Appendix

A. Proof Concentration Bounds

The proof of concentrations bounds of U/V-statistics are standard topics in probability and statistics. We provide proof here for completeness. In addition, we show that the concentration bounds still hold in non-i.i.d. cases when $Q = Q^\pi$ due to a special martingale structure from Bellman equation.

A.1. Proof of Concentration Bound for U-statistics

Assume X is a random variable supported on $\mathcal{X}$. Given some bounded bivariate function[1] $h : \mathcal{X}^2 \to [a, b]$, the U-statistics of h is defined as:

$$U = \frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} h(X_i, X_j)\,,$$

where $X_1, X_2, \cdots, X_n \sim X$ are i.i.d. random variables. It's well known that U-statistics is an unbiased estimation for $\mathbb{E}_{Y,Z\sim X}[h(Y, Z)]$, and we include the concentration property of U-statistics for completeness.

For simplicity, we assume $n = 2k, k \in \mathbb{Z}$ when we discuss the concentration of U-statistics with Hoeffding's inequality.

Theorem A.1 (Hoeffding's Inequality for U-statistics)**.**

$$\mathbb{P}\left[|U - \mathbb{E}[h]| \ge (b-a)\sqrt{\frac{\log \frac{2}{\delta}}{2k}}\right] \le \delta\,.$$

Proof. The proof is originated from Hoeffding (1963), and we restate the original proof here for the completeness.

We first introduce the following notation:

$$V(X_1, X_2, \cdots, X_n) = \frac{1}{k} \sum_{i \in [k]} h(X_{2i-1}, X_{2i})\,. \tag{19}$$

It's easy to see that $\mathbb{E}[V] = \mathbb{E}[h]$, and

$$U = \frac{1}{n!} \sum_{\sigma \in S_n} V(X_{\sigma_1}, X_{\sigma_2}, \cdots, X_{\sigma_n})\,,$$

where S_n is the symmetric group of degree n (i.e. we take the summation over all of the permutation of set $[n]$).

With Chernoff's bound, we can know

$$\mathbb{P}\left[U \ge \delta\right] \le \exp(-\lambda\delta)\mathbb{E}[\exp(\lambda U)], \quad \forall \lambda > 0\,.$$

So we focus on the term $\mathbb{E}\left[\exp(\lambda U)\right]$. With Jensen's inequality, we have:

$$\mathbb{E}[\exp(\lambda U)] = \mathbb{E}\left[\exp\left(\frac{\lambda}{n!}\sum_{\sigma \in S_n} V\left(X_{\sigma_1}, X_{\sigma_2}, \cdots, X_{\sigma_n}\right)\right)\right] \le \frac{1}{n!}\sum_{\sigma \in S_n} \mathbb{E}\left[\exp(\lambda V(X_{\sigma_1}, X_{\sigma_2}, \cdots, X_{\sigma_n}))\right].$$

Thus,

$$\mathbb{P}[U - \mathbb{E}[h] \ge \delta] \le \frac{1}{n!}\sum_{\sigma \in S_n} \mathbb{E}\left[\exp\left(\lambda V(X_{\sigma_1}, X_{\sigma_2}, \cdots, X_{\sigma_n}) - \lambda\mathbb{E}[h] - \lambda\delta\right)\right].$$

[1] U-statistics are not limited to the bivariate functions, however, as the kernel loss we discuss in this paper is a bivariate function, we focus on the bivariate function here.

Notice that, V is sub-Gaussian with variance proxy $\sigma^2 = \frac{(b-a)^2}{4k}$. Thus, with the property of sub-Gaussian random variable,

$$\mathbb{E}\left[\exp\left(\lambda V - \lambda \mathbb{E}[h] - \lambda\delta\right)\right] \leq \exp\left(\frac{\lambda^2(b-a)^2}{8k} - \lambda\delta\right), \quad \forall \lambda > 0.$$

When $\lambda = \frac{4k\delta}{(b-a)^2}$, $\exp(\frac{\lambda^2(b-a)^2}{8k} - \lambda\delta)$ achieves the minimum $\exp(-\frac{2k\delta^2}{(b-a)^2})$. Thus,

$$\mathbb{P}\left[U - \mathbb{E}[h] \geq \delta\right] \leq \frac{1}{n!}\sum_{\sigma \in S_n} \exp\left(-\frac{2k\delta^2}{(b-a)^2}\right) = \exp\left(-\frac{2k\delta^2}{(b-a)^2}\right).$$

Moreover, with the symmetry of U, we have that:

$$\mathbb{P}[|U - \mathbb{E}[h]| \geq \delta] \leq 2\exp\left(-\frac{2k\delta^2}{(b-a)^2}\right),$$

which concludes the proof. □

A.2. Concentration Bounds for V-statistics

We have the following equation for U-statistics and V-statistics

$$\widehat{L}_K^V(Q) = \frac{n-1}{n}\widehat{L}_K^U(Q) + \sum_{i\in[n]} \ell_{\pi,Q}(\tau_i, \tau_i),$$

so we can upper bound $|\widehat{L}_K^V(Q) - L_K(Q)|$ via

$$|\widehat{L}_K^V(Q) - L_K(Q)| \leq \frac{n-1}{n}|\widehat{L}_K^U(Q) - L_K(Q)| + \left|\frac{1}{n^2}\sum_{i\in[n]} \ell_{\pi,Q}(\tau_i,\tau_i) - \frac{1}{n}L_K(Q)\right|.$$

Thus, with the concentration bounds of $\widehat{L}_K^U(Q)$, and the fact that $|\ell_{\pi,Q}(\tau_i,\tau_i)| \leq \ell_{\max}$, and $|L_K(Q)| \leq \ell_{\max}$, we have the desired result.

A.3. Concentration Bounds for Non I.I.D. Samples

In practice, the dataset $\mathcal{D} = \{x_i, r_i, s_i'\}_{1\leq i\leq n}$ can be obtained with a non i.i.d fasion (such as we collect trajectories in the MDP follow by a policy π), which violates the assumption that samples from $\mathcal{D}$ are drawn independently. There are concentration bounds for U-statistics with weakly dependent data. For example, Han (2018) considers the concentration of U-statistics when the data are generated from a Markov Chain under mixing conditions. Here we show that, in the case when $Q = Q^\pi$, the concentration inequality holds without requiring the i.i.d. or any mixing condition, thanks to a martingale structure from the Bellman equation.

Proposition A.1. *Assume the transitions are sampled from the MDP, i.e. $s' \sim \mathcal{P}(\cdot|x)$, $\bar{s}' \sim \mathcal{P}(\cdot|\bar{x})$, then for any joint measure ν of $(x,\bar{x})$, we have the following property for Q^π:*

$$\mathbb{E}_{(x,\bar{x})\sim\nu, s'\sim\mathcal{P}(\cdot|x), \bar{s}'\sim\mathcal{P}(\cdot|\bar{x})}[K(x,\bar{x}) \cdot \widehat{\mathcal{R}}_\pi Q^\pi(x) \cdot \widehat{\mathcal{R}}_\pi Q^\pi(\bar{x})] = 0,$$

For i.i.d. case, $\nu = \mu \times \mu$.

Proof. By the definition of Bellman error for Q-function, we have that

$$\mathbb{E}_{s'\sim\mathcal{P}(\cdot|x)}\left[\widehat{\mathcal{R}}_\pi Q^\pi(x)\right] = \mathbb{E}_{x'\sim\mathcal{P}(\cdot|x)\times\pi(\cdot|s')}\left[\widehat{\mathcal{R}}_\pi Q^\pi(x)\right] = 0.$$

As we can first take expectation w.r.t s' and $\bar{s}'$, we can conclude the proof. □

Theorem A.2 (Concentration Bounds with Non I.I.D. Samples)**.** *Consider a set of random transition pairs* $\{\tau_i\}_{i=1}^n$ *with* $\tau_i = (s_i, a_i, r_i, s_i')$. *Assume the state-action pairs* $(s_i, a_i)_{i=1}^n$ *are drawn from an arbitrary joint distribution, and given* $(s_i, a_i)_{i=1}^n$, *the local rewards and next states* $(r_i, s_i')_{i=1}^n$ *is drawn* independently *from* $r_i = r(s_i, a_i)$ *and* $s_i' \sim \mathcal{P}(\cdot \mid s_i, a_i)$.

Then $\forall \delta \in (0,1)$, *we have the following concentration inequality for* Q^π*:*

$$\mathbb{P}\left[\ \left|\widehat{L}_K^U(Q^\pi)\right| \geq 2\ell_{\max}\sqrt{\frac{\log\frac{2}{\delta}}{n}} \ \right] \leq \delta \,.$$

where $\ell_{\max}$ *is given via Lemma 3.1.*

Proof. First, $\forall\ (x, s'), (\bar{x}, \bar{s}')$ pair, where $s' \sim \mathcal{P}(\cdot|x)$, $\bar{s}' \sim \mathcal{P}(\cdot|\bar{x})$, from the proof of Proposition A.1, we can know

$$\mathbb{E}_{s'\sim\mathcal{P}(\cdot|x), \bar{s}'\sim\mathcal{P}(\cdot|\bar{x})}[K(x,\bar{x}) \cdot \widehat{\mathcal{R}}_\pi Q^\pi(x) \cdot \widehat{\mathcal{R}}_\pi Q^\pi(\bar{x})] = 0 \,.$$

Then we revisit the definition of V in Equation (19):

$$V(X_1, X_2, \cdots, X_n) = \frac{1}{k}\sum_{i\in[k]} h(X_{2i-1}, X_{2i}) \,.$$

For kernel Bellman statistic, $X_i = (x_i, s_i')$, where $s_i' \sim \mathcal{P}(\cdot|x_i)$, and $h(X_i, X_j) = K(x_i, x_j) \cdot \widehat{\mathcal{R}}_\pi Q^\pi(x_i) \cdot \widehat{\mathcal{R}}_\pi Q^\pi(x_j)$. With Proposition A.1, we have that

$$\mathbb{E}_{x_{2i-1}, x_{2i}, s_{2i-1}'\sim\mathcal{P}(\cdot|x_{2i-1}), s_{2i}'\sim\mathcal{P}(\cdot|x_{2i})}\left[K(x_{2i-1}, x_{2i}) \cdot \widehat{\mathcal{R}}_\pi Q^\pi(x_{2i-1}) \cdot \widehat{\mathcal{R}}_\pi Q^\pi(x_{2i}) \mid x_1, \cdots, x_{2i-2}\right] = 0 \,,$$

as the expectation doesn't depend on how we get x_{2i-1} and x_{2i}, but $s_{2i-1}' \sim \mathcal{P}(\cdot|x_{2i-1})$, $s_{2i}' \sim \mathcal{P}(\cdot|x_{2i})$. So we can view V as a summation of bounded martingale differences.

By using the Azuma's inequality for the martingale differences, we can show:

$$\mathbb{E}\left[\exp(\lambda V)\right] \leq \exp\left(\frac{\lambda^2(b-a)^2}{8k}\right) \,.$$

So the Hoeffding-type bound still holds, following the derivation of Appendix A.1. □

Remark We have proved that if the environment is Markovian, we still have the desired Hoeffding-type concentration bound for Q^π, and our algorithms are still valid given non i.i.d. samples. However, in practice, we still need the data collecting process to be ergodic (which is a general assumption (Puterman, 1994)), as we want to estimate Q^π over all of the $\mathcal{S} \times \mathcal{A}$.

Remark If we want to consider any Q function other than Q^π, the non i.i.d x will lead to an additional bias term, which can be difficult to estimate empirically . We leave this as a future work.

Remark Notice that, here we only use the universal upper bound of $\ell_{\pi,Q}(\tau_i, \tau_i)$, so the bound for Q^π is still valid for non i.i.d dataset $\mathcal{D} = \{x_i, r_i, s_i\}_{1\leq i\leq n}$ if we use Hoeffding-type bound for U-statistics.

B. Proof of Lemmas

Lemma 3.1. *Assume the reward function and the kernel function are bounded, i.e.* $\sup_x |r(x)| \leq r_{\max}$, $\sup_{x,\bar{x}} |K(x,\bar{x})| \leq K_{\max}$. *Then we have*

$$\sup_x |Q^\pi(x)| \leq Q_{\max} := \frac{r_{\max}}{1-\gamma} \,,$$
$$\sup_{\tau,\bar{\tau}} |\ell_{\pi,Q^\pi}(\tau,\bar{\tau})| \leq \ell_{\max} := \frac{4K_{\max}r_{\max}^2}{(1-\gamma)^2} \,.$$

Proof. Recall the definition of Q^π, we have

$$Q^\pi(x) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t | s_0 = s, a_0 = a\right] \leq \left(\sum_{t=0}^{\infty}\gamma^t\right) r_{\max} = \frac{r_{\max}}{1-\gamma}, \;\; \forall\, x\,.$$

Recall the definition of $\ell_{\pi,Q^\pi}(\tau,\bar{\tau})$ in (6), we have

$$\begin{aligned}
|\ell_{\pi,Q^\pi}(\tau,\bar{\tau})| &= |(Q(x) - r(x) - \gamma Q(x'))\, K(x,\bar{x})\, (Q(\bar{x}) - r(\bar{x}) - \gamma Q(\bar{x}'))| \\
&\leq \sup_{x,\bar{x}} |K(x,\bar{x})| \cdot \sup_{\tau} \left(Q(x) - r(x) - \gamma Q(x')\right)^2 \\
&\leq K_{\max} \cdot \left(\frac{r_{\max}}{1-\gamma} + r_{\max} + \gamma \cdot \frac{r_{\max}}{1-\gamma}\right)^2 \\
&= \frac{4K_{\max} r_{\max}^2}{(1-\gamma)^2}\,.
\end{aligned}$$

□

C. Accountable Off-Policy Evaluation for Average Reward $(\gamma = 1)$

In this section, we generalize our methods to the average reward setting where $\gamma = 1$. Denote by $M = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma \rangle$ a Markov decision process (MDP), where $\mathcal{S}$ is the state space; $\mathcal{A}$ is the action space; $\mathcal{P}(s'|s,a)$ is the transition probability; $r(s,a)$ is the average immediate reward; $\gamma = 1$ (undiscounted case). The expected reward of a given policy π is

$$\eta^\pi = \lim_{T\to\infty} \frac{1}{T+1} \mathbb{E}_\pi \left[\sum_{t=0}^{T} r_t \right].$$

In the discounted case ($0 < \gamma < 1$), the value function $Q^\pi(s,a)$ is the expected total discounted reward when the initial state s_0 is fixed to be s, and $a \sim \pi(\cdot|s)$: $Q^\pi(s,a) = \mathbb{E}_{\tau\sim\pi_\pi}[\sum_{t=0}^{\infty} \gamma^t r_t | s_0 = s, a_0 = a]$. If the Markov process is ergodic (Puterman, 1994), the expected average reward does not depend on the initial states. In the average case, however, $Q^\pi(s,a)$ measures the *average adjusted* sum of reward:

$$Q^\pi(s,a) := \lim_{T\to\infty} \mathbb{E}_\pi \left[\sum_{t=0}^{T} (r_t - \eta^\pi) \mid s_0 = s, a_0 = a \right],$$

which is referred to as the *adjusted (state-action) value function*.

Under this definition, (Q^π, η^π) is the unique fixed-point solution to the following Bellman equation:

$$Q(s,a) = \mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a), a'\sim\pi(\cdot|s')}[r(s,a) + Q(s',a') - \eta]\,. \tag{20}$$

To simplify notation, we still assume $x = (s,a)$, $\bar{x} = (\bar{s}, \bar{a})$, and $x' := (s', a')$ with $s' \sim \mathcal{P}(\cdot|s,a), a' \sim \pi(\cdot|s')$. Define the *Bellman residual operator* as

$$\mathcal{R}_\pi Q(x) = \mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a), a'\sim\pi(\cdot|s')} \left[r(x) + Q(x') - \eta \right] - Q(x)\,,$$

where Q is an estimation of adjusted value function, and η is an estimation of the expected reward. Note that $\mathcal{R}_\pi Q(x)$ depends on both Q and η, even though it is not indicated explicitly on notation. Given a (Q, η) pair, the kernel loss for $\gamma = 1$ can be defined as

$$L_K(Q,\eta) := \mathbb{E}_{x,\bar{x}\sim\mu}[\mathcal{R}_\pi Q(x) \cdot K(x,\bar{x}) \cdot \mathcal{R}_\pi Q(\bar{x})]\,.$$

Given a set of observed transition pairs $\mathcal{D} = \{\tau_i\}_{i=1}^n$, and we can estimate $L_K(Q,\eta)$ with the following V-statistics:

$$\widehat{L}_K^V(Q,\eta) = \frac{1}{n^2} \sum_{i,j=1}^{n} \ell_{\pi,Q}(\tau_i, \tau_j)\,,$$

where

$$\ell_{\pi,Q}(\tau_i,\tau_j) = \widehat{\mathcal{R}}_\pi Q(x_i) K(x_i, x_j) \widehat{\mathcal{R}}_\pi Q(x_j)\,,$$

and

$$\widehat{\mathcal{R}}_\pi Q(x_i) = r(x_i) + \mathbb{E}_{a_i'\sim\pi(\cdot|s_i')} \left[Q(x_i') \right] - \eta - Q(x_i).$$

Similarly, we can estimate $L_K(Q,\eta)$ via U-statistics:

$$\widehat{L}_K^U(Q,\eta) = \frac{1}{n(n-1)} \sum_{1\le i\ne j\le n} \ell_{\pi,Q}(\tau_i,\tau_j)\,.$$

C.1. Concentration Bounds for U-/V-statistics

Here we derive the concentration bounds of the U-/V-statistics for the average reward case, under the mild assumption that the reward $r(x)$ and $Q(x)$ are bounded.

Lemma C.1. *Assume the reward function, the adjusted (state-action) value function and the kernel function are uniformly bounded, i.e.* $\sup_{s\in\mathcal{S},a\in\mathcal{A}}|r(x)|\le r_{\max}$, $\sup_{s\in\mathcal{S},a\in\mathcal{A}}|Q^{\pi}(x)|\le Q_{\max}$, $\sup_{x,\bar{x}}|K(x,\ \bar{x})|\le K_{\max}$. *Then we have*

$$\sup_{\tau,\bar{\tau}}|\ell_{\pi,Q^{\pi}}(\tau,\bar{\tau})|\le \ell_{\max}:=4K_{\max}(Q_{\max}+r_{\max})^2\,.$$

Proof. By definition, we have

$$\begin{aligned}\sup_{\tau,\bar{\tau}}|\ell_{\pi,Q^{\pi}}(\tau,\bar{\tau})| &\le |r_{\max}+Q_{\max}-(-r_{\max})-(-Q_{\max})|^2\cdot|K(x,\bar{x})|\\ &\le 4K_{\max}(Q_{\max}+r_{\max})^2.\end{aligned}$$

□

With a similar derivation in Appendix A, we can have the same Hoeffding-type bounds for U/V-statistics as that of the discounted case, which can be utilized to construct the confidence interval for η^{π}.

C.2. Confidence Bounds for Average Reward

As our final target is to build the confidence interval for the average reward, we follow a similar idea as the discounted case to obtain a high probability upper bound of the expected reward η^{π} by solving the following optimization problem:

$$\max_{|\eta|\le r_{\max},Q\in\mathcal{F},}\left\{\ \eta\quad\text{s.t.}\quad \widehat{L}_K^V(Q,\eta)\le\lambda_K\right\}\,,$$

where η is a scalar variable and $Q\in\mathcal{F}$ is the adjusted value function, which we want to jointly optimize.

C.3. Optimization in RKHS

Similar to the discounted case, we can use random feature approximation to speed up the optimization. In this case, the optimization reduces to

$$\widehat{\eta}^{+}=\max_{|\eta|\le r_{\max},\theta}\left\{\eta,\quad\text{s.t.}\quad\left(Z\theta+\eta-v\right)^{\top}M\left(Z\theta+\eta-v\right)\le\lambda_K\right\}\,.$$

where the constants are the same as these defined in Section 4.2, except that $Z=X-X'$.

D. Experiments

In this section, we provide the details of the experiments, and some additional experiments for validating the effectiveness of our method.

D.1. Experimental Details

Evaluation Environments We evaluate the proposed algorithms in Section 4 on two continuous control tasks: Inverted-Pendulum and Puck-Mountain.

Inverted-Pendulum is a pendulum that has its center of mass above its pivot point. It has a continuous state space on $\mathbb{R}^4$. We discrete the action space to be $\{-1, -0.3, -0.2, 0, 0.2, 0.3, 1\}$. The pendulum is unstable and would fall over without careful balancing. We train a near optimal policy that can make the pendulum balance for a long horizon using deep Q learnings, and use its softmax function as policies. We set the temperature to be higher for the behavior policies to encourage exploration. We use the implementation from OpenAI Gym (Brockman et al., 2016) and change the dynamic by adding some additional zero mean Gaussian noise to the transition dynamic.

Puck-Mountain is an environment similar to Mountain-Car, except that the goal of the task is to push the puck as high as possible in a local valley whose initial position is at the bottom of the valley. If the ball reaches the top sides of the valley, it will hit a roof and change the speed to its opposite direction with half of its original speeds. The reward was determined by the current velocity and height of the puck. The environment has a $\mathbb{R}^2$ state space, and a discrete action space with 3 possible actions (pushing left, no pushing, and pushing right). We also add zero mean Gaussian perturbations to the transition dynamic to make it stochastic.

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.

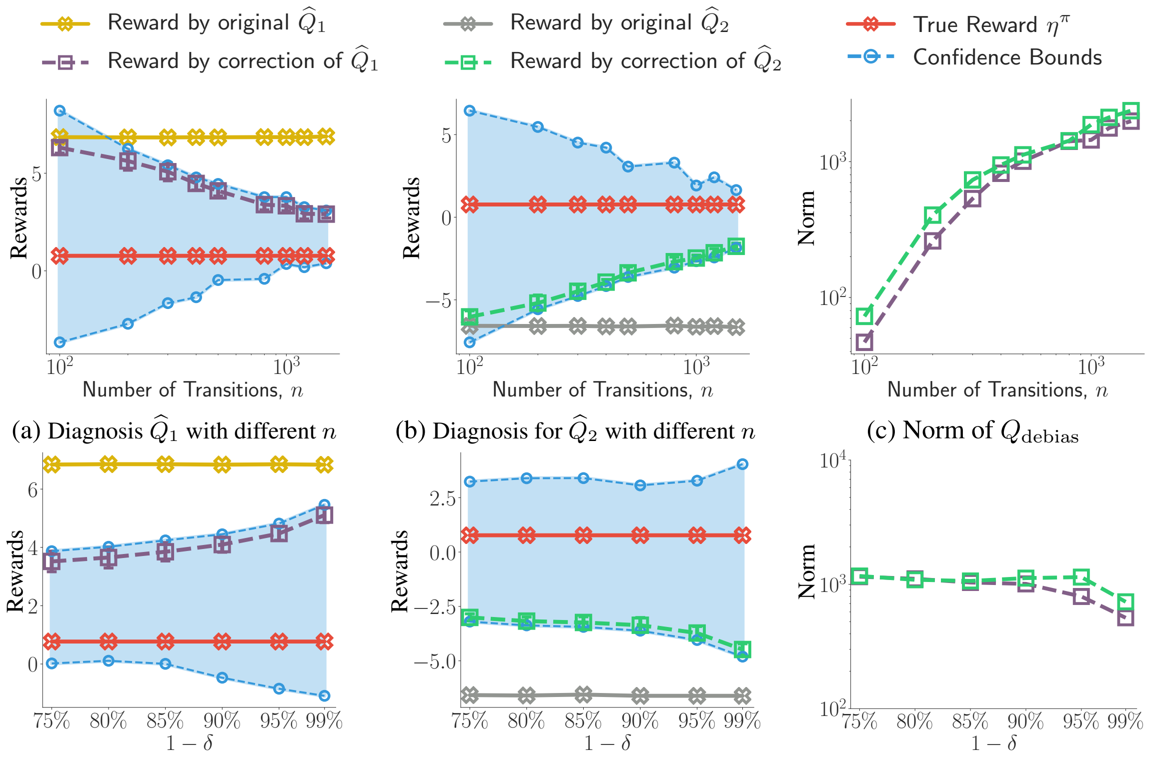
Hyperparameters Selection and Neural Feature Learning For all of our experiments, we use Gaussian RBF kernel $K(x, \bar{x}) = \exp\left(-||x - \bar{x}||_2^2/h^2\right)$ in the kernel Bellman kernel (e.g., for Equation (5)). We evaluate the kernel Bellman loss on a separate batch of training data, and find that we can set the bandwidth to $h = 0.5$, which will give a good solution.

When we parameterize Q function $Q(x) := \theta^\top \Phi(x)$with random Fourier features: $\Phi(x) := [\cos(\mu_i^\top x + b_i)]_{i=1}^m$, where $\mu_i \sim \mathcal{N}(0, \frac{1}{h_0^2}\mathrm{I})$, $b_i \sim$ Uniform$([0, 2\pi])$, and h_0 is a bandwidth parameter. We select the bandwidth h_0 from a candidate set $\Pi = \{h_1, h_2, \ldots, h_k\}$ by finding the smallest lower bound and largest upper bound on a separate validation data. Specifically, for each $h_i \in \Pi$, we follow the procedure of Algorithm 1 to calculate an upper and lower bounds for η^π, and select the lowest lower bound and the largest upper bound as our final bounds. Doing this ensures that our bounds are pessimistic and safe. In our empirical experiments we set $\Pi = \{0.2, 0.5, 0.6, 0.8, 1.0\}$.

Following the similar procedure, we also select a set of neural features (the neural feature map before the last linear layer) on the validation set, which have relatively lower kernel loss when we optimize the neural network. Similarly, we select two different neural features for each environment and use the pessimisitc upper and lower bounds for all of our experiments.

Constructing Existing Estimators in Post-hoc Diagnosis Since we only need to demonstrate the effectiveness of our proposed post-hoc diagnosis process, we simply parameterize Q as a linear function of a small set of random Fourier features ($Q(\cdot) = \theta^\top \Phi(\cdot)$), and estimate θ by minimizing the kernel Bellman loss by running gradient descent for different numbers of iterations. For the experiments in Inverted-Pendulum (Figure 2), $\widehat{Q}_1$ (resp. $\widehat{Q}_2$) are obtained when we run a large (resp. small) number of iterations in training, so that $\widehat{Q}_1$ is relatively accurate while $\widehat{Q}_2$ is not. For Puck-Mountain in Figure 3, the error of both $\widehat{Q}_1$ and $\widehat{Q}_2$ are relatively large.

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



(a) Diagnosis $\widehat{Q}_1$ with different n (b) Diagnosis for $\widehat{Q}_2$ with different n (c) Norm of Q_{debias}

(d) Diagnosis for $\widehat{Q}_1$ with different δ (e) Diagnosis for $\widehat{Q}_2$ with different δ (f) Norm of Q_{debias}

Figure 3. Post-hoc diagnosis on Puck-Mountain. We set the discounted factor $\gamma = 0.95$, the horizon length $T = 100$, number of transitions $n = 500$, failure probability $\delta = 0.10$, temperature of the behavior policy $\tau = 1$, and the feature dimension 10 as default. The rest of the parameters are the same as that in Section 5.

D.2. Additional Experiments

Post-hoc Diagnosis Experiments on Puck-Mountain Figure 3 (a)-(f) show the diagnosis results for two estimations of Q-function ($\widehat{Q}_1$ and $\widehat{Q}_2$) on Puck-Mountain. Here both $\widehat{Q}_1$ and $\widehat{Q}_2$ have relatively large bias, but $\widehat{Q}_1$ tends to overestimate η^π (see Figure 3(a)), while $\widehat{Q}_2$ tends to underestimate η^π (see Figure 3).

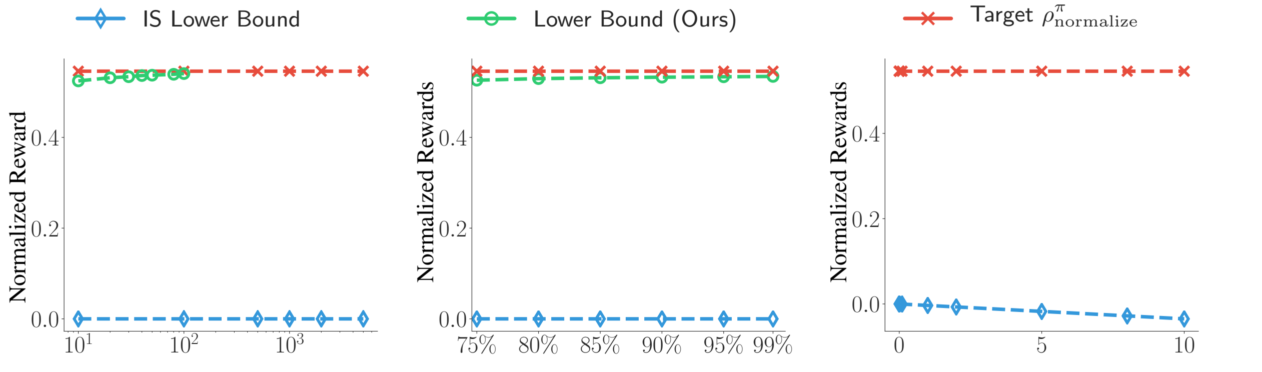
Figure 3(a)-(c) show that as we increase the number of transitions, the norm of the debiasing term Q_{debias} becomes larger. This is because when we have more data, we have a relatively tight confidence bound and we need a more complex debias function to provide good post-hoc correction. Figure 3 (d)-(f) demonstrate the performance of our algorithms when we change the failure probability δ.

Comparison with Thomas et al. (2015a) As a comparison, we implement the method from Thomas et al. (2015a), which uses concentration inequality to construct confidence bounds on an importance sampling (IS) based estimator. Following Thomas et al. (2015a), we assume the expected reward is normalized as follows

$$\rho^\pi_{\text{normalize}} := \frac{\mathbb{E}_\pi\left[\sum_{t=1}^T \gamma^{t-1} r_t\right] - R_{\min}}{R_{\max} - R_{\min}}, \tag{21}$$

where $\sum_{t=1}^T \gamma^{t-1} r_t$ is the discounted return of a trajectory following policy π, and $R_{\max}$ and $R_{\min}$ are the upper and lower bounds on $\sum_{t=1}^T \gamma^{t-1} r_t$; see Thomas et al. (2015a) for the choice of $R_{\max}$ and $R_{\min}$.

Given a set of trajectories $\{(s_t^{(i)}, a_t^{(i)}, r_t^{(i)})_{t=1}^T\}_{1\le i\le N}$ generated by behavior policy π_0, we have



(a) Number of Episodes, N (b) $1-\delta$ (c) Threshold value c

Figure 4. IS-based confidence lower bounds Thomas et al. (2015a) on Inverted-Pendulum following the same setting as that in Section 5. We use the results of random Fourier feature in Figure 1 as our lower bound. For Thomas et al. (2015a), we set the threshold value to be $c = 10^{-5}$ in (a) & (b), and the number of episodes $N = 2000$ in (b) & (c). For our method, we only use $n = 50 \times 20$ number of transition pairs in (b). We report the normalized reward based on the procedure in Thomas et al. (2015a) ($\rho_{\text{normalize}} = (\sum_{t=1}^T \gamma^{t-1} r_t - R_{\min})/(R_{\max} - R_{\min})$).

$$\widehat{\rho}^\pi_{\text{IS}} := \frac{1}{N}\sum_{i=1}^N X_i, \quad \text{with} \quad X_i = \underbrace{R_i}_{\text{return}} \underbrace{\prod_{t=1}^T \frac{\pi(a_t^{(i)}|s_t^{(i)})}{\pi_0(a_t^{(i)}|s_t^{(i)})}}_{\text{importance weight}}, \tag{22}$$

where R_i is reward the i-th trajectory from the data (normalized as shown in Eq (21)). Theorem 1 in Thomas et al. (2015a) provides a concentration inequality for constructing lower bound of $\rho^\pi_{\text{normalize}}$ based on a truncated importance sampling estimator. Let $\{c_i\}_{i=1}^N$ be a set of positive real-valued threshold and $\delta \in (0,1)$ and $Y_i = \min\{X_i, c_i\}$, we have with probability at least $1-\delta$,

$$\rho^\pi_{\text{normalize}} \geq \underbrace{\left(\sum_{i=1}^N \frac{1}{c_i}\right)^{-1} \sum_{i=1}^N \frac{Y_i}{c_i}}_{\text{empirical mean}} - \underbrace{\left(\sum_{i=1}^N \frac{1}{c_i}\right)^{-1} \frac{7N\ln(2/\delta)}{3(N-1)}}_{\text{term that goes to zero as } 1/N \text{ as } N\to\infty} - \underbrace{\left(\sum_{i=1}^N \frac{1}{c_i}\right)^{-1} \sqrt{\frac{\ln(2/\delta)}{N-1}\sum_{i,j=1}^N \left(\frac{Y_i}{c_i} - \frac{Y_j}{c_j}\right)^2}}_{\text{term that goes to zero as } 1/\sqrt{N} \text{ as } N\to\infty}. \tag{23}$$

The RHS provides a lower bound of $\rho^\pi_{\text{normalized}}$ based the empirical trajectories. Following the settings in Thomas et al. (2015a), we set the threshold values to be a constant c, that is, $c_i = c$ for all $i = 1, \ldots, N$.

We evaluate the method on Inverted-Pendulum under the same default settings as our experiments in the paper. Figure 4(a)-(c) show the results of the high confidence lower bound. We can see that the IS lower bounds are almost vacuous (i.e., very close to zero) and it does not help very much when we increase the number of episodes (up to $N = 2000$) or the failure probability δ. This is because there is only small overlap between the target policy π and behavior policy π_0 in this case and when the horizon length is large ($T = 50$), making the IS estimator degenerate. Although we need to point out that to get tighter lower bound with our method, we assume the true Q^π is in the function space $\mathcal{F}$ that we choose, and also assume the horizon length $T \to \infty$ to make the confidence bound provably hold.

E. Discussion on Reproducing Kernel Hilbert Space

We provide proof of Proposition 4.2 and discussion for Section 4.2.

Proof of Proposition 4.2. Consider the Lagrangian of the constrained optimization (10), we have:

$$\begin{aligned} L(Q,\lambda_1,\lambda_2) =& \mathbb{E}_{x\sim\mu_0\times\pi}[Q(x)] - \lambda_1(\widehat{L}_K(Q) - \lambda_K) - \lambda_2(\|Q\|_{\mathcal{H}}^2 - \rho) \\ =& \langle Q, f_0\rangle - \frac{\lambda_1}{n^2}\left(\sum_{i,j}(\langle Q, f_i\rangle - r_i)K(x_i,x_j)(\langle Q, f_j\rangle - r_j)\right) - \lambda_2(\|Q\|_{\mathcal{H}}^2) - C, \end{aligned}$$

where λ_1, λ_2 are Lagrangian multipliers with respect to the two constraints, and C is a constant related to $\lambda_1, \lambda_2, \lambda_K$ and ρ. We rewrite Q into

$$Q = \sum_{k=0}^{n} \alpha_k f_k + Q_\perp,$$

where $Q_\perp$ is in the orthogonal subspace to the linear span of $f_0, f_1, ..., f_n$, that is, $\langle f_i, Q_\perp\rangle = 0,\ \forall i \in [n]$. By decomposin Q into $\sum_{k=0}^{n} \alpha_k f_k$ and $Q_\perp$, we have

$$\begin{aligned} L(Q,\lambda_1,\lambda_2) + C =& \langle Q, f_0\rangle - \frac{\lambda_1}{n^2}\left(\sum_{i,j}(\langle Q, f_i\rangle - r_i)K(x_i,x_j)(\langle Q, f_j\rangle - r_j)\right) - \lambda_2(\|Q\|_{\mathcal{H}}^2) \\ =& \langle \sum_{k=0}^{n}\alpha_k f_k + Q_\perp, f_0\rangle - \frac{\lambda_1}{n^2}\left(\sum_{i,j=1}^{n}(\langle \sum_{k=0}^{n}\alpha_k f_k + Q_\perp, f_i\rangle - r_i)K(x_i,x_j)(\langle \sum_{k=0}^{n}\alpha_k f_k + Q_\perp, f_j\rangle - r_j)\right) \\ & - \lambda_2(\|\sum_{k=0}^{n}\alpha_k f_k\|_{\mathcal{H}}^2 + \|Q_\perp\|_{\mathcal{H}}^2) \\ =& \langle \sum_{k=0}^{n}\alpha_k f_k, f_0\rangle - \frac{\lambda_1}{n^2}\left(\sum_{i,j=1}^{n}(\langle \sum_{k=0}^{n}\alpha_k f_k, f_i\rangle - r_i)K(x_i,x_j)(\langle \sum_{k=0}^{n}\alpha_k f_k, f_j\rangle - r_j)\right) \\ & - \lambda_2(\|\sum_{k=0}^{n}\alpha_k f_k\|_{\mathcal{H}}^2 + \|Q_\perp\|_{\mathcal{H}}^2) \\ \leq& \langle \sum_{k=0}^{n}\alpha_k f_k, f_0\rangle - \frac{\lambda_1}{n^2}\left(\sum_{i,j=1}^{n}(\langle \sum_{k=0}^{n}\alpha_k f_k, f_i\rangle - r_i)K(x_i,x_j)(\langle \sum_{k=0}^{n}\alpha_k f_k, f_j\rangle - r_j)\ell\right) \\ & - \lambda_2(\|\sum_{k=0}^{n}\alpha_k f_k\|_{\mathcal{H}}^2) := L(\alpha,\lambda_1,\lambda_2), \end{aligned}$$

where the optimum Q will have $Q_\perp = 0$ and $L(\alpha, \lambda_1, \lambda_2)$ is the Lagrangian w.r.t. to coefficient α. Collecting the term we can reform the optimization w.r.t. α as

$$\widehat{\eta}^+ := \max_{\{\alpha_i\}_{0\leq i\leq n}} \left\{[c^\top\alpha + \lambda_\eta], \qquad \text{s.t. } \alpha^\top A\alpha + b^\top\alpha + d \leq 0, \qquad \alpha^\top B\alpha \leq \rho.\right\}$$

where $B_{ij} = \langle f_i, f_j\rangle_{\mathcal{H}_{K_0}},\ \forall 0 \leq i, j \leq n$ is the inner product matrix of $\{f_i\}_{0\leq i\leq n}$ under $\mathcal{H}_{K_0}$, c as the first column of B and B_1 be the remaining sub-matrix. Let $M_{ij} = K(x_i, x_j)$ be the kernel matrix, $R_i = r_i$ be the vector of reward from data, then $A = \frac{1}{n^2}B_1MB_1^\top$, $b = -\frac{1}{n^2}B_1MR$ and $d = \frac{1}{n^2}R^\top MR - \lambda_K$. □

Random Feature Approximation Consider the random feature representation of kernel K_0

$$K_0(x, x') = \mathbb{E}_{w\sim\mu}[\phi(x, w)\phi(x', w)], \tag{24}$$

where μ is a distribution of random variable w. Every function f in the RKHS $\mathcal{H}_{K_0}$ associated with K_0 can be represented as

$$f(x) = \mathbb{E}_{w\sim\mu}[\phi(x,w)\alpha_f(w)],$$

whose RKHS norm equals

$$\|f\|^2_{\mathcal{H}_{K_0}} = \mathbb{E}_{w\sim\mu}[\alpha_f(w)^2].$$

To approximate K_0, we draw an i.i.d. sample $\{w_i\}$ from μ to approximate K_0 with

$$\widehat{K_0}(x,x') = \frac{1}{m}\sum_{i=1}^m \phi(x,w_i)\phi(x',w_i). \tag{25}$$

Similarly, each $f \in \mathcal{H}_{\widehat{K_0}}$ can be represented as

$$f(x) = \frac{1}{m}\sum_{i=1}^m \phi(x,w_i)\alpha_i$$

and its corresponding RKHS norm is

$$\|f\|^2_{\mathcal{H}_{\widehat{K_0}}} = \frac{1}{m}\sum_{i=1}^m \alpha_i^2.$$

Denote by $\theta = \alpha/m$, then we have $f(x) = \theta^\top\Phi(x)$ and $\|f\|^2_{\mathcal{H}_{\widehat{K_0}}} = m\,\|\theta\|_2^2$, which is the form used in the paper.

The result below shows that when Q^π is included in $\mathcal{H}_{K_0}$ but may not be included in $\mathcal{H}_{\widehat{K}_0}$, we can still get a provably upper bound by setting the radius of the optimization domain properly large.

Theorem E.1. *Let K_0 be a positive definite kernel with random feature expansion in (24) and $\widehat{K}_0$ defined in (25) with $\{w_i\}_{i=1}^m$ i.i.d. drawn from μ. Assume $\|\phi\|_\infty = \sup_{x,w}|\phi(x,w)| < 0$. Define*

$$\eta^+_{K_0} = \max_{Q\in\mathcal{H}_{K_0}} \{\eta(Q), \quad \textit{s.t.} \quad \widehat{L}_K(Q) \le \lambda, \quad \|Q\|^2_{\mathcal{H}_{K_0}} \le \rho\}. \tag{26}$$

Let Q^* be the optimal solution of (26) and $Q^*(\cdot) = \mathbb{E}_{w\sim\mu}[\phi(\cdot,w)\alpha^*(w)]$. Assume $C := \mathrm{var}_{w\sim\mu}((\alpha^*(w))^2) < \infty$. For $\delta\in(0,1)$, define

$$\eta^+_{\widehat{K_0}} = \max_{Q\in\mathcal{H}_{\widehat{K_0}}} \{\eta(Q), \quad \textit{s.t.} \quad \widehat{L}_K(Q) \le \lambda, \quad \|Q\|^2_{\mathcal{H}_{\widehat{K_0}}} \le \rho'\}. \tag{27}$$

If we set $\rho' \ge \rho + \sqrt{\frac{C}{\delta m}}$, then we have with probability at least $1-2\delta$

$$\eta^+_{K_0} \le \eta^+_{\widehat{K_0}} + \|\phi\|_\infty\sqrt{\frac{\rho}{\delta m}}.$$

Therefore, if Q^π belongs to $\mathcal{H}_{K_0}$ (and hence $\eta^\pi \le \eta^+_{K_0}$), then $\eta^+_{\widehat{K_0}} + \|\phi\|_\infty\sqrt{\frac{\rho}{\delta m}}$ provides a high probability upper bound of η^π.

Proof. Following $Q^*(\cdot) = \mathbb{E}_{w\sim\mu}[\phi(\cdot,w)\alpha^*(w)]$, we have α^* satisfies

$$\|Q^*\|^2_{\mathcal{H}_{K_0}} = \mathbb{E}_{w\sim\mu}\left[(\alpha^*(w))^2\right] \le \rho. \tag{28}$$

Let

$$\widetilde{Q}(\cdot) = \frac{1}{m}\sum_{i=1}^m \phi(\cdot,w_i)\alpha^*(w_i),$$

for which we have

$$\|\widetilde{Q}\|^2_{\mathcal{H}_{\widehat{K_0}}} = \frac{1}{m}\sum_{i=1}^m \alpha^*(w_i)^2.$$

By Chebyshev inequality, we have with probability at least $1-\delta$

$$\|\widetilde{Q}\|_{\mathcal{H}_{\widehat{K}_0}}^2 \leq \|Q^*\|_{\mathcal{H}_{K_0}}^2 + \sqrt{\frac{1}{\delta m}\mathrm{var}_{w\sim\mu}(\alpha^*(w)^2)} \leq \rho + \sqrt{\frac{C}{\delta m}}.$$

If $\rho' \geq \rho + \sqrt{\frac{C}{\delta m}}$, then $\widetilde{Q}$ is included in the optimization set of (27), and hence

$$\eta(\widetilde{Q}) \leq \eta^+_{\widehat{K}_0}. \tag{29}$$

On the other hand, because $\eta(Q)$ is a linear functional, we have

$$\eta(\widetilde{Q}) - \eta(Q^*) = \frac{1}{m}\sum_{i=1}^m \Phi(w_i)\alpha^*(w_i) - \mathbb{E}_{w\sim\mu}[\Phi(w_i)\alpha^*(w)],$$

where $\Phi(w) = \eta(\phi(\cdot, w))$. Note that

$$\mathrm{var}_{w\sim\mu}(\Phi(w)\alpha^*(w)) \leq \|\phi\|_\infty^2 \, \mathbb{E}_{w\sim\mu}\left[(\alpha^*(w))^2\right] \leq \|\phi\|_\infty^2 \, \rho,$$

where we used (28). By Chebyshev inequality, we have with probability at least $1-\delta$

$$|\eta(\widetilde{Q}) - \eta(Q^*)| \leq \sqrt{\frac{\|\phi\|_\infty^2 \rho}{\delta m}}.$$

Combining this with (29), we have, with probability at least $1-2\delta$,

$$\eta^+_{K_0} = \eta(Q^*) = \eta(\widetilde{Q}) \; + \; (\eta(Q^*) - \eta(\widetilde{Q})) \leq \eta^+_{\widehat{K}_0} \; + \; \sqrt{\frac{\|\phi\|_\infty^2 \rho}{\delta m}}.$$

□