CoinDICE: Off-Policy Confidence Interval Estimation

*Bo Dai[1], * Ofir Nachum[1], Yinlam Chow[1]
Lihong Li[1], Csaba Szepesvári[2,3], Dale Schuurmans[1,2]
[1]Google Research, Brain Team [2]University of Alberta [3]DeepMind

Abstract

We study *high-confidence behavior-agnostic off-policy evaluation* in reinforcement learning, where the goal is to estimate a confidence interval on a target policy's value, given only access to a static experience dataset collected by unknown behavior policies. Starting from a function space embedding of the linear program formulation of the Q-function, we obtain an optimization problem with generalized estimating equation constraints. By applying the generalized empirical likelihood method to the resulting Lagrangian, we propose *CoinDICE*, a novel and efficient algorithm for computing confidence intervals. Theoretically, we prove the obtained confidence intervals are valid, in both asymptotic and finite-sample regimes. Empirically, we show in a variety of benchmarks that the confidence interval estimates are tighter and more accurate than existing methods.[2]

1 Introduction

One of the major barriers that hinders the application of reinforcement learning (RL) is the ability to evaluate new policies reliably *before* deployment, a problem generally known as *off-policy evaluation* (OPE). In many real-world domains, *e.g.*, healthcare (Murphy et al., 2001; Gottesman et al., 2018), recommendation (Li et al., 2011; Chen et al., 2019), and education (Mandel et al., 2014), deploying a new policy can be expensive, risky or unsafe. Accordingly, OPE has seen a recent resurgence of research interest, with many methods proposed to estimate the value of a policy (Precup et al., 2000; Dudík et al., 2011; Bottou et al., 2013; Jiang and Li, 2016; Thomas and Brunskill, 2016; Liu et al., 2018; Nachum et al., 2019a; Kallus and Uehara, 2019a,b; Zhang et al., 2020b).

However, the very settings where OPE is necessary usually entail limited data access. In these cases, obtaining knowledge of the uncertainty of the estimate is as important as having a consistent estimator. That is, rather than a *point estimate*, many applications would benefit significantly from having *confidence intervals* on the value of a policy. The problem of estimating these confidence intervals, known as *high-confidence off-policy evaluation* (HCOPE) (Thomas et al., 2015b), is imperative in real-world decision making, where deploying a policy without high-probability safety guarantees can have catastrophic consequences (Thomas, 2015). Most existing high-confidence off-policy evaluation algorithms in RL (Bottou et al., 2013; Thomas et al., 2015a,b; Hanna et al., 2017) construct such intervals using statistical techniques such as concentration inequalities and the bootstrap applied to importance corrected estimates of policy value. The primary challenge with these correction-based approaches is the high variance resulting from multiplying per-step importance ratios in long-horizon problems. Moreover, they typically require full knowledge (or a good estimate) of the behavior policy, which is not easily available in behavior-agnostic OPE settings (Nachum et al., 2019a).

In this work, we propose an algorithm for behavior-agnostic HCOPE. We start from a linear programming formulation of the state-action value function. We show that the value of the policy may be obtained from a Lagrangian optimization problem for generalized estimating equations

*Equal contribution. Email: {bodai, ofirnachum}@google.com.

[2]Open-source code for CoinDICE is available at https://github.com/google-research/dice_rl.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

over data sampled from off-policy distributions. This observation inspires a generalized empirical likelihood approach (Owen, 2001; Broniatowski and Keziou, 2012; Duchi et al., 2016) to confidence interval estimation. These derivations enable us to express high-confidence lower and upper bounds for the policy value as results of minimax optimizations over an arbitrary offline dataset, with the appropriate distribution corrections being implicitly estimated during the optimization. We translate this understanding into a practical estimator, *Confidence Interval DIstribution Correction Estimation* (CoinDICE), and design an efficient algorithm for implementing it. We then justify the asymptotic coverage of these bounds and present non-asymptotic guarantees to characterize finite-sample effects. Notably, CoinDICE is behavior-agnostic and its objective function does not involve any per-step importance ratios, and so the estimator is less susceptible to high-variance gradient updates. We evaluate CoinDICE in a number of settings and show that it provides both tighter confidence interval estimates and more correctly matches the desired statistical coverage compared to existing methods.

2 Preliminaries

For a set W, the set of probability measures over W is denoted by $\mathcal{P}(W)$.[3] We consider a Markov Decision Process (MDP) (Puterman, 2014), $\mathcal{M} = (S, A, T, R, \gamma, \mu_0)$, where S denotes the state space, A denotes the action space, $T : S \times A \to \mathcal{P}(S)$ is the transition probability kernel, $R : S \times A \to \mathcal{P}([0, R_{\max}])$ is a bounded reward kernel, $\gamma \in (0, 1]$ is the discount factor, and μ_0 is the initial state distribution.

A policy, $\pi : S \to \mathcal{P}(A)$, can be used to generate a random trajectory by starting from $s_0 \sim \mu_0(s)$, then following $a_t \sim \pi(s_t)$, $r_t \sim R(s_t, a_t)$ and $s_{t+1} \sim T(s_t, a_t)$ for $t \geqslant 0$. The state- and action-value functions of π are denoted V^π and Q^π, respectively. The policy also induces an occupancy measure, $d^\pi(s,a) := (1-\gamma)\mathbb{E}_\pi\left[\sum_{t\geqslant 0} \gamma^t \mathbf{1}\{s_t = s, a_t = a\}\right]$, the normalized discounted probability of visiting (s, a) in a trajectory generated by π, where $\mathbf{1}\{\cdot\}$ is the indicator function. Finally, the *policy value* is defined as the *normalized* expected reward accumulated along a trajectory:

$$\rho_\pi := (1-\gamma)\,\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t | s_0 \sim \mu_0, a_t \sim \pi(s_t), r_t \sim R(s_t, a_t), s_{t+1} \sim T(s_t, a_t)\right]. \quad (1)$$

We are interested in estimating the policy value and its confidence interval (CI) in the *behavior agnostic off-policy* setting (Nachum et al., 2019a; Zhang et al., 2020a), where interaction with the environment is limited to a static dataset of experience $\mathcal{D} := \{(s, a, s', r)_i\}_{i=1}^n$. Each tuple in $\mathcal{D}$ is generated according to $(s,a) \sim d^{\mathcal{D}}, r \sim R(s,a)$ and $s' \sim T(s,a)$, where $d^{\mathcal{D}}$ is an unknown distribution over $S \times A$, perhaps induced by one or more unknown behavior policies. The initial distribution $\mu_0(s)$ is assumed to be easy to sample from, as is typical in practice. Abusing notation, we denote by $d^{\mathcal{D}}$ both the distribution over (s, a, s', r) and its marginal on (s, a). We use $\mathbb{E}_d[\cdot]$ for the expectation over a given distribution d, and $\mathbb{E}_{\mathcal{D}}[\cdot]$ for its empirical approximation using $\mathcal{D}$.

Following previous work (Sutton et al., 2012; Uehara et al., 2019; Zhang et al., 2020a), for ease of exposition we assume the transitions in $\mathcal{D}$ are *i.i.d.*. However, our results may be extended to fast-mixing, ergodic MDPs, where the the empirical distribution of states along a long trajectory is close to being *i.i.d.* (Antos et al., 2008; Lazaric et al., 2012; Dai et al., 2017; Duchi et al., 2016).

Under mild regularity assumptions, the OPE problem may be formulated as a linear program – referred to as the Q-LP (Nachum et al., 2019b; Nachum and Dai, 2020) – with the following primal and dual forms:

$$\begin{aligned}\min_{Q:S\times A\to\mathbb{R}}\; & (1-\gamma)\,\mathbb{E}_{\mu_0\pi}[Q(s_0,a_0)] \\ \text{s.t.}\; & Q(s,a) \geqslant R(s,a) + \gamma\cdot\mathcal{P}^\pi Q(s,a), \\ & \forall (s,a)\in S\times A,\end{aligned} \quad (2)$$

and

$$\begin{aligned}\max_{d:S\times A\to\mathbb{R}_+}\; & \mathbb{E}_d[r(s,a)] \\ \text{s.t.}\; & d(s,a) = (1-\gamma)\,\mu_0\pi(s,a) + \gamma\cdot\mathcal{P}_*^\pi d(s,a), \\ & \forall (s,a)\in S\times A,\end{aligned} \quad (3)$$

where the operator $\mathcal{P}^\pi$ and its adjoint, $\mathcal{P}_*^\pi$, are defined as

$$\mathcal{P}^\pi Q(s,a) := \mathbb{E}_{s'\sim T(\cdot|s,a), a'\sim\pi(\cdot|s')}[Q(s',a')],$$
$$\mathcal{P}_*^\pi d(s,a) := \pi(a|s)\sum_{\tilde{s},\tilde{a}} T(s|\tilde{s},\tilde{a})\, d(\tilde{s},\tilde{a}).$$

[3] All sets and maps are assumed to satisfy appropriate measurability conditions; which we will omit from below for the sake of reducing clutter.

The optimal solutions of (2) and (3) are the Q-function, Q^{π}, and stationary state-action occupancy, d^{π}, respectively, for policy π; see Nachum et al. (2019b, Theorems 3 & 5) for details as well as extensions to the undiscounted case.

Using the Lagrangian of (2) or (3), we have

$$\rho_{\pi} = \min_{Q} \max_{\tau \geqslant 0} \ (1-\gamma)\, \mathbb{E}_{\mu_0 \pi}\left[Q\left(s_0, a_0\right)\right] + \mathbb{E}_{d^{\mathcal{D}}}\left[\tau\left(s, a\right)\left(R\left(s, a\right) + \gamma Q\left(s', a'\right) - Q\left(s, a\right)\right)\right], \quad (4)$$

where $\tau\left(s, a\right) := \frac{d(s,a)}{d^{\mathcal{D}}(s,a)}$ is the *stationary distribution corrector*. One of the key benefits of the minimax optimization (4) is that both expectations can be immediately approximated by sample averages.[4] In fact, this formulation allows the derivation of several recent behavior-agnostic OPE estimators in a unified manner (Nachum et al., 2019a; Uehara et al., 2019; Zhang et al., 2020a; Nachum and Dai, 2020).

3 CoinDICE

We now develop a new approach to obtaining confidence intervals for OPE. The algorithm, *COnfidence INterval stationary DIstribution Correction Estimation (CoinDICE)*, is derived by combining function space embedding and the previously described Q-LP.

3.1 Function Space Embedding of Constraints

Both the primal and dual forms of the Q-LP contain $|S|\,|A|$ constraints that involve expectations over state transition probabilities. Working directly with these constraints quickly becomes computationally and statistically prohibitive when $|S|\,|A|$ is large or infinite, as with standard LP approaches (De Farias and Van Roy, 2003). Instead, we consider a relaxation that embeds the constraints in a function space:

$$\tilde{\rho}_{\pi} := \max_{d: S\times A \rightarrow \mathbb{R}_+} \mathbb{E}_{d}\left[r\left(s, a\right)\right] \quad \text{s.t.} \ \left\langle \phi, d\right\rangle = \left\langle \phi, \left(1-\gamma\right)\mu_0\pi + \gamma \cdot \mathcal{P}_*^{\pi} d\right\rangle, \quad (5)$$

where $\phi : S \times A \rightarrow \Omega^p \subset \mathbb{R}^p$ is a feature map, and $\left\langle \phi, d\right\rangle := \int \phi\left(s, a\right) d\left(s, a\right) dsda$. By projecting the constraints onto a function space with feature mapping ϕ, we can reduce the number of constraints from $|S|\,|A|$ to p. Note that p may still be infinite. The constraint in (5) can be written as *generalized estimating equations* (Qin and Lawless, 1994; Lam and Zhou, 2017) for the correction ratio $\tau\left(s, a\right)$ over augmented samples $x := \left(s_0, a_0, s, a, r, s', a'\right)$ with $\left(s_0, a_0\right) \sim \mu_0\pi$, $\left(s, a, r, s'\right) \sim d^{\mathcal{D}}$, and $a' \sim \pi(\cdot|s')$,

$$\left\langle \phi, d\right\rangle = \left\langle \phi, \left(1-\gamma\right)\mu_0\pi + \gamma \cdot \mathcal{P}_*^{\pi} d\right\rangle \quad \Leftrightarrow \quad \mathbb{E}_x\left[\Delta\left(x; \tau, \phi\right)\right] = 0, \quad (6)$$

where $\Delta\left(x; \tau, \phi\right) := \left(1-\gamma\right)\phi\left(s_0, a_0\right) + \tau\left(s, a\right)\left(\gamma\phi\left(s', a'\right) - \phi\left(s, a\right)\right)$. The corresponding Lagrangian is

$$\tilde{\rho}_{\pi} = \max_{\tau: S\times A \rightarrow \mathbb{R}_+} \min_{\beta \in \mathbb{R}^p} \mathbb{E}_{d^{\mathcal{D}}}\left[\tau \cdot r\left(s, a\right)\right] + \left\langle \beta, \mathbb{E}_{d^{\mathcal{D}}}\left[\Delta\left(x; \tau, \phi\right)\right]\right\rangle. \quad (7)$$

This embedding approach for the dual Q-LP is closely related to approximation methods for the standard state-value LP (De Farias and Van Roy, 2003; Pazis and Parr, 2011; Lakshminarayanan et al., 2017). The gap between the solutions to (5) and the original dual LP (3) depends on the expressiveness of the feature mapping ϕ. Before stating a theorem that quantifies the error, we first offer a few examples to provide intuition for the role played by ϕ.

Example (Indicator functions): Suppose $p = |S|\,|A|$ is finite and $\phi = \left[\delta_{s,a}\right]_{(s,a)\in S\times A}$, where $\delta_{s,a} \in \{0, 1\}^p$ with $\delta_{s,a} = 1$ at position (s, a) and 0 otherwise. Plugging this feature mapping into (5), we recover the original dual Q-LP (3).

Example (Full-rank basis): Suppose $\Phi \in \mathbb{R}^{p\times p}$ is a full-rank matrix with $p = |S|\,|A|$; furthermore, $\phi(s, a) = \Phi((s, a), \cdot)^{\top}$. Although the constraints in (5) and (3) are different, their solutions are identical. This can be verified by the Lagrangian in Appendix A.

Example (RKHS function mappings): Suppose $\phi\left(s, a\right) := k\left(\left(s, a\right), \cdot\right) \in \mathbb{R}^p$ with $p = \infty$, which forms a reproducing kernel Hilbert space (RKHS) $\mathcal{H}_k$. The LHS and RHS in the constraint of (5) are the kernel embeddings of $d\left(s, a\right)$ and $\left(1-\gamma\right)\mu_0\pi\left(s, a\right) + \gamma \cdot \mathcal{P}_*^{\pi} d\left(s, a\right)$ respectively. The constraint in (5) can then be understood as as a form of distribution matching by comparing

[4] We assume one can sample initial states from μ_0, an assumption that often holds in practice. Then, the data in $\mathcal{D}$ can be treated as being augmented as $\left(s_0, a_0, s, a, r, s', a'\right)$ with $a_0 \sim \pi\left(a|s_0\right), a' \sim \pi\left(a|s'\right)$.

kernel embeddings, rather than element-wise matching as in (3). If the kernel function $k(\cdot,\cdot)$ is characteristic, the embeddings of two distributions will match if and only if the distributions are identical almost surely (Sriperumbudur et al., 2011).

Theorem 1 (Approximation error) *Suppose the constant function $\mathbf{1} \in \mathcal{F}_\phi := \operatorname{span}\{\phi\}$. Then,*

$$0 \leqslant \tilde{\rho}_\pi - \rho_\pi \leqslant 2 \min_\beta \|Q^\pi - \langle \beta, \phi \rangle\|_\infty,$$

where Q^π is the fixed-point solution to the Bellman equation $Q(s,a) = R(s,a) + \gamma \mathcal{P}^\pi Q(s,a)$.

Please refer to Appendix A for the proof. The condition $\mathbf{1} \in \mathcal{F}_\phi$ is standard and is trivial to satisfy. Although the approximation error relies on $\|\cdot\|_\infty$, a sharper bound that relies on a norm taking the state-action distribution into account can also be obtained (De Farias and Van Roy, 2003). We focus on characterizing the uncertainty due to sampling in this paper, so for ease of exposition we will consider a setting where ϕ is sufficiently expressive to make the approximation error zero. If desired, the approximation error in Theorem 1 can be included in the analysis.

Note that, compared to using a characteristic kernel to ensure injectivity for the RKHS embeddings over all distributions (and thus guaranteeing arbitrarily small approximation error), Theorem 1 only requires that Q^π be represented in $\mathcal{F}_\phi$, which is a much weaker condition. In practice, one may also learn the feature mapping ϕ for the projection jointly.

3.2 Off-policy Confidence Interval Estimation

By introducing the function space embedding of the constraints in (5), we have transformed the original point-wise constraints in the Q-LP to generalized estimating equations. This paves the way to applying the generalized empirical likelihood (EL) (Owen, 2001; Broniatowski and Keziou, 2012; Bertail et al., 2014; Duchi et al., 2016) method to estimate a confidence interval on policy value.

Recall that, given a convex, lower-semicontinuous function $f: \mathbb{R}_+ \to \mathbb{R}$ satisfying $f(1) = 0$, the f-divergence between densities p and q on $\mathbb{R}$ is defined as $D_f(P||Q) := \int Q(dx) f\left(\frac{dP(x)}{dQ(x)}\right) dx$.

Given an f-divergence, we propose our main confidence interval estimate based on the following confidence set $C^f_{n,\xi} \subset \mathbb{R}$:

$$C^f_{n,\xi} := \left\{ \tilde{\rho}_\pi(w) = \max_{\tau \geqslant 0} \mathbb{E}_w[\tau \cdot r] \,\middle|\, w \in \mathcal{K}_f, \mathbb{E}_w[\Delta(x;\tau,\phi)] = 0 \right\} \text{ with } \mathcal{K}_f := \left\{ \begin{array}{l} w \in \mathcal{P}^{n-1}(\widehat{p}_n), \\ D_f(w||\widehat{p}_n) \leqslant \frac{\xi}{n} \end{array} \right\}, \tag{8}$$

where $\mathcal{P}^{n-1}(\widehat{p}_n)$ denotes the n-simplex on the support of $\widehat{p}_n$, the empirical distribution over $\mathcal{D}$. It is easy to verify that this set $C^f_{n,\xi} \subset \mathbb{R}$ is convex, since $\tilde{\rho}_\pi(w)$ is a convex function over a convex feasible set. Thus, $C^f_{n,\xi}$ is an interval. In fact, $C^f_{n,\xi}$ is the image of the policy value $\tilde{\rho}_\pi$ on a bounded (in f-divergence) perturbation to w in the neighborhood of the empirical distribution $\widehat{p}_n$.

Intuitively, the confidence interval $C^f_{n,\xi}$ possesses a close relationship to bootstrap estimators. In vanilla bootstrap, one constructs a set of empirical distributions $\{w^i\}_{i=1}^m$ by resampling from the dataset $\mathcal{D}$. Such subsamples are used to form the empirical distribution on $\{\tilde{\rho}(w^i)\}_{i=1}^m$, which provides population statistics for confidence interval estimation. However, this procedure is computationally very expensive, involving m separate optimizations. By contrast, our proposed estimator $C^f_{n,\xi}$ exploits the asymptotic properties of the statistic $\tilde{\rho}_\pi(w)$ to derive a target confidence interval by solving only *two* optimization problems (Section 3.3), a dramatic savings in computational cost.

Before introducing the algorithm for computing $C^f_{n,\xi}$, we establish the first key result that, by choosing $\xi = \chi^{2,1-\alpha}_{(1)}$, $C^f_{n,\xi}$ is asymptotically a $(1-\alpha)$-confidence interval on the policy value, where $\chi^{2,1-\alpha}_{(1)}$ is the $(1-\alpha)$-quantile of the χ^2-distribution with 1 degree of freedom.

Theorem 2 (Informal asymptotic coverage) *Under some mild conditions, if $\mathcal{D}$ contains* i.i.d. *samples and the optimal solution to the Lagrangian of* (5) *is unique, we have*

$$\lim_{n\to\infty} \mathbb{P}\left(\rho_\pi \in C^f_{n,\xi}\right) = \mathbb{P}\left(\chi^2_{(1)} \leqslant \xi\right). \tag{9}$$

Thus, $C^f_{n,\chi^{2,1-\alpha}_{(1)}}$ is an asymptotic $(1-\alpha)$-confidence interval of the value of the policy π.

Please refer to Appendix E.1 for the precise statement and proof of Theorem 2.

Theorem 2 generalizes the result in Duchi et al. (2016) to statistics with generalized estimating equations, maintaining the 1 degree of freedom in the asymptotic $\chi^2_{(1)}$-distribution. One may also apply existing results for EL with generalized estimating equations (e.g., Lam and Zhou, 2017), but these would lead to a limiting distribution of $\chi^2_{(m)}$ with $m \gg 1$ degrees of freedom, resulting in a much looser confidence interval estimate than Theorem 2.

Note that Theorem 2 can also be specialized to multi-armed contextual bandits to achieve a tighter confidence interval estimate in this special case. In particular, for contextual bandits, the stationary distribution constraint in (5), $\mathbb{E}_w\left[\Delta\left(x;\tau,\phi\right)\right]=0$, is no longer needed, and can be replaced by $\mathbb{E}_w\left[\tau-1\right]=0$. Then by the same technique used for MDPs, we can obtain a confidence interval estimate for offline contextual bandits; see details in Appendix C. Interestingly, the resulting confidence interval estimate not only has the same asymptotic coverage as previous work (Karampatziakis et al., 2019), but is also simpler and computationally more efficient.

3.3 Computing the Confidence Interval

Now we provide a distributional robust optimization view of the upper and lower bounds of $C^f_{n,\xi}$.

Theorem 3 (Upper and lower confidence bounds) *Denote the upper and lower confidence bounds of $C^f_{n,\xi}$ by u_n and l_n, respectively:*

$$\left[l_n,u_n\right] = \left[\min_{w\in\mathcal{K}_f}\min_{\beta\in\mathbb{R}^p}\max_{\tau\geqslant 0}\mathbb{E}_w\left[\ell\left(x;\tau,\beta\right)\right],\quad \max_{w\in\mathcal{K}_f}\max_{\tau\geqslant 0}\min_{\beta\in\mathbb{R}^p}\mathbb{E}_w\left[\ell\left(x;\tau,\beta\right)\right]\right], \tag{10}$$

$$= \left[\min_{\beta\in\mathbb{R}^p}\max_{\tau\geqslant 0}\min_{w\in\mathcal{K}_f}\mathbb{E}_w\left[\ell\left(x;\tau,\beta\right)\right],\quad \max_{\tau\geqslant 0}\min_{\beta\in\mathbb{R}^p}\max_{w\in\mathcal{K}_f}\mathbb{E}_w\left[\ell\left(x;\tau,\beta\right)\right]\right], \tag{11}$$

where $\ell\left(x;\tau,\beta\right):=\tau\cdot r+\beta^\top\Delta\left(x;\tau,\phi\right)$. For any $(\tau,\beta,\lambda,\eta)$ that satisfies the constraints in (11), the optimal weights for the upper and lower confidence bounds are

$$w_l = f_*'\left(\frac{\eta-\ell\left(x;\tau,\beta\right)}{\lambda}\right)\quad \text{and} \quad w_u = f_*'\left(\frac{\ell\left(x;\tau,\beta\right)-\eta}{\lambda}\right). \tag{12}$$

respectively. Therefore, the confidence bounds can be simplified as:

$$\begin{bmatrix} l_n \\ u_n \end{bmatrix} = \begin{bmatrix} \min_\beta\max_{\tau\geqslant 0,\lambda\geqslant 0,\eta}\mathbb{E}_\mathcal{D}\left[-\lambda f_*\left(\frac{\eta-\ell(x;\tau,\beta)}{\lambda}\right)+\eta-\lambda\frac{\xi}{n}\right] \\ \max_{\tau\geqslant 0}\min_{\beta,\lambda\geqslant 0,\eta}\mathbb{E}_\mathcal{D}\left[\lambda f_*\left(\frac{\ell(x;\tau,\beta)-\eta}{\lambda}\right)+\eta+\lambda\frac{\xi}{n}\right] \end{bmatrix}. \tag{13}$$

The proof of this result relies on Lagrangian duality and the convexity and concavity of the optimization; it may be found in full detail in Appendix D.1.

As we can see in Theorem 3, by exploiting strong duality properties to move w into the inner most optimizations in (11), the obtained optimization (11) is the distributional robust optimization extenion of the saddle-point problem. The closed-form reweighting scheme is demonstrated in (12). For particular f-divergences, such as the KL- and 2-power divergences, for a fixed (β,τ), the optimal η can be easily computed and the weights w recovered in closed-form. For example, by using $KL\left(w||\widehat{p}_n\right)$, (12) can be used to obtain the updates

$$w_l\left(x\right)=\exp\left(\frac{\eta_l-\ell\left(x;\tau,\beta\right)}{\lambda}\right),\quad w_u\left(x\right)=\exp\left(\frac{\ell\left(x;\tau,\beta\right)-\eta_u}{\lambda}\right), \tag{14}$$

where η_l and η_u provide the normalizing constants. (For closed-form updates of w w.r.t. other f-divergences, please refer to Appendix D.2.) Plug the closed-form of optimal weights into (11), this greatly simplifies the optimization over the data perturbations yielding (13), and estabilishes the connection to the prioritized experiences replay (Schaul et al., 2016), where both reweight the experience data according to their loss, but with different reweighting schemes.

Note that it is straightforward to check that the estimator for u_n in (13) is nonconvex-concave and the estimator for l_n in (13) is nonconcave-convex. Therefore, one could alternatively apply stochastic gradient descent-ascent (SGDA) for to solve (13) and benefit from attractive finite-step convergence guarantees (Lin et al., 2019).

Remark (Practical considerations): As also observed in Namkoong and Duchi (2016), SGDA for (13) could potentially suffer from high variance in both the objective and gradients when λ approaches 0. In Appendix D.3, we exploit several properties of (11), which leads to a computational efficient algorithm, to overcome the numerical issue. Please refer to Appendix D.3 for the details of Algorithm 1 and the practical considerations.

Remark (Joint learning for feature embeddings): The proposed framework also allows for the possibility to learn the features for constraint projection. In particular, consider $\zeta(\cdot,\cdot) := \beta^\top \phi(\cdot,\cdot) : S \times A \to \mathbb{R}$. Note that we could treat the combination $\beta^\top \phi(s,a)$ together as the Lagrange multiplier function for the original Q-LP with *infinitely* many constraints, hence both β and $\phi(\cdot,\cdot)$ could be updated jointly. Although the conditions for asymptotic coverage no longer hold, the finite-sample correction results of the next section are still applicable. This might offer an interesting way to reduce the approximation error introduced by inappropriate feature embeddings of the constraints, while still maintaining calibrated confidence intervals.

4 Finite-sample Analysis

Theorem 2 establishes the asymptotic $(1-\alpha)$-coverage of the confidence interval estimates produced by CoinDICE, ignoring higher-order error terms that vanish as sample size $n \to \infty$. In practice, however, n is always finite, so it is important to quantify these higher-order terms. This section addresses this problem, and presents a finite-sample bound for the estimate of CoinDICE. In the following, we let $\mathcal{F}_\tau$ and $\mathcal{F}_\beta$ be the function classes of τ and β used by CoinDICE.

Theorem 4 (Informal finite-sample correction) *Denote by $d_{\mathcal{F}_\tau}$ and $d_{\mathcal{F}_\beta}$ the finite* `VC`*-dimension of $\mathcal{F}_\tau$ and $\mathcal{F}_\beta$, respectively. Under some mild conditions, when D_f is χ^2-divergence, we have*

$$\mathbb{P}\left(\rho_\pi \in [l_n - \kappa_n, u_n + \kappa_n]\right) \geqslant 1 - 12\exp\left(c_1 + 2\left(d_{\mathcal{F}_\tau} + d_{\mathcal{F}_\beta} - 1\right)\log n - \frac{\xi}{18}\right),$$

where $c_1 = 2c + \log d_{\mathcal{F}_\tau} + \log d_{\mathcal{F}_\beta} + \left(d_{\mathcal{F}_\tau} + d_{\mathcal{F}_\beta} - 1\right)$, $\kappa_n = \frac{11M\xi}{6n} + 2\frac{C_\ell M}{n}\left(1 + 2\sqrt{\frac{\xi}{9n}}\right)$, and (c, M, C_ℓ) are univeral constants.

The precise statement and detailed proof of Theorem 4 can be found in Appendix E.2. The proof relies on empirical Bernstein bounds with a careful analysis of the variance term. Compared to the vanilla sample complexity of $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$, we achieve a faster rate of $\mathcal{O}\left(\frac{1}{n}\right)$ without any additional assumptions on the noise or curvature conditions. The tight sample complexity in Theorem 4 implies that one can construct the $(1-\alpha)$-finite sample confidence interval by optimizing (11) with $\xi = 18\left(\log\frac{\alpha}{12} - c_1 - 2\left(d_{\mathcal{F}_\tau} + d_{\mathcal{F}_\beta} - 1\right)\log n\right)$, and composing with κ_n. However, we observe that this bound can be conservative compared to the asymptotic confidence interval in Theorem 2. Therefore, we will evaluate the asymptotic version of CoinDICE based on Theorem 2 in the experiment.

The conservativeness arises from the use of a union bound. However, we conjecture that the rate is optimal up to a constant. We exploit the VC dimension due to its generality. In fact, the bound can be improved by considering a data-dependent measure, *e.g.*, Rademacher complexity, or by some function class dependent measure, *e.g.*, function norm in RKHS, for specific function approximators.

5 Optimism vs. Pessimism Principle

CoinDICE provide both upper and lower bounds of the target policy's estimated value, which paves the path for applying the principle of optimism (Lattimore and Szepesvári, 2020) or pessimism (Swaminathan and Joachims, 2015) in the face of uncertainty for policy optimization in different learning settings.

Optimism in the face of uncertainty. Optimism in the face of uncertainty leads to *risk-seeking* algorithms, which can be used to balance the exploration/exploitation trade-off. Conceptually, they always treat the environment as the best plausibly possible. This principle has been successfully applied to stochastic bandit problems, leading to many instantiations of UCB algorithms (Lattimore and Szepesvári, 2020). In each round, an action is selected according to the upper confidence bound, and the obtained reward will be used to refine the confidence bound iteratively. When applied to MDPs, this principle inspires many optimistic model-based (Bartlett and Mendelson, 2002; Auer et al., 2009; Strehl et al., 2009; Szita and Szepesvari, 2010; Dann et al., 2017), value-based (Jin et al., 2018), and policy-based algorithms (Cai et al., 2019). Most of these algorithms are not compatible with function approximators.

We can also implement the optimism principle by optimizing the upper bound in CoinDICE iteratively, *i.e.*, $\max_\pi u_{\mathcal{D}}(\pi)$. In t-th iteration, we calculate the gradient of $u_{\mathcal{D}}(\pi^t)$, *i.e.*, $\nabla_\pi u_{\mathcal{D}}(\pi^t)$, based on the existing dataset $\mathcal{D}_t$, then, the policy π_t will be updated by (natural) policy gradient and samples will be collected through the updated policy π_{t+1}. Please refer to Appendix F for the gradient computation and algorithm details.

Pessimism in the face of uncertainty. In offline reinforcement learning (Lange et al., 2012; Fujimoto et al., 2019; Wu et al., 2019; Nachum et al., 2019b), only a fixed set of data from behavior policies is given, a safe optimization criterion is to maximize the worst-case performance among a set of statistically plausible models (Laroche et al., 2019; Kumar et al., 2019; Yu et al., 2020). In contrast to the previous case of online exploration, this is a pessimism principle (Cohen and Hutter, 2020; Buckman et al., 2020) or counterfactual risk minimization (Swaminathan and Joachims, 2015), and highly related to robust MDP (Iyengar, 2005; Nilim and El Ghaoui, 2005; Tamar et al., 2013; Chow et al., 2015).

Different from most of the existing methods where the worst-case performance is characterized by model-based perturbation or ensemble, the proposed CoinDICE provides a lower bound to implement the pessimism principle, *i.e.*, $\max_\pi l_{\mathcal{D}}(\pi)$. Conceptually, we apply the (natural) policy gradient w.r.t. $l_{\mathcal{D}}(\pi^t)$ to update the policy iteratively. Since we are dealing with policy optimization in the offline setting, the dataset $\mathcal{D}$ keeps unchanged. Please refer to Appendix F for the algorithm details.

6 Related Work

Off-policy estimation has been extensively studied in the literature, given its practical importance. Most existing methods are based on the core idea of mportance reweighting to correct for distribution mismatches between the target policy and the off-policy data (Precup et al., 2000; Bottou et al., 2013; Li et al., 2015; Xie et al., 2019). Unfortunately, when applied naively, importance reweighting can result in an excessively high variance, which is known as the "curse of horizon" (Liu et al., 2018). To avoid this drawback, there has been rapidly growing interest in estimating the correction ratio of the *stationary* distribution (e.g., Liu et al., 2018; Nachum et al., 2019a; Uehara et al., 2019; Liu et al., 2019; Zhang et al., 2020a,b). This work is along the same line and thus applicable in long-horizon problems. Other off-policy approaches are also possible, notably model-based (e.g., Fonteneau et al., 2013) and doubly robust methods (Jiang and Li, 2016; Thomas and Brunskill, 2016; Tang et al., 2020; Uehara et al., 2019). These techniques can potentially be combined with our algorithm, which we leave for future investigation.

While most OPE works focus on obtaining accurate *point* estimates, several authors provide ways to quantify the amount of uncertainty in the OPE estimates. In particular, confidence bounds have been developed using the central limit theorem (Bottou et al., 2013), concentration inequalities (Thomas et al., 2015b; Kuzborskij et al., 2020), and nonparametric methods such as the bootstrap (Thomas et al., 2015a; Hanna et al., 2017). In contrast to these works, the CoinDICE is asymptotically pivotal, meaning that there are no hidden quantities we need to estimate, which is based on correcting for the stationary distribution in the *behavior-agnostic* setting, thus avoiding the curse of horizon and broadening the application of the uncertainty estimator. Recently, Jiang and Huang (2020) provide confidence intervals for OPE, but focus on the intervals determined by the *approximation error* induced by a function approximator, while our confidence intervals quantify *statistical error*.

Empirical likelihood (Owen, 2001) is a powerful tool with many applications in statistical inference like econometrics (Chen et al., 2018), and more recently in distributionally robust optimization (Duchi et al., 2016; Lam and Zhou, 2017). EL-based confidence intervals can be used to guide exploration in multiarmed bandits (Honda and Takemura, 2010; Cappé et al., 2013), and for OPE for bandit (Faury et al., 2020; Karampatziakis et al., 2019) and RL (Kallus and Uehara, 2019b). While the work of Kallus and Uehara (2019b) is also based on EL, it differs from the present work in two important ways. First, their focus is on developing an asymptotically efficient OPE *point* estimate, not confidence intervals. Second, they solve for timestep-dependent weights, whereas we only need to solve for timestep-*independent* weights from a system of moment matching equations induced by an underlying ergodic Markov chain.

7 Experiments

We now evaluate the empirical performance of CoinDICE, comparing it to a number of existing confidence interval estimators for OPE based on concentration inequalities. Specifically, given a dataset of logged trajectories, we first use weighted step-wise importance sampling (Precup et al., 2000) to calculate a separate estimate of the target policy value for each trajectory. Then given such a finite sample of estimates, we then use the empirical *Bernstein* inequality (Thomas et al., 2015b) to derive high-confidence lower and upper bounds for the true value. Alternatively, one may also use *Student's t-test* or Efron's bias corrected and accelerated *bootstrap* (Thomas et al., 2015a).

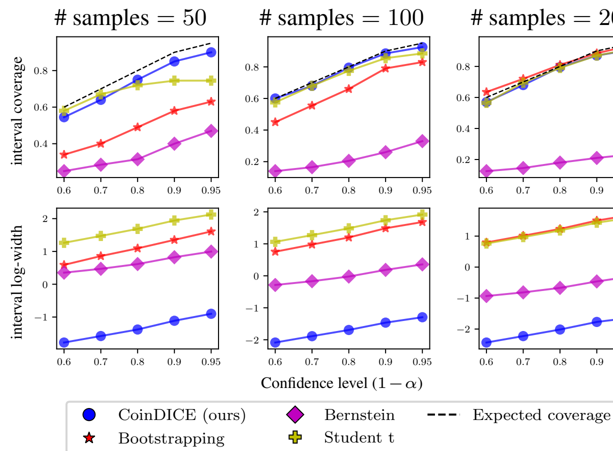

Figure 1: Results of CoinDICE and baseline methods on a simple two-armed bandit. We plot empirical coverage and median log-width (y-axes) of intervals evaluated at a number of desired confidence levels (x-axis), as measured over 200 random trials. We find that CoinDICE achieves more accurate coverage and narrower intervals compared to the baseline confidence interval estimation methods.

We begin with a simple bandit setting, devising a two-armed bandit problem with stochastic payoffs. We define the target policy as a near-optimal policy, which chooses the optimal arm with probability 0.95. We collect off-policy data using a behavior policy which chooses the optimal arm with probability of only 0.55. Our results are presented in Figure 1. We plot the empirical coverage and width of the estimated intervals across different confidence levels. More specifically, each data point in Figure 1 is the result of 200 experiments. In each experiment, we randomly sample a dataset and then compute a confidence interval. The *interval coverage* is then computed as the proportion of intervals out of 200 that contain the true value of the target policy. The *interval* $\log$-*width* is the median of the log of the width of the 200 computed intervals. Figure 1 shows that the intervals produced by CoinDICE achieve an empirical coverage close to the intended coverage. In this simple bandit setting, the coverages of Student's t and bootstrapping are also close to correct, although they suffer more in the low-data regime. Notably, the width of the intervals produced by CoinDICE are especially narrow while maintaining accurate coverage.

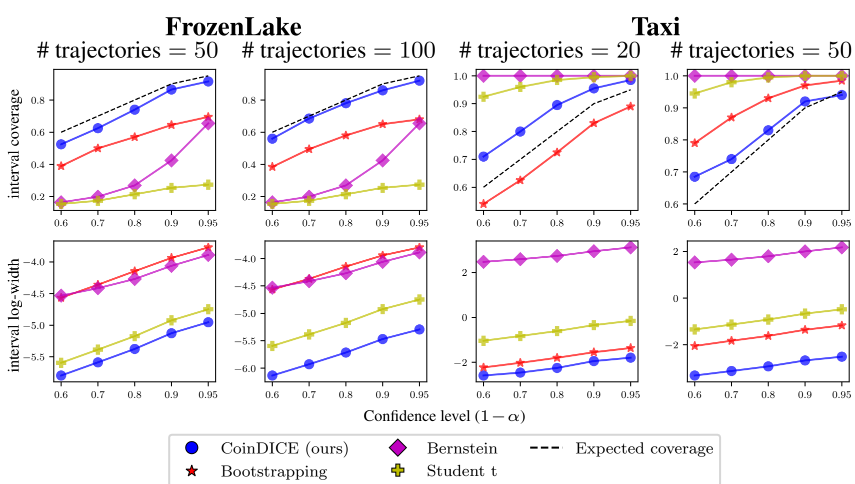

Figure 2: Results of CoinDICE and baseline methods on an infinite-horizon version of FrozenLake and Taxi. In FrozenLake, each dataset consists of trajectories of length 100; in Taxi, each dataset consists of trajectories of length 500.

We now turn to more complicated MDP environments. We use FrozenLake (Brockman et al., 2016), a highly stochastic gridworld environment, and Taxi (Dietterich, 1998), an environment with a moderate state space of 2 000 elements. As in (Liu et al., 2018), we modify these environments to be infinite horizon by randomly resetting the state upon termination. The discount factor is $\gamma = 0.99$. The target policy is taken to be a near-optimal one, while the behavior policy is highly suboptimal. The behavior policy in FrozenLake is the optimal policy with 0.2 white noise, which reduces the policy value dramatically, from 0.74 to 0.24. For the behavior policies in Taxi and Reacher, we follow the same experiment setting for constructing the behavior policies to collect data as in (Nachum et al., 2019a; Liu et al., 2018).

We follow the same evaluation protocol as in the bandit setting, measuring empirical interval coverage and log-width over 200 experimental trials for various dataset sizes and confidence levels. Results are shown in Figure 2. We find a similar conclusion that CoinDICE consistently achieves more accurate coverage and smaller widths than baselines. Notably, the baseline methods' accuracy suffers more significantly compared to the simpler bandit setting described earlier.

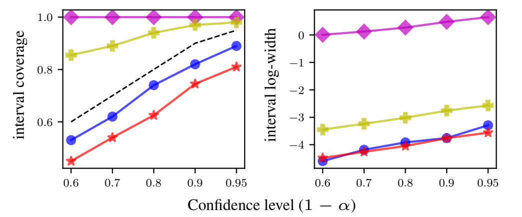

Figure 3: Results of CoinDICE and baseline methods on Reacher (Brockman et al., 2016; Todorov et al., 2012), using 25 trajectories of length 100. Colors and markers are as defined in the legends of previous figures.

Lastly, we evaluate CoinDICE on Reacher (Brockman et al., 2016; Todorov et al., 2012), a continuous control environment. In this setting, we use a one-hidden-layer neural network with ReLU activations. Results are shown in Figure 3. To account for the approximation error of the used neural network, we measure the coverage of CoinDICE with respect to a true value computed as the median of a large ensemble of neural networks trained on the off-policy data. To keep the comparison fair, we measure the coverage of the IS-based baselines with respect to a true value computed as the median of a large number of IS-based point estimates. The results show similar conclusions as before: CoinDICE achieves more accurate coverage than the IS-based methods. Still, we see that CoinDICE coverage suffers in this regime, likely due to optimization difficulties. If the optimum of the Lagrangian is only approximately found, the empirical coverage will inevitably be inexact.

8 Conclusion

In this paper, we have developed CoinDICE, a novel and efficient confidence interval estimator applicable to the *behavior-agnostic offline* setting. The algorithm builds on a few technical components, including a new feature embedded Q-LP, and a generalized empirical likelihood approach to confidence interval estimation. We analyzed the asymptotic coverage of CoinDICE's estimate, and provided an inite-sample bound. On a variety of off-policy benchmarks we empirically compared the new algorithm with several strong baselines and found it to be superior to them.

Broader Impact

This research is fundamental and targets a broad question in reinforcement learning. The ability to reliably assess uncertainty in off-policy evaluation would have significant positive benefits for safety-critical applications of reinforcement learning. Inaccurate uncertainty estimates create the danger of misleading decision makers and could lead to detrimental consequences. However, our primary goal is to improve these estimators and reduce the ultimate risk of deploying reinforcement-learned systems. The techniques are general and do not otherwise target any specific application area.

Acknowledgements

We thank Hanjun Dai, Mengjiao Yang and other members of the Google Brain team for helpful discussions. Csaba Szepesvári gratefully acknowledges funding from the Canada CIFAR AI Chairs Program, Amii and NSERC.

References

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

Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems*, pages 8139–8148, 2019.

P. Auer. Using upper confidence bounds for online learning. In *Proc. 41st Annual Symposium on Foundations of Computer Science*, pages 270–279. IEEE Computer Society Press, Los Alamitos, CA, 2000.

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. In *Advances in neural information processing systems*, pages 89–96, 2009.

P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Patrice Bertail, Emmanuelle Gautherat, and Hugo Harari-Kermadec. Empirical φ_*-divergence minimizers for Hadamard differentiable functionals. In *Topics in Nonparametric Statistics*, pages 21–32. Springer, 2014.

Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *The Journal of Machine Learning Research*, 14(1):3207–3260, 2013.

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

Michel Broniatowski and Amor Keziou. Divergences and duality for estimation and test under moment condition models. *Journal of Statistical Planning and Inference*, 142(9):2554–2573, 2012.

Jacob Buckman, Carles Gelada, and Marc G Bellemare. The importance of pessimism in fixed-dataset policy optimization. *arXiv preprint arXiv:2009.06799*, 2020.

Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. *arXiv preprint arXiv:1912.05830*, 2019.

Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, Gilles Stoltz, et al. Kullback-Leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, 41(3):1516–1541, 2013.

Minmin Chen, Alex Beutel, Paul Covington, Sagar Jain, Francois Belletti, and Ed H Chi. Top-k off-policy correction for a REINFORCE recommender system. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 456–464, 2019.

Xiaohong Chen, Timothy M. Christensen, and Elie Tamer. Monte Carlo confidence sets for identified sets. *Econometrica*, 86(6):1965–2018, 2018.

Yinlam Chow, Aviv Tamar, Shie Mannor, and Marco Pavone. Risk-sensitive and robust decision-making: a cvar optimization approach. In *Advances in Neural Information Processing Systems*, pages 1522–1530, 2015.

Michael K Cohen and Marcus Hutter. Pessimism about unknown unknowns inspires conservatism. In *Conference on Learning Theory*, pages 1344–1373. PMLR, 2020.

Bo Dai, Albert Shaw, Lihong Li, Lin Xiao, Niao He, Zhen Liu, Jianshu Chen, and Le Song. SBEED: Convergent reinforcement learning with nonlinear function approximation. *CoRR*, abs/1712.10285, 2017.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 5713–5723, 2017.

Daniela Pucci De Farias and Benjamin Van Roy. The linear programming approach to approximate dynamic programming. *Operations research*, 51(6):850–865, 2003.

Thomas G. Dietterich. The MAXQ method for hierarchical reinforcement learning. In *Proc. Intl. Conf. Machine Learning*, pages 118–126. Morgan Kaufmann, San Francisco, CA, 1998.

John Duchi, Peter Glynn, and Hongseok Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *arXiv preprint arXiv:1610.03425*, 2016.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on Machine Learning*, pages 1097–1104, 2011. CoRR abs/1103.4601.

Ivar Ekeland and Roger Temam. *Convex analysis and variational problems*, volume 28. SIAM, 1999.

Louis Faury, Ugo Tanielian, Flavian Vasile, Elena Smirnova, and Elvis Dohmatob. Distributionally Robust Counterfactual Risk Minimization. In *AAAI*, 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.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, pages 2052–2062, 2019.

Omer Gottesman, Fredrik Johansson, Joshua Meier, Jack Dent, Donghun Lee, Srivatsan Srinivasan, Linying Zhang, Yi Ding, David Wihl, Xuefeng Peng, Jiayu Yao, Isaac Lage, Christopher Mosch, Li wei H. Lehman, Matthieu Komorowski, Matthieu Komorowski, Aldo Faisal, Leo Anthony Celi, David Sontag, and Finale Doshi-Velez. Evaluating reinforcement learning algorithms in observational health settings, 2018. arXiv:1805.12298.

Josiah P. Hanna, Peter Stone, and Scott Niekum. Bootstrapping with models: Confidence intervals for off-policy evaluation. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pages 4933–4934, 2017.

Junya Honda and Akimichi Takemura. An asymptotically optimal bandit algorithm for bounded support models. In *COLT*, pages 67–79, 2010.

Garud N Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2): 257–280, 2005.

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

Nan Jiang and Lihong Li. Doubly robust off-policy evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 652–661, 2016.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873, 2018.

Nathan Kallus and Masatoshi Uehara. Double reinforcement learning for efficient off-policy evaluation in Markov decision processes. *arXiv preprint arXiv:1908.08526*, 2019a.

Nathan Kallus and Masatoshi Uehara. Intrinsically efficient, stable, and bounded off-policy evaluation for reinforcement learning. In *Advances in Neural Information Processing Systems 32*, pages 3320–3329, 2019b.

Nikos Karampatziakis, John Langford, and Paul Mineiro. Empirical likelihood for contextual bandits. *arXiv preprint arXiv:1906.03323*, 2019.

Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. Stabilizing off-policy Q-learning via bootstrapping error reduction. In *Advances in Neural Information Processing Systems*, pages 11784–11794, 2019.

Ilja Kuzborskij, Claire Vernade, András György, Csaba Szepesvári Confident off-policy evaluation and selection through self-normalized importance weighting. *arXiv preprint arXiv:2006.10460*, 2020.

Chandrashekar Lakshminarayanan, Shalabh Bhatnagar, and Csaba Szepesvari. A linearly relaxed approximate linear program for Markov decision processes. *arXiv preprint arXiv:1704.02544*, 2017.

Henry Lam and Enlu Zhou. The empirical likelihood approach to quantifying uncertainty in sample average approximation. *Operations Research Letters*, 45(4):301–307, 2017.

Sascha Lange, Thomas Gabel, and Martin Riedmiller. Batch reinforcement learning. In *Reinforcement learning*, pages 45–73. Springer, 2012.

Romain Laroche, Paul Trichelair, and Remi Tachet Des Combes. Safe policy improvement with baseline bootstrapping. In *International Conference on Machine Learning*, pages 3652–3661. PMLR, 2019.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Alessandro Lazaric, Mohammad Ghavamzadeh, and Rémi Munos. Finite-sample analysis of least-squares policy iteration. *Journal of Machine Learning Research*, 13(Oct):3041–3074, 2012.

Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 297–306. ACM, 2011.

Lihong Li, Rémi Munos, and Csaba Szepesvári. Toward minimax off-policy value estimation. pages 608–616, 2015.

Tianyi Lin, Chi Jin, and Michael I. Jordan. On gradient descent ascent for nonconvex-concave minimax problems. *CoRR*, abs/1906.00331, 2019.

Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 5356–5366. Curran Associates, Inc., 2018.

Yao Liu, Pierre-Luc Bacon, and Emma Brunskill. Understanding the curse of horizon in off-policy evaluation via conditional importance sampling, 2019. arXiv:1910.06508.

Travis Mandel, Yun-En Liu, Sergey Levine, Emma Brunskill, and Zoran Popovic. Offline policy evaluation across representations with applications to educational games. 2014.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

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

Ofir Nachum and Bo Dai. Reinforcement learning via Fenchel-Rockafellar duality. *arXiv preprint arXiv:2001.01866*, 2020.

Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. DualDICE: Behavior-agnostic estimation of discounted stationary distribution corrections. pages 2315–2325, 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.

Hongseok Namkoong and John C Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *Advances in neural information processing systems*, pages 2208–2216, 2016.

Hongseok Namkoong and John C Duchi. Variance-based regularization with convex objectives. In *Advances in neural information processing systems*, pages 2971–2980, 2017.

Arnab Nilim and Laurent El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

Art B Owen. *Empirical likelihood*. Chapman and Hall/CRC, 2001.

Jason Pazis and Ronald Parr. Non-parametric approximate linear programming for MDPs. In *AAAI*, 2011.

Doina Precup, R. S. Sutton, and S. Singh. Eligibility traces for off-policy policy evaluation. In *Proc. Intl. Conf. Machine Learning*, pages 759–766. Morgan Kaufmann, San Francisco, CA, 2000.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Jin Qin and Jerry Lawless. Empirical likelihood and general estimating equations. *the Annals of Statistics*, pages 300–325, 1994.

R Tyrrell Rockafellar. Augmented lagrange multiplier functions and duality in nonconvex programming. *SIAM Journal on Control*, 12(2):268–285, 1974.

Werner Römisch. Delta method, infinite dimensional. *Wiley StatsRef: Statistics Reference Online*, 2014.

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver Prioritized experience replay. In *Proceedings of the 4th International Conference on Learning Representations*, 2016.

B. Sriperumbudur, K. Fukumizu, and G. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12:2389–2410, 2011.

Alexander L. Strehl, Lihong Li, and Michael L. Littman. Reinforcement learning in finite MDPs: PAC analysis. In *Journal of Machine Learning Research*, 10:2413–2444, 2009.

Richard S Sutton, Csaba Szepesvári, Alborz Geramifard, and Michael P Bowling. Dyna-style planning with linear function approximation and prioritized sweeping. *arXiv preprint arXiv:1206.3285*, 2012.

Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *International Conference on Machine Learning*, pages 814–823, 2015.

Istvan Szita and Csaba Szepesvari. Model-based reinforcement learning with nearly tight exploration complexity bounds. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, page 1031–1038. Omnipress, 2010.

Aviv Tamar, Huan Xu, and Shie Mannor. Scaling up robust MDPs by reinforcement learning. *arXiv preprint arXiv:1306.6189*, 2013.

Ziyang Tang, Yihao Feng, Lihong Li, Dengyong Zhou, and Qiang Liu. Doubly robust bias reduction in infinite horizon off-policy estimation. In *Proceedings of the 8th International Conference on Learning Representations*, 2020.

Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High confidence policy improvement. In *International Conference on Machine Learning*, pages 2380–2388, 2015a.

Philip S Thomas. *Safe reinforcement learning*. PhD thesis, University of Massachusetts Libraries, 2015.

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*, pages 2139–2148, 2016.

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

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, pages 5026–5033. IEEE, 2012.

Masatoshi Uehara, Jiawei Huang, and Nan Jiang. Minimax weight and Q-function learning for off-policy evaluation. *arXiv preprint arXiv:1910.12809*, 2019.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.

Weiran Wang and Miguel A Carreira-Perpinán. Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application. *arXiv preprint arXiv:1309.1541*, 2013.

Yifan Wu, George Tucker, and Ofir Nachum. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*, 2019.

Tengyang Xie, Yifei Ma, and Yu-Xiang Wang. Optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. In *Advances in Neural Information Processing Systems 32*, pages 9665–9675, 2019.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. MOPE: Model-based offline policy optimization. *arXiv preprint arXiv:2005.13239*, 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, 2020b. arXiv:2001.11113.

Appendix

A Approximation Error Analysis

In this section, we provide a complete proof of Theorem 1, quantifying the effect of function embedding of constraints in dual Q-LP. The proof is an adaptation from the standard LP for state-value functions to the case of Q-LP (De Farias and Van Roy, 2003).

We first provide an equivalent reformulation of the primal of the feature embedded LP,

Lemma 5 *The solution defined by*

$$\beta^* = \operatorname*{argmin}_{\beta\in\mathbb{R}^p} \left\{(1-\gamma)\,\mathbb{E}_{\mu_0\pi}\left[\beta^\top\phi\left(s_0,a_0\right)\right] | \beta^\top\phi\left(s,a\right) \geqslant \mathcal{B}_\pi\left(\beta^\top\phi\right)\left(s,a\right),\ \forall\left(s,a\right)\in S\times A\right\},$$

with $\left(\mathcal{B}_\pi Q\right)(s,a) := R(s,a) + \gamma\cdot\mathcal{P}^\pi Q(s,a)$ *is also the solution to*

$$\begin{aligned} \min_{\beta\in\mathbb{R}^p} \quad & \left\|Q^\pi - \beta^\top\phi\right\|_{1,\mu_0\pi} \\ \text{s.t.} \quad & \beta^\top\phi\left(s,a\right) \geqslant \mathcal{B}_\pi\left(\beta^\top\phi\right)\left(s,a\right),\ \forall\left(s,a\right)\in S\times A, \end{aligned} \tag{15}$$

where $\left\|f\right\|_{1,\mu_0\pi} := \int \left|f\left(s,a\right)\right|\mu_0\left(s\right)\pi\left(a|s\right)dsda.$

Proof Recall the fact that $\mathcal{B}_\pi$ is monotonic: given two bounded functions, $\nu_1 \geqslant \nu_2$ implies $\mathcal{B}_\pi\nu_1 \geqslant \mathcal{B}_\pi\nu_2$. Therefore, for any feasible ν, we have $\nu \geqslant \mathcal{B}_\pi\nu \geqslant \mathcal{B}_\pi^2\nu \geqslant \ldots \geqslant \mathcal{B}_\pi^\infty\nu = Q^\pi$, where the convergence to Q^π is due to the contraction property of $\mathcal{B}_\pi$.

Consider a feasible β, we have

$$\left\|Q^\pi - \beta^\top\phi\right\|_{1,\mu_0\pi} = \int \left(\beta^\top\phi\left(s,a\right) - Q^\pi\left(s,a\right)\right)\mu_0\left(s\right)\pi\left(a|s\right)dsda, \tag{16}$$

which implies minimizing $\mathbb{E}_{\mu_0\pi}\left[\beta^\top\phi\right]$ is equivalent to minimizing $\left\|Q^\pi - \beta^\top\phi\right\|_{1,\mu_0\pi}$. ■

Theorem 1 *Suppose the constant function* $\mathbf{1}\in\mathcal{F}_\phi := \operatorname{span}\left\{\phi\right\}$*. Then,*

$$0 \leqslant \tilde{\rho}_\pi - \rho_\pi \leqslant 2\min_\beta\left\|Q^\pi - \langle\beta,\phi\rangle\right\|_\infty,$$

where Q^π *is the fixed-point solution to the Bellman equation* $Q(s,a) = R(s,a) + \gamma\mathcal{P}^\pi Q(s,a)$.

Proof We first show the equivalence between function space embedding of dual Q-LP and the linear approximation of primal Q-LP, which can be easily derived by checking their Lagrangians. Denote

$$\begin{aligned} l\left(d,\beta\right) &:= \mathbb{E}_d\left[r\left(s,a\right)\right] + \beta^\top\left\langle\phi,\left(1-\gamma\right)\mu_0\pi + \gamma\cdot\mathcal{P}_*^\pi d - d\right\rangle \\ &= \left(1-\gamma\right)\mathbb{E}_{\mu_0\pi}\left[\beta^\top\phi\left(s,a\right)\right] + \mathbb{E}_d\left[r\left(s,a\right) + \gamma\cdot\mathcal{P}^\pi\beta^\top\phi\left(s,a\right) - \beta^\top\phi\left(s,a\right)\right] \\ &= \left(1-\gamma\right)\mathbb{E}_{\mu_0\pi}\left[Q_\beta\left(s,a\right)\right] + \mathbb{E}_d\left[r\left(s,a\right) + \gamma\cdot\mathcal{P}^\pi Q_\beta\left(s,a\right) - Q_\beta\left(s,a\right)\right], \end{aligned} \tag{17}$$

where $\beta\in\mathbb{R}^p$ and $Q_\beta(s,a) := \beta^\top\phi(s,a)$. Since the $l(d,\beta)$ is convex-concave w.r.t. (β,d), it is also the Lagrangian of primal Q-LP with linear parametrization, *i.e.*,

$$\begin{aligned} \min_{\beta\in\mathbb{R}^p} \quad & \left(1-\gamma\right)\mathbb{E}_{\mu_0\pi}\left[\beta^\top\phi\left(s_0,a_0\right)\right] \\ \text{s.t.} \quad & \beta^\top\phi\left(s,a\right) \geqslant R\left(s,a\right) + \gamma\cdot\mathcal{P}^\pi\beta^\top\phi\left(s,a\right), \quad \forall\left(s,a\right)\in S\times A. \end{aligned} \tag{18}$$

By Lemma 5, it is equivalent to solving

$$\begin{aligned} \min_{\beta\in\mathbb{R}^p} \quad & \left\|Q^\pi - \beta^\top\phi\right\|_{1,\mu_0\pi} \\ \text{s.t.} \quad & \beta^\top\phi\left(s,a\right) \geqslant \mathcal{B}_\pi\left(\beta^\top\phi\right)\left(s,a\right), \quad \forall\left(s,a\right)\in S\times A. \end{aligned} \tag{19}$$

We now define

$$\left(d^*,\beta^*\right) := \operatorname*{argmax}_{d\geqslant 0}\operatorname*{argmin}_\beta l\left(d,\beta\right),$$

$$\tilde{\beta} := \operatorname*{argmin}_\beta\left\|Q^\pi - \beta^\top\phi\right\|_\infty,$$

$$\epsilon := \left\|Q^\pi - \tilde{\beta}^\top\phi\right\|_\infty,$$

and obtain from strong duality that

$$\mathbb{E}_{d^*}\left[r\left(s,a\right)\right] = \left(1-\gamma\right)\mathbb{E}_{\mu_0\pi}\left[\left(\beta^*\right)^\top\phi\right].$$

Recall the fact $\mathcal{B}_\pi$ is a γ-contraction operator with the norm $\|\cdot\|_\infty$, and we have

$$\left\|\mathcal{B}_\pi\left(\tilde{\beta}^\top\phi\right) - Q^\pi\right\|_\infty \leqslant \gamma\left\|\tilde{\beta}^\top\phi - Q^\pi\right\|_\infty,$$

which implies

$$\mathcal{B}_\pi\left(\tilde{\beta}^\top\phi\right) \leqslant Q^\pi + \gamma\epsilon\mathbf{1}.$$

Now consider a new solution $\left(\tilde{\beta}^\top\phi - c\mathbf{1}\right)$, which must be in $\operatorname{span}\left\{\phi\right\}$ as $\mathbf{1}\in\operatorname{span}\left\{\phi\right\}$. Then,

$$\begin{aligned}
\mathcal{B}_\pi\left(\tilde{\beta}^\top\phi - c\mathbf{1}\right) &= \mathcal{B}_\pi\left(\tilde{\beta}^\top\phi\right) - \gamma c\mathbf{1} \\
&\leqslant Q^\pi + \gamma\epsilon\mathbf{1} - \gamma c\mathbf{1} \\
&\leqslant \tilde{\beta}^\top\phi + \left(1+\gamma\right)\epsilon\mathbf{1} - \gamma c\mathbf{1} \\
&= \tilde{\beta}^\top\phi - c\mathbf{1} + \left(\left(1-\gamma\right)c + \left(1+\gamma\right)\epsilon\right)\mathbf{1}.
\end{aligned}$$

Choose $c = -\left(1+\gamma\right)\epsilon/\left(1-\gamma\right)$, and the above implies $\mathcal{B}_\pi\left(\tilde{\beta}^\top\phi - c\mathbf{1}\right) \leqslant \tilde{\beta}^\top\phi - c\mathbf{1}$. Therefore, there exists some $\bar{\beta}$ such that

$$\bar{\beta}^\top\phi = \tilde{\beta}^\top\phi + \frac{1+\gamma}{1-\gamma}\epsilon\mathbf{1}.$$

Then, we can bound the approximation error

$$\begin{aligned}
\mathbb{E}_{d^*}\left[r\left(s,a\right)\right] - \rho_\pi &= \mathbb{E}_{d^*}\left[r\left(s,a\right)\right] - \left(1-\gamma\right)\mathbb{E}_{\mu_0\pi}\left[Q^\pi\right] \\
&= \left(1-\gamma\right)\mathbb{E}_{\mu_0\pi}\left[\left(\beta^*\right)^\top\phi\right] - \left(1-\gamma\right)\mathbb{E}_{\mu_0\pi}\left[Q^\pi\right] \geqslant 0,
\end{aligned}$$

where the last inequality comes from the fact $\left(1-\gamma\right)\mathbb{E}_{\mu_0\pi}\left[\left(\beta^*\right)^\top\phi\right]$ is the optimal value of a restricted feasible set within linearly representable Q_β.

On the other hand, we bound

$$\begin{aligned}
\left(1-\gamma\right)\mathbb{E}_{\mu_0\pi}\left[\left(\beta^*\right)^\top\phi\right] - \left(1-\gamma\right)\mathbb{E}_{\mu_0\pi}\left[Q^\pi\right] &= \left(1-\gamma\right)\left\|\left(\beta^*\right)^\top\phi - Q_\beta\right\|_{1,\mu_0\pi} \\
&\leqslant \left(1-\gamma\right)\left\|\bar{\beta}^\top\phi - Q_\beta\right\|_{1,\mu_0\pi} \\
&\leqslant \left(1-\gamma\right)\left\|\bar{\beta}^\top\phi - Q_\beta\right\|_\infty \\
&\leqslant \left(1-\gamma\right)\left(\left\|\bar{\beta}^\top\phi - \tilde{\beta}^\top\phi\right\|_\infty + \left\|Q^\pi - \tilde{\beta}^\top\phi\right\|_\infty\right) \\
&\leqslant \left(1-\gamma\right)\left(1 + \frac{1+\gamma}{1-\gamma}\right)\epsilon = 2\epsilon.
\end{aligned}$$

where the third inequality comes from the optimality of (19). ■

Justification of full-rank basis embedding. The effect of full-rank basis embedding in the example in Section 3.1 can be justified straightforwardly. We consider the Lagrangian (17). If the $\phi\in\mathbb{R}^{|S||A|\times|S||A|}$ is full-rank, ϕ^{-1} exists. For arbitrary $Q\in\mathbb{R}^{|S||A|\times 1}$, there exists $\beta = \left(Q\phi^{-1}\right)^\top$, which means there is an one-to-one correspondence between Q and β in Lagrangian. Therefore, in finite state and action MDP, the Lagrangian is not affected by full-rank basis embedding, and therefore, the solution of full-rank basis embedding will be the same as the original LP.

B CoinDICE for Undiscounted and finite-horizon MDPs

In the main text, we consider the CoinDICE for infinite-horizon MDPs with discounted factor $\gamma < 1$. The algorithm can be generalized to undiscounted MDPs with $\gamma = 1$ and finite-horizon MDPs.

Undiscounted MDP. We have the dual form of the Q-LP as

$$\tilde{\rho}_\pi := \left\{ \max_{d:S\times A\to\mathbb{R}_+} \mathbb{E}_d[r(s,a)] \middle| \begin{array}{c} \int d(s,a)\,dsda = 1 \\ d(s,a) = \mathcal{P}_*^\pi d(s,a), \forall (s,a)\in S\times A \end{array} \right\}. \tag{20}$$

Comparing with the (3), we have an extra normalization constraint. Specifically, if $d(s,a)$ is feasible, without the normalization constraint, $c\cdot d(s,a)$ will also be feasible for any $c>0$. Therefore, the optimization could be unbounded.

By change-of-variable $\tau(s,a) = \frac{d^\pi(s,a)}{d^\mathcal{D}(s,a)}$ and feature embeddings of the stationary constraint in (20), we obtain

$$\tilde{\rho}_\pi := \left\{ \max_{\tau:S\times A\to\mathbb{R}_+} \mathbb{E}_{d^\mathcal{D}}[\tau\cdot r(s,a)] \middle| \begin{array}{c} \mathbb{E}_{d^\mathcal{D}}[\tau(s,a)] = 1 \\ \mathbb{E}_{d^\mathcal{D}}[\phi(s',a')(\tau(s',a') - \tau(s,a))] = 0 \end{array} \right\}. \tag{21}$$

Then, the CoinDICE confidence interval is achieved by applying the generalized empirical likelihood to (21), *i.e.*,

$$C_{n,\xi}^f := \left\{ \tilde{\rho}_\pi(w) = \max_{\tau\geqslant 0} \mathbb{E}_w[\tau\cdot r] \middle| \begin{array}{c} w\in\mathcal{K}_f \\ \mathbb{E}_w[\tau-1] = 0 \\ \mathbb{E}_w\left[\bar{\Delta}(x;\tau,\phi)\right] = 0 \end{array} \right\}, \text{ with } \mathcal{K}_f := \left\{ \begin{array}{l} w\in\mathcal{P}^{n-1}(\widehat{p}_n), \\ D_f(w\|\widehat{p}_n) \leqslant \frac{\xi}{n} \end{array} \right\}, \tag{22}$$

where $\bar{\Delta}(x;\tau,\phi) := \phi(s',a')(\tau(s',a') - \tau(s,a))$.

A similar argument of Section 3.3 for discounted MDPs can be applied to (22), resulting in the following confidence interval:

$$C_{n,\xi}^f = [l_n, u_n]$$

with

$$[l_n, u_n] = \left[\min_{\beta\in\mathbb{R}^p,\nu} \max_{\tau\geqslant 0} \min_{w\in\mathcal{K}_f} \mathbb{E}_w[\ell(x;\tau,\beta,\nu)], \quad \max_{\tau\geqslant 0} \min_{\beta\in\mathbb{R}^p,\nu} \max_{w\in\mathcal{K}_f} \mathbb{E}_w[\ell(x;\tau,\beta,\nu)] \right], \tag{23}$$

where $\ell(x;\tau,\beta,\nu) := \tau\cdot r + \beta^\top \Delta(x;\tau,\phi) + \nu - \nu\cdot\tau$.

Remark (Normalization constraint): Although in the discounted MDPs, there is no scaling issue, and thus the normalizaiton constraint is redudant, we still prefer to add the constraint in practice. It does not only bring the benefits in optimization, but also enforce the normalization explicitly and reduce the feasible set, leading to better statistical property.

Finite-horizon MDP. While we mainly focus on infinite-horizon MDPs with a discounted factor, the dual method can be adapted to finite-horizon settings straightforwardly. For example, we have the finite-horizon d-LP as

$$\max_{d_h(s,a):S\times A\to\mathbb{R}_+} \quad \sum_{h=1}^H \mathbb{E}_{d_h}[r_h(s,a)] \tag{24}$$

$$\text{s.t.} \quad d_0(s,a) = \mu_0(s)\pi(a|s), \tag{25}$$

$$d_{h+1}(s,a) = \mathcal{P}_*^\pi d_h(s,a), \ \forall h\in\{1,\ldots,H\}. \tag{26}$$

Upon this finite-horizon formulation, we can derive the finite-horizon CoinDICE following the same technique, *i.e.*.

$$[l_n, u_n] =$$

$$\left[\min_{w\in\mathcal{K}_f} \min_{\beta_{h=1}^H\in\mathbb{R}^p} \max_{\tau_{h=1}^H\geqslant 0} \mathbb{E}_w\left[\ell_H\left(x;\tau_{h=1}^H,\beta_{h=1}^H\right)\right], \max_{w\in\mathcal{K}_f} \max_{\tau_{h=1}^H\geqslant 0} \min_{\beta_{h=1}^H\in\mathbb{R}^p} \mathbb{E}_w\left[\ell_H\left(x;\tau_{h=1}^H,\beta_{h=1}^H\right)\right] \right],$$

where $x := \left\{(s,a,r,s',a',h)_{h=1}^H\right\}$, $\ell_H\left(x;\tau_{h=1}^H,\beta_{h=1}^H\right) := \sum_{h=1}^H \tau_h r_h + \sum_{h=1}^H \beta_h^\top \Delta_h(x;\tau_h,\phi)$, and $\Delta_h(x;\tau_h,\phi) := \tau_h(s,a)\phi(s',a') - \tau_{h+1}(s',a')\phi(s',a')$.

C CoinBandit

MDPs are strictly more general than multi-armed and contextual bandits. Therefore, our estimator can also be specialized accordingly for confidence interval estimation in bandit problems with slight modifications. Without loss of generality, we consider the contextual bandit setting, while the multi-armed bandits can be further reduced from contextual bandit.

Specifically, in the behavior-agnostic contextual bandit setting, the stationary distribution constraint in (5) is no long applicable in bandit setting. We rewrite the policy value as

$$\begin{aligned} \tilde{\rho}_\pi &:= \mathbb{E}_{s\sim\mu^{\mathcal{D}}, a\sim\pi(a|s)}\left[r\left(s,a\right)\right] \\ &= \left\{\max_{\tau: S\times A\to\mathbb{R}_+} \mathbb{E}_{d^{\mathcal{D}}}\left[\tau\cdot r\left(s,a\right)\right] \middle| d^{\mathcal{D}}\cdot\tau = \mu^{\mathcal{D}}\pi, \mathbb{E}_{d^{\mathcal{D}}}\left[\tau\right] = 1\right\}, \end{aligned} \tag{27}$$

where we reload the $\mu^{\mathcal{D}}$ as the contextual distribution, which is unchanged for all policies, $d^{\mathcal{D}}\left(s,a\right) = \mu^{\mathcal{D}}\left(s\right)\pi_b\left(a|s\right)$, $\tau\left(s,a\right) := \frac{\mu^{\mathcal{D}}(s)\pi(a|s)}{\mu^{\mathcal{D}}(s)\pi_b(a|s)}$, and $\phi\left(s,a\right)$ denotes the feature mappings. We keep the normalization constraint to ensure the validation of density ratio empirically.

We apply the same technique to (27), leading to the *CoinBandit* confidence interval estimator

$$C^f_{n,\xi} := \left\{\tilde{\rho}_\pi(w) = \max_{\tau\geqslant 0}\mathbb{E}_w\left[\tau\cdot r\right] \middle| \begin{matrix} w\in\mathcal{K}_f, \mathbb{E}_w\left[\tau-1\right]=0 \\ \mathbb{E}_w\left[\blacktriangle\left(x;\tau,\phi\right)\right]=0\end{matrix}\right\}, \text{ with } \mathcal{K}_f := \left\{\begin{matrix} w\in\mathcal{P}^{n-1}\left(\widehat{p}_n\right), \\ D_f\left(w||\widehat{p}_n\right)\leqslant\frac{\xi}{n}\end{matrix}\right\}, \tag{28}$$

where the $x := \left(s,a,s',a'\right)$ is constructed by $s\sim\mu^{\mathcal{D}}\left(s\right), a\sim\pi\left(a|s\right)$ and $\left(s',a'\right)\sim d^{\mathcal{D}}$, and $\blacktriangle\left(x;\tau,\phi\right) := \phi\left(s,a\right) - \phi\left(s',a'\right)\cdot\tau\left(s',a'\right)$.

Similarly, the interval estimator in CoinBandit (28) can be calculated by solving a minimax optimization.

Remark (Behavior-known contextual bandit): When the behavior policy $\pi_b\left(a|s\right)$ is known, the solution to (27) can be computed in closed-form as $\tau\left(s,a\right) = \frac{\pi(a|s)}{\pi_b(a|s)}$. Then, the CoinBandit reduces to

$$C^f_{n,\xi} := \left\{\tilde{\rho}_\pi(w) = \mathbb{E}_w\left[\tau\cdot r\right] \middle| \begin{matrix} w\in\mathcal{K}_f, \\ \mathbb{E}_w\left[\tau-1\right]=0\end{matrix}\right\}, \text{ with } \mathcal{K}_f := \left\{\begin{matrix} w\in\mathcal{P}^{n-1}\left(\widehat{p}_n\right), \\ D_f\left(w||\widehat{p}_n\right)\leqslant\frac{\xi}{n}\end{matrix}\right\}. \tag{29}$$

Remark (Multi-armed bandit): Furthermore, these estimators (28) and (29) can be further reduced for multi-armed bandit. Specifically, we set all s equivalent, then, the s becomes the dummy variable. The CoinBandit estimators (28) and (29) reduces for the off-policy evaluation in multi-armed bandit. If the action number is finite, we can use tabular representation for $\tau\left(a\right)$, eliminating the approximation error.

Remark (Comparison to Karampatziakis et al. (2019)): Karampatziakis et al. (2019) considers the off-policy contextual bandit confidence interval estimation. Although both CoinBandit and the estimator in Karampatziakis et al. (2019) share the same asymptotic coverage, there are significant differences:

- The estimator in Karampatziakis et al. (2019) is derived from empirical likelihood with reverse KL-divergence, while our CoinBandit is based on generalized empirical likelihood with arbitrary f-divergence.
- More importantly, compared to our CoinBandit, which is applicable for both *behavior-agnostic* and *behavior-known* off-policy setting, the estimator in Karampatziakis et al. (2019) is only valid for behavior-known setting.
- Computationally, the estimator in Karampatziakis et al. (2019) requires an extra statistics, *i.e.*,
$$\left\{\max_w \sum_{i=1}^n \log\left(nw_i\right)\left|\mathbb{E}_w\left[\tau-1\right]=0,\ w\in\mathcal{K}_{-2\log(\cdot)}\right.\right\},$$
while such quantity is not required in CoinBandit, and thus saving the computational cost.
- Statistically, we provide finite sample complexity for CoinBandit in Theorem 4, while such sample complexity is not clear for Karampatziakis et al. (2019).

D Stochastic Confidence Interval Estimation

We analyze the properties of the optimization for the upper and lower bounds and derive the practical algorithm in this section.

D.1 Upper and Lower Confidence Bounds

We first establish the distribution robust optimization representation of the confidence region:

Lemma 6 *Let $\hat{\rho}_\pi(w) = \max_{\tau \geqslant 0} \min_{\beta \in \mathbb{R}^p} \mathbb{E}_w \left[\tau \cdot r + \beta^\top \Delta(x; \tau, \phi)\right]$. The confidence region $C^f_{n,\xi}$ can be represented equivalently as*

$$C^f_{n,\xi} = \left\{\hat{\rho}_\pi(w) \,\middle|\, w \in \mathcal{K}_f\right\}. \tag{30}$$

Proof For any $w \in \mathcal{K}_f$, we rewrite the optimization (8) by its Lagrangian, which will be an estimate of the policy value,

$$\hat{\rho}_\pi(w) = \max_{\tau \geqslant 0} \min_{\beta \in \mathbb{R}^p} \mathbb{E}_w \left[\tau \cdot r + \beta^\top \Delta(x; \tau, \phi)\right]. \tag{31}$$

■

Based on Lemma 6, we can formulate the upper and lower bounds:

Theorem 3 *Denote the upper and lower confidence bounds of $C^f_{n,\xi}$ by u_n and l_n, respectively:*

$$\begin{aligned}
[l_n, u_n] &= \left[\min_{w \in \mathcal{K}_f} \min_{\beta \in \mathbb{R}^p} \max_{\tau \geqslant 0} \mathbb{E}_w \left[\ell(x; \tau, \beta)\right], \quad \max_{w \in \mathcal{K}_f} \max_{\tau \geqslant 0} \min_{\beta \in \mathbb{R}^p} \mathbb{E}_w \left[\ell(x; \tau, \beta)\right]\right], \\
&= \left[\min_{\beta \in \mathbb{R}^p} \max_{\tau \geqslant 0} \min_{w \in \mathcal{K}_f} \mathbb{E}_w \left[\ell(x; \tau, \beta)\right], \quad \max_{\tau \geqslant 0} \min_{\beta \in \mathbb{R}^p} \max_{w \in \mathcal{K}_f} \mathbb{E}_w \left[\ell(x; \tau, \beta)\right]\right],
\end{aligned}$$

where $\ell(x; \tau, \beta) := \tau \cdot r + \beta^\top \Delta(x; \tau, \phi)$. For any $(\tau, \beta, \lambda, \eta)$ that satisfies the constraints in (11), the optimal weights for upper and lower confidence bounds are

$$w_l = f'_*\left(\frac{\eta - \ell(x; \tau, \beta)}{\lambda}\right) \quad \text{and} \quad w_u = f'_*\left(\frac{\ell(x; \tau, \beta) - \eta}{\lambda}\right),$$

respectively. Therefore, the confidence bounds can be simplified as:

$$\begin{bmatrix} l_n \\ u_n \end{bmatrix} = \begin{bmatrix} \min_\beta \max_{\tau \geqslant 0, \lambda \geqslant 0, \eta} \mathbb{E}_{\mathcal{D}} \left[-\lambda f_*\left(\frac{\eta - \ell(x; \tau, \beta)}{\lambda}\right) + \eta - \lambda \frac{\xi}{n}\right] \\ \max_{\tau \geqslant 0} \min_{\beta, \lambda \geqslant 0, \eta} \mathbb{E}_{\mathcal{D}} \left[\lambda f_*\left(\frac{\ell(x; \tau, \beta) - \eta}{\lambda}\right) + \eta + \lambda \frac{\xi}{n}\right] \end{bmatrix}.$$

Proof We first calculate the upper bound u_n using Lemma 6:

$$\begin{aligned}
u_n = \max_{w \in \mathcal{K}_f} \rho_\pi(w) &= \max_{w \in \mathcal{K}_f} \max_{\tau \geqslant 0} \min_{\beta \in \mathbb{R}^p} \mathbb{E}_w \left[\tau \cdot r + \beta^\top \Delta(x; \tau, \phi)\right] \\
&= \max_{\tau \geqslant 0} \max_{w \in \mathcal{K}_f} \min_{\beta \in \mathbb{R}^p} \mathbb{E}_w \left[\tau \cdot r + \beta^\top \Delta(x; \tau, \phi)\right] \qquad (32) \\
&= \max_{\tau \geqslant 0} \min_{\beta \in \mathbb{R}^p} \max_{w \in \mathcal{K}_f} \mathbb{E}_w \left[\tau \cdot r + \beta^\top \Delta(x; \tau, \phi)\right], \qquad (33)
\end{aligned}$$

where the switch between $\max_{w \in \mathcal{K}_f}$ and $\max_{\tau \geqslant 0}$ in (32) is immediate, (33) is due to the fact that the objective is concave w.r.t. β and convex w.r.t. w and τ, separately.

We apply Lagrangian to the inner constrained optimization over w, leading to

$$\begin{aligned}
u_n &= \max_{\tau} \min_{\beta, \lambda \geqslant 0, \eta} \max_{w \geqslant 0} \mathbb{E}_w \left[\tau \cdot r + \beta^\top \Delta(x; \tau, \phi)\right] - \lambda \left(D_f(w \| \widehat{p}_n) - \frac{\xi}{n}\right) + \eta\left(1 - w^\top \mathbf{1}\right) \\
&= \max_{\tau \geqslant 0} \min_{\beta, \lambda \geqslant 0, \eta} \mathbb{E}_{\mathcal{D}} \left[\lambda f_*\left(\frac{\tau \cdot r + \beta^\top \Delta(x; \tau, \phi) - \eta}{\lambda}\right) + \eta + \frac{\lambda \xi}{n}\right], \qquad (34)
\end{aligned}$$

where the last equation comes from the conjugate of f, and for any given $(\tau, \beta, \lambda, \eta)$, the optimal w^* will be

$$w^*_u = f'_*\left(\frac{\tau \cdot r + \beta^\top \Delta(x; \tau, \phi) - \eta}{\lambda}\right).$$

The lower bound l_n may be obtained in a similar fashion:

$$\begin{aligned} l_n = \min_{w\in\mathcal{K}_f} \rho\left(w;\pi\right) &= \min_{w\in\mathcal{K}_f} \max_{\tau\geqslant 0} \min_{\beta\in\mathbb{R}^p} \mathbb{E}_w\left[\tau\cdot r + \beta^\top \Delta\left(x;\tau,\phi\right)\right] \\ &= \min_{w\in\mathcal{K}_f} \min_{\beta\in\mathbb{R}^p} \max_{\tau\geqslant 0} \mathbb{E}_w\left[\tau\cdot r + \beta^\top \Delta\left(x;\tau,\phi\right)\right] \\ &= \min_{\beta\in\mathbb{R}^p} \min_{w\in\mathcal{K}_f} \max_{\tau\geqslant 0} \mathbb{E}_w\left[\tau\cdot r + \beta^\top \Delta\left(x;\tau,\phi\right)\right] \\ &= \min_{\beta\in\mathbb{R}^p} \max_{\tau\geqslant 0} \min_{w\in\mathcal{K}_f} \mathbb{E}_w\left[\tau\cdot r + \beta^\top \Delta\left(x;\tau,\phi\right)\right]. \end{aligned}$$

Again, we consider the Lagrangian

$$\begin{aligned} l_n &= \min_{\beta\in\mathbb{R}^p} \max_{\tau\geqslant 0,\lambda\geqslant 0,\eta} \min_{w\geqslant 0} \mathbb{E}_w\left[\tau\cdot r + \beta^\top \Delta\left(x;\tau,\phi\right)\right] + \lambda\left(D_f\left(w||\widehat{p}_n\right) - \frac{\xi}{n}\right) + \eta\left(1 - w^\top \mathbf{1}\right) \\ &= \min_{\beta} \max_{\tau\geqslant 0,\lambda\geqslant 0,\eta} \mathbb{E}_{\mathcal{D}}\left[-\lambda f_*\left(\frac{\eta - \left(\tau\cdot r + \beta^\top \Delta\left(x;\tau,\phi\right)\right)}{\lambda}\right) + \eta - \frac{\lambda\xi}{n}\right], \end{aligned}$$

and the optimal weight is

$$w_l^* = f_*'\left(\frac{\eta - \left(\tau\cdot r + \beta^\top \Delta\left(x;\tau,\phi\right)\right)}{\lambda}\right).$$

■

D.2 Closed-form Solution for Reweighting

We consider a few examples of f-divergences in Theorem 3, and show how the weights can be efficiently computed, for a given τ and β.

- KL**-divergence.** To satisfy the conditions in Assumption 1, we select $f\left(x\right) = 2x\log x$. Recall the property that for any convex function f and any $\alpha > 0$, the conjugate function of $g(x) = \alpha f(x)$ is equal to $g_*(y) = \alpha f_*(y/\alpha)$. Let f be the standard f-divergence function of KL-divergence KL $\left(w||\widehat{p}_n\right)$, i.e., $f\left(x\right) = 2x\log x$. With $g_*'(y) = f_*'(y/\alpha)$, equation (12) implies that the following upper and lower bounds:

$$\begin{aligned} w_l\left(x\right) &= \exp\left(\frac{\eta_l - \ell\left(x;\tau,\beta\right)}{2\lambda}\right), \quad \eta_l = -\log\sum_{i=1}^n \exp\left(\frac{-\ell\left(x;\tau,\beta\right)}{2\lambda}\right) \\ w_u\left(x\right) &= \exp\left(\frac{\ell\left(x;\tau,\beta\right) - \eta_u}{2\lambda}\right), \quad \eta_u = \log\sum_{i=1}^n \exp\left(\frac{\ell\left(x;\tau,\beta\right)}{2\lambda}\right). \end{aligned}$$

  This can also be verified by plugging the $f\left(x\right) = 2x\log x$ into (12) and considering $w^\top \mathbf{1} = 1$.

- **Reverse KL-divergence.** With the f-divergence function $f\left(x\right) = -\log x$ for the reverse-KL divergence, one has the following upper and lower bounds:

$$\begin{aligned} & w_l\left(x\right) = \lambda\delta\left(\ell\left(x;\tau,\beta\right) > \eta_l\right)\left(\ell\left(x;\tau,\beta\right) - \eta_l\right)^{-1}, \\ & \sum_{i=1}^n \delta\left(\ell\left(x;\tau,\beta\right) > \eta_l\right)\left(\ell\left(x;\tau,\beta\right) - \eta_l\right)^{-1} = \frac{1}{\lambda}, \\ & w_u\left(x\right) = \lambda\delta\left(\eta_u > \ell\left(x;\tau,\beta\right)\right)\left(\eta_u - \ell\left(x;\tau,\beta\right)\right)^{-1}, \\ & \sum_{i=1}^n \delta\left(\eta_u > \ell\left(x;\tau,\beta\right)\right)\left(\eta_u - \ell\left(x;\tau,\beta\right)\right)^{-1} = \frac{1}{\lambda}, \end{aligned}$$

  where $\delta\left(a > b\right) = \begin{cases} 1 & \text{if } a > b \\ 0 & \text{otherwise} \end{cases}$. This is obtained by plugging the $f\left(x\right) = -\log x$ into (12) and considering $w^\top \mathbf{1} = 1$, $w \geqslant 0$ and KKT conditions on the dual variables for $w \geqslant 0$. Unfortunately the reverse KL-divergence does not satisfy the conditions in Assumption 1. Note that this is the standard f-divergence function for empirical likelihood maximization problem, we therefore also include it here for the sake of completeness.

- **χ^2-divergence.** Notice that the standard f-divergence function, i.e., $f(x) = (x-1)^2$, of χ^2-divergence $\chi^2(w\|\widehat{p}_n) := \mathbb{E}_{\widehat{p}_n}\left[\left(\frac{w}{\widehat{p}_n}-1\right)^2\right]$ satisfies the conditions in Assumption 1. Consider the lower bound calculation. Leveraging the closed-form solution of the following ℓ_2 projection problem onto the simplex space $w^\top \mathbf{1} = 1$ and $w \geqslant 0$ (Wang and Carreira-Perpinán, 2013):

$$
\begin{aligned}
&\arg\min_{w:w^\top \mathbf{1}=1, w\geqslant 0} \sum_{i=1}^n w_i \frac{\ell(x_i;\tau,\beta)}{\lambda} + \sum_{i=1}^n \frac{1}{\widehat{p}_{n,i}}(w_i - \widehat{p}_{n,i})^2 \\
=&\sqrt{\widehat{p}_{n,i}} \cdot \arg\min_{v:v^\top \sqrt{\widehat{p}_n}=1, v\geqslant 0} \sum_{i=1}^n \left(v_i - (1-\frac{\ell(x_i;\tau,\beta)}{2\lambda})\cdot\sqrt{\widehat{p}_{n,i}}\right)^2, \ \text{(here we let } v_i = \frac{w_i}{\sqrt{\widehat{p}_{n,i}}}\text{)}
\end{aligned}
$$

the lower bound $w_\ell(x)$ is given by (for any $i \in \{1, 2, \ldots, n\}$)

$$
\begin{aligned}
w_\ell(x_i) &= \sqrt{\widehat{p}_{n,i}} \cdot w^*(x_i) \\
&= \sqrt{\widehat{p}_{n,i}} \cdot \left((1-\frac{\ell(x_i;\tau,\beta)}{2\lambda})\cdot\sqrt{\widehat{p}_{n,i}} + \mathcal{G}_{\widehat{p}_n}\left((1-\frac{\ell(x;\tau,\beta)}{2\lambda})\cdot\sqrt{\widehat{p}_{n,i}}\right)\right)_+,
\end{aligned}
$$

where $\mathcal{G}_{\widehat{p}_n}(y) = \frac{1-\sum_{i=1}^{|\mathcal{S}_{\widehat{p}_n}|} y_i\cdot\sqrt{\widehat{p}_{n,i}}}{\sum_{i=1}^{|\mathcal{S}_{\widehat{p}_n}|}\widehat{p}_{n,i}}$, $\mathcal{S}_{\widehat{p}_n}$ is the set of indices in $\{1,\ldots,n\}$ in which any element j satisfies $y_{(j)} + \frac{1}{\sum_{i=1}^j \widehat{p}_{n,i}}(1-\sum_{i=1}^j y_{(i)}\cdot\sqrt{\widehat{p}_{n,i}}) > 0$. Here $y_{(i)}$ indicates the samples with the i-th largest element of y. Using analogous arguments, by replacing ℓ with $-\ell$ one can also define a similar solution for the upper bound $w_u(x)$. Now suppose $\widehat{p}_{n,i} = \frac{1}{n}, \forall i$. Then, we have

$$
\begin{aligned}
w_l(x_i) &= \sqrt{\frac{1}{n}} \cdot \left((1-\frac{\ell(x_i;\tau,\beta)}{2\lambda})\cdot\sqrt{\frac{1}{n}} + \mathcal{G}_{\frac{1}{n}}\left((1-\frac{\ell(x;\tau,\beta)}{2\lambda})\cdot\sqrt{\frac{1}{n}}\right)\right)_+, \\
w_u(x_i) &= \sqrt{\frac{1}{n}} \cdot \left((1+\frac{\ell(x_i;\tau,\beta)}{2\lambda})\cdot\sqrt{\frac{1}{n}} + \mathcal{G}_{\frac{1}{n}}\left((1+\frac{\ell(x;\tau,\beta)}{2\lambda})\cdot\sqrt{\frac{1}{n}}\right)\right)_+,
\end{aligned}
$$

where $\mathcal{G}_{\frac{1}{n}}(y) = \frac{n-\sum_{i=1}^{|\mathcal{S}_{1/n}|} y_i\cdot\sqrt{n}}{|\mathcal{S}_{1/n}|}$, $\mathcal{S}_{\frac{1}{n}}$ is the set of indices in $\{1,\ldots,n\}$ in which any element j satisfies $y_{(j)} + \frac{1}{j}(n - \sqrt{n}\sum_{i=1}^j y_{(i)}) > 0$. Here $y_{(i)}$ indicates the samples with the i-th largest element of y. This can also be verified by plugging the $f(x) = (x-1)^2$ into (12) and considering $w^\top \mathbf{1} = 1$ and $w \geqslant 0$. In fact, the above can be generalized to the Cressie-Read family with $f(x) = \frac{(x-1)^k - k(x-1)+k-1}{k(k-1)}$.

- **Reverse KL-divergence.** With the f-divergence function $f(x) = -\log x$ for the reverse KL-divergence, one has the following upper and lower bounds:

$$
\begin{aligned}
&w_l(x) = \lambda\delta(\ell(x;\tau,\beta) > \eta_l)(\ell(x;\tau,\beta) - \eta_l)^{-1}, \\
&\sum_{i=1}^n \delta(\ell(x;\tau,\beta) > \eta_l)(\ell(x;\tau,\beta) - \eta_l)^{-1} = \frac{1}{\lambda}, \\
&w_u(x) = \lambda\delta(\eta_u > \ell(x;\tau,\beta))(\eta_u - \ell(x;\tau,\beta))^{-1}, \\
&\sum_{i=1}^n \delta(\eta_u > \ell(x;\tau,\beta))(\eta_u - \ell(x;\tau,\beta))^{-1} = \frac{1}{\lambda},
\end{aligned}
$$

where $\delta(a > b) = \begin{cases} 1 & \text{if } a > b \\ 0 & \text{otherwise} \end{cases}$. This is obtained by plugging the $f(x) = -\log x$ into (12) and considering $w^\top \mathbf{1} = 1$, $w \geqslant 0$ and KKT conditions on the dual variables for $w \geqslant 0$. Unfortunately the reverse KL-divergence does not satisfy the conditions in Assumption 1. Note that this is the standard f-divergence used in the vanilla empirical likelihood, we therefore also include it here for the sake of completeness.

D.3 Practical Algorithm

In (13), we eliminate one level optimization, thus reduce the computational difficulty. Meanwhile, SGDA for (13) could benefit from the attractive finite-step convergence. However, as observed

Algorithm 1 CoinDICE: estimating upper confidence bound using KL-divergence and function approximation.

Inputs: A target policy π, a desired confidence $1-\alpha$, a finite sample dataset $\mathcal{D} := \{(s_0^{(j)}, a_0^{(j)}, s^{(j)}, a^{(j)}, r^{(j)}, s'^{(j)})\}_{j=1}^n$, optimizers $\mathcal{OPT}_\theta$, number of iterations K, T.

Set divergence limit $\xi := \frac{1}{2}\chi_1^{2,1-\alpha}$.
Initialize $\lambda \in \mathbb{R}$, $Q_{\theta_1} : S \times A \to \mathbb{R}$, $\zeta_{\theta_2} : S \times A \to \mathbb{R}$.
for $k = 1, \ldots, K$ **do**
  for $t = 1, \ldots, T$ **do**
    Sample from target policy $a_0^{(j)} \sim \pi(s_0^{(j)})$, $a^{(j)\prime} \sim \pi(s^{(j)\prime})$ for $j = 1, \ldots, n$.
    Compute loss terms:
    $\ell^{(j)} := (1-\gamma)Q_{\theta_1}(s_0^{(j)}, a_0^{(j)}) + \zeta_{\theta_2}(s^{(j)}, a^{(j)}) \cdot (-Q_{\theta_1}(s^{(j)}, a^{(j)}) + r^{(j)} + \gamma Q_{\theta_1}(s^{(j)\prime}, a^{(j)\prime}))$

    Compute loss $\mathcal{L} := \sum_{j=1}^n w^{(j)} \cdot \ell^{(j)}$.
    Update $(\theta_1, \theta_2) \leftarrow \mathcal{OPT}_\theta(\mathcal{L}, \theta_1, \theta_2)$.
  end for
  Update (w, λ) by (35) or (36)
end for

Return $\mathcal{L}$.

in Namkoong and Duchi (2016), when λ approaches 0, SGDA for (13) may suffer from high variance. In this section, we consider two strategies to bypass such difficulty. We take the upper bound as an example, and the lower bound can be handled similarly:

- Instead of using the optimal weights (12), Namkoong and Duchi (2016) suggests to keep (w, λ) in optimization to be updated simultaneously via gradients, *i.e.*, targeting on solving the Lagrangian (33) with SGDA directly. For example, with KL-divergence, this leads to the update of w_u in the t-th iteration as
$$\tilde{w}^{(j)} = \exp\left(\eta_t \ell^{(j)}\right)\left(w^{(j)}\right)^{1-\eta_t\lambda}\left(\frac{1}{n}\right)^{\eta_t\lambda} \quad \text{and} \quad w_u = \frac{\tilde{w}^{(j)}}{\sum_j \tilde{w}^{(j)}}, \tag{35}$$
with stepsize η_t.
- The instability and high variance of solving (13) comes from unboundness of w induced by arbitarry λ during the optimization procedure. In other words, given a fixed (τ, β), if we can keep $w \in \mathcal{K}_f$ satisfied, *i.e.*,
$$\begin{aligned} w_u &= \mathop{\mathrm{argmax}}_{KL(w\|\widehat{p}_n)\leqslant \frac{\xi}{n}} \langle w, \ell\rangle \\ \Rightarrow (w_u, \lambda^*) &= \mathop{\mathrm{argmax}}_{w^\top \mathbf{1}=1, w\geqslant 0}\ \mathop{\mathrm{argmin}}_{\lambda\geqslant 0}\ \langle w, \ell\rangle - \lambda\left(KL\left(w\|\widehat{p}_n\right) - \frac{\xi}{n}\right) \\ \Rightarrow (w_u, \lambda^*) &= \left\{\tilde{w}_{\lambda^*}^{(j)} := \exp\{\frac{\ell^{(j)}}{\lambda^*}\};\quad w_{\lambda^*}^{(j)} := \frac{\tilde{w}_{\lambda^*}^{(j)}}{\sum \tilde{w}_{\lambda^*}^{(j)}} \text{ with } \sum_{j=1}^n w_{\lambda^*}^{(j)} \log w_{\lambda^*}^{(j)} = \frac{\xi}{n}\right\}, \end{aligned} \tag{36}$$
the optimization will be stable.

Moreover, the major computation cost of optimization is updating the w, which is an $\mathcal{O}(n)$ operation. Therefore, we update w less frequently, which corresponds to optimizing the equivalent form (10). Incorporating these techniques into SGDA, we obtain the algorithm in Algorithm 1.

Remark (More regularization for stability): Directly solving a Lagrangian for LP may induce instability, due to lack of curvature. To overcome such difficulty, the augmented Lagrangian method (ALM) (Rockafellar, 1974) is the natural choice. Directly applying the ALM will introduce the regularization $h\left(\mathbb{E}_{\widehat{p}_n}\left[\Delta\left(x; \tau, \phi\right)\right]\right)$ where h denotes some convex function with minimum at zero. Such regularization will not change the optimal solution (τ, β) in (11) and the value $[l_n, u_n]$.

The ALM introduces extra computational cost in optimization since the regularization involves empirical expectations inside a nonlinear function. We exploit alternative regularizations following the spirit of ALM, while circumventing the computational difficulty. Recall the fact that the regularization on dual variable does not change the optimal solution (Nachum et al., 2019b, Theorem 4), *i.e.*

$$\tau^*(s,a) = \left\{ \underset{\tau \geqslant 0}{\operatorname{argmax}} \; \mathbb{E}_{d^{\mathcal{D}}}[\tau \cdot r(s,a)] \,\middle|\, \mathbb{E}_{d^{\mathcal{D}}}[\Delta(x;\tau,\phi)] = 0 \right\} \tag{37}$$

$$= \left\{ \underset{\tau \geqslant 0}{\operatorname{argmax}} \; \mathbb{E}_{d^{\mathcal{D}}}[\tau \cdot r(s,a)] - \alpha \mathbb{E}_p[h(\tau)] \,\middle|\, \mathbb{E}_{d^{\mathcal{D}}}[\Delta(x;\tau,\phi)] = 0 \right\}, \tag{38}$$

where p is some distribution over $S \times A$.

We show the upper bound as an example, and the lower bound can be treated similarly. We have

$$(w_u, \tau^*) = \underset{w \in \mathcal{K}_f}{\operatorname{argmax}} \left\{ \underset{\tau \geqslant 0}{\operatorname{argmax}} \; \mathbb{E}_w[\tau \cdot r(s,a)] \,\middle|\, \mathbb{E}_w[\Delta(x;\tau,\phi)] = 0 \right\}$$

$$= \underset{w \in \mathcal{K}_f}{\operatorname{argmax}} \left\{ \underset{\tau \geqslant 0}{\operatorname{argmax}} \; \mathbb{E}_w[\tau \cdot r(s,a)] - \alpha \mathbb{E}_p[h(\tau)] \,\middle|\, \mathbb{E}_w[\Delta(x;\tau,\phi)] = 0 \right\}, \tag{39}$$

where the equality comes from Nachum et al. (2019b, Theorem 4) and the fact the regularization $\mathbb{E}_p[h(\tau)]$ does not depend on w. Then, we can solve (39) alternatively for (w_u, τ^*) by Lagrangian,

$$\max_{\tau \geqslant 0} \min_{\beta} \max_{w \in \mathcal{K}_f} \; \mathbb{E}_w\left[\tau \cdot r(s,a) + \beta^\top \Delta(x;\tau,\phi)\right] - \alpha \mathbb{E}_p[h(\tau)]. \tag{40}$$

Although the optimal $\tilde{\beta}^*$ to (40) differs from β^*, (w_u, τ^*) are the same. Once we have the (w_u, τ^*), we can recover the original Lagrangian $\tilde{\rho}_\pi(w_u) = \mathbb{E}_{w_u}[\tau \cdot r(s,a)]$, since $\mathbb{E}_{w_u}\left[\beta^{*\top} \Delta(x;\tau^*,\phi)\right] = 0$ in the original Lagrangian $\mathbb{E}_w(\ell(x;\tau^*,\beta^*))$ in (11) due to the KKT condition.

Comparing to the original ALM, the new regularization takes the advantage of ALM while keeps the original computational efficiency.

E Proofs for Statistical Properties

In this section, we provide the detailed proofs for the asymptotic coverage Theorem 2 and the finite-sample correction Theorem 4. For notation simplicity, we use $\sup, \max$ and $\inf, \min$ interchangeably. With a little abuse of notation, we use $\int$ as $\sum$ on discrete domain.

E.1 Asymptotic Coverage

Theorem 2 follows from a result in Duchi et al. (2016). The following notation will be needed:

- $\ell(x;\tau,\beta) = (1-\gamma)\beta^\top \phi(s_0,a_0) + \tau(s,a)\left(r(s,a) + \gamma\beta^\top\phi(s',a') - \beta^\top\phi(s,a)\right)$;
- $\|f\|_1 := \int |f(s,a)|\, d^{\mathcal{D}}(s,a)\, ds da$, and $\|\phi(s,a)\|_2 := \sqrt{\langle \phi, \phi \rangle}$;
- $\|f(s,a)\|_{L^2(d^{\mathcal{D}})} := \mathbb{E}_{d^{\mathcal{D}}}\left[f^2(s,a)\right]^{\frac{1}{2}}$, $\mathcal{H} \subset L^2\left(d^{\mathcal{D}}\right)$, we define $L^\infty(\mathcal{H})$ be the space of bounded linear functionals on $\mathcal{H}$ with $\|L_1 - L_2\|_{\mathcal{H}} := \sup_{h \in \mathcal{H}} |L_1 h - L_2 h|$ for $L_1, L_2 \in L^\infty(\mathcal{H})$;
- $p = \frac{dP}{d\mu}$, with a Lebesgue measure μ, is the Radon-Nikodym derivative. Abusing notation a bit, we use $(D_f(P\|Q), D(p\|q))$, and $(\mathbb{E}_P[\cdot], \mathbb{E}_p(\cdot))$ interchangeably.

Definition 7 *(Duchi et al., 2016,* **Hadamard directionally differentiability)** *Let $\mathcal{Q}$ be the space of signed measures bounded with norm $\|\cdot\|_{\mathcal{H}}$. The functional $T : \mathcal{P} \to \mathbb{R}$ is* Hadamard directionally differentiable *at $P \in \mathcal{P}$ tangentially to $B \subset \mathcal{Q}$ if for all $H \in B$, there exists $dT_p(H) \in \mathbb{R}$ such that for all convergent sequences $t_n \to 0$ and $\|H_n - H\|_{\mathcal{H}} \to 0$ that satisfies $P + t_n H_n \in \mathcal{P}$, the following holds*

$$\frac{T(P + t_n H_n) - T(P)}{t_n} \to dT_P(H), \quad as \quad n \to \infty.$$

We say $T : \mathcal{P} \to \mathbb{R}$ has an influence function *$T^1(x;P) \in \mathbb{R}$ if*

$$dT_P(Q-P) := \int T^1(x;P)\, d(Q-P)(x),$$

and $\mathbb{E}_P\left[T^1(x;P)\right] = 0$.

We consider f in D_f satisfying the following assumption (Duchi et al., 2016),

Assumption 1 (Smoothness of f-divergence) *The function $f : \mathbb{R}_+ \to \mathbb{R}$ is convex, three times differentiable in a neighborhood of 1, $f(1) = f'(1) = 0$ and $f''(1) = 2$.*[5]

Then, the following theorem, which slightly simplifies Duchi et al. (2016, Theorem 10), characterizes the asymptotic coverage of the general uncertainty estimation,

Theorem 8 (General asymptotic coverage) *Let Assumption 1 hold and $\mathcal{H} = \{h(x;\tau,\beta)\}$, where $h(x;\tau,\beta)$ is Lipschitz and the space of (τ,β) is compact. Denote $B \subset \mathcal{Q}$ be such that $\left\|\sqrt{n}\left(\widehat{P}_n - P_0\right) - G\right\|_{\mathcal{H}} \to 0$ with $G \in B$. Assume $T : \mathcal{P} \to \mathbb{R}$ is Hadamard differentiable at P_0 tangentially to B with influence function $T^1(\cdot; P_0)$ and dT_P is defined and continuous on the whole $\mathcal{Q}$, then,*

$$\lim_{n\to\infty} \mathbb{P}\left(T(P_0) \in \left\{T(P) : D_f(P||P_n) \leqslant \frac{\xi}{n}\right\}\right) = \mathbb{P}\left(\chi_1^2 \leqslant \xi\right).$$

Denote the $T(P) = \max_{\tau\geqslant 0}\min_{\beta\in\mathbb{R}^p} \mathbb{E}_P[\ell(x;\tau,\beta)]$ by convexity-concavity, our proof for Theorem 2 will be mainly checking the conditions required by Theorem 8: **i)**, Lipschitz continuity of functions in $\mathcal{H}$, and **ii)** Hadamard differentiability of $T(P)$.

We first specify the regularity assumption for stationary distribution ratio:

Assumption 2 (Stationary ratio regularity) *The target stationary state-action correction rato is bounded: $\|\tau^*\|_\infty \leqslant C_\tau < \infty$, and $\tau^* \in \mathcal{F}_\tau$ where $\mathcal{F}_\tau$ is a convex, compact and bounded RKHS space with bounded kernel function $\|k((\cdot,\cdot),(s,a))\|_{\mathcal{F}_\tau} \leqslant K$.*

The bounded ratio component of Assumption 2 is a standard assumption used in Nachum et al. (2019a); Zhang et al. (2020a); Uehara et al. (2019). The latter part regarding $\mathcal{F}_\tau$ is required for the existence of solutions. In fact, the RKHS assumption $\mathcal{F}_\tau$ is already quite flexible, and it includes deep neural networks by adopting the neural tangent kernels (Arora et al., 2019).

With Assumption 2, we can immediately obtain

$$T(P) = \max_{\tau\in\mathcal{F}_\tau}\min_{\beta\in\mathbb{R}^p} \mathbb{E}_P[\ell(x;\tau,\beta)] = \min_{\beta\in\mathbb{R}^p}\max_{\tau\in\mathcal{F}_\tau} \mathbb{E}_P[\ell(x;\tau,\beta)]$$

by the minimax theorem (Ekeland and Temam, 1999, Proposition 2.1). By this equivalence, we will focus on the $\min$-$\max$ form.

Since $r \in [0, R_{\max}]$, one has for every π that $Q^\pi \leqslant R_{\max}/(1-\gamma)$. Therefore, it is reasonable to assume the following regularity conditions for ϕ:

Assumption 3 (Embedding feature regularity) *There exist some finite constants C_β and C_ϕ, such that $\|\beta\|_2 \leqslant C_\beta$, $\|\phi\|_2 \leqslant C_\phi$. Moreover, $\phi(s,a)$ is L_ϕ-Lipschitz continuous.*

This assumption implies $\left\|\beta^\top\phi\right\|_\infty \leqslant \|\beta\|_2\|\phi\|_2 \leqslant C_\beta C_\phi$ and Lipschitz continuity of $\beta^\top\phi(s,a)$. We define $\mathcal{F}_\beta := \{\beta | \, \|\beta\|_2 \leqslant C_\beta\}$.

Lemma 9 (Lipschitz continuity) *Under Assumptions 2 and 3, function ℓ satisfies $\|\ell(x;\tau,\beta)\|_\infty \leqslant M$ and is C_ℓ-Lipschitz in (τ,β), for some proper finite constants M and C_ℓ.*

Proof We first show the boundedness claim. By the definition of $\ell(x;\tau,\beta)$, one has

$$\begin{aligned}
&\|\ell(x;\tau,\beta)\|_\infty \\
\leqslant\;& (1-\gamma)\left\|\beta^\top\phi\right\|_\infty + \left\|\tau(s,a)\left(r(s,a) + \gamma\beta^\top\phi(s',a') - \beta^\top\phi(s,a)\right)\right\|_\infty \\
\leqslant\;& (1-\gamma)\left\|\beta^\top\phi\right\|_\infty + \|\tau(s,a)\|_\infty\left(r(s,a) + \gamma\beta^\top\phi(s',a') - \beta^\top\phi(s,a)\right) \\
\leqslant\;& (1-\gamma)C_\beta C_\phi + C_\tau\left(R_{\max} + (1+\gamma)C_\beta C_\phi\right) \\
=\;& (C_\tau+1)(1-\gamma)C_\beta C_\phi + C_\tau R_{\max} := M.
\end{aligned}$$

[5]That $f(1) = 0$ is required in the definition of f-divergence. If $f'(1) \neq 0$, one can "lift" it by $\bar{f}(t) = f(t) - f'(1)(t-1)$ so that the new function satisfies $\bar{f}'(1) = 0$. $f''(1) = 2$ is assumed for easier calculation without loss of generality, as discussed in Duchi et al. (2016). For example, one can use $f(t) = 2x\log x - 2(x-1)$ for modified KL-divergence, $f(t) = (x-1)^2$ for χ^2-divergence, and $f(t) = -\log x + (x-1) - \frac{1}{2}(x-1)^2$ for reverse KL-divergence.

We equip $\mathcal{F}_\tau \times \mathcal{F}_\beta$ with the norm

$$\|(\tau, \beta)\| := \|\tau\|_{\mathcal{F}_\tau} + \|\beta\|_2 , \tag{41}$$

Then, we show the Lipschitz continuity of $\ell(x; \tau, \beta)$ in (τ, β),

$$\begin{aligned}
&|\ell(x; \tau_1, \beta_1) - \ell(x; \tau_2, \beta_2)| \\
\leqslant\;& (1-\gamma) \left|\phi(s_0, a_0)^\top (\beta_1 - \beta_2)\right| + \left|\tau_2(s,a)(\beta_1 - \beta_2)^\top (\gamma\phi(s', a') + \phi(s,a))\right| \\
+\;& \left|(\tau_1(s,a) - \tau_2(s,a))\left(r(s,a) + \gamma\beta_1^\top \phi(s', a') - \beta_1^\top \phi(s,a)\right)\right| \\
\leqslant\;& (1-\gamma)((2+\gamma) C_\phi + C_\tau) \|\beta_1 - \beta_2\|_2 + (R_{\max} + (1+\gamma) C_\phi C_\beta) |\tau_1(s,a) - \tau_2(s,a)| , \\
\leqslant\;& (1-\gamma)((2+\gamma) C_\phi + C_\tau) \|\beta_1 - \beta_2\|_2 + (R_{\max} + (1+\gamma) C_\phi C_\beta) K \|\tau_1 - \tau_2\|_{\mathcal{F}_\tau} , \\
\leqslant\;& C_\ell \left(\|\beta_1 - \beta_2\|_2 + \|\tau_1 - \tau_2\|_{\mathcal{F}_\tau}\right) ,
\end{aligned}$$

which implies the $\ell(x; \tau, \beta)$ is C_ℓ-Lipschitz continuous with

$$C_\ell := \max\{(1-\gamma)((2+\gamma) C_\phi + C_\tau, (1+\gamma) C_\phi C_\beta) K\} .$$

■

We now check the Hadamard directional differentiability of $T(P)$. The following proof largely follows Duchi et al. (2016); Römisch (2014).

Lemma 10 (Hadamard Differentiability) *Under Assumptions 2 and 3, the functional* $T(P) = \min_{\beta \in \mathcal{F}_\beta} \max_{\tau \in \mathcal{F}_\tau} \mathbb{E}_P[\ell(x; \tau, \beta)]$ *is Hadamard directionally differentiable on* $\mathcal{P}$ *tangentially to* $B(\mathcal{H}, P_0) \subset L^\infty(\mathcal{H})$ *with derivative*

$$dT_P(H) := \int \ell(x; \tau^*, \beta^*) \, dH(x) ,$$

where $(\beta^*, \tau^*) = \operatorname{argmin}_{\beta \in \mathcal{F}_\beta} \operatorname{argmax}_{\tau \in \mathcal{F}_\tau} \mathbb{E}_{P_0}[\ell(x; \tau, \beta)]$.

Proof For convenience, we define

$$\tilde{H}(\tau, \beta) := \int \ell(x; \tau, \beta) \, dH(x) ,$$

where H is associated with a measure in $\mathcal{Q}$.

We first show the upper bound convergence. For $H_n \in B(\mathcal{H}, P_0)$ with $\|H_n - H\|_{\mathcal{H}} \to 0$, for any sequence $t_n \to 0$, we have

$$\begin{aligned}
&T(P_0 + t_n H_n) - T(P_0) \\
=\;& \min_{\beta \in \mathcal{F}_\beta} \max_{\tau \in \mathcal{F}_\tau} \left(\mathbb{E}_{P_0}[\ell(x; \tau, \beta)] + t_n \tilde{H}_n(\tau, \beta)\right) - \min_{\beta \in \mathcal{F}_\beta} \max_{\tau \in \mathcal{F}_\tau} \mathbb{E}_{P_0}[\ell(x; \tau, \beta)] \\
\leqslant\;& \max_{\tau \in \mathcal{F}_\tau} \left(\mathbb{E}_{P_0}[\ell(x; \tau, \beta^*)] + t_n \tilde{H}_n(\tau, \beta^*)\right) - \mathbb{E}_{P_0}[\ell(x; \tau, \beta^*)] \\
\leqslant\;& \max_{\tau \in \mathcal{F}_\tau} t_n \tilde{H}_n(\tau, \beta^*) .
\end{aligned}$$

Denote $\tau_n^* = \operatorname{argmax}_{\tau \in \mathcal{F}_\tau} \tilde{H}_n(\tau, \beta^*)$, by definition, we have

$$\max_{\tau \in \mathcal{F}_\tau} \tilde{H}_n(\tau, \beta^*) - \max_{\tau \in \mathcal{F}_\tau} \tilde{H}(\tau, \beta^*) \leqslant \tilde{H}_n(\tau_n^*, \beta^*) - \tilde{H}(\tau_n^*, \beta^*) \leqslant \left\|\tilde{H}_n - \tilde{H}\right\|_{\mathcal{H}} \to 0.$$

Therefore, we obtain

$$\limsup_n \frac{1}{t_n} (T(P_0 + t_n H_n) - T(P_0)) \leqslant \tilde{H}(\tau^*, \beta^*) .$$

For the lower bound part, we have

$$\begin{aligned}
&T(P_0 + t_n H_n) \\
=\;& \min_{\beta \in \mathcal{F}_\beta} \left\{\max_{\tau \in \mathcal{F}_\tau} \left(\mathbb{E}_{P_0}[\ell(x; \tau, \beta)] + t_n \tilde{H}_n(\tau, \beta)\right)\right\} \\
=\;& \min_{\beta \in \mathcal{F}_\beta} \left\{\mathbb{E}_{P_0}[\ell(x; \tau_n(\beta), \beta)] + t_n \left(\tilde{H}_n(\tau_n(\beta), \beta) - \tilde{H}(\tau_n(\beta), \beta)\right) + t_n \tilde{H}(\tau_n(\beta), \beta)\right\} \\
\leqslant\;& \min_{\beta \in \mathcal{F}_\beta} \left\{\mathbb{E}_{P_0}[\ell(x; \tau_n(\beta), \beta)] + t_n \left\|\tilde{H}_n - \tilde{H}\right\|_{\mathcal{H}} + t_n \left\|\tilde{H}\right\|_{\mathcal{H}}\right\} \\
\leqslant\;& \min_{\beta \in \mathcal{F}_\beta} \mathbb{E}_{P_0}[\ell(x; \tau_n(\beta), \beta)] + \mathcal{O}(1) \cdot t_n,
\end{aligned}$$

where $\tau_n(\beta) = \text{argmax}_{\tau\in\mathcal{F}_\tau}\left(\mathbb{E}_{P_0}\left[\ell(x;\tau,\beta)\right] + t_n\tilde{H}_n(\tau,\beta)\right)$.

Denote the set of ϵ-ball of solutions w.r.t. P as

$$S_P(\epsilon) := \left\{\beta\in\mathcal{F}_\beta : \max_{\tau\in\mathcal{F}_\tau}\mathbb{E}_P\left[\ell(x;\tau,\beta)\right] \leqslant \min_{\beta\in\mathcal{F}_\beta}\max_{\tau\in\mathcal{F}_\tau}\mathbb{E}_P\left[\ell(x;\tau,\beta)\right] + \epsilon\right\}.$$

Then, $\beta_n^*\in S_{P_0+t_nH_n}(0)$ implies $\beta_n^*\in S_{P_0}(ct_n)$, which in turn implies the sequence of β_n^* has a subsequence $\tilde{\beta}_m^*$ that converges to $\beta^*\in S_{P_0}(0)$.

It is straightforward to check the Lipschitz continuity of $\bar{\ell}(\beta) := \max_\tau\mathbb{E}\left[\ell(x;\tau,\beta)\right]$ as

$$\begin{aligned}
&\left|\bar{\ell}(\beta_1) - \bar{\ell}(\beta_2)\right| \\
\leqslant\ & (1-\gamma)\left\|\beta_1-\beta_2\right\|_2\mathbb{E}_{\mu_0\pi}\left[\left\|\phi_{s_0,a_0}\right\|\right]_2 + \left|\max_{\tau\in\mathcal{F}_\tau}\mathbb{E}\left[\tau\cdot r+\beta_1^\top\Delta\right] - \max_{\tau\in\mathcal{F}_\tau}\mathbb{E}\left[\tau\cdot r+\beta_2^\top\Delta\right]\right| \\
\leqslant\ & (1-\gamma)\left\|\beta_1-\beta_2\right\|_2\mathbb{E}_{\mu_0\pi}\left[\left\|\phi_{s_0,a_0}\right\|\right]_2 + \max_{\tau\in\mathcal{F}}\left|\mathbb{E}\left[\tau\cdot r+\beta_1^\top\Delta\right] - \mathbb{E}\left[\tau\cdot r+\beta_2^\top\Delta\right]\right| \\
\leqslant\ & (1-\gamma)\left\|\beta_1-\beta_2\right\|_2\mathbb{E}_{\mu_0\pi}\left[\left\|\phi_{s_0,a_0}\right\|\right]_2 + \max_{\tau\in\mathcal{F}}\left|\mathbb{E}\left[(\beta_1-\beta_2)^\top\Delta\right]\right| \\
\leqslant\ & \left((1-\gamma)C_\phi + C_\tau(1+\gamma)C_\phi\right)\left\|\beta_1-\beta_2\right\|_2.
\end{aligned}$$

Therefore, with $\tilde{\beta}_n^*\to\beta^*$, we have

$$\lim_m\tilde{\ell}\left(\tilde{\beta}_m^*\right) = \min_\beta\tilde{\ell}(\beta) = T(P_0),$$

and due to the optimality, for any m,

$$\tilde{\ell}\left(\tilde{\beta}_m^*\right) \geqslant \min_\beta\tilde{\ell}(\beta).$$

$$\begin{aligned}
&T(P_0+t_mH_m) - T(P_0) \\
\geqslant\ & \max_{\tau\in\mathcal{F}_\tau}\left\{\mathbb{E}_{P_0}\left[\ell\left(x;\tau,\tilde{\beta}_m^*\right)\right] + t_n\tilde{H}_n\left(\tau,\tilde{\beta}_m^*\right)\right\} - \max_{\tau\in\mathcal{F}_\tau}\mathbb{E}_{P_0}\left[\ell\left(x;\tau,\tilde{\beta}_m^*\right)\right] \\
\geqslant\ & \mathbb{E}_{P_0}\left[\ell\left(x;\tau_m\left(\tilde{\beta}_m^*\right),\tilde{\beta}_m^*\right)\right] + t_n\tilde{H}_n\left(\tau_m\left(\tilde{\beta}_m^*\right),\tilde{\beta}_m^*\right) - \mathbb{E}_{P_0}\left[\ell\left(x;\tau_m\left(\tilde{\beta}_m^*\right),\tilde{\beta}_m^*\right)\right] \\
=\ & t_n\tilde{H}_n\left(\tau_m\left(\tilde{\beta}_m^*\right),\tilde{\beta}_m^*\right),
\end{aligned}$$

where $\tau_m\left(\tilde{\beta}_m^*\right) = \text{argmax}_{\tau\in\mathcal{F}_\tau}\mathbb{E}_{P_0}\left[\ell\left(x;\tau,\tilde{\beta}_m^*\right)\right]$.

Since $\tilde{\beta}_m^*\to\beta^*$, we have $\tau_m\left(\tilde{\beta}_m^*\right)\to\tau^*$, and thus,

$$\begin{aligned}
&\left|\tilde{H}_n\left(\tau_m\left(\tilde{\beta}_m^*\right),\tilde{\beta}_m^*\right) - \tilde{H}(\tau^*,\beta^*)\right| \\
\leqslant\ & \left|\tilde{H}_n\left(\tau_m\left(\tilde{\beta}_m^*\right),\tilde{\beta}_m^*\right) - \tilde{H}\left(\tau_m\left(\tilde{\beta}_m^*\right),\tilde{\beta}_m^*\right)\right| + \left|\tilde{H}\left(\tau_m\left(\tilde{\beta}_m^*\right),\tilde{\beta}_m^*\right) - \tilde{H}(\tau^*,\beta^*)\right| \\
\leqslant\ & \left\|\tilde{H}_n - \tilde{H}\right\|_{\mathcal{H}} + \left|\tilde{H}\left(\tau_m\left(\tilde{\beta}_m^*\right),\tilde{\beta}_m^*\right) - \tilde{H}(\tau^*,\beta^*)\right| \to 0,
\end{aligned}$$

where we use $\ell(\tau,\beta;x)$ is Lipschitz continuous. Therefore, we obtain

$$\liminf_n\frac{1}{t_n}\left(T(P_0+t_nH_n) - T(P_0)\right) \geqslant \tilde{H}(\tau^*,\beta^*).$$

■

Theorem 2 (Asymptotic coverage) Under Assumptions 1, 2, and 3, if $\mathcal{D}$ contains *i.i.d.* samples and the optimal solution to the Lagrangian of (5) is unique, we have

$$\lim_{n\to\infty}\mathbb{P}\left(\rho_\pi\in C_{n,\xi}^f\right) = \mathbb{P}\left(\chi_{(1)}^2\leqslant\xi\right). \tag{42}$$

Therefore, $C^f_{n,\chi^{2,1-\alpha}_{(1)}}$ is an asymptotic $(1-\alpha)$-confidence interval of the value of the policy π.

Proof The proof is to verify the conditions in Theorem 8 hold. By Lemma 6, we can rewrite

$$\mathbb{P}\left(\rho_\pi\in C_{n,\xi}^f\right) = \mathbb{P}\left(\rho_\pi\in\left\{\hat{\rho}_\pi(w)\,\middle|\,w\in\mathcal{K}_f\right\}\right),$$

where, according to the boundedness assumption on β in Assumption 3,

$$\hat{\rho}_\pi(w) = \max_{\tau \geqslant 0} \min_{\beta \in \mathcal{F}_\beta} \mathbb{E}_w\left[\tau \cdot r + \beta^\top \Delta(x; \tau, \phi)\right] = \min_{\beta \in \mathcal{F}_\beta} \max_{\tau \geqslant 0} \mathbb{E}_w\left[\tau \cdot r + \beta^\top \Delta(x; \tau, \phi)\right].$$

With Lemma 9 and Lemma 10, the conditions in Theorem 8 are satisfied. We apply Theorem 8 on the unique optimal solution $(\tau^*, \beta^*) = \mathrm{argmin}_{\beta \in \mathcal{F}_\beta} \mathrm{argmax}_{\tau \geqslant 0} \mathbb{E}_{P_0}\left[\ell(x; \tau, \beta)\right]$. We have dT_P is a linear functional on the space of bounded measures and

$$dT_{P_0}(H) = \int \ell(x; \tau^*, \beta^*)\, dH(x),$$

with the canonical gradient given by $T^1(\cdot; P_0) = \ell(x; \tau^*, \beta^*) - \mathbb{E}_{P_0}\left[\ell(x; \tau^*, \beta^*)\right]$. ■

E.2 Finite-Sample Correction

The previous section considers the asymptotic coverage of CoinDICE. We now analyze the finite-sample effect for the estimator, for the special case $f(x) = (x-1)^2$. Thus, D_f is the χ^2-divergence.

Consider the optimization problem,

$$\max_{w \in \mathbb{R}^n} w^\top z, \quad \text{s.t.} \quad D_f(w \| \widehat{p}_n) \leqslant \frac{\xi}{n}, w \in \mathcal{P}^{n-1}(\widehat{p}_n). \tag{43}$$

The following result will be needed.

Lemma 11 *(Namkoong and Duchi, 2017,* **Theorem 1)** *Let $Z \in [0, M]$ be a random variable, $\sigma^2 = Var(Z)$ and $s_n^2 = \mathbb{E}_{\widehat{p}_n}\left[Z^2\right] - \mathbb{E}_{\widehat{p}_n}\left[Z\right]^2$ as the population and sample variance of Z, respectively. For $\xi \geqslant 0$, we have*

$$\left[\sqrt{\frac{\xi}{n} s_n^2} - \frac{M\xi}{n}\right]_+ \leqslant \max_w \left\{\mathbb{E}_w[Z] \,|\, D_f(w\|\widehat{p}_n) \leqslant \frac{\xi}{n}, w \in \mathcal{P}^{n-1}(\widehat{p}_n)\right\} - \mathbb{E}_{\widehat{p}_n}[Z] \leqslant \sqrt{\frac{\xi}{n} s_n^2}.$$

Moreover, for $n \geqslant \max\left\{2, \frac{M^2}{\sigma^2}\max\{4\sigma, 22\}\right\}$, with probability at least $1 - \exp\left(-\frac{3n\sigma^2}{5M^2}\right)$,

$$\max_w \left\{\mathbb{E}_w[Z] \,|\, D_f(w\|\widehat{p}_n) \leqslant \frac{\xi}{n}, w \in \mathcal{P}^{n-1}(\widehat{p}_n)\right\} = \mathbb{E}_{\widehat{p}_n}[Z] + \sqrt{\frac{\xi}{n} s_n^2}.$$

The follow is the symmetric version of Lemma 11, which can be obtained immediately by negating the random variable Z. For completeness, we give the proof below, which is adapted from Namkoong and Duchi (2017). Recall that the lower bound is obtained by solving the following:

$$\min_{w \in \mathbb{R}^n} w^\top z, \quad \text{s.t.} \quad D_f(w \| \widehat{p}_n) \leqslant \frac{\xi}{n}, \ w \in \mathcal{P}^{n-1}(\widehat{p}_n). \tag{44}$$

Lemma 12 (Lower bound variance representation) *Under the same conditions in Lemma 11, for $\xi \geqslant 0$, we have*

$$\left[\sqrt{\frac{\xi}{n} s_n^2} - \frac{M\xi}{n}\right]_+ \leqslant \mathbb{E}_{\widehat{p}_n}[Z] - \min_w \left\{\mathbb{E}_w[Z] \,|\, D_f(w\|\widehat{p}_n) \leqslant \frac{\xi}{n}, w \in \mathcal{P}^{n-1}(\widehat{p}_n)\right\} \leqslant \sqrt{\frac{\xi}{n} s_n^2}.$$

Moreover, for $n \geqslant \max\left\{2, \frac{M^2}{\sigma^2}\max\{4\sigma, 22\}\right\}$, with probability at least $1 - \exp\left(-\frac{3n\sigma^2}{5M^2}\right)$,

$$\min_w \left\{\mathbb{E}_w[Z] \,|\, D_f(w\|\widehat{p}_n) \leqslant \frac{\xi}{n}, w \in \mathcal{P}^{n-1}(\widehat{p}_n)\right\} = \mathbb{E}_{\widehat{p}_n}[Z] - \sqrt{\frac{\xi}{n} s_n^2}.$$

Proof Denote $u = \frac{1}{n} - w$, we have $u^\top \mathbf{1} = 0$, and the optimization (44) can be written as

$$\bar{z} - \max_u u^\top(z - \bar{z}), \quad \text{s.t.} \quad \|u\|_2^2 \leqslant \frac{\xi}{n}, u^\top \mathbf{1} = 0, u \leqslant \frac{1}{n}, \tag{45}$$

with $\bar{z} = \frac{1}{n}\sum_{i=1}^n z_i$. Obviously, by the Cauchy-Schwartz inequality,

$$u^\top(z - \bar{z}) \leqslant \sqrt{\frac{\xi}{n}} \|z - \bar{z}\|_2,$$

and the equality holds if and only if

$$u_i = \frac{\sqrt{\xi}\,(z-\bar{z})}{n\,\|z-\bar{z}\|_2} = \frac{\sqrt{\xi}\,(z-\bar{z})}{n\sqrt{ns_n^2}}.$$

Given the constraint $u \leqslant \frac{1}{n}$, to achieve the maximum, one needs to ensure

$$\max_i \frac{\sqrt{\xi}\,(z-\bar{z})}{n\sqrt{ns_n^2}} \leqslant 1.$$

If this condition is satisfied, we have

$$\mathbb{E}_{\widehat{p}_n}[Z] - \min_w \left\{\mathbb{E}_w(Z)\,|D_f(w\|\widehat{p}_n) \leqslant \frac{\xi}{n}\right\} \leqslant \sqrt{\frac{\xi}{n}s_n^2}.$$

Since $z \in [0, M]$, we have $|z_i - \bar{z}| \leqslant M$, to ensure the condition, we need $\frac{\xi M^2}{ns_n^2} \leqslant 1 \Leftrightarrow s_n^2 \geqslant \frac{\xi M^2}{n}$. Otherwise, suppose $s_n^2 < \frac{\xi M^2}{n}$, or equivalently $\frac{\xi s_n^2}{n} < \frac{\xi^2 M^2}{n}$, then,

$$\min_w w^\top z \leqslant \mathbb{E}_{\widehat{p}_n}[z] - \left[\sqrt{\frac{\xi}{n}s_n^2} - \frac{M\xi}{n}\right]_+.$$

For the high-probability statement, when $n \geqslant \max\left\{2, \frac{M^2}{\sigma^2}\max\{4\sigma, 22\}\right\}$, and the event $s_n^2 \geqslant \frac{3}{64}\sigma^2$ holds, $s_n^2 \geqslant \frac{\xi M^2}{n}$. Following Maurer and Pontil (2009, Theorem 10), one can bound that

$$\mathbb{P}(|s_n - \sigma| \leqslant a) \leqslant \exp\left(-\frac{na^2}{2M^2}\right).$$

Setting $a = \left(1 - \frac{\sqrt{3}}{8}\right)\sigma$ completes the proof. ■

With Lemma 11 and Lemma 12, we represent the confidence bounds with variance. We resort to an empirical Bernstein bound applied to the function space $\mathcal{F}$ with bounded function $h : \mathcal{X} \to [0, M]$, using empirical ℓ_∞-covering numbers, $\mathcal{N}_\infty(\mathcal{F}, \epsilon, n)$,

Lemma 13 *(Maurer and Pontil, 2009,* **Theorem 6)** *Let $n \geqslant \frac{8M^2}{t}$ and $t \geqslant \log 12$. Then, with probability at least $1 - 6\mathcal{N}_\infty(\mathcal{F}, \epsilon, 2n)\,e^{-t}$, for any $h \in \mathcal{F}$,*

$$\mathbb{E}[h] - \mathbb{E}_{\widehat{p}_n}[h] \leqslant \sqrt{\frac{18Var_{\widehat{p}_n}(h)\,t}{n}} + \frac{15Mt}{n} + 2\left(1 + 2\sqrt{\frac{2t}{n}}\right)\epsilon.$$

Theorem 4 (Finite-sample correction) *Denote by $\mathcal{N}_\infty(\mathcal{F}_\tau, \epsilon, 2n)$ and $\mathcal{N}_\infty(\mathcal{F}_\beta, \epsilon, 2n)$ the ℓ_∞-covering numbers of $\mathcal{F}_\tau$ and $\mathcal{F}_\beta$ with ϵ-ball on $2n$ empirical samples, respectively. Let D_f be χ^2-divergence. Under Assumptions 2 and 3, let $M := (C_\tau + 1)(1-\gamma)C_\beta C_\phi + C_\tau R_{\max}$ and $C_\ell := \max\{(1-\gamma)((2+\gamma)C_\phi + C_\tau, (1+\gamma)C_\phi C_\beta)K\}$, then, we have*

$$\mathbb{P}(\rho_\pi \in [l_u - \zeta_n, u_n + \zeta_n]) \geqslant 1 - 12\mathcal{N}_\infty(\mathcal{F}_\tau, \epsilon, 2n)\,\mathcal{N}_\infty(\mathcal{F}_\beta, \epsilon, 2n)\,e^{-\frac{\xi}{18}},$$

where (l_n, u_n) are the solutions to (11), $\zeta_n = \frac{11M\xi}{6n} + 2\left(1 + 2\sqrt{\frac{\xi}{9n}}\right)C_\ell\epsilon$ and $\xi = \chi^{2,1-\alpha}_{(1)}$.

When the `VC`*-dimensions of $\mathcal{F}_\tau$ and $\mathcal{F}_\beta$ (denoted by $d_{\mathcal{F}_\phi}$ and $d_{\mathcal{F}_\beta}$, respectively) are finite, we have*

$$\mathbb{P}(\rho_\pi \in [l_n - \kappa_n, u_n + \kappa_n]) \geqslant 1 - 12\exp\left(c_1 + 2\left(d_{\mathcal{F}_\tau} + d_{\mathcal{F}_\beta} - 1\right)\log n - \frac{\xi}{18}\right),$$

where $c_1 = 2c + \log d_{\mathcal{F}_\tau} + \log d_{\mathcal{F}_\beta} + \left(d_{\mathcal{F}_\tau} + d_{\mathcal{F}_\beta} - 1\right)$, and $\kappa_n = \frac{11M\xi}{6n} + 2\frac{C_\ell M}{n}\left(1 + 2\sqrt{\frac{\xi}{9n}}\right)$.

Proof We focus on the upper bound, and the lower bound can be bounded in a similar way. Define

$$(\tau^*, \beta^*) := \operatorname*{argmax}_{\tau\in\mathcal{F}_\tau}\operatorname*{argmin}_{\beta}\mathbb{E}_{d^{\mathcal{D}}}[\ell(x;\tau,\beta)]$$

$$\left(\widehat{w}^*, \hat{\tau}^*, \hat{\beta}^*\right) := \operatorname*{argmax}_{w}\operatorname*{argmax}_{\tau\in\mathcal{F}_\tau}\operatorname*{argmin}_{\beta}\mathbb{E}_w[\ell(x;\tau,\beta)].$$

By definition and the optimality of β^*, we have

$$\rho_\pi = \mathbb{E}_{d^{\mathcal{D}}}\left[\ell\left(x;\tau^*,\beta^*\right)\right] \leqslant \mathbb{E}_{d^{\mathcal{D}}}\left[\ell\left(x;\tau^*,\hat{\beta}^*\right)\right]. \tag{46}$$

Applying Lemma 13 and the Lipschitz-continuity of $\ell\left(x;\tau,\beta\right)$ on $\mathcal{F}_\tau\times\mathcal{F}_\beta$, with probability at least $1-6\mathcal{N}_\infty\left(\mathcal{F}_\tau,\epsilon,2n\right)\mathcal{N}_\infty\left(\mathcal{F}_\beta,\epsilon,2n\right)e^{-t}$, we have

$$\begin{aligned}
&\mathbb{E}_{d^{\mathcal{D}}}\left[\ell\left(x;\tau^*,\hat{\beta}^*\right)\right] \\
\leqslant\;& \mathbb{E}_{\widehat{p}_n}\left[\ell\left(x;\tau^*,\hat{\beta}^*\right)\right]+3\sqrt{\frac{2Var_{\widehat{p}_n}\left(\ell\left(x;\tau^*,\hat{\beta}^*\right)\right)t}{n}}+\frac{15Mt}{n}+2\left(1+2\sqrt{\frac{2t}{n}}\right)C_\ell\epsilon \\
\leqslant\;& \max_{D_f(w||\widehat{p}_n)\leqslant\frac{\xi}{n}}\mathbb{E}_w\left[\ell\left(x;\tau^*,\hat{\beta}^*\right)\right]-\left[\sqrt{\frac{\xi Var_{\widehat{p}_n}\left(\ell\left(x;\tau^*,\hat{\beta}^*\right)\right)}{n}}-\frac{M\xi}{n}\right]_+ \\
&+3\sqrt{\frac{2Var_{\widehat{p}_n}\left(\ell\left(x;\tau^*,\hat{\beta}^*\right)\right)t}{n}}+\frac{15Mt}{n}+2\left(1+2\sqrt{\frac{2t}{n}}\right)C_\ell\epsilon \\
\leqslant\;& \max_{D_f(w||\widehat{p}_n)\leqslant\frac{\xi}{n}}\max_{\tau\in\mathcal{F}_\tau}\min_{\beta\in\mathcal{F}_\beta}\mathbb{E}_w\left[\ell\left(x;\tau,\beta\right)\right]+\frac{11}{6n}M\xi+2\left(1+2\sqrt{\frac{2t}{n}}\right)C_\ell\epsilon
\end{aligned}$$

where the second equation comes from Lemma 11 and the third line comes from setting $t\leqslant\frac{\xi}{18}$ and the definition of $\hat{\beta}^*$. Combining this with (46), we may conclude that with probability at least $1-6\mathcal{N}_\infty\left(\mathcal{F}_\tau,\epsilon,2n\right)\mathcal{N}_\infty\left(\mathcal{F}_\beta,\epsilon,2n\right)e^{-\frac{\xi}{18}}$,

$$\rho_\pi\leqslant\max_{D_f(w||\widehat{p}_n)\leqslant\frac{\xi}{n}}\max_{\tau\in\mathcal{F}_\tau}\min_{\beta\in\mathcal{F}_\beta}\mathbb{E}_w\left[\ell\left(x;\tau,\beta\right)\right]+\frac{11M\xi}{6n}+2\left(1+2\sqrt{\frac{\xi}{9n}}\right)C_\ell\epsilon.$$

With the same strategy based on Lemma 12 and Lemma 13, we can also bound the finite-sample lower bound correction that with probability at least $1-6\mathcal{N}_\infty\left(\mathcal{F}_\tau,\epsilon,2n\right)\mathcal{N}_\infty\left(\mathcal{F}_\beta,\epsilon,2n\right)e^{-\frac{\xi}{18}}$,

$$\rho_\pi\geqslant\max_{D_f(w||\widehat{p}_n)\leqslant\frac{\xi}{n}}\max_{\tau\in\mathcal{F}_\tau}\min_{\beta\in\mathcal{F}_\beta}\mathbb{E}_w\left[\ell\left(x;\tau,\beta\right)\right]-\frac{11M\xi}{6n}-2\left(1+2\sqrt{\frac{\xi}{9n}}\right)C_\ell\epsilon.$$

The first part of the theorem is then proved.

For the second part, by van der Vaart and Wellner (1996, Theorem 2.6.7), one can bound $\mathcal{N}\left(\mathcal{F},\epsilon,2n\right)\leqslant c\mathtt{VC}\left(\mathcal{F}\right)\left(\frac{16Mne}{\epsilon}\right)^{\mathtt{VC}(\mathcal{F})-1}$ for some constant c. We set $\epsilon=\frac{M}{n}$ and denote $d_{\mathcal{F}}=\mathtt{VC}\left(\mathcal{F}\right)$. Plugging this into the bound, we obtain

$$\mathbb{P}\left(\rho_\pi\in\left[l_n-\kappa,u_n+\kappa\right]\right)\geqslant 1-12\exp\left(c_1+2\left(d_{\mathcal{F}_\tau}+d_{\mathcal{F}_\beta}-1\right)\log n-\frac{\xi}{18}\right),$$

where c_1 and κ are as given in the theorem statement. ■

F Implementing Principles of Optimism and Pessimism

Based on the discussion in Section 5, the optimism and pessimism principles can be implemented by maximizing $u_{\mathcal{D}}\left(\pi\right)$ and $l_{\mathcal{D}}\left(\pi\right)$, respectively. In this section, we first calculate the gradient $\nabla_\pi u_{\mathcal{D}}\left(\pi\right)$ and $\nabla_\pi l_{\mathcal{D}}\left(\pi\right)$, and elaborate on the algorithm details.

Since we will optimize the policy π, we modify the confidence interval estimator in CoinDICE slightly, so that π is an explicitly parameterized distribution. Concretely, we consider the samples $x:=\left(s_0,s,a,r\right)$ with $s_0\sim\mu_0\left(s\right)$, $\left(s,a,r,s'\right)\sim d^{\mathcal{D}}$, which leads to the corresponding upper and lower bounds with

$$\tilde{\ell}\left(x;\tau,\beta,\pi\right):=\tau\cdot r+\beta^\top\tilde{\Delta}\left(x;\tau,\phi,\pi\right),$$

where $\tilde{\Delta}(x;\tau,\phi,\pi) = (1-\gamma)\mathbb{E}_{\pi(a_0|s_0)}[\phi(s_0,a_0)] + \gamma\mathbb{E}_{\pi(a'|s')}[\phi(s',a')\tau(s,a)] - \phi(s,a)\tau(s,a)$.

Theorem 14 *Given optimal $(\beta_l^*, \tau_l^*, w_l^*)$ and $(\beta_u^*, \tau_u^*, w_u^*)$ for lower and upper bounds, respectively, the gradients of $l_{\mathcal{D}}(\pi)$ and $u_{\mathcal{D}}(\pi)$ can be computed as*

$$\begin{bmatrix}\nabla_\pi l_{\mathcal{D}}(\pi)\\ \\ \nabla_\pi u_{\mathcal{D}}(\pi)\end{bmatrix} = \begin{bmatrix}\mathbb{E}_{w_l^*}\left[(1-\gamma)\mathbb{E}_{a_0\sim\pi}\left[\nabla_\pi\log\pi(a_0|s_0)\,{\beta_l^*}^\top\phi(s_0,a_0)\right] + \gamma\mathbb{E}_{a'\sim\pi(a'|s')}\left[\tau_l^*(s,a)\nabla_\pi\log\pi(a'|s')\,{\beta_l^*}^\top\phi(s',a')\right]\right]\\ \mathbb{E}_{w_u^*}\left[(1-\gamma)\mathbb{E}_{a_0\sim\pi}\left[\nabla_\pi\log\pi(a_0|s_0)\,{\beta_u^*}^\top\phi(s_0,a_0)\right] + \gamma\mathbb{E}_{a'\sim\pi(a'|s')}\left[\tau_u^*(s,a)\nabla_\pi\log\pi(a'|s')\,{\beta_u^*}^\top\phi(s',a')\right]\right]\end{bmatrix} \tag{47}$$

Proof We focus on the computation of $\nabla_\pi u_{\mathcal{D}}(\pi)$ with the optimal $(\beta_u^*, \tau_u^*, w_u^*)$:

$$\begin{aligned}
&\nabla_\pi u_{\mathcal{D}}(\pi) = \mathbb{E}_{w_u^*}\left[\nabla_\pi\tilde{\ell}(x;\tau,\beta)\right]\\
=\;& (1-\gamma)\mathbb{E}_{w_u^*}\nabla_\pi\mathbb{E}_{a_0\sim\pi}\left[{\beta_u^*}^\top\phi(s_0,a_0)\right] + \gamma\mathbb{E}_{w_u^*}\left[\tau_u^*(s,a)\nabla_\pi\mathbb{E}_{a'\sim\pi(a'|s')}\left[{\beta_u^*}^\top\phi(s',a')\right]\right]\\
=\;& (1-\gamma)\mathbb{E}_{w_u^*}\mathbb{E}_{a_0\sim\pi}\left[\nabla_\pi\log\pi(a_0|s_0)\,{\beta_u^*}^\top\phi(s_0,a_0)\right] && (48)\\
&+\gamma\mathbb{E}_{w_u^*}\mathbb{E}_{a'\sim\pi(a'|s')}\left[\tau_u^*(s,a)\nabla_\pi\log\pi(a'|s')\,{\beta_u^*}^\top\phi(s',a')\right]. && (49)
\end{aligned}$$

The case for the lower bound can be obtained similarly:

$$\begin{aligned}
\nabla_\pi l_{\mathcal{D}}(\pi) = (1-\gamma)\mathbb{E}_{w_l^*}\mathbb{E}_{a_0\sim\pi}\left[\nabla_\pi\log\pi(a_0|s_0)\,{\beta_l^*}^\top\phi(s_0,a_0)\right] \qquad (50)\\
+\gamma\mathbb{E}_{w_l^*}\mathbb{E}_{a'\sim\pi(a'|s')}\left[\tau_l^*(s,a)\nabla_\pi\log\pi(a'|s')\,{\beta_l^*}^\top\phi(s',a')\right].
\end{aligned}$$

■

Now, we are ready to apply the policy gradient upon $u_{\mathcal{D}}(\pi)$ or $l_{\mathcal{D}}(\pi)$ to implement the optimism for exploration or pessimism for safe policy improvement, respectively. We illustrate the details in Algorithm 2.

Algorithm 2 CoinDICE-OPT: implementation of optimism/pessimism principle

Inputs: initial policy π_0, a desired confidence $1-\alpha$, a finite sample dataset $\mathcal{D} := \{x^{(j)} = (s_0^{(j)}, s^{(j)}, a^{(j)}, r^{(j)}, s'^{(j)})\}_{j=1}^n$, number of iterations T.
for $t = 1, \ldots, T$ **do**
  Estimate $(\beta_u^*, \tau_u^*, w_u^*)$ via Algorithm 1 for optimism. $\{(\beta_l^*, \tau_l^*, w_l^*)$ for pessimism.$\}$
  Sample $\left\{x^{(j)}\right\}_{j=1}^k \sim \mathcal{D}_t$, $a_0^{(j)} \sim \pi_t(s_0^{(j)})$, $a^{(j)\prime} \sim \pi_t(s^{(j)\prime})$ for $j = 1, \ldots, k$.
  Estimate the stochastic approximation to $\nabla_\pi u_{\mathcal{D}_t}(\pi_t)$ via (49). $\{\nabla_\pi l_{\mathcal{D}_t}(\pi_t)$ via (50) for pessimism.$\}$
  Natural policy gradient update: $\pi_{t+1} = \mathrm{argmin}_\pi - \langle\pi, \nabla_\pi u_{\mathcal{D}_t}(\pi_t)\rangle + \frac{1}{\eta}KL(\pi||\pi_t)$.
  $\{\pi_{t+1} = \mathrm{argmin}_\pi - \langle\pi, \nabla_\pi l_{\mathcal{D}_t}(\pi_t)\rangle + \frac{1}{\eta}KL(\pi||\pi_t)$ for pessimism.$\}$
  Collect samples $\mathcal{E} = \left\{x^{(j)} = (s_0, s, a, r, s')^{(j)}\right\}_{j=1}^m$ by executing π_{t+1}, $\mathcal{D}_{t+1} = \mathcal{D}_t \cup \mathcal{E}$.
  $\{$Skip the data collection step in offline setting.$\}$
end for
Return π_T.