DISTRIBUTIONAL OFF-POLICY EVALUATION WITH BELLMAN RESIDUAL MINIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We consider the problem of distributional off-policy evaluation which serves as the foundation of many distributional reinforcement learning (DRL) algorithms. In contrast to most existing works (that rely on supremum-extended statistical distances), we study the expectation-extended statistical distance for quantifying the distributional Bellman residuals and provide the corresponding theoretical supports. Extending the framework of Bellman residual minimization to DRL, we propose a method called Energy Bellman Residual Minimizer (EBRM) to estimate the return distribution. We establish a finite-sample error bound for the EBRM estimator under the realizability assumption. Additionally, we introduce a variant of our method based on a multi-step bootstrapping procedure to enable multi-step extension. By selecting an appropriate step level, we obtain a better error bound for this variant of EBRM compared to a single-step EBRM, under non-realizability settings. Finally, we demonstrate the superior performance of our method through simulation studies, comparing with other existing methods.

1 INTRODUCTION

In reinforcement learning (RL), the cumulative (discounted) reward, also known as the return, is a crucial quantity for evaluating the performance of a policy. Most existing RL methods focus on only the expectation of the return distribution. In Bellemare et al. (2017a), the focus has been extended to the whole return distribution, and they introduce a distributional RL (DRL) algorithm (hereafter called Categorical algorithm) that achieves a considerably better performance in Atari games than expectation-oriented Deep-Q Networks (Mnih et al., 2015). This has sparked significant interests among the RL community, and was later followed by a series of quantile-based methods including QRDQN, QRTD (Dabney et al., 2018b), IQN (Dabney et al., 2018a), FQF (Yang et al., 2019), EDRL (Rowland et al., 2019) and particle-based methods including MMDRL (Nguyen-Tang et al., 2021), SinkhornDRL (Sun et al., 2022), MD3QN (Zhang et al., 2021). In this paper, we consider the problem of off-policy evaluation in DRL, i.e., estimating the (conditional) return distribution of a target policy based on offline data.

Despite their competitive performances, distributional RL methods are significantly underdeveloped compared with the traditional expectation-based RL, especially in the theoretical development under the offline setting. All aforementioned methods are motivated by supremum-extended distances due to the contraction property (see (4) below), but their algorithms essentially minimize an expectation-extended distance (see (6)), as summarized in the column “Distance Mismatch” of Table 1. This leads to a theory-practice gap. Also, most of these work does not provide any statistical guarantee such as the convergence rate. We note that Rowland et al. (2018) establishes the consistency of their estimator, but no error bound analysis (and convergence rate) is provided. In terms of statistical analysis, a very recent work FLE (Wu et al., 2023) only offers error bound analysis of their estimator for the marginal distribution of return, which is hard to use for policy learning. In addition, their analysis is based on a strong condition called completeness, which in general significantly restricts model choices of return distributions and excludes the non-realizable scenario.

This paper proposes novel estimators, which we call Energy Bellman Residual Minimizer (EBRM), based on the idea of Bellman residual minimization for the conditional distribution of the return. In contrast to existing work, we provide solid theoretical ground for the application of expectation-extended distance in measuring (distributional) Bellman residual. A multi-step extension of our estimator is proposed for non-realizability settings. Our method comes with statistical error bound analyses in both realizable and non-realizable settings. Table 1 provides some key comparisons between our method and some existing works. More details is given in Table 3 in the Appendix D.1. Finally, we summarize our contributions as follows. (1) We provide theoretical foundation of the application of expectation-extended distance for Bellman residual minimization in DRL. See Section 2.3. (2) We develop a novel distributional off-policy evaluation method (EBRM), together with its finite-sample error bound. See Section 3. (3) We develop a multi-step extension of EBRM for non-realizabile settings in Section 4. We also provide corresponding finite-sample error bound under non-realizable settings. (4) Our numerical experiments in Section 5 demonstrate the strong performance of EBRM compared with some baseline methods.

Table 1: Comparison among DRL methods in off-policy evaluation.

Method	**Distance match**	**Statistical error bound**	**Non-realizable**	**Multi-dimension**
Categorical (Bellemare et al., 2017a)	✗	✗	NA	✓
QRTD (Dabney et al., 2018b)	✗	✗	NA	✗
IQN (Dabney et al., 2018a)	✗	✗	NA	✗
FQF (Yang et al., 2019)	✗	✗	NA	✗
EDRL (Rowland et al., 2019)	✗	✗	NA	✗
MMDRL (Nguyen-Tang et al., 2021)	✗	✗	NA	✓
SinkhornDRL (Sun et al., 2022)	✗	✗	NA	✓
MD3QN (Zhang et al., 2021)	✗	✗	NA	✓
FLE (Wu et al., 2023)	✓	✓	NA	✓
EBRM (our method)	✓	✓	✓	✓

2 Off-policy Evaluation Based on Bellman Equation

2.1 Background

We consider an off-policy evaluation (OPE) problem under the framework of infinite-horizon Markov Decision Process (MDP), which is characterized by a state space $\mathcal{S}$, a discrete action space $\mathcal{A}$, and a transition probability $p : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathbb{R}^d \times \mathcal{S})$ with $\mathcal{P}(\mathcal{X})$ denoting the class of probability measures over a generic space $\mathcal{X}$. In other words, p defines a joint distribution of a d-dimensional immediate reward and the next state conditioned on a state-action pair. At each time point, an action is chosen by the agent based on a current state according to a (stochastic) policy, a mapping from $\mathcal{S}$ to $\mathcal{P}(\mathcal{A})$. With the initial state-action pair $(S^{(0)}, A^{(0)})$, a trajectory generated by such an MDP can be written as $\{S^{(t)}, A^{(t)}, R^{(t+1)}\}_{t\geq 0}$. The return variable is defined as $Z := \sum_{t=1}^{\infty} \gamma^{t-1} R^{(t)}$ with $\gamma \in [0, 1)$ being a discount factor, based on which we can evaluate the performance of some target policy π.

Traditional OPE methods are mainly focused on estimating the expectation of return Z under the target policy π, whereas DRL aims to estimate the whole distribution of Z. Letting $\mathcal{L}(X)$ be the probability measure of some random variable (or vector) X, our target is to estimate the collection of return distributions conditioned on different initial state-action pairs $(S^{(0)}, A^{(0)}) = (s, a)$:

$$\Upsilon_\pi(s,a) := \mathcal{L}\bigg(\sum_{t=1}^{\infty} \gamma^{t-1} R^{(t)}\bigg),\ (R^{(t+1)}, S^{(t+1)}) \sim p(\cdot|S^{(t)}, A^{(t)}),\ A^{(t+1)} \sim \pi(\cdot|S^{(t+1)}), \quad (1)$$

collectively written as $\Upsilon_\pi \in \mathcal{P}(\mathbb{R}^d)^{\mathcal{S}\times\mathcal{A}}$. It is analogous to the Q-function in traditional RL, whose evaluation at a state-action pair (s, a) is the expectation of the distribution $\Upsilon_\pi(s, a)$. Our goal in this paper is to use the offline data generated by the behavior policy b to estimate Υ_π.

Similar to most existing DRL methods, our proposal is based on the distributional Bellman equation (Bellemare et al., 2017a). Define the distributional Bellman operator by $\mathcal{T}^\pi : \mathcal{P}(\mathbb{R}^d)^{\mathcal{S}\times\mathcal{A}} \to \mathcal{P}(\mathbb{R}^d)^{\mathcal{S}\times\mathcal{A}}$ such that, for any $\Upsilon \in \mathcal{P}(\mathbb{R}^d)^{\mathcal{S}\times\mathcal{A}}$,

$$(\mathcal{T}^\pi \Upsilon)(s,a) := \int_{\mathbb{R}^d\times\mathcal{S}\times\mathcal{A}} (g_{r,\gamma})_{\#} \Upsilon(s', a') \mathrm{d}\pi(a'|s') \mathrm{d}p(r, s'|s, a), \quad (s,a) \in \mathcal{S}\times\mathcal{A}, \quad (2)$$

where $(g_{r,\gamma})_\# : \mathcal{P}(\mathbb{R}^d) \to \mathcal{P}(\mathbb{R}^d)$ maps the distribution of any random vector X to the distribution of $r + \gamma X$. One can show that Υ_π is the unique solution to the distributional Bellman equation:

$$\mathcal{T}^\pi \Upsilon = \Upsilon. \tag{3}$$

Letting $Z_\pi(s,a)$ be the random vector that follows the distribution $\Upsilon_\pi(s,a)$, one can also express the distributional Bellman equation (3) in a more intuitive way: for all $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$Z_\pi(s,a) \overset{D}{=} R + \gamma Z_\pi(S', A') \quad \text{where} \quad (R, S') \sim p(\cdot|s,a), A' \sim \pi(\cdot|S'),$$

where $\overset{D}{=}$ refers to the equivalence in terms of the underlying distributions. Due to the distributional Bellman equation (3), a sensible approach to find Υ_π is based on minimizing the discrepancy between $\mathcal{T}^\pi\Upsilon$ and Υ with respect to $\Upsilon \in \mathcal{P}(\mathbb{R}^d)^{\mathcal{S}\times\mathcal{A}}$, which will be called *Bellman residual* hereafter. To proceed with this approach, two important issues need to be addressed. First, both $\mathcal{T}^\pi\Upsilon$ and Υ are collections of distributions over $\mathbb{R}^d$, based on which Bellman residual shall be quantified. Second, $\mathcal{T}^\pi$ may not be available and therefore needs to be estimated through data. We will focus on the quantification of Bellman residual first, and defer the proposed estimator of $\mathcal{T}^\pi$ and the formal description of our estimator for Υ_π to Section 3.

2.2 Existing Measures of Bellman Residuals

To quantify the discrepancy between the two sides of the distributional Bellman equation (3), one can use a distance over $\mathcal{P}(\mathbb{R}^d)^{\mathcal{S}\times\mathcal{A}}$. Fixing a state-action pair, one can solely compare two distributions from $\mathcal{P}(\mathbb{R}^d)$. Therefore, a common strategy is to start by selecting a statistical distance $\eta(\cdot,\cdot) : \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d) \to [0,\infty]$, and then define an extended-distance over $\mathcal{P}(\mathbb{R}^d)^{\mathcal{S}\times\mathcal{A}}$ through combining the statistical distances over different state-action pairs. As shown in Table 3 in Appendix D.1, most existing methods (e.g., Bellemare et al., 2017b;a; Nguyen et al., 2020) are based on some *supremum-extended* distance η_∞:

$$\eta_\infty(\Upsilon_1, \Upsilon_2) := \sup_{s,a} \eta\Big\{\Upsilon_1(s,a), \Upsilon_2(s,a)\Big\}. \tag{4}$$

Under various choices of η including Wasserstein-p metric with $1 \le p \le \infty$ (Bellemare et al., 2017a; Dabney et al., 2018b) and maximum mean discrepancy (Nguyen-Tang et al., 2021), it is shown that $\mathcal{T}^\pi$ is a contraction with respect to η_∞. More specifically, $\eta_\infty(\mathcal{T}^\pi\Upsilon_1, \mathcal{T}^\pi\Upsilon_2) \le \gamma^{\beta_0} \cdot \eta_\infty(\Upsilon_1, \Upsilon_2)$ holds for any $\Upsilon_1, \Upsilon_2 \in \mathcal{P}(\mathbb{R}^d)^{\mathcal{S}\times\mathcal{A}}$, where the value of $\beta_0 > 0$ depends on the choice of η. If η_∞ is a metric, then the contractive property implies, for any $\Upsilon \in \mathcal{P}(\mathbb{R}^d)^{\mathcal{S}\times\mathcal{A}}$,

$$\eta_\infty(\Upsilon, \Upsilon_\pi) \le \sum_{k=1}^{\infty} \eta_\infty\big\{(\mathcal{T}^\pi)^{k-1}\Upsilon, (\mathcal{T}^\pi)^k\Upsilon\big\} \le \frac{1}{1-\gamma^{\beta_0}} \cdot \eta_\infty(\Upsilon, \mathcal{T}^\pi\Upsilon). \tag{5}$$

As such, minimizing Bellman residual measured by η_∞ would be a sensible approach for finding Υ_π. However, as surveyed in Appendix D.1, most existing methods in practice essentially minimize an empirical (and approximated) version of the *expectation-extended* distance defined by

$$\bar{\eta}(\Upsilon_1, \Upsilon_2) := \mathbb{E}_{(S,A)\sim b_\mu} \eta\Big\{\Upsilon_1(S,A), \Upsilon_2(S,A)\Big\}, \tag{6}$$

with $(S,A) \sim b_\mu$. Here $b_\mu = \mu \times b$ refer to data distribution over $\mathcal{S} \times \mathcal{A}$ induced by the behavior policy b. With a slight abuse of notation, we will overload the notation b_μ with its density (with respect to some appropriate base measure of $\mathcal{S} \times \mathcal{A}$, e.g., counting measure and Lebesgue measure). We remark that (5) does not hold under $\bar{\eta}$ because η_∞ and $\bar{\eta}$ are not necessarily equivalent for the general state-action space, leading to a theory-practice gap in most methods (Column 1 of Table 1).

2.3 Expectation-extended Distance

Despite the implicit use of expectation-extended distances in some prior works, the corresponding theoretical foundations are not well established. Regarding Bellman residual minimization, a very natural and crucial question is:

> *In terms of an expectation-extended distance, does small Bellman residual of* Υ *lead to closeness between* Υ *and* Υ_π*?*

To proceed, we focus on settings such that the state-action pairs of interest can be well covered by b_μ, as formally stated in the following assumption. Let $q^\pi(s, a|\tilde{s}, \tilde{a})$ be the conditional probability density of the next state-action pair at (s, a) conditional on the current state-action pair at $(\tilde{s}, \tilde{a})$, defined by the transition probability p and the target policy π.

Assumption 1. *There exists $p_{\min} > 0$ and $p_{\max} < \infty$ such that $b_\mu(s, a) \geq p_{\min}$ for all $s, a \in \mathcal{S} \times \mathcal{A}$ and $q^\pi(s, a|\tilde{s}, \tilde{a}) \leq p_{\max}$ for all $(\tilde{s}, \tilde{a}), (\tilde{s}, \tilde{a}) \in \mathcal{S} \times \mathcal{A}$.*

Let $q^{\pi:t}_{b_\mu}(s, a)$ be the probability density (or mass) of $(S^{(t)}, A^{(t)})$ at (s, a), given $(S^{(0)}, A^{(0)}) \sim b_\mu$ and the target policy π. Assumption 1 implies uniformly bounded density ratio, that is $q^{\pi:t}_{b_\mu}(s, a)/b_\mu(s, a) \leq C_{\sup}(< \infty)$ for all $t \in \mathbb{N}$, as proved in Appendix A.1.

In the following Theorem 1 (proved in Appendix A.2), we provide a solid ground for Bellman residual minimization based on expectation-extended distances.

Theorem 1. *Under Assumption 1, if the statistical distance η satisfies translation-invariance, scale-sensitivity of order $\beta_0 > 0$, convexity, and relaxed triangular inequality defined in Appendix A.2.1, then we can bound the inaccuracy:*

$$\bar{\eta}(\Upsilon, \Upsilon_\pi) \leq 2C_{\sup} B_1(\gamma; \beta_0) \cdot \bar{\eta}(\Upsilon, \mathcal{T}^\pi \Upsilon), \tag{7}$$

where $B_1(\gamma; \beta_0) := \frac{1}{2(1-\gamma^{\beta_0})} \sum_{k=1}^{\infty} 4^k \gamma^{(2^{k-1}-1)\beta_0} < \infty$ is an increasing function of $\gamma \in (0, 1)$, and $C_{\sup}$ is defined in (25).

Inequality (7) provides an analogy to Bound (5) for expectation-based distances, answering our prior question positively for some expectation-extended distances. Note that Theorem 1 can be applied to the settings with general state-action space, including continuous one.

In order to take advantage of Theorem 1, we should select a statistical distance that satisfies all the properties stated in Theorem 1. One example is energy distance (Székely & Rizzo, 2013) as proved in Appendix A.3, which is in fact a squared maximum mean discrepancy (Gretton et al., 2012) with kernel $k(\mathbf{x}, \mathbf{y}) = \|\mathbf{x}\| + \|\mathbf{y}\| - \|\mathbf{x} - \mathbf{y}\|$. The energy distance is defined as

$$\mathcal{E}\{\mathcal{L}(\mathbf{X}), \mathcal{L}(\mathbf{Y})\} := 2\mathbb{E}\|\mathbf{X} - \mathbf{Y}\| - \mathbb{E}\|\mathbf{X} - \mathbf{X}'\| - \mathbb{E}\|\mathbf{Y} - \mathbf{Y}'\|, \tag{8}$$

where $\mathbf{X}'$ and $\mathbf{Y}'$ are independent copies of $\mathbf{X}$ and $\mathbf{Y}$ respectively, and $\mathbf{X}, \mathbf{X}', \mathbf{Y}, \mathbf{Y}'$ are independent. In below, we will use energy distance to construct our estimator.

3 Energy Bellman Residual Minimizer

3.1 Estimated Bellman Residual

Despite applicability of Theorem 1 to general state-action space, we will focus on tabular case with finite cardinality $|\mathcal{S} \times \mathcal{A}| < \infty$ for simpler construction of estimation, which enables an in-depth theoretical study under both realizable and non-realizable settings in Sections 3.2 and 4.3. But the reward can be continuous. Our target objective of Bellman residual minimization is

$$\bar{\mathcal{E}}(\Upsilon, \mathcal{T}^\pi \Upsilon) = \sum_{s,a} b_\mu(s, a) \cdot \mathcal{E}\big\{\Upsilon(s, a), \mathcal{T}^\pi \Upsilon(s, a)\big\}, \quad \text{where} \tag{9}$$

$$\begin{aligned}\mathcal{E}\big\{\Upsilon(s, a), \mathcal{T}^\pi \Upsilon(s, a)\big\} = 2\mathbb{E}\|Z_\alpha(s, a) - Z^{(1)}_\beta(s, a)\| - \mathbb{E}\|Z_\alpha(s, a) - Z_\beta(s, a)\| \\ - \mathbb{E}\|Z^{(1)}_\alpha(s, a) - Z^{(1)}_\beta(s, a)\|,\end{aligned}$$

where $Z_\alpha(s, a), Z_\beta(s, a) \sim \Upsilon(s, a)$ and $Z^{(1)}_\alpha(s, a), Z^{(1)}_\beta(s, a) \sim \mathcal{T}^\pi \Upsilon(s, a)$ are all independent. For the tabular case with offline data, we can estimate b_μ and the transition p simply by empirical distributions. That is, given observations $\mathcal{D} = \{(s_i, a_i, r_i, s'_i)\}_{i=1}^N$, we consider

$$\hat{b}_\mu(s, a) := \frac{N(s, a)}{N} \quad \text{where} \quad N(s, a) := \sum_{i=1}^{N} \mathbf{1}\big\{(s_i, a_i) = (s, a)\big\}, \quad \text{and} \tag{10}$$

$$\hat{p}(E|s, a) := \begin{cases} \frac{1}{N(s,a)} \sum_{i:(s_i,a_i)=(s,a)} \delta_{r_i, s'_i}(E) & \text{if } N(s, a) \geq 1, \\ \delta_{\mathbf{0}, s}(E) & \text{if } N(s, a) = 0 \end{cases} \quad \text{for any measurable set } E,$$

where $\delta_{r,s'}$ is the Dirac measure at (r, s'). Based on this, we can estimate $\mathcal{T}^\pi$ for any $\Upsilon \in \mathcal{P}(\mathbb{R}^d)^{\mathcal{S}\times\mathcal{A}}$ by the estimated transition $\hat{p}$ and the target policy π, by replacing p of (2) with $\hat{p}$.

Denoting the conditional expectation by $\tilde{\mathbb{E}}(\cdots) := \mathbb{E}(\cdots|\mathcal{D})$, we can compute

$$\begin{aligned}\mathcal{E}\big\{\Upsilon(s,a), \hat{\mathcal{T}}^\pi\Upsilon(s,a)\big\} = 2\tilde{\mathbb{E}}\|Z_\alpha(s,a) - \hat{Z}_\beta^{(1)}(s,a)\| - \tilde{\mathbb{E}}\|Z_\alpha(s,a) - Z_\beta(s,a)\| \\ - \tilde{\mathbb{E}}\|\hat{Z}_\alpha^{(1)}(s,a) - \hat{Z}_\beta^{(1)}(s,a)\|,\end{aligned} \tag{11}$$

where $Z_\alpha(s,a), Z_\beta(s,a) \sim \Upsilon_\theta(s,a)$ and $\hat{Z}_\alpha^{(1)}(s,a), \hat{Z}_\beta^{(1)}(s,a) \sim \hat{\mathcal{T}}^\pi\Upsilon(s,a)$ are all independent conditioned on the observed data $\mathcal{D}$ that determines $\hat{\mathcal{T}}^\pi$. With the above construction, we can estimate the objective function by

$$\hat{\hat{\mathcal{E}}}(\Upsilon, \hat{\mathcal{T}}^\pi\Upsilon) = \sum_{s,a} \hat{b}_\mu(s,a) \cdot \mathcal{E}\big\{\Upsilon(s,a), \hat{\mathcal{T}}^\pi\Upsilon(s,a)\big\}. \tag{12}$$

Now letting $\{\Upsilon_\theta : \theta \in \Theta\} \subseteq \mathcal{P}(\mathbb{R}^d)^{\mathcal{S}\times\mathcal{A}}$ be the hypothesis class of Υ_π, where each distribution Υ_θ is indexed by an element of candidate space Θ, a special case of which is the parametric case $\Theta \subseteq \mathbb{R}^p$. Then the proposed estimator of Υ_π is $\Upsilon_{\hat{\theta}}$ where

$$\hat{\theta} \in \arg\min_{\theta\in\Theta} \hat{\hat{\mathcal{E}}}(\Upsilon_\theta, \hat{\mathcal{T}}^\pi\Upsilon_\theta). \tag{13}$$

We call our method the *Energy Bellman Residual Minimizer* (EBRM) and summarize it in Algorithm 1. We will refer to the approach here as EBRM-single-step, as opposed to the multi-step extension EBRM-multi-step in Section 4.2.

Algorithm 1 EBRM-single-step

```
Input: $\Theta, \mathcal{D} = \{(s_i, a_i, r_i, s_i')\}_{i=1}^N$
Output: $\hat{\theta}$
Estimate $\hat{b}_\mu$ and $\hat{p}$.                                          ▷ Refer to Equation (10).
Compute $\hat{\theta} = \arg\min_{\theta\in\Theta} \hat{\hat{\mathcal{E}}}(\Upsilon_\theta, \hat{\mathcal{T}}^\pi\Upsilon_\theta)$.       ▷ Refer to Equations (11) and (12).
```

3.2 Statistical Error Bound

In this subsection, we will provide a statistical error bound for EBRM-single-step. As shown in Table 1, most existing distributional OPE methods do not have a finite sample error bound for their estimators. To the best of our knowledge, the only exception is the very recent work named FLE (Wu et al., 2023), which is only able to analyze the marginal distribution of the return instead of conditional distributions of the return on each state-action pair studied in this paper. In passing, we also note that Rowland et al. (2018) also shows the consistency of their estimator, but no error bound analysis (and so convergence rate) is provided. We will first focus on the realizability setting and defer the analysis for the non-realizable case in Section 4.

Assumption 2. *There exists a unique $\theta \in \Theta$ such that $\Upsilon_\pi(s,a) = \Upsilon_\theta(s,a)$ for all $s, a \in \mathcal{S}\times\mathcal{A}$.*

Note that realizability is a generally weaker assumption than the widely-assumed completeness assumption (e.g., used in FLE (Wu et al., 2023)) which states that for all $\theta \in \Theta$, there exist $\theta' \in \Theta$ such that $\mathcal{T}^\pi\Upsilon_\theta = \Upsilon_{\theta'}$, in that it implies realizability due to $\Upsilon_\pi = \lim_{T\to\infty}(\mathcal{T}^\pi)^T\Upsilon_\theta$ under mild conditions. In contrast with non-realizability settings (Section 4), the realizability assumption aligns the minimizer of inaccuracy $\mathcal{E}(\Upsilon, \Upsilon_\pi)$ (best approximation) and the minimizer of Bellman residual, leading to stronger arguments and results.

Additionally, we make several mild assumptions regarding the transition probability p and the candidate space Θ, including the subgaussian rewards. A random variable (vector) $\mathbf{X}$ being subgaussian implies its tail probability decaying as fast as gaussian distribution (e.g., gaussian mixture, bounded random variable), quantified with finite subgaussian norm $\|\mathbf{X}\|_{\psi_2} < \infty$, as explained in Appendix A.4.

Assumption 3. *For any $\theta \in \Theta$, the random element $Z(s,a;\theta)$, which follows $\Upsilon_\theta(s,a)$, has finite expectation with respect to their norms, and the reward distribution are subgaussian, i.e.,*

$$\sup_{\theta\in\Theta}\sup_{s,a} \mathbb{E}\|Z(s,a;\theta)\| < \infty \quad \textit{and} \quad \sup_{s,a} \|R(s,a)\|_{\psi_2} < \infty.$$

Assumption 4. *The offline data $\mathcal{D} = \{(s_i, a_i, r_i, s_i')\}_{i=1}^N$ are iid draws from $b_\mu \times p$.*

Assumption 5. *There exists a metric $\tilde{\eta}$ over $\mathcal{P}(\mathbb{R}^d)^{\mathcal{S}\times\mathcal{A}}$ such that $\mathrm{diam}(\Theta;\tilde{\eta}) := \sup_{\theta_1,\theta_2\in\Theta}\tilde{\eta}(\theta_1,\theta_2) < \infty$, where $\tilde{\eta}(\theta_1,\theta_2) := \tilde{\eta}(\Upsilon_{\theta_1},\Upsilon_{\theta_2})$. For arbitrary $c \in \mathbb{R}^d$, $\gamma_1, \gamma_2 \in [0,1]$, $(s,a), (\tilde{s},\tilde{a}) \in \mathcal{S}\times\mathcal{A}$, letting $Z_i(s,a) \sim \Upsilon_i(s,a)$ be such that $(Z_1(s,a), Z_3(s,a)) \in \mathbb{R}^d \times \mathbb{R}^d$ and $(Z_2(\tilde{s},\tilde{a}), Z_4(\tilde{s},\tilde{a})) \in \mathbb{R}^d \times \mathbb{R}^d$ are mutually independent, $\tilde{\eta}$ should satisfy*

$$\Big|\mathbb{E}\|c + \gamma_1 Z_1(s,a) - \gamma_2 Z_2(\tilde{s},\tilde{a})\| - \mathbb{E}\|c + \gamma_1 Z_3(s,a) - \gamma_2 Z_4(\tilde{s},\tilde{a})\|\Big| \qquad (14)$$
$$\leq \gamma_1 \cdot \tilde{\eta}(\Upsilon_1,\Upsilon_3) + \gamma_2 \cdot \tilde{\eta}(\Upsilon_2,\Upsilon_4).$$

Supremum-extended Wasserstein-1 metric $\mathbb{W}_{1,\infty}$, which is shown to be a metric by Lemma 2 of Bellemare et al. (2017b), is an example that satisfies (14), as proved in Appendix A.5. Then we can obtain the convergence rate $O(\sqrt{\log(N/\delta)/N})$ as follows, with the exact finite-sample error bound demonstrated in Appendix A.6.7. Its proof can be found in Appendix A.6, and its special case for $\Theta \subseteq \mathbb{R}^p$ is covered in Corollary 3 of Appendix A.7.

Theorem 2. (Inaccuracy for realizable scenario) *Under Assumptions 1–5, for any $\delta \in (0,1)$, given large enough sample size $N \geq N(\delta)$, our estimator $\hat{\theta} \in \Theta$ given by (13) satisfies the following bound with probability at least $1-\delta$,*

$$\bar{\mathcal{E}}(\Upsilon_{\hat{\theta}},\Upsilon_\pi) \lesssim \sqrt{\frac{1}{N}\log(\frac{(|\mathcal{S}\times\mathcal{A}|+N)}{\delta})}, \qquad (15)$$

where $N(\delta)$ depends on the complexity of Θ and $\lesssim$ means bounded by the given bound (RHS) multiplied by a positive number that does not depend on N, as defined in Appendix A.6.8.

4 Non-realizable Settings

4.1 Combating Non-realizability With Multi-step Extensions

In the tabular case, most traditional OPE/RL methods do not suffer from model mis-specification and thus realizability always holds. In contrast, in DRL, as our target is to estimate the conditional distribution of return given any state-action pair, which is an infinite-dimensional object, non-realizability could still happen. Hence understanding and analyzing DRL methods for the tabular case under the non-realizable scenario is both important and challenging.

In the previous section under realizability, Theorem 1 played a fundamental role in our analysis. Indeed, Theorem 1 is valid regardless of realizability (Assumption 2), and essentially implies

$$0 \leq \min_{\theta\in\Theta}\bar{\mathcal{E}}(\Upsilon_\theta,\Upsilon_\pi) \leq \bar{\mathcal{E}}(\Upsilon_{\theta_*},\Upsilon_\pi) \leq 2C_{\sup}B_1(\gamma;\beta_0)\cdot\bar{\mathcal{E}}(\Upsilon_{\theta_*},\mathcal{T}^\pi\Upsilon_{\theta_*}), \qquad (16)$$

where $\theta_* := \arg\min_{\theta\in\Theta}\bar{\mathcal{E}}(\Upsilon_\theta,\mathcal{T}^\pi\Upsilon_\theta)$. Violation of Assumption 2 (that is, non-realizability) implies nonzero value of $\bar{\mathcal{E}}(\Upsilon_{\theta_*},\mathcal{T}^\pi\Upsilon_{\theta_*}) > 0$, and so Theorem 1 no longer ensures that θ_* has the smallest inaccuracy among $\theta \in \Theta$. Thus non-realizability may lead to the following mismatch:

$$\tilde{\theta} := \arg\min_{\theta\in\Theta}\bar{\mathcal{E}}(\Upsilon_\theta,\Upsilon_\pi) \neq \arg\min_{\theta\in\Theta}\bar{\mathcal{E}}(\Upsilon_\theta,\mathcal{T}^\pi\Upsilon_\theta) =: \theta_*. \qquad (17)$$

Clearly, this mismatch is not due to sample variability, so it is unrealistic to hope that $\hat{\theta}$ defined by (13) would necessarily converge in probability to $\tilde{\theta}$ as $N \to \infty$.

To solve this issue, we propose a new approach. Temporarily ignoring mathematical rigor, the most important insight is that we can approximate $(\mathcal{T}^\pi)^m\Upsilon \approx \Upsilon_\pi$ with sufficiently large step level $m \in \mathbb{N}$. Thanks to the properties of energy distance, we have the following

$$\sup_{\theta\in\Theta}|\bar{\mathcal{E}}(\Upsilon_\theta,(\mathcal{T}^\pi)^m\Upsilon_\theta) - \bar{\mathcal{E}}(\Upsilon_\theta,\Upsilon_\pi)| \leq C\gamma^m, \quad \text{for some constant } C > 0. \qquad (18)$$

(See Appendix C.2.9 under assumptions of Theorem 3.) As $m \to \infty$, the RHS of (18) shrinks to zero, making m-step Bellman residual $\bar{\mathcal{E}}(\Upsilon_\theta,(\mathcal{T}^\pi)^m\Upsilon_\theta)$ approximate the inaccuracy $\bar{\mathcal{E}}(\Upsilon_\theta,\Upsilon_\pi)$. This leads the two minimizers to be close, as illustrated schematically in Figure 1:

$$\theta_*^{(m)} := \arg\min_{\theta\in\Theta}\bar{\mathcal{E}}(\Upsilon_\theta,(\mathcal{T}^\pi)^m\Upsilon_\theta) \approx \arg\min_{\theta\in\Theta}\bar{\mathcal{E}}(\Upsilon_\theta,\Upsilon_\pi) =: \tilde{\theta} \text{ for large enough } m. \qquad (19)$$

One can intuitively guess that larger step level m is required when the extent of non-realizability is large. Although multi-step idea has been widely employed for the purpose of improving sample efficiency particularly in traditional RL (e.g., Chen et al., 2021), ours is the first approach to use it in DRL for the purpose of overcoming non-realizability, to the best of our knowledge.

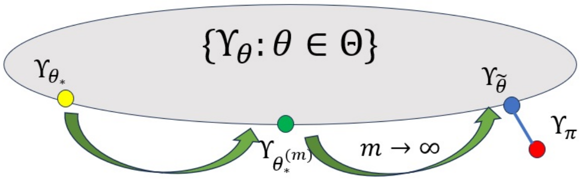
$\{\Upsilon_\theta : \theta \in \Theta\}$
Υ_{θ_*}
$\Upsilon_{\theta_*^{(m)}}$
$m \to \infty$
$\Upsilon_{\tilde\theta}$
Υ_π

Figure 1: Larger m makes $(\mathcal{T}^\pi)^m\Upsilon_\theta \approx \Upsilon_\pi$ in Energy Distance, and thereby leads to $\theta_*^{(m)} \approx \tilde\theta$.

4.2 Bootstrap Operator

Generalizing from definition of $\hat{\mathcal{T}}^\pi$ based on (10), we consider $\hat{Z}^{(m)}(s,a;\theta) \sim (\hat{\mathcal{T}}^\pi)^m\Upsilon_\theta(s,a)$ as the distribution of an m-lengthed trajectories of tuples (s,a,r,s') that is generated under the estimated transition $\hat p$ and the target policy π:

$$\hat{Z}^{(m)}(s,a;\theta) \overset{D}{=} \sum_{t=1}^{m} \gamma^{t-1}\hat{R}^{(t)} + \gamma^m Z(\hat{S}^{(m)}, \hat{A}^{(m)};\theta), \quad \text{where} \tag{20}$$

$$(\hat{R}^{(t)}, \hat{S}^{(t)}) \sim \hat{p}(\cdots|\hat{S}^{(t-1)}, \hat{A}^{(t-1)}) \quad \text{and} \quad \hat{A}^{(t)} \sim \pi(\cdot|\hat{S}^{(t)}) \quad \forall t \geq 1,\ (\hat{S}^{(0)}, \hat{A}^{(0)}) = (s,a).$$

Now we can define the estimated and the population Bellman residual, as well as the inaccuracy function, along with their minimizers as:

$$\hat{F}_m(\theta) := \hat{\bar{\mathcal{E}}}\big(\Upsilon_\theta, (\hat{\mathcal{T}}^\pi)^m\Upsilon_\theta\big), \quad F_m(\theta) := \bar{\mathcal{E}}\big(\Upsilon_\theta, (\mathcal{T}^\pi)^m\Upsilon_\theta\big), \quad F(\theta) := \bar{\mathcal{E}}\big(\Upsilon_\theta, \Upsilon_\pi\big), \tag{21}$$
$$\hat\theta^{(m)} := \arg\min_{\theta\in\Theta} \hat{F}_m(\theta), \quad \theta_*^{(m)} := \arg\min_{\theta\in\Theta} F_m(\theta), \quad \tilde\theta := \arg\min_{\theta\in\Theta} F(\theta).$$

However, the estimation of m-step Bellman operator (20) generally requires computation of N^m trajectories (as discussed in Appendix B.1), which amounts to a heavy computational burden.

To alleviate such burden, we will instead bootstrap $M \ll N^m$ many trajectories by first sampling the initial state-action pairs $(s_i^{(0)}, a_i^{(0)})$ $(1 \leq i \leq M)$ from $\hat{b}_\mu$ and then resampling the subsequent $r_i^{(t+1)}, s_i^{(t+1)} \sim \hat{p}(\cdots|s_i^{(t)}, a_i^{(t)})$ and $a_i^{(t+1)} \sim \pi(\cdot|s_i^{(t+1)})$ for m steps. Let $\hat{p}_m^{(B)}(\cdots|s,a)$ be the empirical probability measure of $(\sum_{t=1}^m \gamma^{t-1} r_i^{(t)}, s_i^{(m)})$ conditioning on $(s_i^{(0)}, a_i^{(0)}) = (s,a)$. We define the *bootstrap operator* as follows, with an abuse of notation $\mathcal{B}_m Z(s,a;\theta) \sim \mathcal{B}_m\Upsilon_\theta(s,a)$,

$$\mathcal{B}_m Z(s,a;\theta) :\overset{D}{=} \sum_{t=1}^{m} \gamma^{t-1}\hat{R}_b^{(t)} + \gamma^m Z(\hat{S}_b^{(m)}, \hat{A}_b^{(m)};\theta), \tag{22}$$

$$\text{where} \quad \Big(\sum_{t=1}^{m}\gamma^{t-1}\hat{R}_b^{(t)}, \hat{S}_b^{(m)}\Big) \sim \hat{p}_m^{(B)}(\cdots|s,a) \quad \text{and} \quad \hat{A}_b^{(m)} \sim \pi(\cdot|\hat{S}_b^{(m)}).$$

Then we can compute our objective function and derive the bootstrap-based multi-step estimator.

$$\hat{F}_m^{(B)}(\theta) := \hat{\bar{\mathcal{E}}}\big(\Upsilon_\theta, \mathcal{B}_m\Upsilon_\theta\big) \quad \text{and} \quad \hat\theta_m^{(B)} := \arg\min_{\theta\in\Theta} \hat{F}_m^{(B)}(\theta). \tag{23}$$

We will refer to this method as EBRM-multi-step, whose procedure is summarized in Algorithm 2.

4.3 Statistical Error Bound

In this section, we develop a theoretical guarantee for $\bar{\mathcal{E}}\big(\Upsilon_{\hat\theta_m^{(B)}}, \Upsilon_{\tilde\theta}\big)$, where $\Upsilon_{\tilde\theta}$ is the best one we can achieve under the non-realizability. To proceed, we need to first deal with the parameter convergence from $\hat\theta_m^{(B)}$ to $\tilde\theta$, which relies on the following assumptions regarding the inaccuracy function $F(\cdot)$ (21), distance $\tilde\eta$ (Assumpion 5), and candidate space Θ.

Algorithm 2 EBRM-multi-step

```
**Input:** $\Theta, \mathcal{D} = \{(s_i, a_i, r_i, s_i')\}_{i=1}^N, m, M$
**Output:** $\hat{\theta}_m^{(B)}$
Estimate $\hat{b}_\mu$ and $\hat{p}$.                                              ▷ Refer to Equation (10).
Randomly generate $M$ tuples of $(\sum_{t=1}^m \gamma^{t-1} r_i^{(t)}, s_i^{(m)})$ $(1 \leq i \leq M)$.
$\hat{\theta}_m^{(B)} = \arg\min_{\theta \in \Theta} \hat{\bar{\mathcal{E}}}(\Upsilon_\theta, \mathcal{B}_m \Upsilon_\theta)$.                    ▷ Refer to Equations (22) and (23).
```

Assumption 6. *The inaccuracy function* (21) $F(\cdot) : \Theta \subset \mathbb{R}^p \to \mathbb{R}$ *has a unique minimizer* $\tilde{\theta}$*, and lower bounded by a polynomial of degree* $q \geq 1$*. That is, for all* $\theta \in \Theta$*, we have* $F(\theta) \geq F(\tilde{\theta}) + c_q \cdot \|\theta - \tilde{\theta}\|^q$ *for some constant* $c_q > 0$*.*

Assumption 7. *The candidate space* Θ *is compact (i.e.,* $\mathrm{diam}(\Theta; \|\cdot\|) < \infty$*). Furthermore, there exists* $L > 0$ *such that*

$$\tilde{\eta}(\theta_1, \theta_2) \leq L\|\theta_1 - \theta_2\| \quad \textit{for} \quad \forall \theta_1, \theta_2 \in \Theta.$$

Assumption 8. $\tilde{\eta}$ *satisfies contractive property, i.e.,* $\tilde{\eta}(\mathcal{T}^\pi \Upsilon_1, \mathcal{T}^\pi \Upsilon_2) \leq \gamma \cdot \tilde{\eta}(\Upsilon_1, \Upsilon_2)$*, where* $\mathcal{T}^\pi$ (2) *may correspond to an arbitrary transition* $p(\cdots|s, a)$*.*

Assumption 6 is used in quantifying the convergence rate. Compactness in Assumption 7 is for ensuring the existence of a minimizer of the estimated objective function (23), which is proved to be continuous in Appendix C.2.11. Compactness can be relaxed to "bounded" under mild conditions. Assumption 8 makes (18) feasible, and thereby shrinks the disparity between Bellman minimizer and the best approximation (19). This is satisfied by $\mathbb{W}_{1,\infty}$ that also satisfies property (14), as proved in Lemma 3 of Bellemare et al. (2017b).

Due to space constraints, we only present a simplified result below (proof in Appendix B.4), and a more detailed version of the finite-sample error bound for a fixed m is given in Appendix B.4.3.

Theorem 3. *Under Assumptions 1, 3–8, letting* $M = \lfloor C_1 \cdot N \rfloor$ *and* $m = \lfloor \frac{1}{4} \log_{(1/\gamma)}(C_2 N / \log N) \rfloor$ *for arbitrary constants* $C_1, C_2 > 0$*, we have the optimal convergence rate of the upper bound*

$$\bar{\mathcal{E}}(\Upsilon_{\hat{\theta}_m^{(B)}}, \Upsilon_{\tilde{\theta}}) \leq \tilde{O}_p\left[\frac{1}{N^{1/(4q)}} \cdot \left\{\log_{\frac{1}{\gamma}}\left(\frac{N}{\log N}\right)\right\}^{2/q}\right],$$

where $\tilde{O}_p$ *indicates the rate of convergence up to logarithmic order.*

The convergence rate of Theorem 3 is the result of the (asymptotically) optimal choice of M and m. In our analysis, we notice a form of bias-variance trade-off in the selection of m, as explained in Appendix B.4.4. Practically, we set $M = N$ which works fine in the simulations of Section 5. A practical rule of m will be discussed in Section 4.4.

Note that the finite-sample error bound in Appendix B.4.3 is applicable to the setting with $m = 1$ and realizability assumption. For instance, assuming that the inaccuracy function $F(\theta)$ is lower-bounded by a quadratic polynomial $q = 2$, it gives us the bound $O[\{\log(N/\delta_1)/N\}^{1/8}]$ under the ideal case where we can ignore the last two sources of inaccuracy specified in Appendix B.4.4, each associated with bootstrap and non-realizability. We can see that it is much slower than the convergence rate $O(\sqrt{\log(N/\delta)/N})$ of Theorem 2, implying that it does not degenerate into Theorem 2. This is fundamentally due to a different proof structure that can be introduced via the application of Theorem 1 in the proof of Theorem 2, as intuitively explained in Appendix C.3.1. As explained earlier in Section 4.1, Theorem 1 can be used effectively to construct convergence of $\hat{\theta}$ under realizability.

4.4 Data-adaptive way of selecting step level

We need to choose m in practice. We will apply Lepski's rule (Lepskii, 1991). Since multi-step construction includes bootstrapping from the observed samples $\mathcal{D} = \{(s_i, a_i, r_i, s_i')\}_{i=1}^N$ (Section 4.2), this enables us to form a confidence interval. Starting from large enough m, we can decrease it until the intersection of the confidence intervals becomes a null set. To elaborate, given the data $\mathcal{D}$, we first generate multiple estimates of $\hat{\theta}_m^{(B)}$ (say $\hat{\theta}_{m,j}^{(B)}$ for $1 \leq j \leq J$), and calculate the disparity from the single-step estimation which has no randomness once $\mathcal{D}$ is given, that is $\hat{\hat{\mathcal{E}}}(\Upsilon_{\hat{\theta}}, \Upsilon_{\hat{\theta}^{(B)}_{m,j}})$ $(1 \leq j \leq J)$. Then we calculate their means and standard deviations, forming a (Mean ± SD) interval for each m. Starting from a large enough value, we decrease m by one or more at a time, and select the m that makes the intersection become a null set $\cap_{k \geq m} I^{(k)} = \varnothing$. If we did not obtain the null set until $\cap_{k \geq 2} I^{(k)} \neq \varnothing$, then we use EBRM-single-step (13) without boostrap. Details are explained in Algorithm 3 of Appendix D.3.1.

5 Experiments

We assume a state space $\mathcal{S} = \{1, 2, \cdots, 30\}$ and an action space $\mathcal{A} = \{-1, 1\}$, each action representing left or right. With the details of the environment in Appendix D.2.1, the initial state distribution and behavior / target policies are

$$S \sim \text{Unif}\{1, 2, \cdots, 30\} \quad \text{and} \quad A \sim b(\cdot|S), \quad \text{where} \quad b(a|s) = 1/2 \quad \text{for} \quad \forall s, a \in \mathcal{S} \times \mathcal{A},$$
$$\pi(-1|s) = 0 \quad \text{and} \quad \pi(1|s) = 1 \quad \text{for} \quad \forall s \in \mathcal{S}. \tag{24}$$

We compare three methods: EBRM, FLE (Wu et al., 2023), and QRTD (Dabney et al., 2018b). Here, we assume realizability where the correct model is known (details in Appendix D.2.2) under two settings (with small and large variances), and the step level m for EBRM is chosen in a data-adaptive way in Section 4.4. With other tuning parameter selections explained in Appendix D.3, we repeated 100 simulations with the given sample size for each case, whose mean and standard deviation (within parenthesis) are recorded in Table 2. EBRM showed the lowest inaccuracy values measured by both $\bar{\mathcal{E}}(\Upsilon_{\hat{\theta}}, \Upsilon_{\pi})$ and $\overline{\mathbb{W}}_1(\Upsilon_{\hat{\theta}}, \Upsilon_{\pi})$, where $\overline{\mathbb{W}}_1$ indicates expectation-extended (6) Wasserstein-1 metric.

We also performed simulations in non-realizable scenarios (Appendix D.2.3) with more variety of sample sizes, and included Wasserstein-1 metric between marginal return distributions (Tables 8–13 of Appendix D.4). In most cases, EBRM showed outstanding performance.

Table 2: Mean $\bar{\mathcal{E}}$-inaccuracy (top) and $\overline{\mathbb{W}}_1$-inaccuracy (bottom) (standard deviation in parenthesis) over 100 simulations under realizability ($\gamma = 0.99$). Smallest inaccuracy values are in boldface.

	Small variance			**Large variance**		
Sample size	2000	5000	10000	2000	5000	10000
EBRM (Ours)	**0.046** (0.060)	**0.019** (0.022)	**0.008** (0.010)	**0.728** (0.920)	**0.301** (0.354)	**0.128** (0.167)
FLE	5.533 (6.448)	2.385 (2.883)	1.220 (1.618)	24.603 (25.768)	14.482 (16.101)	6.528 (7.814)
QRTD	48.679 (34.323)	46.032 (30.909)	49.402 (34.617)	105.274 (11.728)	75.173 (21.515)	70.483 (33.965)

	Small variance			**Large variance**		
Sample size	2000	5000	10000	2000	5000	10000
EBRM (Ours)	**1.339** (0.651)	**0.985** (0.388)	**0.782** (0.227)	**21.221** (10.337)	**15.532** (6.117)	**12.371** (3.595)
FLE	12.374 (7.843)	8.036 (5.091)	5.694 (3.773)	101.232 (58.586)	79.628 (46.772)	53.745 (33.948)
QRTD	56.739 (23.716)	54.397 (22.259)	57.145 (24.314)	274.405 (11.003)	236.383 (22.376)	223.537 (38.935)

6 Conclusion

In this paper, we justify the use of expectation-extended distances for Bellman residual minimization in DRL under general state-action space, based on which we propose a distributional OPE method called EBRM. We establish finite sample error bounds of the proposed estimator for the tabular case with or without realizability assumption. One interesting future direction is to extend EBRM to non-tabular case via linear MDP (e.g., Lazic et al., 2020; Bradtke & Barto, 1996), as we will briefly discuss in Appendix C.3.2.

REFERENCES

Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International conference on machine learning*, pp. 449–458. PMLR, 2017a.

Marc G Bellemare, Ivo Danihelka, Will Dabney, Shakir Mohamed, Balaji Lakshminarayanan, Stephan Hoyer, and Rémi Munos. The cramer distance as a solution to biased wasserstein gradients. *arXiv preprint arXiv:1705.10743*, 2017b.

Steven J Bradtke and Andrew G Barto. Linear least-squares algorithms for temporal difference learning. *Machine learning*, 22:33–57, 1996.

Zaiwei Chen, Siva Theja Maguluri, Sanjay Shakkottai, and Karthikeyan Shanmugam. Finite-sample analysis of off-policy td-learning via generalized bellman operators. *Advances in Neural Information Processing Systems*, 34:21440–21452, 2021.

Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. In *International conference on machine learning*, pp. 1096–1105. PMLR, 2018a.

Will Dabney, Mark Rowland, Marc Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018b.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Jonas Moritz Kohler and Aurelien Lucchi. Sub-sampled cubic regularization for non-convex optimization. In *International Conference on Machine Learning*, pp. 1895–1904. PMLR, 2017.

Nevena Lazic, Dong Yin, Mehrdad Farajtabar, Nir Levine, Dilan Gorur, Chris Harris, and Dale Schuurmans. A maximum-entropy approach to off-policy evaluation in average-reward mdps. *Advances in Neural Information Processing Systems*, 33:12461–12471, 2020.

OV Lepskii. On a problem of adaptive estimation in gaussian white noise. *Theory of Probability & Its Applications*, 35(3):454–466, 1991.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

Thanh Tang Nguyen, Sunil Gupta, and Svetha Venkatesh. Distributional reinforcement learning with maximum mean discrepancy. *Association for the Advancement of Artificial Intelligence (AAAI)*, 2020.

Thanh Nguyen-Tang, Sunil Gupta, and Svetha Venkatesh. Distributional reinforcement learning via moment matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 9144–9152, 2021.

Mark Rowland, Marc Bellemare, Will Dabney, Rémi Munos, and Yee Whye Teh. An analysis of categorical distributional reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 29–37. PMLR, 2018.

Mark Rowland, Robert Dadashi, Saurabh Kumar, Rémi Munos, Marc G Bellemare, and Will Dabney. Statistics and samples in distributional reinforcement learning. In *International Conference on Machine Learning*, pp. 5528–5536. PMLR, 2019.

Bodhisattva Sen. A gentle introduction to empirical process theory and applications. *Lecture Notes, Columbia University*, 11:28–29, 2018.

Yi Su, Pavithra Srinath, and Akshay Krishnamurthy. Adaptive estimator selection for off-policy evaluation. In *International Conference on Machine Learning*, pp. 9196–9205. PMLR, 2020.

Ke Sun, Yingnan Zhao, Yi Liu, Wulong Liu, Bei Jiang, and Linglong Kong. Distributional reinforcement learning via sinkhorn iterations. *arXiv preprint arXiv:2202.00769*, 2022.

Gábor J Székely and Maria L Rizzo. Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference*, 143(8):1249–1272, 2013.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Jiayi Wang, Raymond KW Wong, and Xiaoke Zhang. Low-rank covariance function estimation for multidimensional functional data. *Journal of the American Statistical Association*, 117(538): 809–822, 2022.

Runzhe Wu, Masatoshi Uehara, and Wen Sun. Distributional offline policy evaluation with predictive error guarantees. *arXiv preprint arXiv:2302.09456*, 2023.

Derek Yang, Li Zhao, Zichuan Lin, Tao Qin, Jiang Bian, and Tie-Yan Liu. Fully parameterized quantile function for distributional reinforcement learning. *Advances in neural information processing systems*, 32, 2019.

Pushi Zhang, Xiaoyu Chen, Li Zhao, Wei Xiong, Tao Qin, and Tie-Yan Liu. Distributional reinforcement learning for multi-dimensional reward functions. *Advances in Neural Information Processing Systems*, 34:1519–1529, 2021.