}$$

where $\delta_a := |a' - a|$, $\delta_b := |b' - b|$ and $c := \max\{|a/b|, |a'/b'|\}$.

Proof. If $\delta_b > 2b/3$, the RHS of Eq. (56) is larger than $2c$, while the LHS is always no larger than $2c$. On the contrary, if $\delta_b \leq 2b/3$, take the absolute values of both sides of

$$\frac{a'}{b'} - \frac{a}{b} = \frac{a'-a}{b'} + \frac{a}{b}\left(\frac{1}{1+(b'-b)/b} - 1\right) \tag{57}$$

and obtain

$$\left|\frac{a'}{b'} - \frac{a}{b}\right| = \frac{\delta_a}{b'} + \frac{|a|}{b}\left|\frac{1}{1+(b'-b)/b} - 1\right|. \tag{58}$$

Since $b' \geq b - \delta_b \geq b/3$ and $|(1+x)^{-1} - 1| \leq 3|x|$ for $|x| \leq 2/3$, we further have

$$\left|\frac{a'}{b'} - \frac{a}{b}\right| = \frac{3\delta_a}{b} + \frac{3a\delta_b}{b^2}. \tag{59}$$

The desired result is obtained by bounding the second term in the RHS with $a/b \leq c$. □

Corollary E.1. *Let $\psi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{>0}$ be 1-Lipschitz. Under the same setting as Lemma E.4, we have*

$$\left|\frac{a'}{\psi(b')} - \frac{a}{\psi(b)}\right| \leq \frac{3}{\psi(b)}\left(\delta_a + \widetilde{c}\delta_b\right), \tag{60}$$

where $\widetilde{c} := \max\{a/\psi(b), a'/\psi(b')\}$.

Proof. It follows from $|\psi(b') - \psi(b)| \leq \delta_b$. □

E.3.2 Concentration of Empirical Norms

Lemma E.5. *Let $g, g_1, \dots, g_m$ be i.i.d. random variables with $\|g\|_2 := \{\mathbb{E}[g^2]\}^{1/2}$ and $\|g\|_\infty := \inf_{c\geq 0}\{\mathbb{P}(|g| \leq c) = 1\}$. Then, $\widehat{\nu} := \{\frac{1}{m}\sum_{i=1} g_i^2\}^{1/2}$ satisfies*

$$\mathbb{P}\left\{\widehat{\nu} - \|g\|_2 \leq \|g\|_\infty \sqrt{\frac{\ln(1/\delta)}{2m}}\right\} \geq 1-\delta, \tag{61}$$

$$\mathbb{P}\left\{\widehat{\nu} - \|g\|_2 \geq -\|g\|_\infty \sqrt{\frac{8\ln(1/\delta)}{3m}}\right\} \geq 1-\delta. \tag{62}$$

Proof. The one-side Bennett's inequality implies, with probability $1-\delta$,

$$\widehat{\nu}^2 - \|g\|_2^2 \leq \sqrt{\frac{2\mathbb{V}[g^2]\ln(1/\delta)}{m}} + \frac{\|g\|_\infty^2 \ln(1/\delta)}{3m} \tag{63}$$

$$\leq 2\|g\|_2\|g\|_\infty \sqrt{\frac{\ln(1/\delta)}{2m}} + \frac{\|g\|_\infty^2\ln(1/\delta)}{3m}, \qquad \because \sqrt{\mathbb{V}[g^2]} \leq \|g^2\|_2 \leq \|g\|_\infty\|g\|_2 \tag{64}$$

resulting in

$$\widehat{\nu}^2 \leq \left(\|g\|_2 + \|g\|_\infty\sqrt{\frac{\ln(1/\delta)}{2m}}\right)^2. \tag{65}$$

This proves the first result. Similarly, the other side of Bennett's inequality implies with probability $1-\delta$,

$$\widehat{\nu}^2 - \|g\|_2^2 \geq -2\|g\|_2\|g\|_\infty\sqrt{\frac{\ln(1/\delta)}{2m}} - \frac{\|g\|_\infty^2\ln(1/\delta)}{3m}, \tag{66}$$

resulting in

$$\widehat{\nu}^2 + \frac{5\|g\|_\infty^2\ln(1/\delta)}{6m} \geq \left(\|g\|_2 - \|g\|_\infty\sqrt{\frac{\ln(1/\delta)}{2m}}\right)^2 \tag{67}$$

This proves the second result using $\sqrt{5/6} + \sqrt{1/2} \leq \sqrt{8/3}$. □

E.3.3 Concentration of Empirical Objective

Lemma E.6. *With probability $1-\delta$, we have*

$$|\widehat{R}_j^{\pi,q,\psi}(f) - R^{\pi,q,\psi}(f)| \leq 10C_Q'\|f\|_\infty\sqrt{\frac{\ln(4/\delta)}{m_j}} \tag{68}$$

for all $q \in [-C_Q, C_Q]^{\mathcal{S}\times\mathcal{A}}$ and $f \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$.

Proof. Let $\delta_{\mathrm{num}} := |\widehat{\mathbb{E}}_{\mathcal{D}_j}[e_q^\pi(z)\,f(s,a)] - \langle q - \mathcal{B}^\pi q, f\rangle|$ and $\delta_{\mathrm{den}} := |\{\widehat{\mathbb{E}}_{\mathcal{D}_j}[f^2(s,a)]\}^{1/2} - \|f\|_2|$. By Corollary E.1, we have

$$|\widehat{R}_j^{\pi,q,\psi}(f) - R^{\pi,q,\psi}(f)| \leq \frac{3}{\psi(\|f\|_2)}\left(\delta_{\mathrm{num}} + C_Q'\delta_{\mathrm{den}}\right). \tag{69}$$

We bound δ_{num} and δ_{num} separately. By Hoeffding's inequality (Maurer and Pontil, 2009) with $\|e_q^\pi f\|_\infty \leq C_Q'\|f\|_\infty$, δ_{num} is bounded as

$$\delta_{\mathrm{num}} \leq C_Q'\|f\|_\infty\sqrt{\frac{2\ln(4/\delta)}{m_j}} \tag{70}$$

with probability $1-\delta/2$. Moreover, By Lemma E.5 with $g = f(s,a)$, δ_{den} is bounded as

$$\delta_{\text{den}} \leq \|f\|_\infty \sqrt{\frac{8\ln(4/\delta)}{3m_j}} \tag{71}$$

with probability $1-\delta/2$. Plugging them back to Eq. (69) and bounding the resulting expression from above with $\psi(\|f\|_2) \geq 1$, we obtain the desired result using $3(\sqrt{2}+\sqrt{8/3}) \leq 10$. □