OFFLINE POLICY INTERVAL ESTIMATION WITHOUT SUFFICIENT EXPLORATION OR REALIZABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the problem of offline policy evaluation (OPE), where the goal is to estimate the value of given decision-making policy without interacting with the actual environment. In particular, we consider the interval-based OPE, where the output is an interval rather than a point, indicating the uncertainty of the evaluation. The interval-based estimation is especially important in OPE since, when the data coverage is insufficient relative to the complexity of the environmental model, any OPE method can be biased even with infinite sample size. In this paper, we characterize such irreducible biases in terms of the discrepancy between the target policy and the data-sampling distribution, and show that the marginal-importance-sampling (MIS) estimator achieves the minimax bias with an appropriate importance-weight function. Motivated with this result, we then propose a new interval-based MIS estimator that asymptotically achieves the minimax bias.

1 INTRODUCTION

The offline policy evaluation (OPE) is the art of estimating the value of given decision-making policies based on offline datasets without interacting with the actual environment. Since the interaction with the environment is often infeasible or expensive in many real-world applications, it is better to evaluate the value offline rather than online.

In the literature, it is understood from theoretical perspectives that there are two fundamental conditions for OPE to be successful: *sufficient exploration*, the coverage of the data-sampling distribution over the state-action space relative to the target policy, and *realizability*, the knowledge of correct environmental model with bounded complexity. In particular, if neither of these two conditions are met in a certain manner, it is known that OPE is never sample efficient, i.e., it takes prohibitively large sample to make the estimation reasonably accurate (Wang et al., 2020; Zanette, 2021). In practice, given a problem instance of OPE, consisting of an environment and a dataset, it is difficult to confirm that these conditions hold or to modify the problem instance so that these conditions hold, making the existing theoretical guarantees less practical. Towards practical OPE, we set our research objective to develop a theoretically-sound value estimator without assuming these two conditions.

Towards our objective, we first analyze the statistical performance of OPE methods when the two assumptions do not hold (Section 4). The key quantity is the information-theoretic worst-case bias of the value estimator (Eq. (5)) and its minimum termed the *minimax bias* (Eq. (6)), which is positive when there exist multiple indistinguishable environments, given only a problem instance of OPE. In fact, we show that the minimax bias can be non-zero if we do not assume the two conditions (Corollary 4.2). It suggests that, without the two assumptions, there exists a problem instance that any point-based value estimator is not reliable.

Given the existence of irreducible bias, we propose an alternative formulation of offline policy evaluation called *minimax-bias offline policy interval estimation (minimax-bias OPI)*, where the objective is to estimate the shortest possible interval containing the true value, instead of a point estimate (Section 5). Since our characterization of the minimax bias allows us to define the optimal interval (Definition 5.1), the minimax-bias OPI is formulated as a problem to estimate the optimal interval (Problem 5.1).

We provide a theoretical foundation to solve the minimax-bias OPI based on the marginal importance sampling estimator (Section 6). The key result is that the optimal importance weight mini-

mizing the distributional Bellman residual (DBR) allows us to construct an approximately optimal interval (Theorem 6.3). This illustrates that our problem setting is well-posed and can be solved under realistic assumptions if we can solve the minimization of DBR. Accordingly, we develop a novel algorithm in Section 7 to find the best importance weight function, which results in an interval estimator applicable even if the two fundamental conditions do not hold (Theorem 7.7).

Before proceeding to these results, we introduce basic mathematical notation in the rest of this section, review the related work in Section 2, and introduce the useful OPE-specific notation in Section 3.

Mathematical notation. Let I denote the identity operator and let $a \vee b \coloneqq \max\{a, b\}$ and $a \wedge b \coloneqq \min\{a, b\}$ denote the maximum and minimum operators for $a, b \in \mathbb{R}$, respectively.

Let $\mathcal{X}$ be a metric space with Borel algebra Σ. Let $\mathscr{B}(\mathcal{X})$ and $\mathscr{C}(\mathcal{X})$ be the spaces of the real-valued measurable bounded functions and the continuous functions on $\mathcal{X}$, respectively, both of which is equipped with the uniform norm $\|f\|_\infty \coloneqq \sup_{x \in \mathcal{X}} |f(x)|$. Let $\mathscr{M}(\mathcal{X})$ denote the space of the finite signed measures on the same space $\mathcal{X}$, equipped with the total variation (TV) norm $\|P\|_{\mathrm{TV}} \coloneqq \sup_{E_+, E_- \in \Sigma}\{P(E_+) - P(E_-)\}$. In particular, let $\delta_x \in \mathscr{M}(\mathcal{X})$, $x \in \mathcal{X}$, denote Dirac's delta measure. For any $f \in \mathscr{B}(\mathcal{X})$ and any $P \in \mathscr{M}(\mathcal{X})$, let $\langle f, P \rangle \coloneqq \int f(x) dP(x)$ be a shorthand for the (signed) expectation of f with respect to P. Let $\odot$ denote the importance-weighting operation given by $d(f \odot P)(x) \coloneqq f(x) dP(x)$, $f \in \mathscr{B}(\mathcal{X})$, $P \in \mathscr{M}(\mathcal{X})$. Let $L^1(P)$ be the space of the functions integrable with respect to $P \in \mathscr{M}(\mathcal{X})$, i.e., $\|f \odot P\|_{\mathrm{TV}} < \infty$.

Let $\mathscr{L}(\mathcal{V})$ denote the set of the bounded linear operator on a normed vector space $\mathcal{V}$. For any $A \in \mathscr{L}(\mathscr{M}(\mathcal{X}))$, let $A^* \in \mathscr{L}(\mathscr{B}(\mathcal{X}))$ denote the conjugate operator such that $\langle A^* f, P \rangle = \langle f, AP \rangle$ for $f \in \mathscr{B}(\mathcal{X})$ and $P \in \mathscr{M}(\mathcal{X})$.

2 Related Work

The problem of estimating the interval containing the true value has been known as *offline policy interval estimation* (OPI). This section reviews the existing studies on OPI by dividing the previous OPI methods into two categories: non-asymptotic and asymptotic methods (see Table 1 for the summary of comparison). We also discuss our contribution to the literature.

The non-asymptotic methods typically put their emphasis on the validity of the interval with any finite sample size, where intervals are valid if they contain the true value $J(\pi)$. For instance, Feng et al. (2020; 2021) compute intervals that contain the true policy value with high probability, under the realizability of the policy Q-functions q_π. Jiang and Huang (2020) also proposed an interval estimator with validity under more relaxed realizability condition that either the policy Q-function q_π or the marginal density ratio function w_π is realizable. One limitation of this approach is the theoretical understanding on the tightness of the interval is often unclear or partial. Another limitation of this approach is that they tend to require the realizability with *known complexity*. This requirement is not desirable for practical use; if we used a too complex hypothesis class such as a reproducing kernel Hilbert space with infinite radius, the resultant interval would be trivial, and thus, non-informative.

The asymptotic methods focus on the asymptotically dominant term of the uncertainty in the large sample limit, which typically allows us to theoretically understand their behavior, especially the tightness, in depth. For instance, Kallus and Uehara (2020); Shi et al. (2021) gave confidence interval estimators that achieve the efficiency lower bound. The bootstrap estimators (Hao et al., 2021) also enable us to compute the asymptotically exact confidence intervals in a more flexible manner. One major limitation is that they assume both the sufficient exploration and the realizability conditions of q_π and w_π hold, which can be hardly validated in real-world applications. These assumptions are essential to their analyses because they focus on estimation of the asymptotic *variance* of order $O(n^{-1/2})$, assuming that the *bias* is negligible. Therefore, these methods are not applicable to our setting where the asymptotic *bias* of order $O(1)$ dominates the asymptotic *variance*.

In this study, we take the asymptotic approach, but with a focus on the estimation of the *bias* rather than the *variance*, because the bias is dominant in our setting where the sufficient exploration and the realizability do not hold at all. Our contributions are threefold. First, we characterize the theoretical lower bound of the asymptotic bias through the asymptotic analysis, which serves as a theoretical

Table 1: Comparison of OPI methods. q_π and w_π denote the Q-function and the marginal density ratio function, respectively, $w_\pi^\sharp$ denotes a generalization of w_π for the insufficient exploration setting.

Method	Assumptions				Guarantee
	Asymptotic	Exploration	Realizability	Complexity	
BONDIC (Feng et al., 2021)	—	—	q_π	Known	Valid
MVI (Jiang and Huang, 2020)	—	—	q_π or w_π	Known	Valid
DRL (Kallus and Uehara, 2020) D2OPE (Shi et al., 2021)	Yes	Yes	q_π and w_π	—	Efficient
Ours	Yes	—	— $w_\pi^\sharp$	—	Valid Optimal

foundation of OPE without sufficient exploration or realizability assumptions. Second, without the two assumptions, we develop an interval estimation method that outputs an asymptotically valid interval, that is, an interval that contains the true value in the large sample limit. Third, under the realizability condition of the generalized marginal density ratio function $w_\pi^\sharp$, we show that the estimated interval is optimal.

3 Preliminaries

We first introduce our formulation of reinforcement learning and offline policy evaluation. Then, we introduce two fundamental concepts in RL, a Q-function and an occupancy measure, along with shorthand notation for them.

Offline policy evaluation. Let $\mathcal{X} \coloneqq \mathcal{S} \times \mathcal{A}$ be the state-action space of the system with $|\mathcal{X}| < \infty$.[1] Let $\mathcal{M} \coloneqq (\iota, T, R)$ be the Markov decision process (MDP) of environment on $\mathcal{X}$, where $\iota \in \mathscr{M}(\mathcal{S})$ is the initial state distribution, $T : \mathcal{X} \to \mathscr{M}(\mathcal{S})$ is the transition dynamics and $R : \mathcal{X} \to \mathscr{M}([-1, 1])$ is the conditional reward distribution. Let $\pi : \mathcal{S} \to \mathscr{M}(\mathcal{A})$ be the target policy. Then, the *value* $J(\pi)$ of π with respect to $\mathcal{M}$ is given by the γ-discounted expected average reward

$$J(\pi) \equiv J_{\mathcal{M}}(\pi) \coloneqq \mathbb{E}^{\mathcal{M},\pi}\left[(1-\gamma)\sum_{t=1}^{\infty}\gamma^{t-1} r_t\right],$$

where $\gamma \in (0, 1)$ is a discounting factor and $\mathbb{E}^{\mathcal{M},\pi}$ denotes the expectation with respect to the Markov chain generated with $a_t \sim \pi(s_t), r_t \sim R(s_t, a_t), s_{t+1} \sim T(s_t, a_t)$ for all $t \geq 1$ and $s_1 \sim \iota$.

In offline policy evaluation, we are given a dataset $\mathcal{D} \coloneqq (\mathcal{D}_\iota, \mathcal{D}_{T,R})$ as input, where $\mathcal{D}_\iota \coloneqq \{s_{\iota,j}\}_{j=1}^n$ is a set of initial states and $\mathcal{D}_{T,R} \coloneqq \{(x_i, s'_i, r_i)\}_{i=1}^n$ is a set of transition records sampled from

$$\mathrm{d}G_{\mathcal{M},\beta}(\mathcal{D}) \coloneqq \prod_{j=1}^{n} \mathrm{d}\iota(s_{\iota,j}) \cdot \prod_{i=1}^{n} \mathrm{d}\beta(x_i)\mathrm{d}T(s'_i|x_i)\mathrm{d}R(r_i|x_i),$$

where $\beta \in \mathscr{M}(\mathcal{X})$ is an arbitrary state-action-sampling distribution.[2] Then, an instance of the offline policy evaluation (OPE) is identified by the quadruple $\mathcal{P} \coloneqq (\mathcal{M}, \beta, \pi, \gamma)$ and formalized as follows.

Problem 3.1 (Offline policy evaluation, OPE)**.** *Given* $(\mathcal{D}, \pi, \gamma)$ *where* $\mathcal{D} \sim G_{\mathcal{M},\beta}$*, estimate* $J(\pi)$*.*

Q-function and occupancy measure. Let $\iota_\pi \in \mathscr{M}(\mathcal{X})$ and $T_\pi \in \mathscr{L}(\mathscr{M}(\mathcal{X}))$ be the initial state-action distribution and the state-action transition operator associated with π such that $\mathrm{d}\iota_\pi(s, a) \coloneqq \mathrm{d}\iota(s)\mathrm{d}\pi(a|s)$ and $\mathrm{d}(T_\pi P)(s, a) \coloneqq \int \mathrm{d}T(s|x)\mathrm{d}\pi(a|s)\mathrm{d}P(x)$ for $s \in \mathcal{S}$, $a \in \mathcal{A}$ and $P \in \mathscr{M}(\mathcal{X})$,

[1] Note that there is no explicit dependency on $|\mathcal{X}|$ in all of our analysis. $\mathcal{X}$ can be generalized to a compact Hausdorff space in exchange for the simplicity of the exposition.

[2] For simplicity, we assume the sample sizes of $\mathcal{D}_\iota$ and $\mathcal{D}_{T,R}$ are the same. The generalization with different sample sizes is possible with minor modification.

respectively. Also let $\rho \in \mathscr{B}(\mathcal{X})$ be the expected reward function such that $\rho(x) \coloneqq \int r \mathrm{d}R(r|x)$ for $x \in \mathcal{X}$. Then, the value $J(\pi)$ is rewritten as

$$J(\pi) = \langle \rho, \Gamma_\pi \iota_\pi \rangle = \langle \Gamma_\pi^* \rho, \iota_\pi \rangle = \langle \rho, \mu_\pi \rangle = \langle q_\pi, \iota_\pi \rangle, \tag{1}$$

where $\Gamma_\pi \coloneqq (1-\gamma) \sum_{t=1}^\infty (\gamma T_\pi)^{t-1} \in \mathscr{L}(\mathscr{M}(\mathcal{X}))$ is the accumulation operator, $\mu_\pi \coloneqq \Gamma_\pi \iota_\pi \in \mathscr{M}(\mathcal{X})$ is the normalized occupancy measure of π (henceforth *the occupancy measure*), and $q_\pi \coloneqq \Gamma_\pi^* \rho \in \mathscr{B}(\mathcal{X})$ is the normalized Q-function of π (henceforth *the Q-function*). Note that we have $\|q\|_\infty \leq 1$ and $\|\mu_\pi\|_{\mathrm{TV}} = 1$ thanks to the normalization.

Two Bellman equations. One of the essential difficulties of OPE lies in the fact the direct estimation of the accumulation operator Γ_π (and hence μ_π and q_π) is intractable due to the infinite sum. The Bellman equation is useful to mitigate this problem. Here, we introduce two variants of the Bellman equation, the *functional and distributional Bellman equations*, given by

$$\rho = \Delta_\pi^* q_\pi, \qquad \iota_\pi = \Delta_\pi \mu_\pi, \tag{2}$$

where $\Delta_\pi \coloneqq \Gamma_\pi^{-1} = (I - \gamma T_\pi)/(1-\gamma)$ is the difference operator. Note that, in the Bellman equations, both q_π and μ_π are *uniquely* characterized via more directly estimatable quantities (ρ, T_π) and (ι_π, T_π), respectively.

When approximating the Q-function or the occupancy measure by a model, the learned model does not completely comply with the Bellman equation. Such a discrepancy can be used to measure the approximation error. In specific, to measure the approximation error of w to w_π, we employ the distributional Bellman residual (DBR),

$$\mathcal{R}_\pi(w) \coloneqq \iota_\pi - \Delta_\pi(w \odot \beta) \in \mathscr{M}(\mathcal{X}),$$

which takes zero if and only if $w = w_\pi$ according to the Bellman equation (2).

Empirical estimates. Finally, we introduce the empirical estimates of $(\iota_\pi, T_\pi, \rho, \beta)$ based on the dataset $\mathcal{D}$ as follows. For all $P \in \mathscr{M}(\mathcal{X})$ and $x \in \mathcal{X}$,

$$\hat{\iota}_\pi \coloneqq \frac{1}{n} \sum_{j=1}^n \delta_{x_{\iota,j}}, \quad \hat{T}_\pi P \coloneqq \sum_{i=1}^n \frac{\delta_{x'_i}}{N(x_i)} P(\{x_i\}), \quad \hat{\rho}(x) \coloneqq \frac{1}{N(x)} \sum_{i: x_i = x} r_i, \quad \hat{\beta} \coloneqq \frac{1}{n} \sum_{i=1}^n \delta_{x_i}, \tag{3}$$

where $N(x) \coloneqq 1 \vee |\{i \,:\, x_i = x\}|$ is the data-counting function (with the zero-division safeguard) and $x_{\iota,i} \coloneqq (s_{\iota,i}, a_{\iota,i})$ and $x'_i \coloneqq (s'_i, a'_i)$ are the state-action pairs associated with additional samples $a_{\iota,i} \sim \pi(s_{\iota,i})$ and $a'_i \sim \pi(s'_i)$, respectively.

Throughout this paper, we employ the conventional marginal importance sampling (MIS) estimator (Liu et al., 2018; Xie et al., 2019) to estimate the value in offline. The MIS estimator associated with a weight function $w \in \mathscr{B}(\mathcal{X})$ is given by

$$\hat{J}(w) \coloneqq \langle \hat{\rho},\ w \odot \hat{\beta} \rangle. \tag{4}$$

The MIS estimator is justified if the weight function w is equal to the marginal density $w_\pi \coloneqq \frac{\mathrm{d}\mu_\pi}{\mathrm{d}\beta}$ (assuming it exists) since, in that case, the MIS estimator is unbiased, $\mathbb{E}[\hat{J}(w_\pi)] = J(\pi)$, according to (1).

Note, however, that w_π does not exist when the exploration is insufficient, $\beta \not\gg \mu_\pi$, and the unbiasedness cannot be guaranteed in general. Two natural questions thus arise: Does the MIS estimator still enjoy any theoretical guarantee in such a general setting? If so, what is the best weight function w? In short, the answer to the first question is affirmative and the answer to the second question is one of the main contributions of this work.

4 Irreducible Bias in Offline Policy Evaluation

In this section, we theoretically analyze the statistical performance of OPE methods without sufficient exploration or realizability assumptions. As a result, we will show that any OPE method must

incur an irreducible bias that never disapears even when the sample size goes infinity. Given such a negative result, we instead propose a novel problem setting called the *minimax-bias OPI*, where the goal is to estimate the interval that contains the true value and is as short as possible. The proposed problem setting is expected to be solved without sufficient exploration or realizability assumptions, and thus, will be of practical use.

To study the statistical performance of OPE methods, we introduce the notion of the *minimax bias* of the point-based estimators. Let $\hat{J}$ be *any* random variable representing a point-based OPE estimator. Then, the information-theoretic worst-case bias of $\hat{J}$ is given by

$$\epsilon[\hat{J}] \equiv \epsilon[\hat{J}; \mathcal{P}] \coloneqq \sup_{(\mathcal{M}',\beta)\sim(\mathcal{M},\beta)} \left| J_{\mathcal{M}'}(\pi) - \mathbb{E}\hat{J} \right|, \tag{5}$$

where the equivalence $\sim$ is defined by the equality with respect to the corresponding distributions of the dataset, *i.e.*, $G_{\mathcal{M}',\beta} = G_{\mathcal{M},\beta}$. If there exist equivalent environments $\mathcal{M}$ and $\mathcal{M}'$ that result in the different policy values $J_{\mathcal{M}}(\pi) \neq J_{\mathcal{M}'}(\pi)$ yet indistinguishable from the dataset, the worst-case bias $\epsilon[\hat{J}]$ is inevitable without an additional source of information, i.e., domain knowledge. The minimax bias is then defined as the minimum possible worst-case bias of OPE,

$$\epsilon_\star(\pi) \equiv \epsilon_\star(\pi; \mathcal{P}) \coloneqq \inf_{\hat{J}} \epsilon[\hat{J}; \mathcal{P}], \tag{6}$$

which can be thought of as a characteristic of the problem $\mathcal{P}$ indicating its hardness in terms of the irreducible uncertainty even with infinitely large sample. In fact, there exists the unique $\hat{J}$ achieving the infimum and we refer to it as the optimal point estimator $J_\star(\pi)$. Our main objective is to understand the minimax bias in various settings.

To this end, we introduce a novel concept, the projection of the occupancy measure μ_π with respect to β. Let Π_β be the projection operator onto the support of β such that $\Pi_\beta P = \chi_\beta \odot P$, $P \in \mathscr{M}(\mathcal{X})$ and $\chi_\beta(x) = \mathbf{1}\{x \in \operatorname{supp}\beta\}$.

Definition 4.1 (Projected occupancy measure and its importance weight)**.** *We refer to*

$$\mu_\pi^\sharp \coloneqq (1-\gamma)\Pi_\beta \sum_{t=0}^{\infty} (\gamma T_\pi \Pi_\beta)^t \iota_\pi \tag{7}$$

as the projected occupancy measure of π*. Correspondingly, we also refer to*

$$w_\pi^\sharp \coloneqq \frac{\mathrm{d}\mu_\pi^\sharp}{\mathrm{d}\beta}$$

as the projected importance weight of π *with respect to* β*.*

Note that $w_\pi^\sharp$ can be thought of as an extension of w_π in the sense it is always well-defined and $w_\pi^\sharp = w_\pi$ whenever w_π exists. On the other hand, $\mu_\pi^\sharp$ can be thought of as the *known component* of μ_π since it is always identifiable given $G_{\mathcal{M},\beta}$, thanks to the projection Π_β, and $\mu_\pi^\sharp = \mu_\pi$ whenever μ_π is also identifiable.

We now present our main result, which discovers close relationships between the minimax bias $\epsilon_\star(\pi)$, the MIS estimator $\hat{J}(w)$, the DBR $\mathcal{R}_\pi(w)$ and the projected importance weight $w_\pi^\sharp$.

Theorem 4.1. *For all* $w \in \mathscr{B}(\mathcal{X})$*, we have*

$$\epsilon_\star(\pi) \leq \epsilon\left[\hat{J}(w)\right] \leq \|\mathcal{R}_\pi(w)\|_{\mathrm{TV}} \leq \epsilon_\star(\pi) + \frac{1+\gamma}{1-\gamma}\|w - w_\pi^\sharp\|_{L^1(\beta)}. \tag{8}$$

Proof (sketch). The most nontrivial part is the last inequality, which follows from the constructive proof of

$$\epsilon_\star(\pi) \geq \|\mathcal{R}_\pi(w_\pi^\sharp)\|_{\mathrm{TV}}. \tag{9}$$

In particular, we construct two worst-case environments $\mathcal{M}_\pm$ under the constraint $(\mathcal{M}_\pm, \beta) \sim (\mathcal{M}, \beta)$. Roughly speaking, the environments are constructed to have a special state $\perp$ in the under-explored region of $\mathcal{X}$ absorbing all the transition to that region, and have the extreme reward there, *i.e.*, $\rho(\perp) = \pm 1$. With this explicit construction of the environments, we can give an analytic lower bound of $\epsilon_\star(\pi)$, which coincides with $\|\mathcal{R}_\pi(w_\pi^\sharp)\|_{\mathrm{TV}}$. See Section B for the complete proof. $\square$

An immediate consequence of Theorem 4.1 is that it tells us when and how the minimax bias is positive. To see this, let $\mu_\pi^{\not\sharp} \coloneqq \mu_\pi - \mu_\pi^\sharp$ be the projection residual of μ_π.

Corollary 4.2. *We have* $\epsilon_\star(\pi) = \|\Delta_\pi \mu_\pi^{\not\sharp}\|_{\mathrm{TV}}$ *and thus*

$$\|\mu_\pi^{\not\sharp}\|_{\mathrm{TV}} \le \epsilon_\star(\pi) \le \frac{1+\gamma}{1-\gamma}\|\mu_\pi^{\not\sharp}\|_{\mathrm{TV}}.$$

In other words, the minimax bias is zero if and only if the projection residual $\mu_\pi^{\not\sharp}$ is zero, or equivalently, if μ_π is absolutely continuous with respect to the data distribution β. Moreover, the size of the minimax bias is proportional to the size of the projection residual $\mu_\pi^{\not\sharp}$. It thus formally asserts the limitation of the point-based estimators in the insufficient exploration settings.

In summary, any OPE methods must be biased in the worst case whenever the exploration is insufficient, $\mu_\pi \not\ll \beta$, motivating the interval-based approach.

5 Problem Setup: Minimax-Bias OPI

As discussed above, the point-based estimator suffers from irreducible bias, suggesting a hardness in Problem 3.1 under realistic assumptions. Given such a limitation, we propose an alternative problem setting called the *minimax-bias OPI* so that we can solve it under realistic assumptions. Since there exists a bias, our idea is to estimate the value by an *interval* that contains the true value, instead of a *point*.

Let us first define the target of the estimation, which we call the *optimal interval*. As discussed earlier, since $J_\star(\pi)$ and $\epsilon_\star(\pi)$ are the best possible point estimator and its error guarantee, respectively, the optimal interval can be naturally formulated as follows.

Definition 5.1 (Optimal interval)**.** *The following is referred to as the optimal interval:*

$$\mathcal{I}_\star(\pi) \equiv \mathcal{I}_\star(\pi; \mathcal{P}) \coloneqq [J_\star(\pi) - \epsilon_\star(\pi), J_\star(\pi) + \epsilon_\star(\pi)].$$

Then, the problem of offline policy interval estimation (OPI), an uncertainty-aware interval extension of Problem 3.1, is formalized as follows.

Problem 5.1 (Minimax-bias OPI)**.** *Estimate* $\mathcal{I}_\star(\pi)$ *based on* $(\mathcal{D}, \pi, \gamma)$*, where* $\mathcal{D} \sim G_{\mathcal{M},\beta}$*.*

Towards estimating the optimal interval, let us introduce two approximate correspondences. Definition 5.2 is stronger than Definition 5.3.

Definition 5.2. *We refer to the interval* $\mathcal{I}$ *satisfying* $d_H(\mathcal{I}, \mathcal{I}_\star(\pi)) \le \epsilon$*, where* $d_H(\cdot,\cdot)$ *is the Hausdorff distance, as* ϵ*-approximately optimal. Moreover, a sequence of intervals* $\{\mathcal{I}_n\}_{n\ge1}$ *is said to be asymptotically (approximately) optimal if it converges to an (approximately) optimal interval.*

Definition 5.3 (Valid interval)**.** *An interval* $\mathcal{I} \subset \mathbb{R}$ *is said to be valid if* $\mathcal{I} \supset \mathcal{I}_\star(\pi)$*. Moreover, a sequence of intervals* $\{\mathcal{I}_n\}_{n\ge1}$ *is said to be asymptotically valid if its lower limit* $\lim_{k\to\infty}\bigcap_{n\ge k}\mathcal{I}_n$ *is valid.*

6 Theoretical Foundation of Minimax-Bias OPI

We provide a theoretical foundation to solve Problem 5.1, mainly referring to Theorem 4.1.

First, Theorem 4.1 implies that the MIS estimator $\hat{J}(w)$ with the projected importance weight $w = w_\pi^\sharp$ is optimal in the sense it achieves the minimax bias. More generally:

Corollary 6.1. *There exists* $w \in L^1(\beta)$ *such that* $\epsilon[\hat{J}(w)] = \epsilon_\star(\pi)$*.*

This motivates us to seek for the best weight function w within the MIS framework. The following corollary shows the next significant implications, that the optimal point-based and interval-based OPE is achieved with combining the MIS estimator and the minimization of DBR.

Corollary 6.2. *Let* $w_\star$ *be a minimizer of DBR in a compact hypothesis class* $\mathcal{W} \subset \mathscr{B}(\mathcal{X})$*,*

$$w_\star \in \operatorname*{argmin}_{w\in\mathcal{W}} \|\mathcal{R}_\pi(w)\|_{\mathrm{TV}}. \tag{10}$$

Then, we have

$$\left|\mathbb{E}\hat{J}(w_\star) - J_\star(\pi)\right| \leq \frac{1+\gamma}{1-\gamma}\epsilon_{\mathcal{W}}, \tag{11}$$

and

$$|\|\mathcal{R}_\pi(w_\star)\|_{\mathrm{TV}} - \epsilon_\star(\pi)| \leq \frac{1+\gamma}{1-\gamma}\epsilon_{\mathcal{W}}, \tag{12}$$

where $\epsilon_{\mathcal{W}} \coloneqq \min_{w\in\mathcal{W}} \|w - w_\pi^\sharp\|_{L^1(\beta)}$ *is the realizability error of* $\mathcal{W}$.

In other words, given $\mathcal{W}$ is expressive enough to approximate $w_\pi^\sharp$ well and thus $\epsilon_{\mathcal{W}}$ is negligible, the optimal point estimator $J_\star(\pi)$ and its uncertainty $\epsilon_\star(\pi)$ can be estimated with a solution $w_\star$ to (10) and its objective value $\|\mathcal{R}_\pi(w_\star)\|_{\mathrm{TV}}$, respectively. These observations naturally lead us to the following proxy to the optimal interval

$$\mathcal{I}(\pi;\mathcal{W}) \coloneqq \left[\mathbb{E}\hat{J}(w_\star) - \|\mathcal{R}_\pi(w_\star)\|_{\mathrm{TV}},\ \mathbb{E}\hat{J}(w_\star) + \|\mathcal{R}_\pi(w_\star)\|_{\mathrm{TV}}\right]. \tag{13}$$

In fact, it satisfies two desirable properties: the validity and the approximate optimality.

Theorem 6.3. *The interval* $\mathcal{I}(\pi;\mathcal{W})$ *is valid and* $2\frac{1+\gamma}{1-\gamma}\epsilon_{\mathcal{W}}$*-approximately optimal.*

7 Estimation of Optimal Interval

As discussed in the previous section, the estimation of the optimal interval $\mathcal{I}_\star(\pi)$ is reduced to the minimization of the TV norm $\|\mathcal{R}_\pi(w)\|_{\mathrm{TV}}$. However, even the estimation of the exact TV norm is notorious for its difficulty (e.g., see Section 5 in Sriperumbudur et al. (2012)), let alone the minimization. This motivate us to develop new variational approximations for the TV norm. In particular, we newly introduce two approximations of $\|\mathcal{R}_\pi(w)\|_{\mathrm{TV}}$, each of which is suitable for the evaluation and the optimization of the objective, respectively.

7.1 Evaluating the objective

The first approximation reduces the evaluation of the TV norm to a conventional regression problem. To see this, let us begin with the approximation formula for the TV norm of general measures.

Let $\mathcal{F}$ be a *universal function approximator* on $\mathcal{X}$, *i.e.*, a set of functions dense in $\mathscr{C}(\mathcal{X})$, such as the reproducing kernel Hilbert space (RKHS) with a universal kernel (Sriperumbudur et al., 2010) or a set of neural networks (Hornik et al., 1989).

Proposition 7.1. *For all* $P \in \mathscr{M}(\mathcal{X})$*, we have*

$$\|P\|_{\mathrm{TV}} = \sup_{f\in\mathcal{F}} \langle [\![f]\!],\ P\rangle, \tag{14}$$

where $[\![f]\!](x) \coloneqq \max\{-1, \min\{1, f(x)\}\}$ *denotes the clipping of* $f(x)$ *to* $[-1, 1]$.

The proof is relegated to Section C. Note that solving the supremum (14) is possible via the empirical risk minimization framework, given an empirical approximation $\hat{P}_n$ of P. More specifically, one may adopt structural risk minimization (SRM) principle (Vapnik, 1999) to deal with the unbounded complexity of $\mathcal{F}$.

Letting $P = \mathcal{R}_\pi(w)$, we immediately get the following special case.

Corollary 7.2. *For all* $w \in \mathscr{B}(\mathcal{X})$*, we have*

$$\|\mathcal{R}_\pi(w)\|_{\mathrm{TV}} = \sup_{f\in\mathcal{F}} \langle [\![f]\!],\ \mathcal{R}_\pi(w)\rangle. \tag{15}$$

In fact, as mentioned above, the RHS of (15) is straightforwardly estimated via the SRM principle under additional mild assumptions on $\mathcal{F}$. See Lemma D.1 for the formal statement. Specifically, we employ the empirical estimate of the DBR,

$$\hat{\mathcal{R}}_\pi(w) \coloneqq \hat{\iota}_\pi - \hat{\Delta}_\pi(w \odot \hat{\beta}).$$

The resultant evaluation procedure is summarized in Algorithm A.1. By its derivation, we can formally show that Algorithm A.1 is consistent. The proof is straightforward and thus omitted.

Theorem 7.3. *Let* `EvaluateDBR`$(\mathcal{D}, \mathcal{F}, w)$ *be the output of Algorithm A.1. Then, for all* $w \in \mathcal{W}$*, we have* `EvaluateDBR`$(\mathcal{D}, \mathcal{F}, w) \to \|\mathcal{R}_\pi(w)\|_{\mathrm{TV}}$ *in probability as* $n \to \infty$.

7.2 MINIMIZING THE OBJECTIVE

We now turn to the minimization of $\|\mathcal{R}_\pi(w)\|_{\mathrm{TV}}$ with respect to $w \in \mathcal{W}$. The previous variational formula (15) is not straightforwardly usable for this purpose since it ends up with the saddle-point problem, which we found is too unstable in practice. To mitigate this issue, we introduce a minimization-based approximation of the TV norm.

To this end, we first introduce the convolution of the TV norm with the maximum mean discrepancy (MMD) (Sriperumbudur et al., 2009). Here, the MMD of a measure $P \in \mathscr{M}(\mathcal{X})$ is given by $\mathrm{MMD}_\kappa(P) := \sqrt{\langle \kappa,\ P^{\otimes 2}\rangle}$, where $\kappa : \mathcal{X}^2 \to \mathbb{R}$ is a bounded universal kernel on $\mathcal{X}$ and $P^{\otimes 2}$ denotes the product measure of P on $\mathcal{X}^2$.

Definition 7.1 (Convolution norm)**.** *For all $P \in \mathscr{M}(\mathcal{X})$ and $u \geq 1$, we refer to*

$$\|P\|_u := \inf_{Q \ll P} \{u\,\mathrm{MMD}_\kappa(P - Q) + \|Q\|_{\mathrm{TV}}\} \tag{16}$$

as the u-convolution norm of P.

The following proposition shows the u-convolution norm is a reasonable approximation of the TV norm and admits a sample-based estimation unlike the TV norm.

Proposition 7.4. *For all $P \in \mathscr{M}(\mathcal{X})$, we have*

$$\|P\|_{\mathrm{TV}} = \lim_{u \to \infty} \|P\|_u. \tag{17}$$

Moreover, if P is a probability measure, for all $\delta \in (0, 1)$, we have

$$\|\hat{P}_n - P\|_u = O\left(\sqrt{\frac{u^2 + \ln(1/\delta)}{n}}\right) \tag{18}$$

with probability $\geq 1 - \delta$, where $\hat{P}_n := \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ is the empirical distribution of an n-sample $(x_1, ..., x_n)$ independently drawn from P, $n \geq 1$.

Proof. The key of the proof is Lemma E.1, which gives the dual representation of the convolution norm,

$$\|P\|_u = \sup_{\substack{\|f\|_{\mathcal{H}} \leq u \\ \|f\|_\infty \leq 1}} \langle f,\ P\rangle,$$

where $\mathcal{H}$ is the RKHS generated by κ. Then, the density of the universal RKHS in $\mathscr{C}(\mathcal{X})$ implies that $\lim_{u\to\infty} \|P\|_u = \sup_{\|f\|_\infty \leq 1} \langle f,\ P\rangle = \|P\|_{\mathrm{TV}}$, which proves the first statement. The second statement also follows by the uniform law of large number,

$$\|\hat{P}_n - P\|_u = \sup_{\substack{\|f\|_{\mathcal{H}} \leq u \\ \|f\|_\infty \leq 1}} \langle f,\ \hat{P}_n - P\rangle \leq 2u\mathfrak{R}_n(\mathcal{H}) + O\left(\sqrt{\frac{\ln(1/\delta)}{n}}\right),$$

where $\mathfrak{R}_n(\mathcal{H})$ denotes the Rademacher complexity of the unit ball of $\mathcal{H}$, which is of order $O(n^{-1/2})$. □

As a corollary, we obtain the following approximation formula for $\|\mathcal{R}_\pi(w)\|_{\mathrm{TV}}$.

Corollary 7.5. *For all $w \in \mathscr{B}(\mathcal{X})$, we have*

$$\|\hat{\mathcal{R}}_\pi(w)\|_u \to \|\mathcal{R}_\pi(w)\|_{\mathrm{TV}} \tag{19}$$

as $u \to \infty$ and $n/u^2 \to \infty$.

In the practical implementation, one may change the variable in the LHS of (19) from the measure $Q \in \mathscr{M}(\mathcal{X})$ to the function $g \in \mathscr{B}(\mathcal{X})$,

$$\begin{aligned}\|\hat{\mathcal{R}}_\pi(w)\|_u &= \inf_{Q \ll \hat{\mathcal{R}}_\pi(w)} \left\{u\,\mathrm{MMD}_\kappa(\hat{\mathcal{R}}_\pi(w) - Q) + \|Q\|_{\mathrm{TV}}\right\} \\ &= \min_{g \in \mathscr{B}(\mathcal{X})} \left\{u\,\mathrm{MMD}_\kappa(\hat{\mathcal{R}}_\pi(w) - g \odot \hat{\eta}_\pi) + \langle |g|,\ \hat{\eta}_\pi\rangle\right\},\end{aligned} \tag{20}$$

where $\hat{\eta}_\pi \coloneqq (\hat{\iota}_\pi + \hat{T}_\pi \hat{\beta} + \hat{\beta})/3$. Here, we have exploited the transitivity of the absolute continuity with $Q \ll \hat{\mathcal{R}}_\pi(w) \ll \hat{\eta}_\pi$. Note that the minimization with respect to g is tractable since g is only evaluated on the support of $\hat{\eta}_\pi$, which is a finite set. Moreover, the convolution-based formula (20) is stable with the minimization with respect to $w \in \mathcal{W}$ resulting in the joint minimization problem of (w, g), unlike the regression-based formula (15) resulting in the saddle-point problem.

Since the objective of (20) is lower bounded and convex with respect to (w, g), the minimizer

$$\hat{w}_u \coloneqq \operatorname*{argmin}_{w \in \mathcal{W}} \|\hat{\mathcal{R}}_\pi(w)\|_u \tag{21}$$

is computed with any convex optimization algorithms. The hyperparameter u is then chosen from a predefined grid $\mathcal{U}$ so that the TV norm of DBR is minimized. Specifically, one may choose it as the logarithmically-even grid $\mathcal{U} \coloneqq \{2^k\}_{k=0}^{k_{\max}}$, where the upper limit $k_{\max} \coloneqq \lfloor \log_2 \sqrt{\min\{n, m\}} \rfloor$ is determined according to the order of empirical approximation error (18). The entire procedure of the weight estimation is summarized in Algorithm A.2. By its derivation, we can formally show the consistency of Algorithm A.2. The proof is relegated to Section F.

Theorem 7.6. *Let* `OptimizeDBR`$(\mathcal{D}, \mathcal{F}, \mathcal{W}, \kappa)$ *be the output of Algorithm A.2. Then,* `OptimizeDBR`$(\mathcal{D}, \mathcal{F}, \mathcal{W}, \kappa) \to w_\star$ *in probability as* $n \to \infty$.

Finally, we present our OPI method in Algorithm A.3, called the *minimax optimal interval estimation (MOI)*, which is straightforwardly derived from Algorithm A.2 and the equation (13). We can also guarantee the validity and the approximate consistency of MOI in the asymptotic sense. The proof is straightforward and thus omitted.

Theorem 7.7. *Let* `MOI`$(\mathcal{D}, \mathcal{F}, \mathcal{W}, \kappa)$ *be the output of Algorithm A.3. Then,* `MOI`$(\mathcal{D}, \mathcal{F}, \mathcal{W}, \kappa)$ *is asymptotically valid and* $2\frac{1+\gamma}{1-\gamma}\epsilon_{\mathcal{W}}$*-approximately optimal in probability.*

8 Conclusion

In this paper, we have studied OPI without the sufficient exploration and the realizability conditions. In particular, we have pointed out the existence of the irreducible bias in such a general setting, and correspondingly, introduced a novel formulation of the interval-based OPE. We have then revealed the connection between the conventional MIS estimator and the irreducible bias, which is eventually utilized to construct the proposed method, the minimax optimal interval estimator (MOI), and to prove its optimality.

One of the major limitations of the proposed method is its model agnosticity, lying at the opposite end to the model-based approach, e.g., Yu et al. (2020), that depends on the full correctness of the model. It is left for future work to extend and combine these methods to be applicable to partially correct models.

References

Feng, Y., Ren, T., Tang, Z., and Liu, Q. (2020). Accountable off-policy evaluation with kernel bellman statistics. In *International Conference on Machine Learning*, pages 3102–3111. PMLR.

Feng, Y., Tang, Z., Zhang, N., and Liu, Q. (2021). Non-asymptotic confidence intervals of off-policy evaluation: Primal and dual bounds. *arXiv preprint arXiv:2103.05741*.

Hao, B., Ji, X., Duan, Y., Lu, H., Szepesvari, C., and Wang, M. (2021). Bootstrapping fitted q-evaluation for off-policy inference. In *International Conference on Machine Learning*, pages 4074–4084. PMLR.

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366.

Jiang, N. and Huang, J. (2020). Minimax value interval for off-policy evaluation and policy optimization. *Advances in Neural Information Processing Systems*, 33:2747–2758.

Kallus, N. and Uehara, M. (2020). Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *Journal of Machine Learning Research*, 21(167).

Liu, Q., Li, L., Tang, Z., and Zhou, D. (2018). Breaking the curse of horizon: Infinite-horizon off-policy estimation. *Advances in Neural Information Processing Systems*, 31.

Shi, C., Wan, R., Chernozhukov, V., and Song, R. (2021). Deeply-debiased off-policy interval estimation. In *International Conference on Machine Learning*, pages 9580–9591. PMLR.

Sriperumbudur, B., Fukumizu, K., and Lanckriet, G. (2010). On the relation between universality, characteristic kernels and rkhs embedding of measures. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 773–780. JMLR Workshop and Conference Proceedings.

Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Schölkopf, B., and Lanckriet, G. R. (2009). On integral probability metrics,\phi-divergences and binary classification. *arXiv preprint arXiv:0901.2698*.

Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Schölkopf, B., and Lanckriet, G. R. (2012). On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599.

Vapnik, V. (1999). *The nature of statistical learning theory*. Springer science & business media.

Wang, R., Foster, D. P., and Kakade, S. M. (2020). What are the statistical limits of offline rl with linear function approximation? *arXiv preprint arXiv:2010.11895*.

Xie, T., Ma, Y., and Wang, Y.-X. (2019). Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. *Advances in Neural Information Processing Systems*, 32.

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

Zanette, A. (2021). Exponential lower bounds for batch reinforcement learning: Batch rl can be exponentially harder than online rl. In *International Conference on Machine Learning*, pages 12287–12297. PMLR.

Algorithm A.1: Evaluation of DBR

Input: Dataset $\mathcal{D}$, universal function approximator $\mathcal{F}$, density-ratio estimate w;
Output: Estimate of $\|\mathcal{R}_\pi(w)\|_{\mathrm{TV}}$;

Solve the regression problem $\hat{L} \approx \sup_{f \in \mathcal{F}} \langle [\![f]\!],\ \mathcal{R}_\pi(w) \rangle$; `// See Lemma D.1`

return $\hat{L}$;

Algorithm A.2: Optimization of DBR

Input: Dataset $\mathcal{D}$, universal approximator $\mathcal{F}$, hypothesis class $\mathcal{W}$, kernel κ;
Output: Approximate minimizer of $\|\mathcal{R}_\pi(w)\|_{\mathrm{TV}}$;

for $u \in \mathcal{U}$ **do**

  $\hat{w}_u \coloneqq \mathrm{argmin}_{w \in \mathcal{W}} \|\hat{\mathcal{R}}_\pi(w)\|_u$;

  $\hat{\epsilon}_u \coloneqq \mathtt{EvaluateDBR}(\mathcal{D}, \mathcal{F}, \hat{w}_u)$; `// Algorithm A.1`

$\hat{u} \coloneqq \mathrm{argmin}_{u \in \mathcal{U}} \hat{\epsilon}_u$;

return $\hat{w}_{\hat{u}}$;

A Algorithms

B Proof of Theorem 4.1

We first show the first two inequality.

Lemma B.1. *For all $w \in \mathscr{B}(\mathcal{X})$,*

$$\epsilon_\star(\pi) \le \epsilon\left[\hat{J}(w)\right] \le \|\mathcal{R}_\pi(w)\|_{\mathrm{TV}} \tag{22}$$

Proof. The first inequality is trivial. To show the second one, observe

$$\begin{aligned}\left|J_{\mathcal{M}}(\pi) - \mathbb{E}\hat{J}(w)\right| &= |\langle \rho,\ \Gamma_\pi \iota_\pi - w \odot \beta \rangle| \\ &= |\langle q_\pi,\ \iota_\pi - \Delta_\pi(w \odot \beta) \rangle| \\ &= |\langle q_\pi,\ \mathcal{R}_\pi(w) \rangle| \\ &\le \|\mathcal{R}_\pi(w)\|_{\mathrm{TV}},\end{aligned}$$

where the last inequality is owing to $\|q_\pi\|_\infty \le 1$. Since the RHS is independent of $\mathcal{M}$ given $G_{\mathcal{M},\beta}$, we thus have

$$\epsilon\left[\hat{J}(w)\right] = \sup_{\mathcal{M}' : G_{\mathcal{M}',\beta} = G_{\mathcal{M},\beta}} \left|J_{\mathcal{M}'}(\pi) - \mathbb{E}\hat{J}(w)\right| \le \|\mathcal{R}_\pi(w)\|_{\mathrm{TV}}. \tag{23}$$

□

Now, to prove the last inequality, we prepare two extreme, yet indistinguishable environments $\mathcal{M}_\pm \coloneqq (\iota, \tilde{T}, R_\pm)$.

Let $\tilde{T}$ be an arbitrary state-transition operator indistinguishable from T, which will be determined later. Also let $\tilde{T}_\pi$ be the state-action transition operator associated with $\tilde{T}$ and π such that $\mathrm{d}\tilde{T}_\pi(s,a|x) = \mathrm{d}\tilde{T}(s|x)\mathrm{d}\pi(a|s)$ for $x, (s,a) \in \mathcal{X}$. Let $\tilde{\mu}_\pi \coloneqq (1-\gamma)(I - \gamma\tilde{T}_\pi)^{-1}\iota_\pi$ be the common occupancy measure of $\mathcal{M}_\pm$ induced with $\tilde{T}$, and $\tilde{\mu}_\pi|_{\not\ll\beta}$ be the singular component of $\tilde{\mu}_\pi$ with respect to β. Let $\mathcal{X}_0$ be a set separating $\tilde{\mu}_\pi|_{\not\ll\beta}$ from β and $\mathcal{X}_\beta \coloneqq \mathcal{X} \setminus \mathcal{X}_0$ be its complement.[3] For convenience, let $\Pi_\beta, \Pi_0 : \mathscr{M}(\mathcal{X}) \to \mathscr{M}(\mathcal{X})$ denote the projections of measure onto $\mathcal{X}_\beta$ and $\mathcal{X}_0$, respectively, given by $\Pi_\beta \coloneqq \chi_\beta \odot P$, and $\Pi_0 P \coloneqq (1-\chi_\beta) \odot P$ where χ_β being the indicator

[3]That is, $\{\mathcal{X}_0, \mathcal{X}_\beta\}$ is a partition of $\mathcal{X}$ such that $\tilde{\mu}_\pi|_{\not\ll\beta}(E) = 0$ for all measurable $E \subset \mathcal{X}_\beta$ and $\beta(E) = 0$ for all measurable $E \subset \mathcal{X}_0$.

Algorithm A.3: Minimax Optimal Interval Estimation (MOI)

Input: Dataset $\mathcal{D}$, universal approximator $\mathcal{F}$, hypothesis class $\mathcal{W}$, kernel κ;
Output: Estimate of the optimal interval $\mathcal{I}_\star(\pi)$;

```
ŵ ≔ OptimizeDBR(D, F, W, κ) ;                    // Algorithm A.2
ε̂ ≔ EvaluateDBR(D, F, ŵ) ;                       // Algorithm A.1
return [Ĵ(ŵ) − ε̂, J(ŵ) + ε̂];
```

function of $\mathcal{X}_\beta$ such that $\chi_\beta(x) = 1$ if $x \in \mathcal{X}_\beta$ and $\chi_0(x) = 0$ otherwise. Note that $\tilde{\mu}_\pi|_{\not\ll\beta} = \Pi_0\tilde{\mu}_\pi$ by construction.

Finally, put $R_\pm(x) = \delta_{\pm 1}$ for $x \in \mathcal{X}_0$ and $R_\pm(x) = R(x)$ otherwise, which is necessary for the indistinguishability of $R_\pm$, and denote the associated expected reward by $\rho_\pm(x) \coloneqq \int r dR_\pm(r|x)$, $x \in \mathcal{X}$. Then, we have

$$J_+(\pi) - J_-(\pi) = \langle \rho_+ - \rho_-, \tilde{\mu}_\pi \rangle = 2\|\tilde{\mu}_\pi|_{\not\ll\beta}\|_{\mathrm{TV}}, \tag{24}$$

where $J_\pm(\pi)$ are the policy values with respect to $\mathcal{M}_\pm$.

Now, the following lemma connects the RHS of (24) with DBR.

Lemma B.2. *There exist $\tilde{T}$ indistinguishable from T such that*

$$\|\tilde{\mu}_\pi|_{\not\ll\beta}\|_{\mathrm{TV}} = \|\mathcal{R}_\pi(w_\pi^\sharp)\|_{\mathrm{TV}}. \tag{25}$$

Proof. The proof is constructive. Consider an expanded state space $\mathcal{S} \leftarrow \mathcal{S} \cup \{\perp\}$, where $\perp$ denotes an absorbing state of $\tilde{T}$. Accordingly, let $\mathcal{X}_0 \leftarrow \mathcal{X}_0 \cup (\{\perp\} \times \mathcal{A})$. Now, put $\tilde{T} \coloneqq \tilde{T}|_{\ll\beta} + \tilde{T}|_{\not\ll\beta}$ such that $\tilde{T}|_{\ll\beta}$ is the restriction of T onto $\mathcal{X}_\beta$ and $\tilde{T}|_{\not\ll\beta}$ is the absorbing transition, respectively given by

$$\tilde{T}|_{\ll\beta} = T\Pi_\beta, \qquad \tilde{T}|_{\not\ll\beta}P = P(\mathcal{X}_0)\,\delta_\perp, \qquad P \in \mathscr{M}(\mathcal{X}).$$

Here, $\tilde{T}|_{\ll\beta}$ is corresponding to the known component of $\tilde{T}$, which necessarily coincides with that of T by definition, and $\tilde{T}|_{\not\ll\beta}$ is corresponding to the unknown component of $\tilde{T}$. Note that $\tilde{T}$ is a proper transition operator indistinguishable from T.

Let $\tilde{T}_\pi|_{\ll\beta}$ and $\tilde{T}_\pi|_{\not\ll\beta}$ be the state-action transition operator associated with T_β and T_0, respectively, such that $\mathrm{d}\tilde{T}_\pi|_{\ll\beta}(s,a|x) \coloneqq \mathrm{d}\tilde{T}|_{\ll\beta}(s|x)\mathrm{d}\pi(a|s)$ and $\mathrm{d}\tilde{T}_\pi|_{\not\ll\beta}(s,a|x) \coloneqq \mathrm{d}\tilde{T}|_{\not\ll\beta}(s|x)\mathrm{d}\pi(a|s)$ for $x, (s,a) \in \mathcal{X}$. Then, we have

$$\begin{aligned}
(1-\gamma)\iota_\pi &= (I - \gamma\tilde{T}_\pi)\tilde{\mu}_\pi \\
&= (I - \gamma\tilde{T}_\pi|_{\ll\beta} - \gamma\tilde{T}_\pi|_{\not\ll\beta})\tilde{\mu}_\pi \\
&= (I - \gamma\tilde{T}_\pi|_{\ll\beta})\tilde{\mu}_\pi - \gamma\tilde{\mu}_\pi(\mathcal{X}_0)\delta_{\perp,\pi} \\
&= (I - \gamma\tilde{T}_\pi|_{\ll\beta})\tilde{\mu}_\pi - \gamma\|\tilde{\mu}_\pi|_{\not\ll\beta}\|_{\mathrm{TV}}\delta_{\perp,\pi},
\end{aligned}$$

which implies, with $\tilde{P} \coloneqq (I - \gamma\tilde{T}_\pi|_{\ll\beta})^{-1}\iota_\pi$,

$$\begin{aligned}
\tilde{\mu}_\pi &= (1-\gamma)\tilde{P} + \gamma\|\tilde{\mu}_\pi|_{\not\ll\beta}\|_{\mathrm{TV}}(I - \gamma\tilde{T}_\pi|_{\ll\beta})^{-1}\delta_{\perp,\pi} \\
&= (1-\gamma)\tilde{P} + \gamma\|\tilde{\mu}_\pi|_{\not\ll\beta}\|_{\mathrm{TV}}\delta_{\perp,\pi}. \qquad (\because \tilde{T}_\pi|_{\ll\beta}\delta_{\perp,\pi} = 0)
\end{aligned}$$

Measuring the volumes on $\mathcal{X}_0$, we further get

$$\|\tilde{\mu}_\pi|_{\not\ll\beta}\|_{\mathrm{TV}} = (1-\gamma)\tilde{P}(\mathcal{X}_0) + \gamma\|\tilde{\mu}_\pi|_{\not\ll\beta}\|_{\mathrm{TV}},$$

which yields

$$\tilde{P}(\mathcal{X}_0) = \|\tilde{\mu}_\pi|_{\not\ll\beta}\|_{\mathrm{TV}}. \tag{26}$$

On the other hand, since $w_\pi^\sharp = (1-\gamma)\mathrm{d}(\Pi_\beta \tilde{P})/\mathrm{d}\beta$, we have

$$\begin{aligned}
\|\mathcal{R}_\pi(w_\pi^\sharp)\|_{\mathrm{TV}} &= \left\|\iota_\pi - (1-\gamma)\Delta_\pi \Pi_\beta \tilde{P}\right\|_{\mathrm{TV}} \\
&= \left\|\iota_\pi - (\Pi_\beta - \gamma \tilde{T}_\pi|_{\ll\beta})\tilde{P}\right\|_{\mathrm{TV}} && \left(\because T_\pi \Pi_\beta = \tilde{T}_\pi|_{\ll\beta}\right) \\
&= \left\|\iota_\pi - \Pi_\beta \iota_\pi + \gamma \Pi_0 T_\pi|_{\ll\beta}\tilde{P}\right\|_{\mathrm{TV}} && \left(\because \tilde{T}_\pi|_{\ll\beta} = \Pi_\beta \tilde{T}_\pi|_{\ll\beta} + \Pi_0 \tilde{T}_\pi|_{\ll\beta}\right) \\
&= \left\|\Pi_0\left(\iota_\pi + \gamma T_\pi|_{\ll\beta}\tilde{P}\right)\right\|_{\mathrm{TV}} \\
&= \left\|\Pi_0 \tilde{P}\right\|_{\mathrm{TV}} && \left(\because (I - \gamma \tilde{T}_\pi|_{\ll\beta})\tilde{P} = \iota_\pi\right) \\
&= \tilde{P}(\mathcal{X}_0).
\end{aligned}$$

Combining it with (26), we get the desired result. □

Plugging (25) into (24), we have

$$\|\mathcal{R}_\pi(w_\pi^\sharp)\|_{\mathrm{TV}} = \frac{1}{2}\left\{J_+(\pi) - J_-(\pi)\right\} \tag{27}$$

with a specific configuration of $\tilde{T}$. Since $\mathcal{M}_\pm$ are indistinguishable from one another, any estimators must incur the bias of at least a half of the difference $J_+(\pi) - J_-(\pi)$ in the worst case, i.e., $\mathcal{R}_\pi(w_\pi^\sharp) \le \epsilon_\star(\pi)$. This proves the last inequality of (8) via the triangle inequality,

$$\begin{aligned}
\|\mathcal{R}_\pi(w)\|_{\mathrm{TV}} &\le \|\mathcal{R}_\pi(w_\pi^\sharp)\|_{\mathrm{TV}} + \|\Delta_\pi(w - w_\pi^\sharp)\odot\beta\|_{\mathrm{TV}} \\
&\le \epsilon_\star(\pi) + \frac{1+\gamma}{1-\gamma}\|w - w_\pi^\sharp\|_{L^1(\beta)}.
\end{aligned}$$

and thus concludes the proof of Proposition 4.1.

C Proof of Proposition 7.1

First, we introduce a saddle-point formulation of the TV norm. Let $\mathcal{F}_1 \coloneqq \{f \in \mathcal{F} : \|f\|_\infty \le 1\}$ be the intersection of $\mathcal{F}$ with the unit ball of $\mathscr{B}(\mathcal{X})$. Lemma C.1 gives a general variational formula of the TV norm.

Lemma C.1. *For all $P \in \mathscr{M}(\mathcal{X})$, we have*

$$\|P\|_{\mathrm{TV}} = \sup_{f\in\mathcal{F}_1}\langle f,\ P\rangle. \tag{28}$$

Proof. Let us denote the unit ball of $\mathscr{B}(\mathcal{X})$ with $\mathcal{U}_1 \coloneqq \{f \in \mathscr{B}(\mathcal{X}) : \|f\|_\infty \le 1\}$. As an instance of the integral probability metrics (IPM), the TV norm is known to be written as

$$\|P\|_{\mathrm{TV}} = \sup_{g\in\mathcal{U}_1}\langle g,\ P\rangle. \tag{29}$$

Now fix $g \in \mathcal{U}_1$ such that $|\|P\|_{\mathrm{TV}} - \langle g,\ P\rangle| \le c$ for a positive constant $c > 0$. Since $\mathcal{F}_1$ is dense in $\mathscr{C}(\mathcal{X}) \cap \mathcal{U}_1$, it is also dense in $L^1(P) \cap \mathcal{U}_1$ and therefore there exists $f \in \mathcal{F}_1$ such that $\|f - g\|_{L^1(P)} \le c$. Then it follows that

$$|\|P\|_{\mathrm{TV}} - \langle f,\ P\rangle| \le |\|P\|_{\mathrm{TV}} - \langle g,\ P\rangle| + \|f - g\|_{L^1(P)} \le 2c. \tag{30}$$

Since $c > 0$ can be arbitrarily small, we finally have

$$\|P\|_{\mathrm{TV}} \le \sup_{f\in\mathcal{F}_1}\langle f,\ P\rangle \le \sup_{g\in\mathcal{U}_1}\langle g,\ P\rangle \le \|P\|_{\mathrm{TV}}, \tag{31}$$

which proves the desired result. □

Since the supremum in (28) is taken over a constrained domain $\mathcal{F}_1$, its computation is not necessarily tractable in general. The following lemma is useful to make the domain unconstrained.

Lemma C.2. *Let $\Psi : \mathscr{B}(\mathcal{X}) \to \mathbb{R}_{\geq 0}$ be a penalty function such that $\Psi(f) = 0$ if $f \in \mathcal{F}_1$. Then, for all $P \in \mathscr{M}(\mathcal{X})$, we have*

$$\sup_{f \in \mathcal{F}_1} \langle f,\ P \rangle = \sup_{f \in \mathcal{F}} \left\{ \langle [\![f]\!],\ P \rangle - \Psi(f) \right\}, \tag{32}$$

where $[\![f]\!](x) \coloneqq \max\{-1, \min\{1, f(x)\}\}$ denotes the clipping of $f(x)$ to $[-1, 1]$.

Proof. Slightly extending (31), we have

$$\begin{aligned}
\|P\|_{\mathrm{TV}} &\leq \sup_{f \in \mathcal{F}_1} \langle f,\ P \rangle \\
&= \sup_{f \in \mathcal{F}_1} \left\{ \langle f,\ P \rangle - \Psi(f) \right\} \\
&\leq \sup_{f \in \mathcal{F}} \left\{ \langle [\![f]\!],\ P \rangle - \Psi(f) \right\} && (\because \text{domain expansion}) \\
&\leq \sup_{g \in \mathcal{U}_1} \langle g,\ P \rangle && (\because [\![f]\!] \in \mathcal{U}_1,\ \Psi(f) \geq 0) \\
&\leq \|P\|_{\mathrm{TV}}.
\end{aligned}$$

This yields the desired claim. □

Finally, we obtain the desired result (Proposition 7.1) by taking the penalty function as the trivial one, $\Psi(f) = 0$ for all $f \in \mathcal{F}$. Note, however, the penalty function is useful for the optimization viewpoint. In particular, if we take

$$\Psi(f) = \langle (|f| - 1)_+,\ P \rangle,$$

where $(g)_+(x) \coloneqq \max\{0, g(x)\}$ denotes the positive part of a function $g \in \mathscr{B}(\mathcal{X})$, then the resultant objective of the supremum in (32) becomes concave without altering the value of the supremum.

D Structural Risk Minimization for Regression Problem (15)

The following theorem shows that the SRM principle is effective to solve (15).

Lemma D.1. *Suppose $\mathcal{F}$ admits the following decomposition*

$$\mathcal{F} = \bigcup_{\ell=1}^{\infty} \mathcal{F}_\ell \tag{33}$$

such that $\mathcal{F}_\ell$ is increasing as $\ell \to \infty$ and $\ln \mathcal{N}(\mathcal{F}_\ell, \epsilon) \leq \ell \ln(1/\epsilon)$ for all $\ell \geq 1$, where $\mathcal{N}(\mathcal{F}_\ell, \epsilon)$ is the metric entropy of $\mathcal{F}_\ell$ with respect to $\| \cdot \|_\infty$.

Now define

$$L \coloneqq \sup_{f \in \mathcal{F}} \langle [\![f]\!],\ \mathcal{R}_\pi(w) \rangle, \tag{34}$$

$$\hat{L}_k \coloneqq \sup_{f \in \mathcal{F}_k} \langle [\![f]\!],\ \hat{\mathcal{R}}_\pi(w) \rangle, \qquad k \geq 1. \tag{35}$$

Then, $\hat{L}_k$ converges in probability to L as $k \to \infty$ and $n/(k \ln n) \to \infty$.

Proof. Define

$$L_k \coloneqq \sup f \in \mathcal{F}_k \langle [\![f]\!],\ \mathcal{R}_\pi(w) \rangle$$

for all $k \geq 1$. Define $\beta'_\pi, \hat{\beta}'_\pi \in \mathscr{M}(\mathcal{X}^2)$ by

$$\mathrm{d}\beta'_\pi(x, x') \coloneqq \mathrm{d}\beta(x) \mathrm{d}T_\pi(x'|x), \qquad \mathrm{d}\hat{\beta}'_\pi(x, x') \coloneqq \mathrm{d}\hat{\beta}(x) \mathrm{d}\hat{T}_\pi(x'|x).$$

First, we follow the standard argument of the generalization error analysis to show $\hat{L}_k \to L_k$. Observe that, for all $k \geq 1$ and $\epsilon > 0$, with probability $\geq 1 - \delta$,

$$
\begin{aligned}
&|L_k - \hat{L}_k| \\
&\leq \sup_{f \in \mathcal{F}_k} |\langle [\![f]\!],\ \hat{\mathcal{R}}_\pi(w) - \mathcal{R}_\pi(w)\rangle| \\
&\leq (1-\gamma) \sup_{f \in \mathcal{F}_k} |\langle [\![f]\!],\ \hat{\iota}_\pi - \iota_\pi\rangle| + \gamma \sup_{f \in \mathcal{F}_k} |\langle w \otimes [\![f]\!],\ \hat{\beta}'_\pi - \beta'_\pi\rangle| + \sup_{f \in \mathcal{F}_k} |\langle w[\![f]\!],\ \hat{\beta} - \beta\rangle|, \\
&= (1-\gamma) \sup_{f \in \mathcal{F}_{k,1}} |\langle f,\ \hat{\iota}_\pi - \iota_\pi\rangle| + \gamma W \sup_{f \in \mathcal{F}_{k,2}} |\langle f,\ \hat{\beta}'_\pi - \beta'_\pi\rangle| + W \sup_{f \in \mathcal{F}_{k,3}} |\langle f,\ \hat{\beta} - \beta\rangle|, \\
&\leq C\epsilon + (1-\gamma)\sqrt{\frac{2\ln \mathcal{N}(\mathcal{F}_{k,1}, \epsilon)/\delta}{n}} + \gamma W \sqrt{\frac{2\ln \mathcal{N}(\mathcal{F}_{k,2}, \epsilon)/\delta}{n}} + W\sqrt{\frac{2\ln \mathcal{N}(\mathcal{F}_{k,3}, \epsilon)/\delta}{n}}
\end{aligned}
$$

where $W \coloneqq \|w\|_\infty$, $C \coloneqq 1 - \gamma + (1+\gamma)W$, $\mathcal{F}_{k,1} \coloneqq \{[\![f]\!] : f \in \mathcal{F}_k\}$, $\mathcal{F}_{k,2} \coloneqq \{w \otimes [\![f]\!]/W : f \in \mathcal{F}_k\}$ and $\mathcal{F}_{k,3} \coloneqq \{w[\![f]\!]/W : f \in \mathcal{F}_k\}$. Since $\mathcal{N}(\mathcal{F}_{k,m}, \epsilon) \leq k\ln(1/\epsilon)$ for $m = 1, 2, 3$ due to the Lipschitz continuity of $f \mapsto [\![f]\!]$, we have

$$
|L_k - \hat{L}_k| = O\left(C\sqrt{\frac{k\ln(n/\delta)}{n}}\right)
$$

taking $\epsilon = 1/\sqrt{n}$. This proves $\hat{L}_k \to L_k$ in probability. Since $L_k \to L$ by definition, we also get $\hat{L}_k \to L$. This completes the proof. □

E Dual Representation of Convolution Norm

The following lemma is used to prove the dual representation of the u-convolution norm in Proposition 7.4.

Lemma E.1. *Let $\|f\|_A$ and $\|f\|_B$ be arbitrary norms of $f \in \mathscr{B}(\mathcal{X})$. Also let $\|f\|_{A\vee B} \coloneqq \|f\|_A \vee \|f\|_B$ be the norm defined by the maximum of these. Then, for all $P \in \mathscr{M}(\mathcal{X})$, we have*

$$
\|P\|_{(A\vee B)^*} = \inf_{Q \ll P} \left\{\|P - Q\|_{A^*} + \|Q\|_{B^*}\right\},
$$

where we denote the dual norm of $\|\cdot\|_X$ on $\mathscr{B}(\mathcal{X})$ by $\|P\|_{X^*} \coloneqq \sup_{\|f\|_X \leq 1}\langle f,\ P\rangle$, $P \in \mathscr{M}(\mathcal{X})$.

Proof. Observe

$$
\begin{aligned}
\|P\|_{(A\vee B)^*} &= \sup_{\|f\|_A \leq 1, \|f\|_B \leq 1} \langle f,\ P\rangle \\
&= \sup_{\|f\|_A \leq 1, \|g\|_B \leq 1} \inf_{Q \ll P} \{\langle f,\ P\rangle + \langle g - f,\ Q\rangle\} \qquad (\because Q \text{ as a Lagrange multiplier}) \\
&= \sup_{\|f\|_A \leq 1, \|g\|_B \leq 1} \inf_{Q \ll P} \{\langle f,\ P - Q\rangle + \langle g,\ Q\rangle\} \\
&= \inf_{Q \ll P} \left\{\sup_{\|f\|_A \leq 1} \langle f,\ P - Q\rangle + \sup_{\|g\|_B \leq 1} \langle g,\ Q\rangle\right\} \qquad (\because \text{Sion's minimax theorem}) \\
&= \inf_{Q \ll P} \left\{\|P - Q\|_{A^*} + \|Q\|_{B^*}\right\}.
\end{aligned}
$$

Here, to invoke Sion's minimax theorem, we have exploited the fact that $\mathcal{X}$ is finite set and thus f and g take values in a compact space. □

F Proof of Theorem 7.6

Note that Theorem 7.3 combined with the compactness of $\mathcal{W}$ and the continuity of $w \mapsto \|\mathcal{R}_\pi(w)\|_{\mathrm{TV}}$ implies that $\texttt{EvaluateDBR}(\mathcal{D}, \mathcal{F}, w)$ converges to $\|\mathcal{R}_\pi(w)\|_{\mathrm{TV}}$ uniformly on $w \in \mathcal{W}$. Thus, it suffice to show the following lemma.

Lemma F.1. *We have*

$$\min_{u \in \mathcal{U}} \|\mathcal{R}_\pi(\hat{w}_u)\|_{\mathrm{TV}} \to \min_{w \in \mathcal{W}} \|\mathcal{R}_\pi(w)\|_{\mathrm{TV}}.$$

Proof. Note also that Corollary 7.5 combined with the compactness of $\mathcal{W}$ and the continuity of $w \mapsto \|\mathcal{R}_\pi(w)\|_u$ implies

$$\|\hat{\mathcal{R}}_\pi(w)\|_u \to \|\mathcal{R}_\pi(w)\|_{\mathrm{TV}}$$

uniformly for all $w \in \mathcal{W}$, under suitable asymptotics of u and n as in Corollary 7.5. In other words, for all $c > 0$ and $\delta \in (0, 1)$, there exists $u_0 \geq 1$ and $p_0 > 0$ such that, for all $u \geq u_0$ and $n \geq p_0 u^2$ such that $\sup_{w \in \mathcal{W}} |\|\hat{\mathcal{R}}_\pi(w)\|_u - \|\mathcal{R}_\pi(w)\|_{\mathrm{TV}}| \leq c$ with probability $\geq 1 - \delta$. Therefore, taking such a pair (u, n) satisfying $u \in \mathcal{U}$ (which exists by the definition of $\mathcal{U}$), we have

$$\begin{aligned}
\min_{w \in \mathcal{W}} \|\mathcal{R}_\pi(w)\|_{\mathrm{TV}} &\leq \min_{u' \in \mathcal{U}} \|\mathcal{R}_\pi(\hat{w}_{u'})\|_{\mathrm{TV}} && (\because \text{restriction of domain}) \\
&\leq c + \|\hat{\mathcal{R}}_\pi(\hat{w}_u)\|_u \\
&\leq c + \|\hat{\mathcal{R}}_\pi(w_\star)\|_u && (\because \text{definition of } \hat{w}_u) \\
&\leq 2c + \|\hat{\mathcal{R}}_\pi(w_\star)\|_{\mathrm{TV}} \\
&= 2c + \min_{w \in \mathcal{W}} \|\mathcal{R}_\pi(w)\|_{\mathrm{TV}}
\end{aligned}$$

with probability $\geq 1 - \delta$. Since $c > 0$ can be arbitrary small, we have just proved the desired result. □