NON-ASYMPTOTIC CONFIDENCE INTERVALS OF OFF-POLICY EVALUATION: PRIMAL AND DUAL BOUNDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Off-policy evaluation remains a central challenge in applying reinforcement learning to real-world observational data. Constructing confidence intervals of off-policy estimators can be even more challenging. However, in high-stakes areas such as medical treatment, it is essential to provide rigorous confidence bounds, not just a point estimation, for safety concerns. In this work, we propose a novel approach to constructing non-asymptotic confidence intervals of off-policy evaluation. Our algorithm is highly flexible and can be applied to infinite-horizon reinforcement learning problems with behavior-agnostic data, consisting of dependent transition pairs collected in an arbitrary way. The key idea of our method is taking an optimization-based approach to confidence bounds through a new primal-dual lens, and leveraging a new tight concentration inequality applicable to dependent data. We further present empirical results that clearly demonstrate the advantages of our approach over existing methods.

1 INTRODUCTION

Off-policy evaluation (OPE) seeks to estimate the expected reward of a target policy in reinforcement learnings (RL) from observational data collected under different policies (e.g., Murphy et al., 2001; Fonteneau et al., 2013; Jiang & Li, 2016; Liu et al., 2018). OPE plays a central role in applying reinforcement learning (RL) with only observational data and has found important applications in areas such as medicine, self-driving, where interactive "on-policy" data is expensive or even infeasible to collect. A critical challenge in OPE is the uncertainty estimation, as having reliable confidence bounds is essential for making high-stakes decisions. In this work, we aim to tackle this problem by providing non-asymptotic confidence intervals of the expected value of the target policy. Our method allows us to rigorously quantify the uncertainty of the prediction and hence avoid the dangerous case of being overconfident in making costly and/or irreversible decisions.

However, off-policy evaluation per se has remained a key technical challenge in the literature (e.g., Precup, 2000; Thomas & Brunskill, 2016; Jiang & Li, 2016; Liu et al., 2018), let alone gaining rigorous confidence estimation of it. This is especially true when 1) the underlying RL problem is long or infinite horizon, and 2) the data are collected under arbitrary and unknown algorithms (a.k.a. behavior-agnostic). As a consequence, the collected data can exhibit arbitrary dependency structure, which makes constructing rigorous non-asymptotic confidence bounds particularly challenging. Traditionally, the only approach to provide non-asymptotic confidence bounds in OPE is to combine importance sampling with concentration inequalities (e.g., Thomas et al., 2015a), which, however, tends to degenerate for long/infinite horizon problems (Liu et al., 2018). Furthermore, neither can this approach be applied to the behavior-agnostic settings, nor can it effectively handle the complicated time dependency structure inside individual trajectories. Instead, it requires to use a large number of independently collected trajectories drawn under known policies.

In this work, we provide a practical approach for **B**ehavior-agnostic, **O**ff-policy, **I**nfinite-horizon, **N**on-asymptotic, **C**onfidence intervals based on arbitrarily **D**ependent data (**BONDIC**). Our method is motivated by a recently proposed optimization-based (or variational) approach to estimate OPE confidence bounds (Feng et al., 2020), which leverages a tail bound of kernel Bellman statistics (Feng et al., 2019). Our approach achieves a new bound that is both an order-of-magnitude tighter and computationally much more efficient than that of Feng et al. (2020). Our improvements are based on two pillars 1) developing a new primal-dual perspective on the non-asymptotic OPE confidence

bounds, which is connected to a body of recent works on infinite-horizon value estimation (Liu et al., 2018; Nachum et al., 2019a; Tang et al., 2020; Mousavi et al., 2020); and 2) offering a new tight concentration inequality on the kernel Bellman statistics that applies to behavior-agnostic off-policy data with arbitrary dependency between transition pairs. Empirically, we demonstrate that our method can provide reliable and tight bounds on a variety of well-established benchmarks.

Related Work Besides the aforementioned approach based on the combination of IS and concentration inequalities (e.g., Thomas et al., 2015a), bootstrapping methods have also been widely used in off-policy estimation (e.g., White & White, 2010; Hanna et al., 2017; Kostrikov & Nachum, 2020). But the latter is limited to asymptotic bounds. Alternatively, Bayesian methods (e.g. Engel et al., 2005) offers a different way to estimate the uncertainty in RL, but fails to guarantee frequentist coverage. In addition, Distributed RL (Bellemare et al., 2017) seeks to quantify the intrinsic uncertainties inside the Markov decision process, which is orthogonal to the estimation of uncertainty that we consider.

Our work reveals new aspects on the recent advances in behavior-agnostic infinite-horizon OPE, including Liu et al. (2018); Feng et al. (2019); Tang et al. (2020); Mousavi et al. (2020), as well as the DICE-family (e.g., Nachum et al., 2019a; Zhang et al., 2020a;b). Recently, a minimax aspect of the infinite-horizon OPE in the *infinite data region* is covered in Tang et al. (2020); Uehara et al. (2020); Jiang & Huang (2020). Our framework is a substantial extension of it to the *non-asymptotic region*, and therefore is of both theoretical and practical significance.

Outline For the rest of the paper, we start with the problem statement in Section 2 , and an overview on the two dual approaches to infinite-horizon OPE that are tightly connected to our method in Section 3. We then present our main approach in Section 4 and perform empirical studies in Section 5. The proof and an abundance of additional discussions can be found in Appendix.

2 Background, Data Assumption, Problem Setting

Consider an agent acting in an unknown environment. At each time step t, the agent observes the current state s_t in a state space $\mathcal{S}$, takes an action $a_t \sim \pi(\cdot \mid s_t)$ in an action space $\mathcal{A}$ according to a *given* policy π; then, the agent receives a reward r_t and the state transits to $s'_t = s_{t+1}$, following an *unknown* transition/reward distribution $(r_t, s_{t+1}) \sim \mathsf{P}(\cdot \mid s_t, a_t)$. Assume the initial state s_0 is drawn from an *known* initial distribution D_0. Let $\gamma \in (0, 1)$ be a discount factor. In this setting, the expected reward of π is defined as $J_\pi := \mathbb{E}_\pi \left[\sum_{t=0}^{T} \gamma^t r_t \mid s_0 \sim \mathsf{D}_0\right]$, which is the expected total discounted rewards when we execute π starting from D_0 for T steps. In this work, we consider the *infinite-horizon* case with $T \to +\infty$.

Our goal is to provide an interval estimation of J_π for a general and challenging setting with significantly released constraints on the data. In particular, we assume the data is *behavior-agnostic* and *off-policy*, which means that the data can be collected from multiple experiments, each of which can execute a mix of arbitrary, unknown policies, or even follow a non-fixed policy. More concretely, suppose that the model P is unknown, and we have a set of transition pairs $\hat{\mathsf{D}}_n = (s_i, a_i, r_i, s'_i)_{i=1}^n$ collected from previous experiments in a sequential order, such that for each data point i, the (r_i, s'_i) is drawn from the model $\mathsf{P}(\cdot \mid s_i, a_i)$, while (s_i, a_i) is generated with an arbitrary black box given the previous data points. We formalize both the data assumption and goal as below.

Assumption 2.1 (**Data Assumption**). *Assume the data $\hat{\mathsf{D}}_n = (s_i, a_i, r_i, s'_i)_{i=1}^n$ is drawn from an arbitrary joint distribution, such that for each $i = 1, \ldots, n$, conditional on $\hat{\mathsf{D}}_{<i} := (s_j, a_j, r_j, s'_j)_{j<i} \cup (s_i, a_i)$, the local rewards and next states (r_i, s'_i) is drawn from $\mathsf{P}(\cdot \mid s_i, a_i)$.*

Goal *Given a confidence level $\delta \in (0, 1)$, we want to construct an interval $[\hat{J}^-, \hat{J}^+] \subset \mathbb{R}$ based on the data $\hat{\mathsf{D}}_n$, such that $\Pr(J_\pi \in [\hat{J}^-, \hat{J}^+]) \geq 1 - \delta$, where $\Pr(\cdot)$ is w.r.t. the randomness of the data.*

The partial ordering on the data points is introduced to accommodate the case that s_{i+1} equals s'_j for some $j \leq i$. The data assumption only requires that (r_i, s'_i) is generated from $\mathsf{P}(\cdot \mid s_i, a_i)$, and imposes no constraints on how (s_i, a_i) is generated. This then provides great flexibility in terms of the data collection process. In particular, we do not require $(s_i, a_i)_{i=1}^n$ to be independent as always assumed in recent works (Liu et al., 2018; Mousavi et al., 2020).

A crucial fact is that our data assumption actually implies a martingale structure on the empirical Bellman residual operator of the Q-function, As we will show in Section 4.1, this enables us to derive a key concentration inequality underpinning our non-asymptotic confidence bounds.

Here, we summarize a few notations that will simplify the presentation in the rest of work. First of all, we append each (s_i, a_i, r_i, s_i') with an action $a_i' \sim \pi(\cdot \mid s_i')$ following s_i'. This can be done for free as long as π is given. Also, we write $x_i = (s_i, a_i)$, $x_i' = (s_i', a_i')$, and $y_i = (x_i', r_i) = (s_i', a_i', r_i)$. In this way, the observed data can be written as pairs of $\{x_i, y_i\}_{i=1}^n$. Correspondingly, define $\mathcal{X} = \mathcal{S} \times \mathcal{A}$ to be the state-action space. Then, we can denote $\mathsf{P}_\pi(y \mid x) = \mathsf{P}(s', r \mid x)\pi(a' \mid s')$. Under these notations, Assumption 2.1 is equivalent to saying that $y_i \sim \mathsf{P}_\pi(\cdot \mid x_i)$ given $\hat{\mathsf{D}}_{<i}$. We equalize the data $\hat{\mathsf{D}}_n$ with its empirical measure $\hat{\mathsf{D}}_n = \sum_{i=1}^n \delta_{x_i, y_i}/n$, where δ is the Delta measure.

3 Two dual approaches to Infinite-horizon off-policy estimation

The deficiency of the traditional IS methods on long-/infinite-horizon RL problems (a.k.a. *the curse of horizon* (Liu et al., 2018)) has motivated a line of work on developing efficient infinite-horizon value estimation (e.g., Liu et al., 2018; Feng et al., 2019; Nachum et al., 2019a; Zhang et al., 2020a; Mousavi et al., 2020; Tang et al., 2020). The main idea is to transform the value estimation problem into estimating either the *Q-function* or the *visitation distribution (or its related density ratio)* of the policy π. This section introduces and reinterprets these two tightly connected methods, which serves to lay out a foundation for our main confidence bounds from a primal and dual perspective.

Given a policy π, its Q-function is defined as $q_\pi(x) = \mathbb{E}_\pi\left[\sum_{t=0}^\infty \gamma^t r_t \mid x_0 = x\right]$, where the expectation is taken when we execute π initialized from a fixed state-action pair $(s_0, a_0) = x_0 = x$. Let $\mathsf{D}_{\pi,t}$ be the distribution of $(x_t, y_t) = (s_t, a_t, s_t', a_t', r_t)$ when executing policy π starting from $s_0 \sim \mathsf{D}_0$ for t steps. The visitation distribution of π is defined as $\mathsf{D}_\pi = \sum_{t=0}^\infty \gamma^t \mathsf{D}_{\pi,t}$. Note that D_π integrates to $1/(1-\gamma)$, while we treat it as a probability measure in the notation.

The expected reward $J_\pi$can be expressed using either q_π or D_π as follows:

$$J_\pi := \mathbb{E}_\pi\left[\sum_{t=0}^\infty \gamma^t r_t\right] = \mathbb{E}_{r\sim\mathsf{D}_\pi}[r] = \mathbb{E}_{x\sim\mathsf{D}_{\pi,0}}[q_\pi(x)], \tag{1}$$

where $r \sim \mathsf{D}_\pi$ (resp. $x \sim \mathsf{D}_{\pi,0}$) denotes sampling from the r-(resp. x-) marginal distribution of D_π (resp. $\mathsf{D}_{\pi,0}$). Eq. (1) plays a key role in the infinite-horizon value estimation by transforming the estimation of J_π into estimating either q_π or D_π.

Value Estimation via Q Function Because $\mathsf{D}_{\pi,0}(x) = \mathsf{D}_0(s)\pi(a|s)$ is known, we can estimate J_π by $\mathbb{E}_{x\sim\mathsf{D}_{\pi,0}}[\hat{q}(x)]$ with any estimation $\hat{q}$ of the true Q-function q_π; the expectation under $x \sim \mathsf{D}_{\pi,0}$ can be estimated to any accuracy with Monte Carlo. To estimate q_π, we consider the empirical and expected Bellman residual operator:

$$\hat{\mathbf{R}}q(x,y) = q(x) - \gamma q(x') - r, \qquad \mathbf{R}_\pi q(x) = \mathbb{E}_{y\sim\mathsf{P}_\pi(\cdot|x)}\left[\hat{\mathbf{R}}q(x,y)\right]. \tag{2}$$

It is well-known that q_π is the unique solution of the *Bellman equation* $\mathbf{R}_\pi q = 0$. Since $y_i \sim \mathsf{P}_\pi(\cdot|x_i)$ for each data point in $\hat{\mathsf{D}}_n$, if $q = q_\pi$, then $\hat{\mathbf{R}}q(x_i, y_i)$, $i = 1, \ldots, n$ are all *zero-mean* random variables. Let ω be any function from $\mathcal{X}$ to $\mathbb{R}$, then $\sum_i \hat{\mathbf{R}}q(x_i, y_i)\omega(x_i)$ also has zero mean. This motivates the following *functional* Bellman loss (Feng et al., 2019; 2020; Xie & Jiang, 2020),

$$L_{\mathcal{W}}(q;\ \hat{\mathsf{D}}_n) := \sup_{\omega\in\mathcal{W}}\left\{\frac{1}{n}\sum_{i=1}^n \hat{\mathbf{R}}q(x_i,y_i)\omega(x_i)\right\}, \tag{3}$$

where $\mathcal{W}$ is a set of functions $\omega\colon \mathcal{X} \to \mathbb{R}$. To ensure that the sup is finite, $\mathcal{W}$ is typically set to be an unit ball of some normed function space $\mathcal{W}_o$, such that $\mathcal{W} = \{\omega \in \mathcal{W}_o\colon \|\omega\|_{\mathcal{W}_o} \le 1\}$. Feng et al. (2019) considers the simple case when $\mathcal{W}$ is taken to be the unit ball $\mathcal{K}$ of the reproducing kernel Hilbert space (RKHS) with a positive definite kernel $k\colon \mathcal{X}\times\mathcal{X} \to \mathbb{R}$, in which case the loss has a simple closed form solution:

$$L_{\mathcal{K}}(q;\ \hat{\mathsf{D}}_n) = \sqrt{\frac{1}{n^2}\sum_{ij=1}^n \hat{\mathbf{R}}q(x_i,y_i)k(x_i,x_j)\hat{\mathbf{R}}q(x_j,y_j)}. \tag{4}$$

Note that the RHS of Eq. (4) is the *square root* of the kernel Bellman V-statistics in Feng et al. (2019). Feng et al. (2019) showed that, when the support of data distribution $\hat{\mathsf{D}}_n$ covers the whole space (which will mostly require an infinite data size) and k is an integrally strictly positive definite kernel, $L_{\mathcal{K}}(q;\ \hat{\mathsf{D}}_n) = 0$ iff $q = q_\pi$. Therefore, one can estimate q_π by minimizing $L_{\mathcal{K}}(q, \hat{\mathsf{D}}_n)$.

Remark The empirical Bellman residual operator $\hat{\mathbf{R}}$ can be extended to $\hat{\mathbf{R}}q(x,y) = q(x) - r - \gamma \frac{1}{m}\sum_{\ell=1}^{m} q(s', a'_\ell)$, where $\{a'_\ell\}_{i=1}^m$ are i.i.d. drawn from $\pi(\cdot|s')$. As m increases, this gives a lower variance estimation of $\mathbf{R}_\pi q$. If $m = +\infty$, we have $\hat{\mathbf{R}}q(x,y) = q(x) - r - \gamma \mathbb{E}_{a'\sim\pi(\cdot\ |\ s')}[q(s',a')]$, which coincides with the operator used in the expected SARSA (Sutton & Barto, 1998). In fact, without any modification, all results in this work can be applied to $\hat{\mathbf{R}}q$ for any m.

Value Estimation via Visitation Distribution Another way to estimate J_π in Eq. (1) is to approximate D_π with a weighted empirical measure of the data (Liu et al., 2018; Nachum et al., 2019a; Mousavi et al., 2020; Zhang et al., 2020a). The key idea is to assign an importance weight $\omega(x_i)$ to each data point in $\hat{\mathsf{D}}_n$. We can choose the function $\omega\colon \mathcal{X} \to \mathbb{R}$ properly such that D_π and hence J_π can be approximated by the ω-weighted empirical measure of $\hat{\mathsf{D}}_n$ as follows:

$$J_\pi \approx \hat{J}_\omega := \mathbb{E}_{\hat{\mathsf{D}}_n^\omega}[r] = \frac{1}{n}\sum_{i=1}^{n} \omega(x_i) r_i, \qquad \mathsf{D}_\pi \approx \hat{\mathsf{D}}_n^\omega := \frac{1}{n}\sum_{i=1}^{n} \omega(x_i)\delta_{x_i,y_i}. \tag{5}$$

Intuitively, ω can be viewed as the density ratio between D_π and $\hat{\mathsf{D}}_n$, although the empirical measure $\hat{\mathsf{D}}_n$ may not have well-defined density. Liu et al. (2018); Mousavi et al. (2020) proposed to estimate ω by minimizing a discrepancy measure between $\hat{\mathsf{D}}_n^\omega$ and D_π. To see this, note that $\mathsf{D} = \mathsf{D}_\pi$ if and only if $\Delta(\mathsf{D}, q) = 0$ for any function q, where

$$\begin{aligned}\Delta(\mathsf{D}, q) &= \mathbb{E}_{\mathsf{D}}[\gamma q(x') - q(x)] - \mathbb{E}_{\mathsf{D}_\pi}[\gamma q(x') - q(x)] \\ &= \mathbb{E}_{\mathsf{D}}[\gamma q(x') - q(x)] + \mathbb{E}_{\mathsf{D}_{\pi,0}}[q(x)],\end{aligned} \tag{6}$$

using the fact that $\mathbb{E}_{\mathsf{D}_\pi}[\gamma q(x') - q(x)] = -\mathbb{E}_{\mathsf{D}_{\pi,0}}[q(x)]$ (Theorem 1, Liu et al., 2018). Also note that the RHS of Eq. (6) can be practically calculated given any D and q without knowing D_π. Let $\mathcal{Q}$ be a set of functions $q\colon \mathcal{X} \to \mathbb{R}$. One can define the following loss for ω:

$$I_{\mathcal{Q}}(\omega;\ \hat{\mathsf{D}}_n) = \sup_{q\in\mathcal{Q}} \left\{ \Delta(\hat{\mathsf{D}}_n^\omega, q) \right\}. \tag{7}$$

Similar to $L_{\mathcal{W}}(q;\ \hat{\mathsf{D}}_n)$, when $\mathcal{Q}$ is a ball in RKHS, $I_{\mathcal{Q}}(\omega;\ \hat{\mathsf{D}}_n)$ also has a bilinear closed form analogous to Eq. (4); see Mousavi et al. (2020) and Appendix F. As we show in Section 4, $I_{\mathcal{Q}}(\omega;\ \hat{\mathsf{D}}_n)$ and $L_{\mathcal{W}}(q;\ \hat{\mathsf{D}}_n)$ are connected to the primal and dual views of our confidence bounds, respectively.

4 Main Approach

Let $\mathcal{Q}$ be a large enough function set including the true Q-function q_π, that is, $q_\pi \in \mathcal{Q}$. Following Feng et al. (2020), a confidence interval $[\hat{J}^-_{\mathcal{Q},\mathcal{W}}, \hat{J}^+_{\mathcal{Q},\mathcal{W}}]$ of J_π can be constructed as follows:

$$\hat{J}^+_{\mathcal{Q},\mathcal{W}} = \sup_{q\in\mathcal{Q}} \left\{ \mathbb{E}_{\mathsf{D}_{\pi,0}}[q] \quad s.t. \quad L_{\mathcal{W}}(q;\ \hat{\mathsf{D}}_n) \le \varepsilon_n \right\}, \tag{8}$$

and $\hat{J}^-_{\mathcal{Q},\mathcal{W}}$ is defined in a similar way by replacing $\max$ on $q \in \mathcal{Q}$ with $\min$.

The idea here is to seek the extreme q function with the largest (resp. smallest) expected values in set $\mathcal{F} := \mathcal{Q} \cap \{q\colon L_{\mathcal{K}}(q;\ \hat{\mathsf{D}}_n) \le \varepsilon_n\}$. Therefore, Eq. (8) would be a $1-\delta$ confidence interval if q_π is included in $\mathcal{F}$ with at least probability $1-\delta$, which is ensured when $q_\pi \in \mathcal{Q}$ and

$$\Pr(L_{\mathcal{W}}(q_\pi;\ \hat{\mathsf{D}}_n) \le \varepsilon_n) \ge 1 - \delta\,. \tag{9}$$

Feng et al. (2020) showed that in the RKHS case when $\mathcal{W} = \mathcal{K}$, Eq. (9) can be achieved with

$$\varepsilon_n = \sqrt{2c_{q_\pi,k}\left(\frac{n-1}{n}\sqrt{\frac{\log(1/\delta)}{n}} + \frac{1}{n}\right)}, \tag{10}$$

when n is an even number, where $c_{q_\pi,k} = \sup_{x,y} \hat{\mathbf{R}} q_\pi(x,y)^2 k(x,x)$. This was proved using Hoeffding's inequality for U-statistics (Hoeffding, 1963). To solve Eq. (8) efficiently, Feng et al. (2020) took $\mathcal{Q}$ to be a ball in RKHS with random feature approximation. Unfortunately, this method as described by Eq. (8)-(10) has two major disadvantages:

1) Bound Needs to Be Tightened (Section 4.1) The bound of $\varepsilon_n = O(n^{-1/4})$ in Eq. (10) is sub-optimal in rate. In Section 4.1, we improve it by an $\varepsilon_n = O(n^{-1/2})$ bound under the mild Assumption 2.1, which gets rid of the independence requirement between the transition pairs. Our tightened bound is achieved by firstly noting a Martingale structure on the empirical Bellman operator under Assumption 2.1, and then applying an inequality in Pinelis (1992).

2) Dependence on Global Optimization (Section 4.2) The bound in Eq. (8) is guaranteed to be a $1-\delta$ confidence bound only when the maximization in Eq. (8) is solved to global optimality. With a large n, this leads to a high computational cost, even when choosing $\mathcal{Q}$ as the RKHS. Feng et al. (2020) solved Eq. (8) approximately using a random feature technique, but this method suffers from a gap between the theory and practice. In Section 4.2, we address this problem by presenting a dual form of Eq. (8), which sidesteps solving the challenging global optimization in Eq. (8). Moreover, the dual form enables us to better analyze the tightness of the confidence interval and issues regarding the choices of $\mathcal{Q}$ and $\mathcal{W}$.

4.1 Tight Concentration Inequality

In this section, we explain our method to improve the bound in Eq. (10) by giving a tighter concentration inequality for the kernel Bellman loss in Eq. (4). We introduce the following *semi-expected* kernel Bellman loss:

$$L^*_{\mathcal{K}}(q;\ \hat{\mathsf{D}}_n) = \sqrt{\frac{1}{n^2}\sum_{ij=1}^{n} \mathbf{R}_\pi q(x_i) k(x_i,x_j) \mathbf{R}_\pi q(x_j)}\,, \tag{11}$$

in which we replace the empirical Bellman residual operator $\hat{\mathbf{R}}q$ in Eq. (3) with its expected counterpart $\mathbf{R}_\pi q$, but still take the empirical average over $\{x_i\}_{i=1}^n$ in $\hat{\mathsf{D}}_n$. For a more general function set $\mathcal{W}$, we can similarly define $L^*_{\mathcal{W}}(q;\ \hat{\mathsf{D}}_n)$ by replacing $\hat{\mathbf{R}}q$ with $\mathbf{R}_\pi q$ in Eq. (3). Obviously, we have $L^*_{\mathcal{W}}(q;\ \hat{\mathsf{D}}_n) = 0$ when $q = q_\pi$.

Theorem 4.1 below shows that $L_{\mathcal{K}}(q;\ \hat{\mathsf{D}}_n)$ concentrates around $L^*_{\mathcal{K}}(q;\ \hat{\mathsf{D}}_n)$ with an $O(n^{-1/2})$ error under Assumption 2.1. As Assumption 2.1 does not impose any (weak) independence between $\{x_i\}$, we cannot establish that $L_{\mathcal{K}}(q;\ \hat{\mathsf{D}}_n)$ concentrates around its mean $\mathbb{E}_{\hat{\mathsf{D}}_n}[L_{\mathcal{K}}(q;\ \hat{\mathsf{D}}_n)]$ (which is a *full* expected kernel bellman loss), without introducing further assumptions.

Theorem 4.1. *Assume $\mathcal{K}$ is the unit ball of RKHS with a positive definite kernel $k(\cdot,\cdot)$. Let $c_{q,k} := \sup_{x,y}(\hat{\mathbf{R}}q(x,y) - \mathbf{R}_\pi q(x))^2 k(x,x) < \infty$. Under Assumption 2.1, for any $\delta \in (0,1)$, with at least probability $1-\delta$, we have*

$$\left| L_{\mathcal{K}}(q;\ \hat{\mathsf{D}}_n) - L^*_{\mathcal{K}}(q;\ \hat{\mathsf{D}}_n) \right| \le \sqrt{\frac{2c_{q,k}\log(2/\delta)}{n}}\,. \tag{12}$$

In particular, when $q = q_\pi$, we have $c_{q_\pi,k} = \sup_{x,y}(\hat{\mathbf{R}}q_\pi(x,y))^2 k(x,x)$, and

$$L_{\mathcal{K}}(q_\pi;\ \hat{\mathsf{D}}_n) \le \sqrt{\frac{2c_{q_\pi,k}\log(2/\delta)}{n}}\,. \tag{13}$$

Intuitively, to see why we can expect an $O(n^{-1/2})$ bound, note that $L_{\mathcal{K}}(q, \hat{\mathsf{D}}_n)$ consists of the square root of the product of two $\hat{\mathbf{R}}q$ terms, each of which contributes an $O(n^{-1/2})$ error w.r.t. $\mathbf{R}_\pi q$.

Technically, the proof of this leverages a special property of RKHS and applies a Hoeffding-like inequality on the Hilbert spaces as in Pinelis (1992) (see Appendix B). For other more general function sets $\mathcal{W}$, we establish in Appendix E a similar bound by using Rademacher complexity, although it yields a less tight bound than Eq. (12) when $\mathcal{W} = \mathcal{K}$.

At a first glance, it may seem surprising that the concentration bound is able to hold even without any independence assumption between $\{x_i\}$. An easy way to make sense of this is by recognizing that the randomness in y_i conditional on x_i is aggregated through averaging, even when $\{x_i\}$ are deterministic. Formally, the proof is based on a key observation: Assumption 2.1 ensures that $Z_i := \hat{\mathbf{R}}q(x_i, y_i) - \mathbf{R}_\pi q(x_i)$, $i = 1, \ldots, n$ forms a *martingale difference* sequence w.r.t. $\{\hat{\mathsf{D}}_{<i}\colon\ \forall i = 1, \ldots, n\}$, in the sense that $\mathbb{E}[Z_i \mid \hat{\mathsf{D}}_{<i}] = 0, \forall i$. See Appendix B for details.

4.2 Dual Confidence Bounds

We derive a dual form of Eq. (8) that sidesteps the need for solving the challenging global optimization in Eq. (8). To do so, let us plug the definition of $L_{\mathcal{W}}(q;\ \hat{\mathsf{D}}_n)$ into Eq. (3) and introduce a Lagrange multiplier:

$$\hat{J}^{+}_{\mathcal{Q},\mathcal{W}} = \sup_{q\in\mathcal{Q}} \inf_{h\in\mathcal{W}} \inf_{\lambda\geq 0} \mathbb{E}_{\mathsf{D}_{\pi,0}}[q] - \lambda \left(\frac{1}{n}\sum_{i=1}^{n} h(x_i)\hat{\mathbf{R}}q(x_i, y_i) - \varepsilon_n \right) \tag{14}$$

$$= \sup_{q\in\mathcal{Q}} \inf_{\omega\in\mathcal{W}_o} \left\{ \mathbb{E}_{\mathsf{D}_{\pi,0}}[q] - \frac{1}{n}\sum_{i=1}^{n} \omega(x_i)\hat{\mathbf{R}}q(x_i) + \varepsilon_n \|\omega\|_{\mathcal{W}_o} \right\}, \tag{15}$$

where we use $\omega(x) = \lambda h(x)$. Exchanging the order of min/max and some further derivation yields the following main result.

Theorem 4.2. *I) Let $\mathcal{W}$ be the unit ball of a normed function space $\mathcal{W}_o$. We have*

$$\begin{aligned} \hat{J}^{+}_{\mathcal{Q},\mathcal{W}} &\leq \hat{F}^{+}_{\mathcal{Q}}(\omega) := \mathbb{E}_{\hat{\mathsf{D}}^{\omega}_n}[r] + I_{\mathcal{Q}}(\omega;\ \hat{\mathsf{D}}_n) + \varepsilon_n \|\omega\|_{\mathcal{W}_o}, \quad \forall \omega \in \mathcal{W}_o, \\ \hat{J}^{-}_{\mathcal{Q},\mathcal{W}} &\geq \hat{F}^{-}_{\mathcal{Q}}(\omega) := \mathbb{E}_{\hat{\mathsf{D}}^{\omega}_n}[r] - I_{-\mathcal{Q}}(\omega;\ \hat{\mathsf{D}}_n) - \varepsilon_n \|\omega\|_{\mathcal{W}_o}, \quad \forall \omega \in \mathcal{W}_o, \end{aligned} \tag{16}$$

where $-\mathcal{Q} = \{-q\colon q \in \mathcal{Q}\}$ and $I_{-\mathcal{Q}}(\omega;\ \hat{\mathsf{D}}_n) = I_{\mathcal{Q}}(\omega;\ \hat{\mathsf{D}}_n)$ if $\mathcal{Q} = -\mathcal{Q}$. Further, we have $\hat{J}^{+}_{\mathcal{Q},\mathcal{W}} = \inf_{\omega\in\mathcal{W}_o} \hat{F}^{+}_{\mathcal{Q}}(\omega)$ and $\hat{J}^{-}_{\mathcal{Q},\mathcal{W}} = \sup_{\omega\in\mathcal{W}_o} \hat{F}^{-}_{\mathcal{Q}}(\omega)$ if $\mathcal{Q}$ is convex and there exists a $q \in \mathcal{Q}$ that satisfies the strict feasibility condition that $L_{\mathcal{W}}(q;\ \hat{\mathsf{D}}_n) < \varepsilon_n$.

II) For $\hat{\mathsf{D}}_n$ and $\delta \in (0, 1)$, assume $\mathcal{W}_o$ and $\varepsilon_n \in \mathbb{R}$ satisfy Eq. (9) (e.g., via Theorem 4.1). Then for any function set $\mathcal{Q}$ with $q_\pi \in \mathcal{Q}$, and any function $\omega_+, \omega_- \in \mathcal{W}_o$ (the choice of $\mathcal{Q}$, ω_+, ω_- can depend on $\hat{\mathsf{D}}_n$ arbitrarily), we have

$$\Pr\left(J_\pi \in \left[\hat{F}^{-}_{\mathcal{Q}}(\omega_-),\ \ \hat{F}^{+}_{\mathcal{Q}}(\omega_+) \right] \right) \geq 1 - \delta. \tag{17}$$

Theorem 4.2 transforms the original bound in Eq. (8), framed in terms of q and $L_{\mathcal{W}}(q;\ \hat{\mathsf{D}}_n)$, into a form that involves the density-ratio ω and the related loss $I_{\mathcal{Q}}(\omega;\ \hat{\mathsf{D}}_n)$. The bounds in Eq. (16) can be interpreted as assigning an error bar around the ω-based estimator $\hat{J}_\omega = \mathbb{E}_{\hat{\mathsf{D}}^{\omega}_n}[r]$ in Eq. (5), with the error bar of $I_{\pm\mathcal{Q}}(\omega;\ \hat{\mathsf{D}}_n) + \varepsilon_n \|\omega\|_{\mathcal{W}_o}$. Specifically, the first term $I_{\pm\mathcal{Q}}(\omega;\ \hat{\mathsf{D}}_n)$ measures the discrepancy between $\hat{\mathsf{D}}^{\omega}_n$ and D_π as discussed in Eq. (7), whereas the second term captures the randomness in the empirical Bellman residual operator $\hat{\mathbf{R}}q_\pi$.

Compared with Eq. (8), the global maximization on $q \in \mathcal{Q}$ is now transformed inside the $I_{\mathcal{Q}}(\omega; \hat{\mathsf{D}}_n)$ term, which yields a simple closed form solution in the RKHS case (see Appendix F). In practice, we can optimize ω_+ and ω_- to obtain the tightest possible bound (and hence recover the primal bound) by minimizing/maximizing $\hat{F}^{+}_{\mathcal{Q}}(\omega)$ and $\hat{F}^{-}_{\mathcal{Q}}(\omega)$, but it is not necessary to solve the optimization to global optimality. When $\mathcal{W}_o$ is an RKHS, by the standard finite representer theorem (Scholkopf & Smola, 2018), the optimization on ω reduces to a finite dimensional optimization, which can be sped up with any favourable approximation techniques. We elaborate on this in Appendix D.

Length of the Confidence Interval The form in Eq. (16) also makes it much easier to analyze the tightness of the confidence interval. Suppose $\omega = \omega_+ = \omega_-$ and $\mathcal{Q} = -\mathcal{Q}$, the length of the optimal confidence interval is

$$\text{length}([\hat{J}^{-}_{\mathcal{Q},\mathcal{W}},\ \ \hat{J}^{+}_{\mathcal{Q},\mathcal{W}}]) = \min_{\omega\in\mathcal{W}_o} \left\{ 2I_{\mathcal{Q}}(\omega;\ \hat{\mathsf{D}}_n) + 2\varepsilon_n \|\omega\|_{\mathcal{W}_o} \right\}.$$

Given ε_n is $O(n^{-1/2})$, we can make the overall length of the optimal confidence interval also $O(n^{-1/2})$, as long as $\mathcal{W}_o$ is rich enough to include a *good* density ratio estimator ω^* that has an $O(n^{-1/2})$ loss $I_{\mathcal{Q}}(\omega^*;\ \hat{\mathsf{D}}_n)$ and bounded norm $\|\omega\|_{\mathcal{W}_o}$.

We can expect to achieve $I_{\mathcal{Q}}(\omega^*;\ \hat{\mathsf{D}}_n) = O(n^{-1/2})$, when (1) $\mathcal{Q}$ has an $O(n^{-1/2})$ sequential Rademacher complexity (Rakhlin et al., 2015) (e.g., a finite ball in RKHS); and (2) $\hat{\mathsf{D}}_n$ is collected following a Markov chain with strong mixing condition and weakly converges to some limit distribution D_∞ with support is $\mathcal{X}$, and therefore we can define ω^* as the density ratio between D_π and D_∞. Refer to Appendix C for more discussions. Indeed, our experiments show that the lengths of practically constructed confidence intervals do tend to decay with an $O(n^{-1/2})$ rate.

Choice of $\mathcal{W}$ and $\mathcal{Q}$ To ensure the concentration inequality in Theorem 4.1 is valid, the choice of $\mathcal{W}_o$ cannot depend on the data $\hat{\mathsf{D}}_n$. Therefore, we should use a separate holdout data given a data-dependent $\mathcal{W}_o$. In contrast, *the choice of $\mathcal{Q}$ can depend on the data $\hat{\mathsf{D}}_n$ arbitrarily*, since it is a part of the optimization bound Eq. (8) but not in the tail bound Eq. (9). In this light, one can construct the best possible $\mathcal{Q}$ by exploiting the data information in the most favourable way. For example, we can construct an estimator of $\hat{q} \approx q_\pi$ based on any state-of-the-art method (e.g., Q-learning or model-based methods), and set $\mathcal{Q}$ to be a ball centering around $\hat{q}$ such that $q_\pi - \hat{q} \in \mathcal{Q}$. This enables *post-hoc analysis* based on prior information on q_π, as suggested in Feng et al. (2020).

Mis-specification of $\mathcal{Q}$ and Oracle Upper/Lower Estimates Our result relies on the assumption that $q_\pi \in \mathcal{Q}$. However, as with other statistical estimation problems, there exists no provably way to empirically verify the correctness of model assumptions such as $q_\pi \in \mathcal{Q}$. Because empirical data only reveals the information of the unknown function (in our case q_π) on a finite number data points, but no conclusion can be made on the unseeing data points without imposing certain smoothness assumption. Typically, what we can do is the opposite: reject $q_\pi \in \mathcal{Q}$ when the Bellman loss $L_{\mathcal{W}}(q;\ \hat{\mathsf{D}}_n)$ of all q in $\mathcal{Q}$ is larger than the threshold ε_n.

We highlight that, even without verifying $q_\pi \in \mathcal{Q}$, our method can still be viewed as a confidence interval of a best possible (oracle) upper and lower estimation given the data $\hat{\mathsf{D}}_n$ *plus* the assumption that $q_\pi \in \mathcal{Q}$, defined as

$$\hat{J}^+_{\mathcal{Q},*} = \max_{q\in\mathcal{Q}} \left\{ \mathbb{E}_{\mathsf{D}_{\pi,0}}[q] \quad s.t. \quad \hat{\mathbf{R}}q(x_i, y_i) = \hat{\mathbf{R}}q_\pi(x_i, y_i), \quad \forall i = 1, \dots, n \right\} . \tag{18}$$

In fact, it is impossible to derive empirical upper bounds lower than $\hat{J}^+_{\mathcal{Q},*}$, as there is no way to distinguish q and q_π if $\hat{\mathbf{R}}q(x_i, y_i) = \hat{\mathbf{R}}q_\pi(x_i, y_i)$ for all i. But our interval $[\hat{J}_{\mathcal{Q},\mathcal{K}}, \hat{J}^+_{\mathcal{Q},\mathcal{K}}]$ provides a $1-\delta$ confidence outer bound of $[\hat{J}^-_{\mathcal{Q},*}, \hat{J}^+_{\mathcal{Q},*}]$ once Eq. (9) holds, regardless if $q_\pi \in \mathcal{Q}$ holds or not. Hence, it is of independent interest to further explore the dual form of Eq. (18), which is another starting point for deriving our bound. We have more discussion in Appendix G.

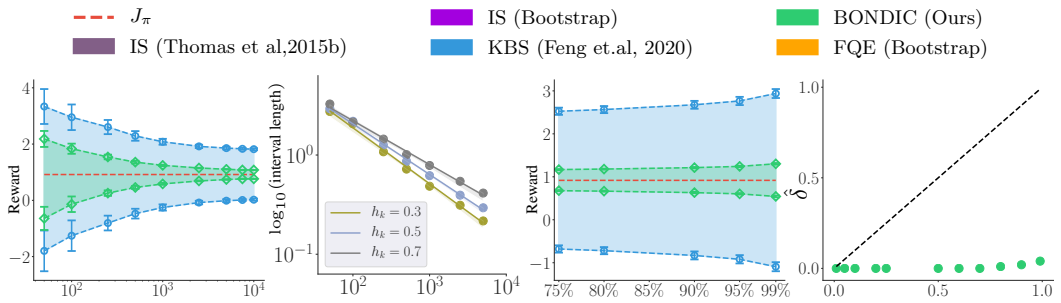
Lastly, we argue that it is important to include the $\mathcal{Q}$ in the bound. Proposition G.1 in Appendix shows that removing the $q \in \mathcal{Q}$ constraint in Eq. (18) would lead to an infinite upper bound, unless the $\{s_i, s_i'\}_{i=1}^n$ from $\hat{\mathsf{D}}_n$ almost surely covers the whole state space $\mathcal{S}$ in the sense that $\Pr_{s\sim\mathsf{D}_0}(s \in \{s_i, s_i'\}_{i=1}^n) = 1$.

5 Experiments

We compare our method with a variety of existing algorithms for obtaining asymptotic and non-asymptotic bounds on a number of benchmarks. We find our method can provide confidence interval that correctly covers the true expected reward (and hence safe) across the multiple examples we tested. In contrasted, the IS and bootstrapping based methods often fail to capture the true values (and hence unsafe). We also verify that the length of our confidence interval decays with an $O(n^{-1/2})$ rate approximately as the data size n increases, suggesting that our method is asymptotically consistent.

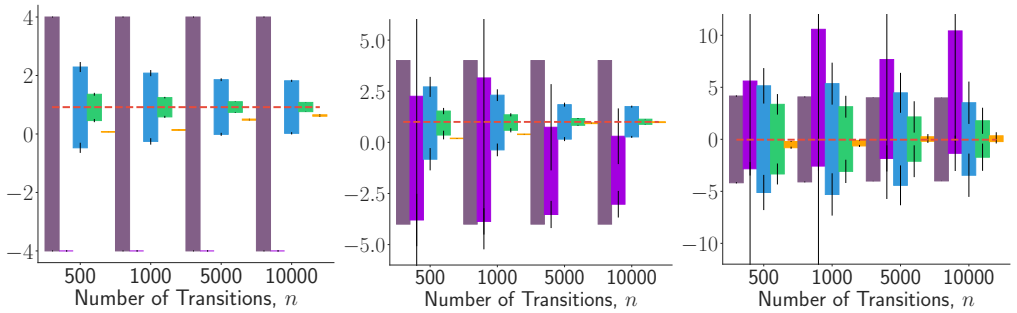
Environments and Dataset Construction We test our method on three environments: Inverted-Pendulum and CartPole from OpenAI Gym (Brockman et al., 2016), and a Type-1 Diabetes medical treatment simulator.[1] We follow a similar procedure as Feng et al. (2020) to construct behavior and target policies. Check Appendix H for more details on environments and data collection procedure.

[1] https://github.com/jxx123/simglucose.



(a) Transitions, n (b) Transitions, n (c) $1-\delta$ (d) δ

Figure 1: Results in Inverted-Pendulum. We set discounted factor $\gamma = 0.95$, the horizon length $H = 50$, number of transitions $n = 5000$, failure probability $\delta = 0.1$, temperature of the behavior policy as default. The slopes of the three lines in (b) from the top to bottom are $\{-0.44, -0.51, -0.56\}$, and each line represents the results using different bandwidth h_k for the kernel k of $\mathcal{W}$.



(a) Pendulum (b) CartPole (c) Type 1 diabetes

Figure 2: Evaluation results on different environment. Y-axis in all figures denotes the policy reward. We set number of transitions $n = 5000$, failure probability $\delta = 0.1$ as default.

Algorithm Settings We test the bound in Eq. (16). Throughout the experiment, we always set $\mathcal{W} = \mathcal{K}$, the unit ball of RKHS with kernel k, and $\mathcal{Q} = r_{\mathcal{Q}}\tilde{\mathcal{K}}$, the ball of radius $r_{\mathcal{Q}}$ in an RKHS with another kernel $\tilde{k}$. We take both kernels to be Gaussian RBF kernel and choose $r_{\mathcal{Q}}$ and bandwith as shown in Appendix H. We optimize ω using the method described in Appendix H. Once ω_+ and ω_- are found, we evaluate the bound in Eq. (16) exactly to ensure the theoretical guarantee holds.

Baseline Algorithms We compare our method with IS-based with high confidence bound (Thomas et al., 2015b) and bootstrapping methods for both fitted Q-evaluation and IS estimators (Kostrikov & Nachum, 2020). Specifically, we use the confidence bounds in Theorem 1 of Thomas et al. (2015b), and the bootstrapping procedure in Kostrikov & Nachum (2020).

Results Figure 1 shows the results on Inverted-Pendulum. We can see that our method is much tighter than Feng et al. (2020), which is obvious since we use a tighter ε_n. We can see that we provide increasingly tight bounds as the data size n increases, and the length of the interval decays with an $O(n^{-1/2})$ rate approximately as shown in Figure 1 (b). Figure 1 (d) plots the frequency that the interval fails to capture the true value in a total 100 random trials (denoted as $\hat{\delta}$). We can see that $\hat{\delta}$ remains close to zero even when δ is large. Part of the reason of this conservative results is because we are using conservative choices of the radius $r_{\mathcal{Q}}$ and coefficient $c_{q_\pi,k}$ in Eq. (13).

In Figure 2 we compare different algorithms on different domains. Our method is safe (always captures the true value) and tight (compared with IS-nonasymptotic bounds), while the other methods are either too loose or often fail to capture the ground truth.

6 Conclusion

We develop a dual approach to construct high confidence bounds for off-policy evaluation with an improved rate over Feng et al. (2020). Our method can handle dependent data, and does not require a global optimization to get a valid bound. Empirical results demonstrate that our bounds is tight and valid compared with a range of existing baseline. Future directions include leveraging our bounds for policy optimization and safe exploration.

REFERENCES

Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International Conference on Machine Learning*, pp. 449–458, 2017.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

Yaakov Engel, Shie Mannor, and Ron Meir. Reinforcement learning with gaussian processes. In *Proceedings of the 22nd international conference on Machine learning*, pp. 201–208, 2005.

Yihao Feng, Lihong Li, and Qiang Liu. A kernel loss for solving the bellman equation. In *Advances in Neural Information Processing Systems*, pp. 15456–15467, 2019.

Yihao Feng, Tongzheng Ren, Ziyang Tang, and Qiang Liu. Accountable off-policy evaluation with kernel bellman statistics. In *International Conference on Machine Learning*, 2020.

Raphael Fonteneau, Susan A. Murphy, Louis Wehenkel, and Damien Ernst. Batch mode reinforcement learning based on the synthesis of artificial trajectories. *Annals of Operations Research*, 208(1): 383–416, 2013.

Josiah P Hanna, Peter Stone, and Scott Niekum. Bootstrapping with models: Confidence intervals for off-policy evaluation. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Nan Jiang and Jiawei Huang. Minimax confidence interval for off-policy evaluation and policy optimization. *arXiv preprint arXiv:2002.02081*, 2020.

Nan Jiang and Lihong Li. Doubly robust off-policy evaluation for reinforcement learning. In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 652–661, 2016.

Ilya Kostrikov and Ofir Nachum. Statistical bootstrapping for uncertainty estimation in off-policy evaluation. *arXiv preprint arXiv:2007.13609*, 2020.

Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems*, pp. 5356–5366, 2018.

Ali Mousavi, Lihong Li, Qiang Liu, and Denny Zhou. Black-box off-policy estimation for infinite-horizon reinforcement learning. In *International Conference on Learning Representations*, 2020.

Susan A. Murphy, Mark van der Laan, and James M. Robins. Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, 96(456):1410–1423, 2001.

Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. In *Advances in Neural Information Processing Systems*, pp. 2318–2328, 2019a.

Ofir Nachum, Bo Dai, Ilya Kostrikov, Yinlam Chow, Lihong Li, and Dale Schuurmans. Algaedice: Policy gradient from arbitrary experience. *arXiv preprint arXiv:1912.02074*, 2019b.

Iosif Pinelis. An approach to inequalities for the distributions of infinite-dimensional martingales. In *Probability in Banach Spaces, 8: Proceedings of the Eighth International Conference*, pp. 128–134. Springer, 1992.

Doina Precup. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, pp. 80, 2000.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Sequential complexities and uniform martingale laws of large numbers. *Probability Theory and Related Fields*, 161(1-2):111–153, 2015.

Lorenzo Rosasco, Mikhail Belkin, and Ernesto De Vito. On learning with integral operators. *Journal of Machine Learning Research*, 11(2), 2010.

Bernhard Scholkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. Adaptive Computation and Machine Learning series, 2018.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A hilbert space embedding for distributions. In *Algorithmic learning theory*, pp. 13–31. Springer, 2007.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, March 1998. ISBN 0-262-19398-1.

Ziyang Tang, Yihao Feng, Lihong Li, Dengyong Zhou, and Qiang Liu. Doubly robust bias reduction in infinite horizon off-policy estimation. In *International Conference on Learning Representations (ICLR)*, 2020.

Philip S. Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, pp. 2139–2148, 2016.

Philip S. Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High confidence policy improvement. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 2380–2388, 2015a.

Philip S Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015b.

Masatoshi Uehara, Jiawei Huang, and Nan Jiang. Minimax weight and q-function learning for off-policy evaluation. *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Martha White and Adam White. Interval estimation for reinforcement-learning algorithms in continuous-state domains. In *Advances in Neural Information Processing Systems*, pp. 2433–2441, 2010.

Tengyang Xie and Nan Jiang. Q* approximation schemes for batch reinforcement learning: A theoretical comparison. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2020.

Ruiyi Zhang, Bo Dai, Lihong Li, and Dale Schuurmans. Gendice: Generalized offline estimation of stationary values. In *International Conference on Learning Representations*, 2020a.

Shangtong Zhang, Bo Liu, and Shimon Whiteson. Gradientdice: Rethinking generalized offline estimation of stationary values. In *International Conference on Machine Learning*, 2020b.

A PROOF OF THE DUAL BOUND IN THEOREM 4.2

Proof. Introducing a Lagrange multiplier, the bound in (8) is equivalent to

$$\begin{aligned}\hat{J}^{+}_{\mathcal{Q},\mathcal{W}} &= \max_{q\in\mathcal{Q}} \min_{\lambda\geq 0} \left\{ \mathbb{E}_{\mathsf{D}_{\pi,0}}[q] \;-\; \lambda \left(\max_{h\in\mathcal{W}} \frac{1}{n}\sum_{i=1}^{n} h(x_i)\hat{\mathbf{R}}q(x_i,y_i) - \varepsilon_n \right) \right\} \\ &= \max_{q\in\mathcal{Q}} \min_{\lambda\geq 0} \min_{h\in\mathcal{W}} \left\{ \mathbb{E}_{\mathsf{D}_{\pi,0}}[q] \;-\; \lambda \left(\frac{1}{n}\sum_{i=1}^{n} h(x_i)\hat{\mathbf{R}}q(x_i,y_i) - \varepsilon_n \right) \right\} \\ &= \max_{q\in\mathcal{Q}} \min_{\omega\in\mathcal{W}_o} \left\{ \mathbb{E}_{\mathsf{D}_{\pi,0}}[q] \;-\; \frac{1}{n}\sum_{i=1}^{n} \omega(x_i)\hat{\mathbf{R}}q(x_i,y_i) + \varepsilon_n \left\|\omega\right\|_{\mathcal{W}_o} \right\},\end{aligned}$$

where we use $\omega = \lambda h(x)$, such that λ is replaced by $\|\omega\|_{\mathcal{W}_o}$. Define

$$\begin{aligned}M(q,\ \omega;\ \hat{\mathsf{D}}_n) &= \mathbb{E}_{\mathsf{D}_{\pi,0}}[q] \;-\frac{1}{n}\sum_{i=1}^{n} \omega(x_i)\hat{\mathbf{R}}q(x_i,y_i) + \varepsilon_n \left\|\omega\right\|_{\mathcal{W}_o} \\ &= \mathbb{E}_{\hat{\mathsf{D}}^{\omega}_n}[r] + \Delta(\hat{\mathsf{D}}^{\omega}_n,\ q) + \varepsilon_n \left\|\omega\right\|_{\mathcal{W}_o}.\end{aligned}$$

Then we have

$$\begin{aligned}\max_{q\in\mathcal{Q}} M(q,\ \omega;\ \hat{\mathsf{D}}_n) &= \mathbb{E}_{\hat{\mathsf{D}}^{\omega}_n}[r] + \max_{q\in\mathcal{Q}} \Delta(\hat{\mathsf{D}}^{\omega}_n,\ q) + \varepsilon_n \left\|\omega\right\|_{\mathcal{W}_o} \\ &= \mathbb{E}_{\hat{\mathsf{D}}^{\omega}_n}[r] + I_{\mathcal{Q}}(\omega;\ \hat{\mathsf{D}}_n) + \varepsilon_n \left\|\omega\right\|_{\mathcal{W}_o} \\ &= \hat{F}^{+}_{\mathcal{Q}}(\omega).\end{aligned}$$

Therefore,

$$\begin{aligned}\hat{J}^{+}_{\mathcal{Q},\mathcal{W}} &= \max_{q\in\mathcal{Q}} \min_{\omega\in\mathcal{W}_o} M(q,\ \omega;\ \hat{\mathsf{D}}_n) \\ &\leq \min_{\omega\in\mathcal{W}_o} \max_{q\in\mathcal{Q}} M(q,\ \omega;\ \hat{\mathsf{D}}_n) \\ &= \min_{\omega\in\mathcal{W}_o} \hat{F}^{+}_{\mathcal{Q}}(\omega).\end{aligned}$$

The lower bound follows analogously. The strong duality holds when the Slater's condition is satisfied, which amounts to saying that the primal problem in (8) is convex and strictly feasible, which requires that $\mathcal{Q}$ is convex and there exists at least one solution $q \in \mathcal{Q}$ that satisfy that constraint strictly, that is, $L_{\mathcal{W}}(q;\ \hat{\mathsf{D}}_n) < \varepsilon_n$; note that the objective function $\mathcal{Q}$ is linear on q and the constraint function $L_{\mathcal{W}}(q;\ \hat{\mathsf{D}}_n)$ is always convex on q (since it is the sup a set of linear functions on q following (3)).

□

B PROOF OF CONCENTRATION BOUND IN THEOREM 4.1

Our proof require the following Hoeffding inequality on Hilbert spaces by Pinelis (Theorem 3, 1992); see also Section 2.4 of Rosasco et al. (2010).

Lemma B.1. *(Theorem 3, Pinelis, 1992) Let $\mathcal{H}$ be a Hilbert space and $\{f_i\}_{i=1}^n$ is a Martingale sequence in $\mathcal{H}$ that satisfies $\sup_i \|f_i\|_{\mathcal{H}} \leq \sigma$ almost surely. We have for any $\epsilon > 0$,*

$$\Pr\left(\left\| \frac{1}{n}\sum_{i=1}^{n} f_i \right\|_{\mathcal{H}} \geq \epsilon \right) \leq 2\exp\left(-\frac{\epsilon^2}{2\sigma^2} \right).$$

Therefore, with probability at least $1-\delta$, we have $\left\|\frac{1}{n}\sum_{i=1}^{n} f_i\right\|_{\mathcal{H}} \leq \sqrt{\frac{2\sigma^2\log(2/\delta)}{n}}$.

Lemma B.2. *Let $k(x,x')$ be a positive definite kernel whose RKHS is $\mathcal{H}_k$. Define*

$$f_i(\cdot) = \hat{\mathbf{R}}q(x_i,y_i)k(x_i,\cdot) - \mathbf{R}_{\pi}q(x_i)k(x_i,\cdot).$$

Then $\{f_i\}_{i=1}^n$ is a Martingale difference sequence in $\mathcal{H}_k$ w.r.t. $T_{<i} := (x_j, y_j)_{j<i} \cup (x_i)$. That is, $\mathbb{E}\left[f_{i+1}(\cdot) \mid T_{<i}\right] = 0$. In addition,

$$\left\| \frac{1}{n} \sum_{i=1}^n f_i \right\|_{\mathcal{H}_k}^2 = \frac{1}{n^2} \sum_{ij=1}^n \left(\hat{\mathbf{R}} q(x_i, y_i) - \mathbf{R}_\pi q(x_i) \right) k(x_i, x_j) \left(\hat{\mathbf{R}} q(x_j, y_j) - \mathbf{R}_\pi q(x_j) \right),$$

and $\|f_i\|_{\mathcal{H}_k}^2 \leq c_{q,k}$ for $\forall i = 1, \ldots, n$.

Proof of Theorem 4.1. Following Lemma B.1 and (B.2), since $\{f_i\}_{i=1}^n$ is a Martingale difference sequence in $\mathcal{H}_k$ with $\|f_i\|_{\mathcal{H}_k} \leq c_{q,k}$ almost surely, we have with probability at least $1 - \delta$,

$$\frac{1}{n^2} \sum_{ij=1}^n \left(\hat{\mathbf{R}} q(x_i, y_i) - \mathbf{R}_\pi q(x_i) \right) k(x_i, x_j) \left(\hat{\mathbf{R}} q(x_j, y_j) - \mathbf{R}_\pi q(x_j) \right) = \left\| \frac{1}{n} \sum_{i=1}^n f_i \right\|_{\mathcal{H}_k}^2 \leq \frac{2 c_{q,k} \log(2/\delta)}{n}.$$

Using Lemma B.3 below, we have

$$\left| L_{\mathcal{K}}(q;\ \hat{\mathsf{D}}_n) - L_{\mathcal{K}}^*(q;\ \hat{\mathsf{D}}_n) \right| \leq \left\| \frac{1}{n} \sum_{i=1}^n f_i \right\|_{\mathcal{H}_k} \leq \sqrt{\frac{2 c_{q,k} \log(2/\delta)}{n}}.$$

This completes the proof. □

Lemma B.3. *Assume $k(x, x')$ is a positive definite kernel. We have*

$$\left| L_{\mathcal{K}}(q;\ \hat{\mathsf{D}}_n) - L_{\mathcal{K}}^*(q;\ \hat{\mathsf{D}}_n) \right|^2 \leq \left(\hat{\mathbf{R}} q(x_i, y_i) - \mathbf{R}_\pi q(x_i) \right) k(x_i, x_j) \left(\hat{\mathbf{R}} q(x_j, y_j) - \mathbf{R}_\pi q(x_j) \right).$$

Proof. Define

$$\hat{g}(\cdot) = \frac{1}{n} \sum_{i=1}^n \hat{\mathbf{R}} q(x_i, y_i) k(x_i, \cdot) \qquad g(\cdot) = \frac{1}{n} \sum_{i=1}^n \mathbf{R}_\pi q(x_i) k(x_i, \cdot).$$

Then we have

$$\|\hat{g}\|_{\mathcal{H}_k}^2 = \frac{1}{n^2} \sum_{ij=1}^n \hat{\mathbf{R}} q(x_i, y_i) k(x_i, x_j) \hat{\mathbf{R}} q(x_i, y_i) = \hat{L}_{\mathcal{K}}(q;\ \hat{\mathsf{D}}_n),$$

$$\|g\|_{\mathcal{H}_k}^2 = \frac{1}{n^2} \sum_{ij=1}^n \mathbf{R}_\pi q(x_i) k(x_i, x_j) \mathbf{R}_\pi q(x_i) = L_{\mathcal{K}}^*(q;\ \hat{\mathsf{D}}_n),$$

$$\|\hat{g} - g\|_{\mathcal{H}_k}^2 = \frac{1}{n^2} \sum_{ij=1}^n \left(\hat{\mathbf{R}} q(x_i, y_i) - \mathbf{R}_\pi q(x_i) \right) k(x_i, x_j) \left(\hat{\mathbf{R}} q(x_j, y_j) - \mathbf{R}_\pi q(x_j) \right).$$

The result then follows the triangle inequality $\left| \|\hat{g}\|_{\mathcal{H}_k} - \|g\|_{\mathcal{H}_k} \right| \leq \|\hat{g} - g\|_{\mathcal{H}_k}$. □

C More on the Tightness of the Confidence Interval

The benefit of having both upper and lower bounds is that we can empirically access the tightness of the bound by checking the length of the interval $[\hat{F}_{\mathcal{Q}}^-(\omega_-), \hat{F}_{\mathcal{Q}}^+(\omega_+)]$. However, from the theoretical perspective, it is desirable to know *a priori* that the length of the interval will decrease with a fast rate as the data size n increases. We now show that this is the case if $\mathcal{W}_o$ is chosen to be sufficiently rich so that it includes a $\omega \in \mathcal{W}_o$ such that $\hat{\mathsf{D}}_n^\omega \approx \mathsf{D}_\pi$.

Theorem C.1. *Assume $\mathcal{W}_o$ is sufficiently rich to include a "good" ω^* in $\mathcal{W}_o$ with $\hat{\mathsf{D}}_n^{\omega^*} \approx \mathsf{D}_\pi$ in that*

$$\sup_{q \in \mathcal{Q}} \left| \mathbb{E}_{\hat{\mathsf{D}}_n^{\omega^*}} \left[\hat{\mathbf{R}} q(x; x', r) \right] - \mathbb{E}_{\mathsf{D}_\pi} \left[\hat{\mathbf{R}} q(x; x', r) \right] \right| \leq \frac{c}{n^\alpha}, \tag{19}$$

where c and α are two positive coefficients. Then we have

$$\max \left\{ \hat{J}_{\mathcal{Q},\mathcal{W}}^+ - J_\pi,\ \ J_\pi - \hat{J}_{\mathcal{Q},\mathcal{W}}^- \right\} \leq \frac{c}{n^\alpha} + \varepsilon_n \|\omega^*\|_{\mathcal{W}_o}.$$

Assumption (19) holds if $\hat{\mathsf{D}}_n$ weakly converges to some limit discussion $\hat{\mathsf{D}}_\infty$ whose support is $\mathcal{X}$, for which we can define $\omega^*(x) = \mathsf{D}_\pi(x)/\mathsf{D}_\infty(x)$. If $\mathcal{Q}$ is a finite ball in RKHS, then we can achieve (19) with $\alpha = 1/2$, and yields the overall bound of rate $O(n^{-1/2})$. For more general function classes, α depends on the martingale Rademacher complexity of function set $\hat{\mathbf{R}}\mathcal{Q} = \{\mathbf{R}q(x,y) \colon q \in \mathcal{Q}\}$ Rakhlin et al. (2015). In our empirical reults, we observe that the gap of the practically constructed bounds tend to follow the $O(n^{-1/2})$ rate.

Proof. Note that

$$J_\pi = \mathbb{E}_{\mathsf{D}_\pi}[r] = \mathbb{E}_{\mathsf{D}_\pi}[r],$$

and

$$I_{\mathcal{Q}}(\omega;\ \hat{\mathsf{D}}_n) = \sup_{q\in\mathcal{Q}} \left\{ \mathbb{E}_{\hat{\mathsf{D}}_n^\omega}[\gamma q(x') - q(x)] - \mathbb{E}_{\mathsf{D}_\pi}[\gamma q(x') - q(x)] \right\}.$$

Because $\omega^* \in \mathcal{W}$, we have

$$\begin{aligned}
\hat{J}^+_{\mathcal{W},\mathcal{Q}} - J_\pi &\le \hat{F}^+_{\mathcal{Q}}(\omega^*) - J_\pi \\
&= \mathbb{E}_{\hat{\mathsf{D}}_n^\omega}[r] - \mathbb{E}_{\mathsf{D}_\pi}[r] + I_{\mathcal{Q}}(\omega_\pi;\ \hat{\mathsf{D}}_n) + \varepsilon_n \|\omega^*\|_{\mathcal{W}_o} \\
&= \sup_{q\in\mathcal{Q}} \left\{ \mathbb{E}_{\hat{\mathsf{D}}_n^\omega}\left[\hat{\mathbf{R}}q(x,y)\right] - \mathbb{E}_{\mathsf{D}_\pi}\left[\hat{\mathbf{R}}q(x,y)\right] \right\} + \varepsilon_n \|\omega^*\|_{\mathcal{W}_o} \\
&\le \frac{c}{n^\alpha} + \varepsilon_n \|\omega^*\|_{\mathcal{W}_o}.
\end{aligned}$$

The case of lower bound follows similarly. □

D Optimization on $\mathcal{W}_o$

Consider the optimization of ω in $\mathcal{W}_o$

$$\hat{F}^+_{\mathcal{Q}}(\omega) := \frac{1}{n}\sum_{i=1}^n r_i\omega(x_i) + I_{\mathcal{Q}}(\omega;\ \hat{\mathsf{D}}_n) + \|\omega\|_{\mathcal{W}_o}\sqrt{\frac{2c_{q_\pi,k}\log(2/\delta)}{n}} \tag{20}$$

Assume $\mathcal{W}_o$ is the RKHS of kernel $k(x,\bar{x})$, that is, $\mathcal{W}_o = \mathcal{H}_k$. By the finite representer theorem of RKHS (Smola et al., 2007). the optimization of ω in RKHS $\mathcal{H}_k$ can be reduced to a finite dimensional optimization problem. Specifically, the optimal solution of (20) can be written into a form of $\omega(x) = \sum_{i=1}^n k(x,x_i)\alpha_i$ with $\|\omega\|^2_{\mathcal{H}_k} = \sum_{i,j=1}^n k(x_i,x_j)\alpha_i\alpha_j$ for some vector $\boldsymbol{\alpha} := [\alpha_i]_{i=1}^n \in \mathbb{R}^n$. Write $\boldsymbol{K} = [k(x_i,x_j)]_{i,j=1}^n$ and $\boldsymbol{r} = [r_i]_{i=1}^n$. The optimization of ω reduces to a finite dimensional optimization on $\boldsymbol{\alpha}$:

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^n} \frac{1}{n}\boldsymbol{r}^\top\boldsymbol{K}\boldsymbol{\alpha} + I_{\mathcal{Q}}(\boldsymbol{K}\boldsymbol{\alpha};\ \hat{\mathsf{D}}_n) + \sqrt{\boldsymbol{\alpha}\boldsymbol{K}\boldsymbol{\alpha}}\sqrt{\frac{2c_{q_\pi,k}\log(2/\delta)}{n}},$$

where

$$I_{\mathcal{Q}}(\boldsymbol{K}\boldsymbol{\alpha};\ \hat{\mathsf{D}}_n) = \max_{q\in\mathcal{Q}} \left\{ \mathbb{E}_{\mathsf{D}_{\pi,0}}[q] + \frac{1}{n}(\hat{\mathbf{T}}q)^\top\boldsymbol{K}\boldsymbol{\alpha} \right\},$$

and $\hat{\mathbf{T}}q = [\gamma q(x_i') - q(x_i)]_{i=1}^n$. When $\mathcal{Q}$ is RKHS, we can calculate $I_{\mathcal{Q}}(\boldsymbol{K}\boldsymbol{\alpha};\ \hat{\mathsf{D}}_n)$ using (22) in section F.

This computation can be still expensive when n is large. Fortunately, our confidence bound holds for any ω; better ω only gives tighter bounds, but it is not necessary to find the global optimal ω. Therefore, one can use any approximation algorithm to find ω, which provides a trade-off of tightness and computational cost. We discuss two methods:

1) Approximating $\boldsymbol{\alpha}$ The length of $\boldsymbol{\alpha}$ can be too large when n is large. To address this, we assume $\alpha_i = g(x_i,\ \theta)$, where g is any parametric function (such as a neural network) with parameter θ which can be much lower dimensional than $\boldsymbol{\alpha}$. We can then optimize θ with stochastic gradient descent, by approximating all the data averaging $\frac{1}{n}\sum_{i=1}^n(\cdot)$ with averages over small mini-batches; this would introduce biases in gradient estimation, but it is not an issue when the goal is only to get a reasonable approximation.

2) Replacing kernel k Assume the kernel k yields a random feature expansion: $k(x,\bar{x}) = \mathbb{E}_{\beta\sim\pi}[\phi(x,\beta)\phi(\bar{x},\beta)]$, where $\phi(x,\beta)$ is a feature map with parameter β and π is a distribution of β. We draw $\{\beta_i\}_{i=1}^m$ i.i.d. from π, where m is taken to be much smaller than n. We replace k with $\hat{k}(x,\bar{x}) = \frac{1}{m}\sum_{i=1}^m \phi(x,\beta_i)\phi(\bar{x},\beta_i)$ and $\mathcal{H}_k$ with $\mathcal{H}_{\hat{k}}$, That is, we consider to solve

$$\hat{J}^+_{\mathcal{Q},\mathcal{W}} = \min_{\omega\in\mathcal{H}_{\hat{k}}} \left\{ \hat{F}^+_{\mathcal{Q}}(\omega) := \frac{1}{n}\sum_{i=1}^n r_i\omega(x_i) + I_{\mathcal{Q}}(\omega;\ \hat{\mathsf{D}}_n) + \|\omega\|_{\mathcal{H}_{\hat{k}}} \sqrt{\frac{2c_{q_\pi,\hat{k}}\log(2/\delta)}{n}} \right\}.$$

It is known that any function ω in $\mathcal{H}_{\hat{k}}$ can be represented as $\omega(x) = \frac{1}{m}\sum_{i=1}^m w_i\phi(x,\beta_i)$, for some $\boldsymbol{w} = [w_i]_{i=1}^m \in \mathbb{R}^m$ and satisfies $\|\omega\|^2_{\mathcal{H}_{\hat{k}}} = \frac{1}{m}\sum_{i=1}^m w_i^2$. In this way, the problem reduces to optimizing an m-dimensional vector $\boldsymbol{w}$, which can be solved by standard convex optimization techniques.

E Concentration Inequalities of General Functional Bellman Losses

When $\mathcal{K}$ is a general function set, one can still obtain a general concentration bound using Rademacher complexity. Define $\hat{\mathbf{R}}q\circ\mathcal{W} := \{h(x,y) = \hat{\mathbf{R}}q(x,y)\omega(x)\colon \omega\in\mathcal{W}\}$. Using the standard derivation in Radermacher complexity theory in conjunction with Martingale theory (Rakhlin et al., 2015), we have

$$\sup_{\omega\in\mathcal{W}} \left\{ \frac{1}{n}\sum_{i=1}^n (\hat{\mathbf{R}}q(x_i,y_i) - \mathbf{R}_\pi q(x_i))\omega(x_i) \right\} \le 2Rad(\hat{\mathbf{R}}q\circ\mathcal{W}) + \sqrt{\frac{2c_q\log(2/\delta)}{n}},$$

where $Rad(\hat{\mathbf{R}}q\circ\mathcal{K})$ is the sequential Radermacher complexity as defined in (Rakhlin et al., 2015). A triangle inequality yields

$$|\ L_k(q;\ \hat{\mathsf{D}}_n) - L_k(q;\ \hat{\mathsf{D}}_n)\ | \le \sup_{\omega\in\mathcal{W}} \left\{ \frac{1}{n}\sum_{i=1}^n (\hat{\mathbf{R}}q(x_i,y_i) - \mathbf{R}_\pi q(x_i))\omega(x_i) \right\}$$

Therefore,

$$|\ L_{\mathcal{W}}(q;\ \hat{\mathsf{D}}_n) - L_{\mathcal{W}}(q;\ \hat{\mathsf{D}}_n)\ | \le 2Rad(\hat{\mathbf{R}}q\circ\mathcal{W}) + \sqrt{\frac{2c_q\log(2/\delta)}{n}}, \tag{21}$$

where $c_{q,\mathcal{W}} = \sup_{\omega\in\mathcal{W}}\sup_{x,y}(\hat{\mathbf{R}}q(x,y) - \mathbf{R}_\pi q(x))^2\omega(x)^2$. When $\mathcal{W}$ equals the unit ball $\mathcal{K}$ of the RKHS related to kernel k, we have $c_{q,k} = c_{q,\mathcal{W}}$, and hence this bound is strictly worse than the bound in Theorem 4.1.

F Closed Form of $I_{\mathcal{Q}}(\omega;\ \hat{\mathsf{D}}_n)$ when $\mathcal{Q}$ is RKHS

Similar to $L_{\mathcal{K}}(q;\ \hat{\mathsf{D}}_n)$, when $\mathcal{Q}$ is taken to be the unit ball $\tilde{\mathcal{K}}$ of the RKHS of a positive definite kernel $\tilde{k}(x,\bar{x})$, (7) can be expressed into a bilinear closed form shown in Mousavi et al. (2020):

$$I_{\mathcal{Q}}(\omega;\ \hat{\mathsf{D}}_n)^2 = A - 2B + C, \tag{22}$$

where

$$\begin{aligned} A &= \mathbb{E}_{(x,\bar{x})\sim\mathsf{D}_{\pi,0}\times\mathsf{D}_{\pi,0}}\left[k(x,\bar{x})\right] \\ B &= \mathbb{E}_{(x,\bar{x})\sim\hat{\mathsf{D}}^\omega_n\times\mathsf{D}_{\pi,0}}\left[\hat{\mathbf{T}}^x_\pi k(x,\bar{x})\right] \\ C &= \mathbb{E}_{(x,\bar{x})\sim\hat{\mathsf{D}}^\omega_n\times\hat{\mathsf{D}}^\omega_n}\left[\hat{\mathbf{T}}^x_\pi\hat{\mathbf{T}}^{\bar{x}}_\pi k(x,\bar{x})\right], \end{aligned}$$

were $\hat{\mathbf{T}}_\pi f(x) = \gamma f(x') - f(x)$; in $\hat{\mathbf{T}}^x_\pi\hat{\mathbf{T}}^{\bar{x}}_\pi k(x,\bar{x})$, we apply $\hat{\mathbf{T}}^{\bar{x}}_\pi$ and $\hat{\mathbf{T}}^x_\pi$ in a sequential order by treating k as a function of $\bar{x}$ and then of x.

G MORE ON THE ORACLE BOUND AND ITS DUAL FORM

The oracle bound (18) provides another starting point for deriving optimization-based confidence bounds. We derive its due form here. Using Lagrangian multiplier, the optimization in (18) can be rewritten into

$$\hat{J}^{+}_{\mathcal{Q},*} = \max_{q\in\mathcal{Q}} \min_{\omega} M(q,\omega;\ \hat{\mathsf{D}}_n), \tag{23}$$

$$\text{where} \quad M_*(q,\omega;\ \hat{\mathsf{D}}_n) = \mathbb{E}_{\mathsf{D}_{\pi,0}}[q] - \frac{1}{n}\sum_{i=1}^{n}\omega(x_i)\left(\hat{\mathbf{R}}q(x_i,y_i) - \hat{\mathbf{R}}q_\pi(x_i,y_i)\right),$$

where ω now serves as the Lagrangian multiplier. By the weak duality, we have

$$J^*_{\mathcal{Q},+} \le \hat{F}^{+}_{\mathcal{Q},*}(\omega) := \underbrace{\mathbb{E}_{\hat{\mathsf{D}}^{\omega}_n}[r] + I_{\mathcal{Q}}(\omega;\ \hat{\mathsf{D}}_n)}_{known} + \underbrace{R(\omega,\ q_\pi)}_{unknown}, \quad \forall\omega.$$

and

$$R(\omega, q_\pi) = \frac{1}{n}\sum_{i=1}^{n}\omega(x_i)\hat{\mathbf{R}}q_\pi(x_i).$$

The derivation follows similarly for the lower bound. So for any $\omega \in \mathcal{W}_o$, we have $[\hat{J}^{-}_{\mathcal{Q},*},\ \ \hat{J}^{+}_{\mathcal{Q},*}] \subseteq [\hat{F}^{-}_{\mathcal{Q},*}(\omega),\ \ \hat{F}^{+}_{\mathcal{Q},*}(\omega)]$.

Here the first two terms of $\hat{F}^{+}_{\mathcal{Q},*}(\omega)$ can be empirically estimated (it is the same as the first two terms of (16)), but the third term $R(\omega, q_\pi)$ depends on the unknown q_π and hence need to be further upper bounded.

Our method can be viewed as constraining ω in $\mathcal{W}$, which is assumed to be the unit ball of $\mathcal{W}_o$, and applying a worst case bound:

$$\begin{aligned}
\hat{F}^{+}_{\mathcal{Q},*}(\omega) &:= \mathbb{E}_{\hat{\mathsf{D}}^{\omega}_n}[r] + I_{\mathcal{Q}}(\omega;\ \hat{\mathsf{D}}_n) + R(\omega,\ q_\pi), \quad \forall\omega\in\mathcal{W}_o \\
&\le \mathbb{E}_{\hat{\mathsf{D}}^{\omega}_n}[r] + I_{\mathcal{Q}}(\omega;\ \hat{\mathsf{D}}_n) + \|w\|_{\mathcal{W}_o}\sup_{h\in\mathcal{W}} R(h,\ q_\pi), \quad \forall\omega\in\mathcal{W}_o \\
&\le \mathbb{E}_{\hat{\mathsf{D}}^{\omega}_n}[r] + I_{\mathcal{Q}}(\omega;\ \hat{\mathsf{D}}_n) + \|w\|_{\mathcal{W}_o} L_{\mathcal{W}}(q_\pi, \hat{\mathsf{D}}_n), \quad \forall\omega\in\mathcal{W}_o \\
&\overset{w.p.1-\delta}{\le} \mathbb{E}_{\hat{\mathsf{D}}^{\omega}_n}[r] + I_{\mathcal{Q}}(\omega;\ \hat{\mathsf{D}}_n) + \epsilon\,\|w\|_{\mathcal{W}_o}, \quad \forall\omega\in\mathcal{W}_o \\
&= \hat{F}^{+}_{\mathcal{Q}}(\omega).
\end{aligned}$$

where the last step applies the high probability bound that $\Pr(L_{\mathcal{W}}(q_\pi, \hat{\mathsf{D}}_n) \le \varepsilon) \ge 1-\delta$. Similar derivation on the lower bound counterpart gives

$$\Pr\left(\left[\hat{F}^{+}_{\mathcal{Q},*}(\omega),\ \ \hat{F}^{+}_{\mathcal{Q},*}(\omega)\right] \subseteq \left[\hat{F}^{-}_{\mathcal{Q}}(\omega), \hat{F}^{+}_{\mathcal{Q}}(\omega)\right]\right) \ge 1-\delta.$$

Therefore, our confidence bound $[\hat{F}^{-}_{\mathcal{Q}}(\omega), \hat{F}^{+}_{\mathcal{Q}}(\omega)]$ is a $1-\delta$ confidence outer bound of the oracle bound $[\hat{J}^{-}_{\mathcal{Q},*},\ \ \hat{J}^{+}_{\mathcal{Q},*}] \subseteq [\hat{F}^{-}_{\mathcal{Q},*}(\omega),\ \ \hat{F}^{+}_{\mathcal{Q},*}(\omega)]$.

Introducing $\mathcal{Q}$ is necessarily Our method does not require any independence assumption between the transition pairs, the trade-off is that that we have to assume that q_π falls into a function set $\mathcal{Q}$ that imposes certain smoothness assumption. This is necessary because the data only provide information regarding q_π on a finite number of points, and q_π can be arbitrarily non-smooth outside of the data points, and hence no reasonable upper/lower bound can be obtained without any smoothness condition that extend the information on the data points to other points in the domain.

Proposition G.1. *Unless* $\Pr_{s\sim\mathsf{D}_{\pi,0}}(s \notin \{s_i, s_i'\}_{i=1}^n) = 0$*, for any* $u \in \mathbb{R}$*, there exists a function* $q\colon \mathcal{S}\times\mathcal{A}\to\mathbb{R}$*, such that*

$$\mathbb{E}_{\mathsf{D}_{\pi,0}}[q] = u, \quad \hat{\mathbf{R}}q(x_i,y_i) = \hat{\mathbf{R}}q_\pi(x_i,y_i),\ \ \forall i = 1,\dots,n.$$

Proof. Let $\mathcal{Q}_{\text{null}}$ be the set of functions that are zero on $\{s_i, s_i'\}_{i=1}^n$, that is,

$$\mathcal{Q}_{\text{null}} = \{g \colon \mathcal{S} \times \mathcal{A} \to \mathbb{R} \colon \quad g(s,a) = 0, \ \ \forall s \in \{s_i, s_i'\}_{i=1}^n, \quad a \in \mathcal{A}\}.$$

Then we have

$$\hat{\mathbf{R}}_\pi(q_\pi + g)(x_i, y_i) = \hat{\mathbf{R}}_\pi q_\pi(x_i, y_i), \qquad \forall i = 1, \ldots, n.$$

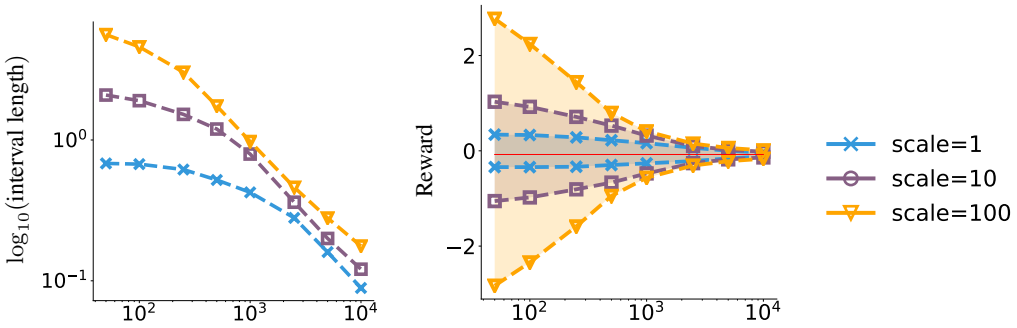
and

$$\mathbb{E}_{\mathsf{D}_{\pi,0}}[q_\pi + g] = \mathbb{E}_{\mathsf{D}_{\pi},0}[q_\pi] + \mathbb{E}_{\mathsf{D}_{\pi,0}}[g] = J_\pi + \mathbb{E}_{\mathsf{D}_{\pi,0}}[g].$$

Taking $g(s,a) = z\mathbb{I}(s \notin \{s_i, s_i'\}_{i=1}^n)$, where z is any real number. Then we have

$$\mathbb{E}_{\mathsf{D}_{\pi,0}}[q_\pi + g] = J_\pi + z\mathrm{Pr}_{s \sim \mathsf{D}_{\pi,0}}(s \notin \{s_i, s_i'\}_{i=1}^n).$$

Because $\mathrm{Pr}_{s \sim \mathsf{D}_{\pi,0}}(s \notin \{s_i, s_i'\}_{i=1}^n). \neq 0$, we can take z to be arbitrary value to make $\mathbb{E}_{\mathsf{D}_{\pi,0}}[q_\pi + g]$ to take arbitrary value. $\square$



(a) Number of transitions, n (b) Number of transitions n

Figure 3: Ablation study on the radius $r_{\mathcal{Q}}$ of the function class $\mathcal{Q}$.

H Ablation Study and Experimental Details

H.1 Ablation Study

We investigate the sensitivity of hyper-parameters in our algorithm. The hyper-parameter mainly depends on how we choose our function class $\mathcal{Q}$ and $\mathcal{W}$. We also study the changing of different behavior policy (or $\hat{\mathsf{D}}_n$) will affect the performance.

Radius of $\mathcal{Q}$ In practice, we choose $\mathcal{Q}$ to be an RKHS with norm $r_{\mathcal{Q}}$, or more concretely, we let

$$\mathcal{Q} = r_{\mathcal{Q}}\tilde{\mathcal{K}} = \{r_{\mathcal{Q}} f : f \in \tilde{\mathcal{K}}\},$$

where $\tilde{\mathcal{K}}$ is the unit ball RKHS. Ideally, we hope $r_{\mathcal{Q}} \geq \|q_\pi\|_{\tilde{\mathcal{K}}}$ so that $q_\pi \in \mathcal{Q}$.

Since it is hard to analyze the behavior of the algorithm when q_π is unknown, we consider a synthetic environment where the true q_π is known. This is done by enforcing a deterministic reward function r for this environment as the reverse Bellman error of q_π:

$$r(x) = q_\pi(x) - \gamma \mathbb{E}_{x' \sim \mathsf{P}_\pi(\cdot|x)}[q_\pi(x')].$$

In this way we can get access to the true RKHS norm of q_π:

$$\rho^* = \|q_\pi\|_{\tilde{\mathcal{K}}}.$$

For simplicity, we set the state and action both lies in a one dimension real number and set a Gaussian policy similar as all other settings.

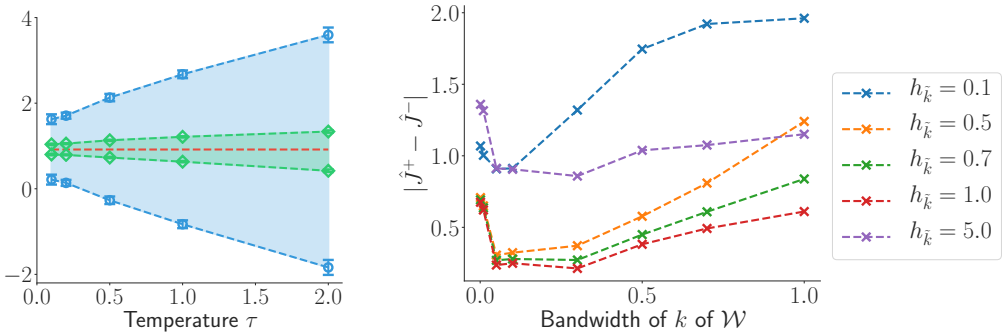
Figure 3 shows three different curves that set $r_{\mathcal{Q}}$ to be different scale of ρ^*, with $r_{\mathcal{Q}}$ equals to ρ^*, $10\rho^*$ and $100\rho^*$ respectively. We can see from the curve that the tightness of the bound is affected by different scale of radius when the number of samples is very small. However, as the number of samples grow (e.g., greater than 2000 in our experiment) the length of the bounds become less sensitive to the changing of the predefined norm of $\mathcal{Q}$.

Similarity between Behavior policy and Target Policy We study the performance of changing temperature of the behavior policy. Not surprisingly, we can see that the more close the behavior policy to the target policy(with temperature $\tau = 0.1$), the tighter our confidence interval will be, which is observed in Figure 4 (a).

Bandwidth of RBF kernels We study the sensitivity by changing different bandwidth in kernel k and $\tilde{k}$ for $\mathcal{W}$ and $\mathcal{Q}$ respectively. We output the length of the gap in Inverted-Pendulum environment in different combination of bandwidth pair in Figure 4(b). As we can see that as long as we set the bandwidth in a reasonable region (e.g., we set the bandwidth of k in $[0.1, 0.5]$, the bandwidth of $\tilde{k}$ in $[0.5, 3]$), we can always get relatively tighter confidence bounds.

H.2 Experimental Details

Environments and Dataset Construction We test our method on three environments: Inverted-Pendulum and CartPole from OpenAI Gym (Brockman et al., 2016), and a Type-1 Diabetes



(a) Ablation study when we changing τ

(b) Ablation study on the choice of bandwidth.

Figure 4: Ablation study on InvertedPendulum. Fig (a): Confidence bounds when we change the temperature τ of the behavior policies. Fig (b): The lengths of our confidence interval when we change both the bandwidth of the kernels of $\mathcal{Q}$ and $\mathcal{W}$. Here $h_{\tilde{k}}$ denotes the bandwidth of kernel $\tilde{k}$ in $\mathcal{Q}$.

medical treatment simulator. For Inverted-Pendulum we discretize the action space to be $\{-1, -0.3, -0.2, 0, 0.2, 0.3, 1\}$.

We follow a similar setup as Feng et al. (2020) to construct behavior and target policies. For Inverted-Pendulum and CartPole, we constraint our policy class to be a softmax policy and use PPO (Schulman et al., 2017) to train a good policy π and set it as our target policy. For Diabetes simulator, we use PPO to train a Gaussian policy(deterministic action function plus a Gaussian noise) as our target policy π. We fix $\gamma = 0.95$ and set horizon length $H = 50$ for InvertedPendulum, $H = 100$ for CartPole, and $H = 50$ for Diabetes simulator.

Policy Construction We use the open source implementation[2] of deep Q-learning to train a 32×32 MLP parameterized Q-function to converge. We then use the softmax policy of the learned Q-function with different temperatures as policies. We set a default temperature $\tau = 0.1$ (to make it more deterministic) for the target policy π. For behavior policies, we set the temperature $\tau = 1$ as default. We also study the performance of our method under behavior policies with different temperatures to demonstrate the effectiveness of our method under behavior agnostic settings.

Algorithm Settings We test the bound in (16)-(17). Throughout the experiment, we always set $\mathcal{W} = \mathcal{K}$, the unit ball of RKHS with kernel $k(\cdot, \cdot)$. We set $\mathcal{Q} = r_{\mathcal{Q}}\tilde{\mathcal{K}}$, the zero-centered ball of radius $r_{\mathcal{Q}}$ in an RKHS with kernel $\tilde{k}(\cdot, \cdot)$. We take both k and $\tilde{k}$ to be Gaussian RBF kernel. The bandwidth of k and $\tilde{k}$ are selected to make sure the function Bellman loss is not large on a validation set. The radius is selected to be sufficiently large to ensure that q_π is included in $\mathcal{Q}$. To ensure a sufficiently large radius, we use the data to approximate a $\hat{q}$ so that its functional Bellman loss is small than ϵ_n. Then we set $r_{\mathcal{Q}} = 10 * \|\hat{q}\|_{\tilde{\mathcal{K}}}$. We optimize ω using the random feature approximation method described in Appendix D. Once ω_+ and ω_- is found, we evaluate the bound in (16) exactly, to ensure the theoretical guarantee holds.

[2]https://github.com/openai/baselines.

I MORE RELATED WORKS

Trajectory-Based Importance Sampling A classical approach for high confidence off-policy estimation is based on importance sampling and concentration inequalities (Thomas et al., 2015a;b). However, it requires strong assumption on the data and does not work well on infinite-horizon problems due to the explosion of high variance and the dependency between transition pairs. Specifically, IS requires to collect a set of independent trajectories $\{s_t^i, a_t^i, r_t^i\}_{t=1}^T$, $i = 1, \ldots, m$, upto a finite horizon length T, from unrolling a known policy π_0 (called behavior policy), and estimate J_π by a weighted average

$$\hat{J}^{IS} = \frac{1}{mT} \sum_{i=1}^{m} \sum_{t=1}^{T} \omega_t^m r_t^m, \quad \text{where} \quad \omega_t^m = \prod_{\tau \leq t} \frac{\pi(a_\tau^m | s_\tau^m)}{\pi_0(a_\tau^m | s_\tau^m)}. \tag{24}$$

Since the importance weights are product over time, it tends to cause an explosion in variance. One can construct non-asymptotic confidence bounds using standard concentration inequalities on (24). However, due to the dependency between transition pairs inside each trajectory, the concentration bounds can only be applied on the trajectory and decay with a $O(1/\sqrt{m})$ rate. There seems to exist no efficient algorithms to construct non-asymptotic IS-based confidence bounds that decay with the number $n = mT$ of transitions pairs, rather than the number of independent trajectories.

Asymptotic Bootstrapping Bounds Another types of confidence bounds approaches leverage the statistical procedure of bootstrapping (e.g., White & White, 2010; Hanna et al., 2017; Kostrikov & Nachum, 2020). A main drawback of standard bootstrapping methods is the need of i.i.d. assumption on data which is hard to use in transition level, where the transition data is often draw from a highly dependent Markovian structure. White & White (2010) considers moving block bootstrap methods to adapt the Markovian dependent data, however, their methods assume Q-function lies in a linear function class which is hard to capture the true q_π. Moreover, the estimation of bootstrapping is time consuming due to it requires a large number of bootstrap resamples.

Behavior-agnostic OPE We consider the behavior-agnostic OPE which is studied in the recent literature(e.g., Nachum et al., 2019a; Zhang et al., 2020a; Nachum et al., 2019b; Mousavi et al., 2020; Zhang et al., 2020b). However, most of the existing works consider point estimation, which is foundmentally different from our goal in this paper. It is worth to note that Mousavi et al. (2020) gives a theoretical confidence bounds betwen the point estimation and the ground truth, with the requirement of calculating Rademacher complexity and the i.i.d assumption on the transition data.

Duality between Q-function and Density Function Our method is a natural extension of Feng et al. (2020), which inspired by a line of recent works focusing on the duality relation between value and density ratio functions (e.g., Tang et al., 2020; Uehara et al., 2020; Jiang & Huang, 2020). Moreover, Jiang & Huang (2020) can be viewed as a special case of our method, where they assume that we can obtain infinite number of data and do not take the randomness of the data into account.