
Distributional Offline Policy Evaluation with Predictive Error Guarantees

Runzhe Wu¹ Masatoshi Uehara¹ Wen Sun¹

Abstract

We study the problem of estimating the distribution of the return of a policy using an offline dataset that is not generated from the policy, i.e., distributional offline policy evaluation (OPE). We propose an algorithm called Fitted Likelihood Estimation (FLE), which conducts a sequence of Maximum Likelihood Estimation (MLE) and has the flexibility of integrating any state-of-the-art probabilistic generative models as long as it can be trained via MLE. FLE can be used for both finite-horizon and infinite-horizon discounted settings where rewards can be multi-dimensional vectors. Our theoretical results show that for both finite-horizon and infinite-horizon discounted settings, FLE can learn distributions that are close to the ground truth under total variation distance and Wasserstein distance, respectively. Our theoretical results hold under the conditions that the offline data covers the test policy’s traces and that the supervised learning MLE procedures succeed. Experimentally, we demonstrate the performance of FLE with two generative models, Gaussian mixture models and diffusion models. For the multi-dimensional reward setting, FLE with diffusion models is capable of estimating the complicated distribution of the return of a test policy.

1. Introduction

Traditional Reinforcement Learning (RL) focuses on studying the expected behaviors of a learning agent. However, modeling the expected behavior is not enough for many interesting applications. For instance, when estimating the value of a new medical treatment, instead of just predicting its expected value, we may be interested in estimating the variance of the value as well. For a self-driving car whose goal is to reach a destination as soon as possible, in addition

to predicting the expected traveling time, we may be interested in estimating the tails of the distribution of traveling time so that customers can prepare for worst-case situations. Other risk-sensitive applications in finance and control often require one to model beyond the expectation as well.

In this work, we study how to estimate the distribution of the return of a policy in Markov Decision Processes (MDPs) using only an offline dataset that is not necessarily generated from the test policy (i.e., distributional offline policy evaluation). Estimating distributions of returns has been studied in the setting called distributional RL (Bellemare et al., 2017), where most existing works focus on solving the regular RL problem, i.e., finding a policy that maximizes the expected return by treating the task of predicting additional information beyond the mean as an auxiliary task. Empirically, it is believed that this auxiliary task helps representation learning which in turn leads to better empirical performance. Instead of focusing on this auxiliary loss perspective, we aim to design distributional OPE algorithms, which can accurately estimate the distribution of returns with provable guarantees. We are also interested in the setting where the one-step reward could be *multi-dimensional* (i.e., multi-objective RL), and the state/action spaces could be large or even continuous. This requires us to design new algorithms that can leverage rich function approximation (e.g., state-of-art probabilistic generative models).

Our algorithm, *Fitted Likelihood Estimation* (FLE), is inspired by the classic OPE algorithm Fitted Q Evaluation (FQE) (Munos & Szepesvári, 2008). Given a test policy and an offline dataset, FLE iteratively calls a supervised learning oracle — Maximum Likelihood Estimation (MLE) in this case, to fit a conditional distribution to approximate a target distribution constructed using the distribution learned from the previous iteration. At the end of the training procedure, it outputs an estimator which approximates the true distribution of the return of the test policy. Our algorithm is simple: like FQE, it decomposes the distributional OPE problem into a sequence of supervised learning problems (in this case, MLE). Thus it has great flexibility to leverage any state-of-art probabilistic generative models as long as it can be trained via MLE. Such flexibility is important, especially when we have large state/action spaces, and reward vectors coming from complicated high-dimensional distributions. FLE naturally works for both finite-horizon

¹Department of Computer Science, Cornell University, Ithaca, NY, USA. Correspondence to: Runzhe Wu <rw646@cornell.edu>.

setting and infinite-horizon discounted setting.

Theoretically, we prove that our algorithm, FLE, can learn an accurate estimator of the return distribution for both finite-horizon MDPs and infinite-horizon discounted MDPs, under the assumptions that (1) *MLE can achieve good in-distribution generalization bounds (i.e., supervised learning succeeds)*, and (2) *the offline state-action distribution covers the test policy’s state-action distribution*. The first condition is well studied in statistical learning theory, and in practice, the state-of-the-art probabilistic generative models trained via MLE (e.g., FLOW models (Dinh et al., 2014) and Diffusion models (Sohl-Dickstein et al., 2015)) indeed also exhibit amazing generalization ability. The second condition is necessary for offline RL and is widely used in the regular offline RL literature (e.g., Munos & Szepesvári (2008)). In other words, our analysis is modular: it simply transfers the supervised learning MLE in-distribution generalization bounds to a bound of distributional OPE. The accuracy of the estimator computed by FLE is measured under total variation distance and p -Wasserstein distance, for finite-horizon setting and infinite-horizon discounted setting, respectively. To complete the picture, we further provide concrete examples showing that MLE can provably have small in-distribution generalization errors. To the best of our knowledge, this is the first PAC (Probably Approximately Correct) learning algorithm for distributional OPE with general function approximation.

Finally, we demonstrate our approach on a rich observation combination lock MDP where it has a latent structure with the observations being high-dimensional and continuous (Misra et al., 2020; Agarwal et al., 2020a; Zhang et al., 2022b). We consider the setting where the reward comes from complicated multi-dimensional continuous distributions (thus existing algorithms such as quantile-regression TD (Dabney et al., 2018) do not directly apply here). We demonstrate the flexibility of our approach by using two generative models in FLE: the classic Gaussian mixture model and state-of-the-art diffusion model (Ho et al., 2020).

1.1. Related Works

Distributional RL. Quantile regression TD (Dabney et al., 2018) is one of the common approaches for distributional OPE. A very recent work (Rowland et al., 2023) demonstrates that quantile regression TD can converge to the TD fixed point solution of which the existence is proved under an ℓ_∞ -style norm (i.e., sup over all states). Rowland et al. (2023) do not consider the sample complexity of OPE and the impact of learning from off-policy samples, and their convergence analysis is asymptotic. Also, quantile regression TD only works for scalar rewards. Another popular approach is categorical TD (Bellemare et al., 2017), where one explicitly discretizes the return space. However, for

high-dimensional rewards, explicitly discretizing the return space evenly can suffer the curse of dimensionality and fail to capture some low-dimensional structures in the data distribution. Moreover, there is no convergence or sample complexity analysis of the categorical algorithm for OPE. Another direction in distributional RL concentrates on estimating cumulative distribution functions (CDFs) instead of densities (Zhang et al., 2022a; Prashanth & Bhat, 2022). In addition, there are also methods based on generative models that aim to effectively represent continuous return distributions (Freirich et al., 2019; Doan et al., 2018; Li & Faisal, 2021). We discuss some closely related works below.

Ma et al. (2021) studied distributional offline policy optimization. They focused on tabular MDPs with scalar rewards, and their algorithm can learn a pessimistic estimate of the true inverse CDF of the return. Keramati et al. (2020) also uses the distributional RL framework to optimistically estimate the CVaR value of a policy’s return. Their analysis also only applies to tabular MDPs with scalar rewards. In contrast, we focus on distributional OPE with general function approximation beyond tabular or linear formats and MDPs with multi-dimensional rewards.

Zhang et al. (2021) also consider learning from vector-valued rewards. They propose a practical algorithm that minimizes the Maximum Mean Discrepancy (MMD) without a sample complexity analysis. In contrast, we use MLE to minimize total variation distance, and our error bound is based on total variation distance. Note that a small total variation distance implies a small MMD but not vice versa, which implies that our results are stronger.

Huang et al. (2021; 2022) explore return distribution estimation for contextual bandits and MDPs using off-policy data. They focus on learning CDFs with an estimator that leverages importance sampling and learns the transition and reward of the underlying MDP to reduce variance while maintaining unbiasedness. However, their estimator can incur exponential error in the worst case due to importance sampling. Moreover, they measure estimation error using the ℓ_∞ norm on CDFs, which is upper bounded by total variation distance but not the other way around. They further showed how to estimate a range of risk functionals via the estimated distribution. Notably, our method is also applicable to risk assessment, as shown in Remark 4.8.

Offline policy evaluation. Fitted Q evaluation (FQE) (Munos & Szepesvári, 2008; Ernst et al., 2005) is one of the most classic OPE algorithms. Many alternative approaches have been recently proposed, such as minimax algorithms (Yang et al., 2020; Feng et al., 2019; Uehara et al., 2020). Somewhat surprisingly, algorithms based on FQE are often robust and achieve stronger empirical performance in various benchmark tasks (Fu et al., 2021; Chang et al., 2022). Our proposed algorithm can be understood as a direct gener-

alization of FQE to the distributional setting. Note sequential importance sampling approaches (Jiang & Li, 2016; Precup et al., 2000) in regular RL have been applied to estimate distributions (Chandak et al., 2021). However, these methods suffer from the curse of the horizon, i.e., the variance necessarily grows exponentially in the horizon.

2. Preliminaries

In this section, we introduce the setup of the Markov decision process and the offline policy evaluation.

Notations. We define $\Delta(\mathcal{S})$ as the set of all distributions over a set \mathcal{S} . For any $a, b \in \mathbb{R}$, we denote $[a, b] = \{x \in \mathbb{R} : a \leq x \leq b\}$. For any integer N , we denote $[N]$ as the set of integers between 1 and N inclusively. Given two distributions P_1 and P_2 on a set \mathcal{S} , we denote d_{tv} as the total variation distance between the two distributions, i.e., $d_{tv}(P_1, P_2) = \|P_1 - P_2\|_1/2$. We denote $d_{w,p}$ as the p -Wasserstein distance, i.e., $d_{w,p}(P_1, P_2) = (\inf_{c \in \mathcal{C}} \mathbb{E}_{x,y \sim c} \|x - y\|^p)^{1/p}$ where \mathcal{C} denotes the set of all couplings of P_1 and P_2 . We note that d_{tv} dominates $d_{w,p}$ when the support is bounded (see Lemma C.6 for details):

$$d_{w,p}^p(P_1, P_2) \leq \text{diam}^p(\mathcal{S}) \cdot d_{tv}(P_1, P_2) \quad (1)$$

where $\text{diam}(\mathcal{S}) = \sup_{x,y \in \mathcal{S}} \|x - y\|$ is the diameter of \mathcal{S} .

2.1. Finite-Horizon MDPs

We consider a finite-horizon MDP with a vector-valued reward function, which is a tuple $M(\mathcal{X}, \mathcal{A}, r, P, H, \mu)$ where \mathcal{X} and \mathcal{A} are the state and action spaces, respectively, P is the transition kernel, r is the reward function, i.e., $r(x, a) \in \Delta([0, 1]^d)$ where $d \in \mathbb{Z}^+$, H is the length of each episode, and $\mu \in \Delta(\mathcal{X})$ is the initial state distribution. A policy is a mapping $\pi : \mathcal{X} \rightarrow \Delta(\mathcal{A})$. We denote $z \in [0, H]^d$ as the accumulative reward vector across H steps, i.e., $z = \sum_{h=1}^H r_h$. Note that z is a random vector whose distribution is determined by a policy π and the MDP. We denote $Z^\pi \in \Delta([0, H]^d)$ as the distribution¹ of the random variable z under policy π . In this paper, we are interested in estimating Z^π using offline data. We also define conditional distributions $Z_h^\pi(x, a) \in \Delta([0, H]^d)$ which is the distribution of the return under policy π starting with state action $(x_h, a_h) := (x, a)$ at time step h . It is easy to see that $Z^\pi = \mathbb{E}_{x \sim \mu, a \sim \pi(x)} [Z_1^\pi(x, a)]$. We define d_h^π as the state-action distribution induced by policy π at time step h , and $d^\pi = \sum_{h=1}^H d_h^\pi / H$ as the average state-action distribution induced by π .

We denote the distributional Bellman operator (Morimura et al., 2012) associated with π as \mathcal{T}^π , which maps a condi-

tional distribution to another conditional distribution: given a state-action conditional distribution $f \in \mathcal{X} \times \mathcal{A} \mapsto \Delta([0, H]^d)$, we have $\mathcal{T}^\pi f \in \mathcal{X} \times \mathcal{A} \mapsto \Delta([0, H]^d)$, such that for any (x, a, z) :

$$\begin{aligned} [\mathcal{T}^\pi f](z | x, a) \\ = \mathbb{E}_{r \sim r(s,a), x' \sim P(x,a), a' \sim \pi(x)} [f(z - r | x', a')]. \end{aligned}$$

We can verify that $\mathcal{T}^\pi Z_{h+1}^\pi = Z_h^\pi$ for all h .

2.2. Discounted Infinite-Horizon MDPs

The discounted infinite-horizon MDP is a tuple $M(\mathcal{X}, \mathcal{A}, r, P, \gamma, \mu)$. The return vector is defined as $z = \sum_{h=1}^{\infty} \gamma^{h-1} r_h$. We call $\gamma \in (0, 1)$ the discount factor. The distribution of return z is thus $Z^\pi \in \Delta([0, (1 - \gamma)^{-1}]^d)$. We also define the conditional distribution $\bar{Z}^\pi(x, a) \in \Delta([0, (1 - \gamma)^{-1}]^d)$ which is the distribution of the return under policy π starting with state action (x, a) . It is easy to see that $Z^\pi = \mathbb{E}_{x \sim \mu, a \sim \pi(x)} [\bar{Z}^\pi(x, a)]$. The state-action distribution of a given policy π is also defined in a discounted way: $d^\pi = (1 - \gamma)^{-1} \sum_{h=1}^{\infty} \gamma^{h-1} d_h^\pi$ where d_h^π is the state-action distribution induced by π at time step h . The distributional Bellman operator maps a state-action conditional distribution $f \in \mathcal{X} \times \mathcal{A} \mapsto ([0, (1 - \gamma)^{-1}]^d)$ to $\mathcal{T}^\pi f \in \mathcal{X} \times \mathcal{A} \mapsto ([0, (1 - \gamma)^{-1}]^d)$ for which

$$\begin{aligned} [\mathcal{T}^\pi f](z | x, a) \\ = \mathbb{E}_{r \sim r(x,a), x' \sim P(x,a), a' \sim \pi(x)} \left[f \left(\frac{z - r}{\gamma} \mid x', a' \right) \right]. \end{aligned}$$

for any (x, a, z) . We can verify that \bar{Z}^π is a fixed point of the distributional Bellman operator, i.e., $\mathcal{T}^\pi \bar{Z}^\pi = \bar{Z}^\pi$.

2.3. Offline Policy Evaluation Setup

We consider estimating the distribution Z^π using offline data which does not come from π (i.e., off-policy setting). We assume we have a dataset $\mathcal{D} = \{x_i, a_i, r_i, x'_i\}_{i=1}^n$ that contains i.i.d. tuples, such that $x, a \sim \rho \in \Delta(\mathcal{X} \times \mathcal{A})$, $s' \sim P(\cdot | s, a)$, and $r \sim r(s, a)$. For finite-horizon MDPs, we randomly and evenly split \mathcal{D} into H subsets, $\mathcal{D}_1, \dots, \mathcal{D}_H$, for the convenience of analysis. Each subset contains n/H samples. For infinite-horizon MDPs, we split it into T subsets in the same way. Here T is the number of iterations which we will define later.

We consider learning distribution Z^π via general function approximation. For finite-horizon MDPs, we denote \mathcal{F}_h as a function class that contains state-action conditional distributions, i.e., $\mathcal{F}_h \subset \mathcal{X} \times \mathcal{A} \mapsto \Delta([0, H]^d)$, which will be used to learn Z_h^π . For infinite-horizon MDPs, we assume a function class $\mathcal{F} \subset \mathcal{X} \times \mathcal{A} \mapsto \Delta([0, (1 - \gamma)^{-1}]^d)$.

¹Formally, they are called probability density functions in the continuous setting and probability mass functions in discrete settings, which are different from cumulative distribution functions.

3. Fitted Likelihood Estimation

In this section, we present our algorithm — *Fitted Likelihood Estimation* (FLE) for distributional OPE. Algorithm 1 is for finite-horizon MDPs, and Algorithm 2 is for infinite-horizon MDPs.

Algorithm 1 takes the offline dataset $\mathcal{D} = \{\mathcal{D}_h\}_{h=1}^H$ and the function class $\{\mathcal{F}_h\}_{h=1}^H$ as inputs and iteratively performs Maximum likelihood estimation (MLE) starting from H to time step $h = 1$. For a particular time step h , given \hat{f}_{h+1} which is learned from the previous iteration, FLE treats $\mathcal{T}^\pi \hat{f}_{h+1}$ as the target distribution to fit. To learn $\mathcal{T}^\pi \hat{f}_{h+1}$, it first generates samples from it (Line 6), which is doable as long as we can generate samples from the conditional distribution $\hat{f}_{h+1}(\cdot|x, a)$ given any (x, a) . Once we generate samples from $\mathcal{T}^\pi \hat{f}_{h+1}$, we fit \hat{f}_h to estimate $\mathcal{T}^\pi \hat{f}_{h+1}$ by MLE (Line 13). The algorithm returns \hat{f}_1 to approximate Z_1^π . To estimate Z^π , we can compute $\mathbb{E}_{x \sim \mu, a \sim \pi(x)} \hat{f}_1(x, a)$, recalling μ is the initial state distribution.

Algorithm 2 is quite similar to Algorithm 1 but is for infinite-horizon MDPs, and it has two distinctions. First, we introduce the discount factor γ . Second, compared to Algorithm 1 where we perform MLE in a backward manner (from $h = H$ to 1), here we repeatedly apply MLE in a time-independent way. Particularly, it treats $\mathcal{T}^\pi f_{t-1}$ as the target distribution to fit by MLE at round t . To finally estimate Z^π , we can compute $\mathbb{E}_{x \sim \mu, a \sim \pi(x)} \hat{f}_T(x, a)$.

To implement either algorithm, we need a function f that has the following two properties: (1) it can generate samples given any state-action pair, i.e., $z \sim f(\cdot|x, a)$, and (2) given any triple (x, a, z) we can evaluate the conditional likelihood, i.e., we can compute $f(z|x, a)$. Such function approximation is widely available in practice, including discrete histogram-based models, Gaussian mixture models, Flow models (Dinh et al., 2014), and diffusion model (Sohl-Dickstein et al., 2015). Indeed, in our experiment, we implement FLE with Gaussian mixture models and diffusion models (Ho et al., 2020), both of which are optimized via MLE.

Regarding computation, the main bottleneck is the MLE step (Line 13 and 10). While we present it with a $\arg \max$ oracle, in both practice and theory, an approximation optimization oracle is enough. In theory, as we will demonstrate, as long as we can find some \hat{f}_h that exhibits good in-distribution generalization bound (i.e., $\mathbb{E}_{x, a \sim \rho} d_{tv}(\hat{f}_h(x, a), [\mathcal{T}^\pi \hat{f}_{h+1}](x, a))$ or $\mathbb{E}_{x, a \sim \rho} d_{tv}(\hat{f}_t(x, a), [\mathcal{T}^\pi \hat{f}_{t-1}](x, a))$ is small), then we can guarantee to have an accurate estimator for Z^π . Note that here ρ is the training distribution for MLE, thus we care about in-distribution generalization. Thus our approach is truly a reduction to supervised learning: as long as the supervised learning procedure (in this case, MLE) learns a model with good in-distribution generalization performance, we

Algorithm 1 Fitted Likelihood Estimation (FLE) for finite-horizon MDPs

```

1: Input: dataset  $\{\mathcal{D}_h\}_{h=1}^H$  and function classes  $\{\mathcal{F}_h\}_{h=1}^H$ 
2: for  $h = H, H - 1, \dots, 1$  do
3:    $\mathcal{D}'_h = \emptyset$ 
4:   for  $x, a, r, x' \in \mathcal{D}_h$  do
5:     if  $h < H$  then
6:        $a' \sim \pi(x'), y \sim \hat{f}_{h+1}(\cdot|x', a')$ 
7:       Set  $z = r + y$ 
8:     else
9:       Set  $z = r$ 
10:    end if
11:     $\mathcal{D}'_h = \mathcal{D}'_h \cup \{(x, a, z)\}$ 
12:  end for
13:   $\hat{f}_h = \arg \max_{f \in \mathcal{F}_h} \sum_{(x, a, z) \in \mathcal{D}'_h} \log f(z|x, a)$ 
14: end for
    
```

Algorithm 2 Fitted Likelihood Estimation (FLE) for infinite-horizon MDPs

```

1: Input: dataset  $\{\mathcal{D}_t\}_{t=1}^T$  and function classes  $\mathcal{F}$ 
2: for  $t = 1, 2, \dots, T$  do
3:    $\mathcal{D}'_t = \emptyset$ 
4:   for  $x, a, r, x' \in \mathcal{D}_t$  do
5:      $a' \sim \pi(x')$ 
6:      $y \sim \hat{f}_{t-1}(\cdot|x', a')$ 
7:      $z = r + \gamma y$ 
8:      $\mathcal{D}'_t = \mathcal{D}'_t \cup \{(x, a, z)\}$ 
9:   end for
10:   $\hat{f}_t = \arg \max_{f \in \mathcal{F}} \sum_{(x, a, z) \in \mathcal{D}'_t} \log f(z|x, a)$ 
11: end for
    
```

can guarantee good prediction performance for FLE. Any advancements from training generative models via MLE (e.g., better training heuristics and better models) thus can immediately lead to improvement in distributional OPE.

Remark 3.1 (Comparison to prior models). The categorical algorithm (Bellemare et al., 2017) works by minimizing the cross-entropy loss between the (projected) target distribution and the parametric distribution, which is equivalent to maximizing the likelihood of the parametric model.

Remark 3.2 (FQE as a special instance). When reward is only a scalar, and we use fixed-variance Gaussian distribution $f(\cdot|x, a) := \mathcal{N}(g(x, a), \sigma^2)$ where $g : \mathcal{X} \times \mathcal{A} \mapsto [0, H]$, and $\sigma > 0$ is a fixed (not learnable) parameter, MLE becomes a least square oracle, and FLE reduces to FQE — the classic offline policy evaluation algorithm.

4. Theoretical Analysis

In this section, we present the theoretical guarantees of FLE. As a warm-up, we start by analyzing the performance of FLE for the finite-horizon setting (Section 4.1) where we

bound the prediction error using total variation distance. Then we study the guarantees for the infinite-horizon discounted scenario in Section 4.2 where the prediction error is measured under p -Wasserstein distance. Note that from Equation (1), TD distance dominates p -Wasserstein distance, which indicates that our guarantee for the finite horizon setting is stronger. This shows an interesting difference between the two settings. In addition, we present two concrete examples (tabular MDPs and linear quadratic regulators) in Appendix B. All proofs can be found in Appendix D.

4.1. Finite Horizon

We start by stating the key assumption for OPE, which concerns the overlap between π 's distribution and the offline distribution ρ .

Assumption 4.1 (Coverage). We assume there exists a constant C such that for all $h \in [H]$ the following holds

$$\sup_{\substack{f_h \in \mathcal{F}_h \\ f_{h+1} \in \mathcal{F}_{h+1}}} \frac{\mathbb{E}_{x,a \sim d_h^\pi} d_{tv}^2(f_h(x,a), [\mathcal{T}^\pi f_{h+1}](x,a))}{\mathbb{E}_{x,a \sim \rho} d_{tv}^2(f_h(x,a), [\mathcal{T}^\pi f_{h+1}](x,a))} \leq C.$$

The data coverage assumption is necessary for off-policy learning. Assumption 4.1 incorporates the function class into the definition of data coverage and is always no larger than the usual density ratio-based coverage definition, i.e., $\sup_{h,x,a} d_h^\pi(x,a)/\rho(x,a)$ which is a classic coverage measure in offline RL literature (e.g., Munos & Szepesvári (2008)). This type of refined coverage is used in the regular RL setting (Xie et al., 2021; Uehara & Sun, 2021).

Next, we present the theoretical guarantee of our approach under the *assumption that the MLE can achieve good supervised learning-style in-distribution generalization bound*. Recall that in each iteration of our algorithm, we perform MLE to learn a function \hat{f}_h to approximate the target $\mathcal{T}^\pi \hat{f}_{h+1}$ under the training data from ρ . By supervised learning style in-distribution generalization error, we mean the divergence d_{tv} between \hat{f}_h and the target $\mathcal{T}^\pi \hat{f}_{h+1}$ under the *training distribution* ρ . Such an in-distribution generalization bound for MLE is widely studied in statistical learning theory literature (Van de Geer, 2000; Zhang, 2006), and used in RL literature (e.g., Agarwal et al. (2020b); Uehara et al. (2021); Zhan et al. (2022)). The following theorem demonstrates a reduction framework: as long as supervised learning MLE works, our estimator of Z^π is accurate.

Theorem 4.2. *Under Assumption 4.1, suppose we have a sequence of functions $\hat{f}_1, \dots, \hat{f}_H : \mathcal{X} \times \mathcal{A} \mapsto \Delta([0, H]^d)$ and a sequence of values $\zeta_1, \dots, \zeta_H \in \mathbb{R}$ such that*

$$\left(\mathbb{E}_{x,a \sim \rho} d_{tv}^2(\hat{f}_h(x,a), [\mathcal{T}^\pi \hat{f}_{h+1}](x,a)) \right)^{1/2} \leq \zeta_h$$

holds for all $h \in [H]$. Let our estimator $\hat{f} :=$

$\mathbb{E}_{x \sim \mu, a \sim \pi(x)} \hat{f}_1(x,a)$. Then we have

$$d_{tv}(\hat{f}, Z^\pi) \leq \sqrt{C} \sum_{h=1}^H \zeta_h.$$

Here recall that C is the coverage definition. Thus the above theorem demonstrates that when ρ covers d^π (i.e., $C < \infty$), small supervised learning errors (i.e., ζ_h) imply small prediction error for distributional OPE.

Now to complete the picture, we provide some sufficient conditions where MLE can achieve small in-distribution generalization errors. The first condition is stated below.

Assumption 4.3 (Bellman completeness). We assume the following holds:

$$\max_{h \in [H], f \in \mathcal{F}_{h+1}} \min_{g \in \mathcal{F}_h} \mathbb{E}_{x,a \sim \rho} d_{tv}(g(x,a), [\mathcal{T}^\pi f](x,a)) = 0.$$

We call the LHS of the above inequality *inherent (distributional) Bellman error*.

This condition ensures that in each call of MLE in our algorithm, the function class \mathcal{F}_h contains the target $\mathcal{T}^\pi \hat{f}_{h+1}$. It is possible to relax this condition to a setting where the inherent Bellman error is bounded by a small number δ (i.e., for MLE, this corresponds to agnostic learning where the hypothesis class may not contain the target, which is also a well-studied problem in statistical learning theory (Van de Geer, 2000)). Here we mainly focus on the $\delta = 0$ case.

The Bellman completeness assumption (or, more generally, inherent Bellman error being small) is standard in offline RL literature (Munos & Szepesvári, 2008). Indeed, in the regular RL setting, when learning with off-policy data, without such a Bellman completeness condition, algorithms such as TD learning or value iteration-based approaches (e.g., FQE) can diverge (Tsitsiklis & Van Roy, 1996), and the TD fixed solution can be arbitrarily bad in terms of approximating the true value (e.g., Munos (2003); Scherrer (2010); Kolter (2011)). Since distributional RL generalizes regular RL, to prove convergence and provide an explicit sample complexity, we also need such a Bellman completeness condition.

The second condition is the bounded complexity of \mathcal{F}_h . A simple case is when \mathcal{F} is discrete where the standard statistical complexity of \mathcal{F} is $\ln(|\mathcal{F}_h|)$. We show the following result for MLE's in-distribution generalization error.

Lemma 4.4. *Assume $|\mathcal{F}_h| < \infty$. For FLE (Algorithm 1), under Assumption 4.3, MLEs have the following guarantee:*

$$\mathbb{E}_{x,a \sim \rho} d_{tv}^2(\hat{f}_h(x,a), [\mathcal{T}^\pi \hat{f}_{h+1}](x,a)) \leq \frac{4H}{n} \log(|\mathcal{F}_h|H/\delta)$$

for all $h \in [H]$ with probability at least $1 - \delta$.

For infinite hypothesis classes, we use bracketing number (Van de Geer, 2000) to quantify the statistical complexities.

Definition 4.5 (Bracketing number). Consider a function class \mathcal{F} that maps \mathcal{X} to \mathbb{R} . Given two functions l and u , the bracket $[l, u]$ is the set of all functions $f \in \mathcal{F}$ with $l(x) \leq f(x) \leq u(x)$ for all $x \in \mathcal{X}$. An ϵ -bracket is a bracket $[l, u]$ with $\|l - u\| \leq \epsilon$. The bracketing number of \mathcal{F} w.r.t. the metric $\|\cdot\|$ denoted by $N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|)$ is the minimum number of ϵ -brackets needed to cover \mathcal{F} .

We can bound MLE’s generalization error using the bracket number of \mathcal{F} .

Lemma 4.6. *For FLE (Algorithm 1), under Assumption 4.3, we have*

$$\begin{aligned} \mathbb{E}_{x,a \sim \rho} d_{tv}^2 \left(\hat{f}_h(x, a), [\mathcal{T}^\pi \hat{f}_{h+1}](x, a) \right) \\ \leq \frac{10H}{n} \log \left(N_{[]} \left((nH^d)^{-1}, \mathcal{F}_h, \|\cdot\|_\infty \right) H/\delta \right) \end{aligned}$$

for all $h \in [H]$ with probability at least $1 - \delta$.

It is noteworthy that the logarithm of the bracketing number is small in many common scenarios. We offer several examples in Section B. Previous studies have also extensively examined it (e.g., Van der Vaart (2000)).

With the generalization bounds of MLE, via Theorem 4.2, we can derive the following specific error bound for FLE.

Corollary 4.7. *Under Assumption 4.1 and 4.3, for FLE (Algorithm 1), with probability at least $1 - \delta$, we have*

$$d_{tv} \left(\hat{f}, Z^\pi \right) \leq \sqrt{C} \sum_{h=1}^H \sqrt{\frac{4H}{n} \log(|\mathcal{F}_h| H/\delta)}$$

when $|\mathcal{F}_h| < \infty$ for all $h \in [H]$, and

$$\begin{aligned} d_{tv} \left(\hat{f}, Z^\pi \right) \\ \leq \sqrt{C} \sum_{h=1}^H \sqrt{\frac{10H}{n} \log \left(N_{[]} \left((nH^d)^{-1}, \mathcal{F}_h, \|\cdot\|_\infty \right) H/\delta \right)}. \end{aligned}$$

for infinite function class \mathcal{F}_h .

Overall, our theory indicates that if we can train accurate distributions (e.g., generative models) via supervised learning (i.e., MLE here), we automatically have good predictive performance on estimating Z^π . This provides great flexibility for designing special algorithms.

Remark 4.8 (Offline CVaR Estimation). As a simple application, FLE can derive an estimator for the CVaR of the return under the test policy π . This is doable because CVaR is Lipschitz with respect to distributions in total variation distance, and thus our results can be directly transferred. See Appendix A for details. Essentially, any quantity that is Lipschitz with respect to distributions in total variation distance can be estimated using our method and the error bound of FLE directly applies.

4.2. Infinite Horizon

Next we introduce the theoretical guarantees of FLE for infinite horizon MDPs. Although the idea is similar, there is an obstacle: we can no longer obtain guarantees in terms of the total variation distance. This is perhaps not surprising considering that the distributional Bellman operator for discounted setting is *not* contractive in total variation distance (Bellemare et al., 2017). Fortunately, we found the Bellman operator is contractive under the Wasserstein distance measure. Note that the contractive result we established under Wasserstein distance is different from previous works (Bellemare et al., 2017; 2023; Zhang et al., 2021) in that these previous works consider the *supremum* Wasserstein distance: $\sup_{x,a} d_{w,p}$, while our contractive property is measured under an *average* Wasserstein distance: $(\mathbb{E}_{x,a \sim d^\pi} d_{w,p}^{2p})^{1/(2p)}$ which is critical to get a sample complexity bound for distributional OPE. More formally, the following lemma summarizes the contractive property.

Lemma 4.9. *The distributional Bellman operator is $\gamma^{1-1/(2p)}$ -contractive under the metric $(\mathbb{E}_{x,a \sim d^\pi} d_{w,p}^{2p})^{1/(2p)}$, i.e., for any $f, f' \in \mathcal{X} \times \mathcal{A} \mapsto [0, (1 - \gamma)^{-1}]^d$, it holds that*

$$\begin{aligned} \left(\mathbb{E}_{x,a \sim d^\pi} d_{w,p}^{2p} \left([\mathcal{T}^\pi f](x, a), [\mathcal{T}^\pi f'](x, a) \right) \right)^{\frac{1}{2p}} \\ \leq \gamma^{1-\frac{1}{2p}} \cdot \left(\mathbb{E}_{x,a \sim d^\pi} d_{w,p}^{2p} \left(f(x, a), f'(x, a) \right) \right)^{\frac{1}{2p}}. \end{aligned}$$

We note that the contractive result in $\sup_{x,a} d_{w,p}$ does not imply the result in the above lemma, thus not directly applicable to the OPE setting.

Due to the dominance of total variation distance over Wasserstein distance on bounded sets (see (1)), MLE’s estimation error under total variation distance can be converted to Wasserstein distance. This allows us to derive theoretical guarantees for FLE under Wasserstein distance. To that end, we start again with the coverage assumption that is similar to Assumption 4.1. Note that we have replaced the total variation distance with the Wasserstein distance.

Assumption 4.10 (Coverage). We assume there exists a constant C such that the following holds

$$\sup_{f, f' \in \mathcal{F}} \frac{\mathbb{E}_{x,a \sim d^\pi} d_{w,p}^{2p} \left(f(x, a), [\mathcal{T}^\pi f'](x, a) \right)}{\mathbb{E}_{x,a \sim \rho} d_{w,p}^{2p} \left(f(x, a), [\mathcal{T}^\pi f'](x, a) \right)} \leq C.$$

As similar to Theorem 4.2, the following theorem states that as long as the supervised learning is accurate, our estimator of Z^π will be accurate as well under p -Wasserstein distance.

Theorem 4.11. *Under Assumption 4.10, suppose we have a sequence of functions $\hat{f}_1, \dots, \hat{f}_T : \mathcal{X} \times \mathcal{A} \mapsto \Delta([0, (1 -$*

$\gamma)^{-1}]^d)$ and an upper bound $\zeta \in \mathbb{R}$ such that

$$\left(\mathbb{E}_{x,a \sim \rho} d_{w,p}^{2p} \left(\hat{f}_t(x,a), [\mathcal{T}^\pi \hat{f}_{t-1}](x,a) \right) \right)^{\frac{1}{2p}} \leq \zeta$$

holds for all $t \in [T]$. Let our estimator $\hat{f} := \mathbb{E}_{x \sim \mu, a \sim \pi(x)} \hat{f}_T(x,a)$. Then we have, for all $p \geq 1$,

$$d_{w,p}(\hat{f}, Z^\pi) \leq \frac{2C^{\frac{1}{2p}}}{(1-\gamma)^{\frac{3}{2}}} \cdot \zeta + \frac{\sqrt{d} \cdot \gamma^{\frac{T}{2}}}{(1-\gamma)^{\frac{3}{2}}}. \quad (2)$$

The upper bound in (2) is actually a simplified version as we aim to present a cleaner result. For a more refined upper bound that has detailed p -dependent terms, please refer to Theorem D.2 in the appendix. For the first additive term in (2), we will later demonstrate that the ζ obtained from MLE depends on p^{-1} at an exponential rate. The second term is insignificant as it converges to zero at the rate of $\gamma^{T/2}$.

To proceed, we introduce the Bellman completeness assumption for infinite-horizon MDPs, a key condition for MLE to achieve small in-distribution generalization errors.

Assumption 4.12 (Bellman completeness). We assume the following holds:

$$\max_{f \in \mathcal{F}} \min_{g \in \mathcal{F}} \mathbb{E}_{x,a \sim \rho} d_{w,p}(g(x,a), [\mathcal{T}^\pi f](x,a)) = 0.$$

Similar to the previous result, when Bellman completeness holds and the function class has bounded complexity, MLE achieves small generalization error, as the following shows.

Lemma 4.13. For FLE (Algorithm 2), under Assumption 4.12, by applying MLEs we have, for all $t \in [T]$,

$$\begin{aligned} \mathbb{E}_{x,a \sim \rho} d_{w,p}^{2p} \left(\hat{f}_t(x,a), [\mathcal{T}^\pi \hat{f}_{t-1}](x,a) \right) \\ \leq \left(\frac{\sqrt{d}}{1-\gamma} \right)^{2p} \frac{4T}{n} \log(|\mathcal{F}|T/\delta) \end{aligned}$$

when $|\mathcal{F}| < \infty$, and

$$\begin{aligned} \mathbb{E}_{x,a \sim \rho} d_{w,p}^{2p} \left(\hat{f}_t(x,a), [\mathcal{T}^\pi \hat{f}_{t-1}](x,a) \right) \\ \leq \left(\frac{\sqrt{d}}{1-\gamma} \right)^{2p} \frac{10T}{n} \log \left(N_{[]} \left(\frac{(1-\gamma)^d}{n}, \mathcal{F}, \|\cdot\|_\infty \right) T/\delta \right) \end{aligned}$$

when $|\mathcal{F}| = \infty$, with probability at least $1 - \delta$.

The multiplicative term T in the upper bounds above comes from the data splitting (recall that we have split the dataset \mathcal{D} into T subsets: $\mathcal{D}_1, \dots, \mathcal{D}_T$). A more careful analysis may be able to get rid of it, leading to a slightly better polynomial dependence on the effective horizon $1/(1-\gamma)$ in the final sample complexity bound. We leave this for future work.

In view of the above result, to derive the specific error bound of FLE, we need to choose an appropriate T to make a good balance. The T we choose is of the logarithmic order. It is shown in the corollary below.

Corollary 4.14. We define

$$\iota = \begin{cases} \log(|\mathcal{F}|/\delta), & \text{if } |\mathcal{F}| < \infty; \\ \log \left(N_{[]} \left(\frac{(1-\gamma)^d}{n}, \mathcal{F}, \|\cdot\|_\infty \right) / \delta \right), & \text{if } |\mathcal{F}| = \infty. \end{cases}$$

Then under Assumption 4.10 and 4.12, for FLE (Algorithm 2), if we pick

$$T = \log \left(C^{\frac{1}{2p}} \cdot \iota^{\frac{1}{2p}} \cdot \left(1 - \gamma^{\frac{1}{2}}\right)^{-1} \cdot n^{-\frac{1}{2p}} \right) / \log \left(\gamma^{1 - \frac{1}{2p}} \right)$$

then with probability at least $1 - \delta$, we have

$$d_{w,p}(\hat{f}, Z^\pi) \leq \tilde{O} \left(\frac{C^{\frac{1}{2p}} \cdot \iota^{\frac{1}{2p}} \cdot \sqrt{d}}{(1-\gamma)^{\frac{5}{2}}} \cdot n^{-\frac{1}{2p}} \right)$$

where $\hat{f} := \mathbb{E}_{x \sim \mu, a \sim \pi(x)} \hat{f}_T(x,a)$.

The above upper bound depends on $n^{-1/(2p)}$, which seems unsatisfactory, especially when p is large. However, we believe that it is actually tight since the previous study has shown that the minimax rate of estimating $d_{w,p}$ using i.i.d samples from the given distribution is around $O(n^{-1/(2p)})$ (Singh & Póczos, 2018). More formally, given a distribution Q and n i.i.d samples from Q , any algorithm that maps the n i.i.d samples to a distribution \hat{Q} , must have $d_{w,p}(\hat{Q}, Q) = \tilde{\Omega}(n^{-1/(2p)})$ in the worst case. Note that distributional OPE is strictly harder than this problem.

5. Simulation

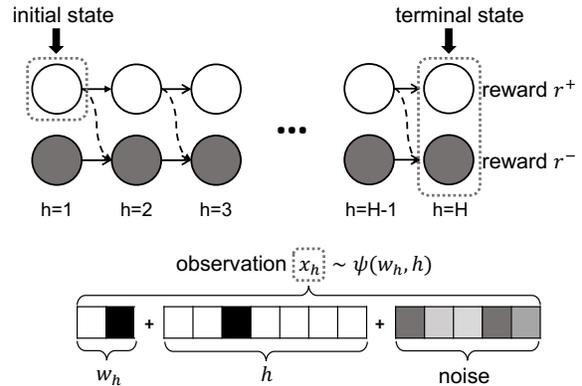


Figure 1. Visualization of the combination lock. The dotted lines denote transiting from good states (white) to bad states (gray). Once the agent transits to a bad state, it stays there forever. The observation is composed of three parts: one-hot encoding of the latent state w_h , one-hot encoding of the step h , and random noise.

In this section, we show the empirical performance of two instances of FLE: *GMM-FLE* and *Diff-FLE*. The GMM-FLE uses conditional Gaussian mixture models for \mathcal{F} , for which the weights and the mean and covariance of Gaussians are all learnable. For Diff-FLE, we model the distribution $f(\cdot | x, a)$ as a conditional diffusion probabilistic model (Sohl-Dickstein et al., 2015). The implementation is based on DDPM (Ho et al., 2020). We elaborate on other components of the experiments below. See Appendix E for implementation details and a full list of results.

The combination lock environment. The combination lock consists of two chains. One of the chains is good, while the other is bad. The agent wants to stay on the good chain, for which the only approach is to take the unique optimal action at all time steps. See Figure 1 for an illustration. Mathematically, the combination lock is a finite-horizon MDP of horizon H . There are two latent states $w_h \in \{0, 1\}$. At any time step $h \in [H]$, there is only one optimal action a_h^* among A actions. If the agent is in the latent state $w_h = 0$ and takes a_h^* , it transits to $w_{h+1} = 0$, and otherwise transits to $w_{h+1} = 1$. If it is already in $w_h = 1$, no matter what action it takes, it transits to $w_{h+1} = 1$. When $h = H$, it receives a random reward r^+ if $w_H = 0$; otherwise, it gets r^- . The agent cannot observe the latent state w_h directly. Instead, the observation it receives, $\psi(w_h, h)$, is the concatenation of one-hot coding of the latent state w_h and the current time step h , appended with Gaussian noise. This environment has been used in prior works (Misra et al., 2020; Zhang et al., 2022b) where it was shown that standard deep RL methods struggle due to the challenges from exploration and high-dimensional observation.

Test policy. The test policy is stochastic: it takes a random action with probability ϵ and takes the optimal policy otherwise. In all experiments, we set $\epsilon = 1/7$.

Offline data generation. The offline dataset is generated uniformly. Specifically, for each time step $h \in [H]$ and each latent state $w_h \in \{0, 1\}$, we first randomly sample 10000 observable state $\phi(w_h)$. Then for each of them, we uniformly randomly sample action and perform one step simulation. It is clear that the offline data distribution here satisfies the coverage assumption (Assumption 4.1).

5.1. One-Dimensional Reward

To compare to classic methods such as the categorical algorithm (Bellemare et al., 2017) and quantile TD (Dabney et al., 2018), we first run experiments with a 1-d reward. Specifically, we have $r^+ \sim \mathcal{N}(1, 0.1^2)$ and $r^- \sim \mathcal{N}(-1, 0.1^2)$. The horizon is $H = 20$.

The categorical algorithm discretizes the range $[-1.5, 1.5]$ using 100 atoms. For quantile TD, we set the number of quantiles to 100 as well. The GMM-FLE uses 10 atomic

Gaussian distributions, although eventually, only two are significant. See Append E for a detailed description of implementations. We plot the PDFs $\mathbb{E}_{x \sim \psi(0, h)} \hat{f}_h(x, a_h^*)$ (here 0 denotes the good latent state in h) learned by different methods in Figure 2, at three different time steps. As we can see, GMM-FLE in general fits the ground truth the best.

We also compute the approximated d_{tv} between the learned distribution and the true one. Ideally, we want to compute $d_{tv}(\mathbb{E}_{x \sim \psi(0, h)} \hat{f}_h(x, a_h^*), \mathbb{E}_{x \sim \psi(0, h)} Z_h^\pi(x, a_h^*))$. However, since obtaining the density of certain models is impossible (e.g., Diff-FLE) and certain other models have only discrete supports, we use an approximated version: we sample $20k$ points from each distribution, construct two histograms, and calculate d_{tv} between the two histograms. The results are shown in Table 1. Again, GMM-FLE achieves the smallest total variation distance. This intuitively makes sense since the ground truth return is a mixture of Gaussians. Moreover, we notice that GMM-FLE, Diff-FLE, and categorical algorithms achieve significantly better performance than the quantile regression TD algorithm. This perhaps is not surprising because our theory has provided performance guarantees for those three algorithms under d_{tv} (recall that the categorical algorithm can be roughly considered a specification of FLE, see Remark 3.1), while it is unclear if quantile regression TD can achieve similar guarantees in this setting.

| h | Cate Alg | Quan Alg | Diff-FLE | GMM-FLE |
|-----|-------------------|-------------------|-------------------|-------------------------------------|
| 1 | 0.071 \pm 0.015 | 0.603 \pm 0.011 | 0.292 \pm 0.073 | 0.039 \pm 0.004 |
| 10 | 0.079 \pm 0.017 | 0.494 \pm 0.018 | 0.234 \pm 0.043 | 0.044 \pm 0.012 |
| 19 | 0.078 \pm 0.011 | 0.167 \pm 0.019 | 0.109 \pm 0.031 | 0.018 \pm 0.008 |

Table 1. Approximated d_{tv} between $\mathbb{E}_{x \sim \psi(0, h)} \hat{f}_h(x, a_h^*)$ and $\mathbb{E}_{x \sim \psi(0, h)} Z_h^\pi(x, a_h^*)$ in the 1-d case. The means and standard errors are computed via five independent runs.

5.2. Two-Dimensional Reward

We also conducted experiments on two-dimensional rewards where r^+ is sampled from a ring in \mathbb{R}^2 of radius 2 and r^- follows a Gaussian centered at the origin. The horizon is $H = 10$. The categorical algorithm discretizes the range $[-4, 4]^2$ into 30 atoms per dimension (totaling 900 atoms). Although the 2-d version of the categorical algorithm is not introduced in the original paper (Bellemare et al., 2017), the extension is intuitive. The GMM-FLE employs 30 atomic Gaussian distributions, but only up to six prove significant in the end. We note that extending quantile regression TD to multi-dimensional rewards is not straightforward.

We plotted the 2-d visualization of the learned distribution in Figure 3 and computed the approximated TV distance using the same method as in the 1-d case, which is shown in Table 2. Diff-FLE achieves the smallest TV error (Table 2) and captures the correlation among dimensions (i.e., see Figure 3 where Diff-FLE captures the ring structures in

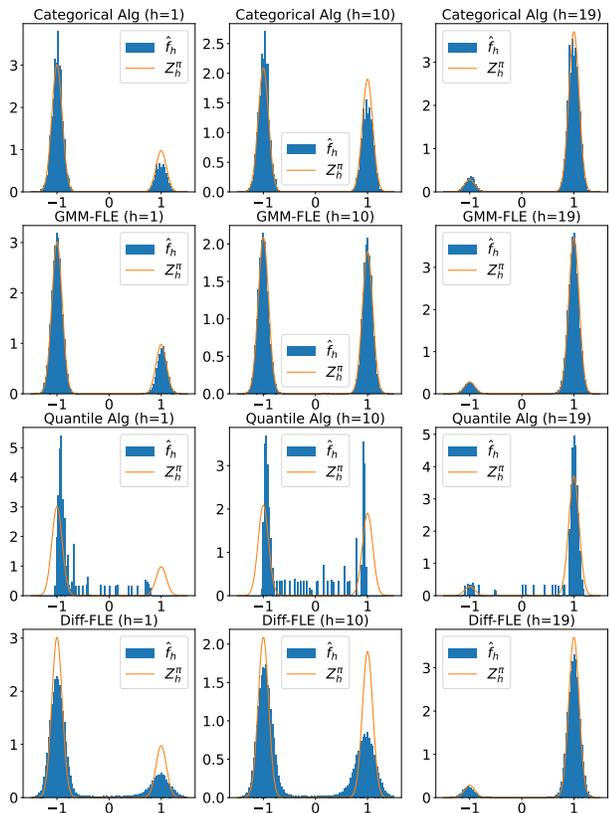


Figure 2. Plots of $\mathbb{E}_{x \sim \psi(0, h)} \hat{f}_h(x, a_h^*)$ and $\mathbb{E}_{x \sim \psi(0, h)} Z_h^\pi(x, a_h^*)$. The histograms are generated via 50k samples.

all steps). However, the GMM-FLE also doesn't perform well since it is hard for vanilla GMM with a finite number of mixtures to capture a ring-like data distribution. The two-dimensional categorical algorithm performed badly as well, even though it uses a larger number of atoms (recall that for the 1-d case it only uses 100 atoms and already achieves excellent performance), implying that it suffers from the curse of dimensionality statistically, i.e., explicitly discretizing the 2-d return space evenly can fail to capture the underlying data structure (e.g., in our ring example, data actually approximately lives in a sub-manifold). Moreover, the training is also significantly slower. In our implementation, we found that running the 2-d categorical algorithm with 100^2 atoms is about 100 times slower than running the 1-d algorithm with 100 atoms, while the training time of

| h | CATE ALG | DIFF-FLE | GMM-FLE |
|-----|-------------------|-------------------------------------|-------------------|
| 1 | 0.483 ± 0.003 | 0.357 ± 0.031 | 0.438 ± 0.008 |
| 5 | 0.466 ± 0.001 | 0.310 ± 0.019 | 0.493 ± 0.050 |
| 9 | 0.453 ± 0.001 | 0.207 ± 0.014 | 0.502 ± 0.094 |

Table 2. Approximated d_{tv} between $\mathbb{E}_{x \sim \psi(0, h)} \hat{f}_h(x, a_h^*)$ and $\mathbb{E}_{x \sim \psi(0, h)} Z_h^\pi(x, a_h^*)$ in the 2-d case. The means and standard errors are computed via five independent runs.

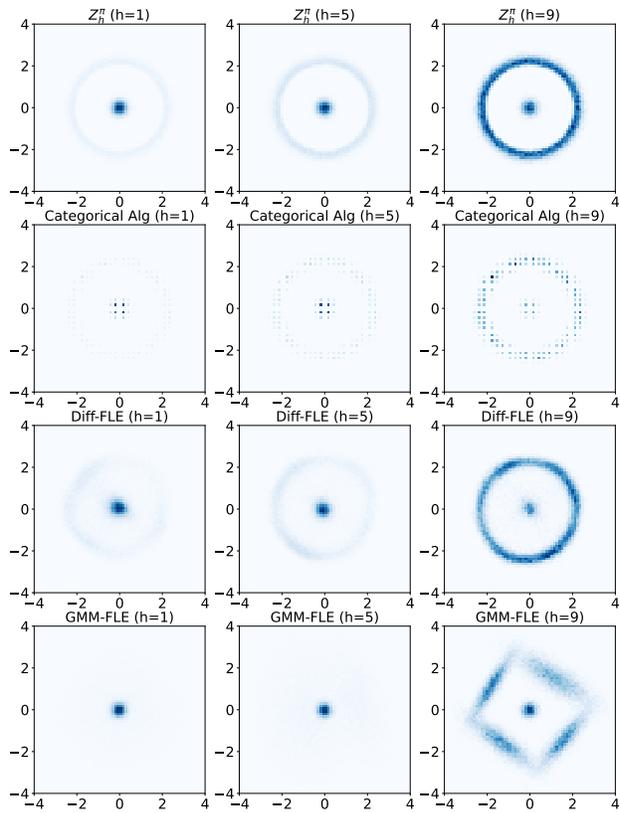


Figure 3. Plots of $\mathbb{E}_{x \sim \psi(0, h)} \hat{f}_h(x, a_h^*)$ (generated via 50k samples), and the ground truth $\mathbb{E}_{x \sim \psi(0, h)} Z_h^\pi(x, a_h^*)$ (top row).

Diff-FLE and GMM-FLE does not change too much.

6. Discussion and Future Work

We proposed Fitted Likelihood Estimation (FLE), a simple algorithm for distributional OPE with multi-dimensional rewards. FLE conducts a sequence of MLEs and can incorporate any state-of-the-art generative models trained via MLE. Thus, FLE is scalable to the setting where reward vectors are high-dimensional. Theoretically, we showed that the learned distribution is accurate under total variation distance and p -Wasserstein distance for the finite-horizon and infinite-horizon discounted setting, respectively. In practice, we demonstrated its flexibility in utilizing generative models such as GMMs and diffusion models.

Our work may offer several promising avenues for future research in distributional RL. One immediate direction is to adapt our algorithms to the policy optimization. Another direction is the development of more efficient algorithms that can work in more complex environments.

Acknowledgement

WS acknowledges funding support from NSF IIS-2154711.

References

- Agarwal, A., Henaff, M., Kakade, S., and Sun, W. Pcp: Policy cover directed exploration for provable policy gradient learning. *Advances in neural information processing systems*, 33:13399–13412, 2020a.
- Agarwal, A., Kakade, S., Krishnamurthy, A., and Sun, W. Flambe: Structural complexity and representation learning of low rank mdps. *Advances in neural information processing systems*, 33:20095–20107, 2020b.
- Bellemare, M. G., Dabney, W., and Munos, R. A distributional perspective on reinforcement learning. In *International Conference on Machine Learning*, pp. 449–458. PMLR, 2017.
- Bellemare, M. G., Dabney, W., and Rowland, M. *Distributional Reinforcement Learning*. MIT Press, 2023. <http://www.distributional-rl.org>.
- Chandak, Y., Niekum, S., da Silva, B., Learned-Miller, E., Brunskill, E., and Thomas, P. S. Universal off-policy evaluation. *Advances in Neural Information Processing Systems*, 34:27475–27490, 2021.
- Chang, J., Wang, K., Kallus, N., and Sun, W. Learning bellman complete representations for offline policy evaluation. In *International Conference on Machine Learning*, pp. 2938–2971. PMLR, 2022.
- Dabney, W., Rowland, M., Bellemare, M., and Munos, R. Distributional reinforcement learning with quantile regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.
- Dinh, L., Krueger, D., and Bengio, Y. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.
- Doan, T., Mazouze, B., and Lyle, C. Gan q-learning. *arXiv preprint arXiv:1805.04874*, 2018.
- Ernst, D., Geurts, P., and Wehenkel, L. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6, 2005.
- Feng, Y., Li, L., and Liu, Q. A kernel loss for solving the bellman equation. *Advances in Neural Information Processing Systems*, 32, 2019.
- Freirich, D., Shimkin, T., Meir, R., and Tamar, A. Distributional multivariate policy evaluation and exploration with the bellman gan. In *International Conference on Machine Learning*, pp. 1983–1992. PMLR, 2019.
- Fu, J., Norouzi, M., Nachum, O., Tucker, G., Wang, Z., Novikov, A., Yang, M., Zhang, M. R., Chen, Y., Kumar, A., et al. Benchmarks for deep off-policy evaluation. *arXiv preprint arXiv:2103.16596*, 2021.
- Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.
- Huang, A., Leqi, L., Lipton, Z., and Azizzadenesheli, K. Off-policy risk assessment in contextual bandits. *Advances in Neural Information Processing Systems*, 34:23714–23726, 2021.
- Huang, A., Leqi, L., Lipton, Z., and Azizzadenesheli, K. Off-policy risk assessment for markov decision processes. In *International Conference on Artificial Intelligence and Statistics*, pp. 5022–5050. PMLR, 2022.
- Jiang, N. and Li, L. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 652–661. PMLR, 2016.
- Keramati, R., Dann, C., Tamkin, A., and Brunskill, E. Being optimistic to be conservative: Quickly learning a cvar policy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 4436–4443, 2020.
- Kolter, J. The fixed points of off-policy td. *Advances in Neural Information Processing Systems*, 24, 2011.
- Levin, D. A. and Peres, Y. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.
- Li, L. and Faisal, A. A. Bayesian distributional policy gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8429–8437, 2021.
- Ma, Y., Jayaraman, D., and Bastani, O. Conservative offline distributional reinforcement learning. *Advances in Neural Information Processing Systems*, 34:19235–19247, 2021.
- Misra, D., Henaff, M., Krishnamurthy, A., and Langford, J. Kinematic state abstraction and provably efficient rich-observation reinforcement learning. In *International conference on machine learning*, pp. 6961–6971. PMLR, 2020.
- Morimura, T., Sugiyama, M., Kashima, H., Hachiya, H., and Tanaka, T. Parametric return density estimation for reinforcement learning. *arXiv preprint arXiv:1203.3497*, 2012.
- Munos, R. Error bounds for approximate policy iteration. In *ICML*, volume 3, pp. 560–567. Citeseer, 2003.
- Munos, R. and Szepesvári, C. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9 (5), 2008.
- Prashanth, L. and Bhat, S. P. A wasserstein distance approach for concentration of empirical risk estimates. *The Journal of Machine Learning Research*, 23(1):10830–10890, 2022.

- Precup, D., Sutton, R. S., and Singh, S. P. Eligibility traces for off-policy policy evaluation. In *ICML*, 2000.
- Rowland, M., Munos, R., Azar, M. G., Tang, Y., Ostrovski, G., Harutyunyan, A., Tuyls, K., Bellemare, M. G., and Dabney, W. An analysis of quantile temporal-difference learning. *arXiv preprint arXiv:2301.04462*, 2023.
- Scherrer, B. Should one compute the temporal difference fix point or minimize the bellman residual? the unified oblique projection view. *arXiv preprint arXiv:1011.4362*, 2010.
- Singh, S. and Póczos, B. Minimax distribution estimation in wasserstein distance. *arXiv preprint arXiv:1802.08855*, 2018.
- Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015.
- Tsitsiklis, J. and Van Roy, B. Analysis of temporal-difference learning with function approximation. *Advances in neural information processing systems*, 9, 1996.
- Uehara, M. and Sun, W. Pessimistic model-based offline reinforcement learning under partial coverage. *arXiv preprint arXiv:2107.06226*, 2021.
- Uehara, M., Huang, J., and Jiang, N. Minimax weight and q-function learning for off-policy evaluation. In *International Conference on Machine Learning*, pp. 9659–9668. PMLR, 2020.
- Uehara, M., Zhang, X., and Sun, W. Representation learning for online and offline rl in low-rank mdps. In *International Conference on Learning Representations*, 2021.
- Van de Geer, S. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.
- Van der Vaart, A. W. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.
- Villani, C. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2021.
- Villani, C. et al. *Optimal transport: old and new*, volume 338. Springer, 2009.
- Xie, T., Cheng, C.-A., Jiang, N., Mineiro, P., and Agarwal, A. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34:6683–6694, 2021.
- Yang, M., Nachum, O., Dai, B., Li, L., and Schuurmans, D. Off-policy evaluation via the regularized lagrangian. *Advances in Neural Information Processing Systems*, 33: 6551–6561, 2020.
- Zhan, W., Uehara, M., Sun, W., and Lee, J. D. Pac reinforcement learning for predictive state representations. *arXiv preprint arXiv:2207.05738*, 2022.
- Zhang, P., Chen, X., Zhao, L., Xiong, W., Qin, T., and Liu, T.-Y. Distributional reinforcement learning for multi-dimensional reward functions. *Advances in Neural Information Processing Systems*, 34:1519–1529, 2021.
- Zhang, Q., Makur, A., and Azizzadenesheli, K. Functional linear regression of cdfs. *arXiv preprint arXiv:2205.14545*, 2022a.
- Zhang, T. From ϵ -entropy to kl-entropy: Analysis of minimum information complexity density estimation. *The Annals of Statistics*, 34(5):2180–2210, 2006.
- Zhang, X., Song, Y., Uehara, M., Wang, M., Agarwal, A., and Sun, W. Efficient reinforcement learning in block mdps: A model-free representation learning approach. In *International Conference on Machine Learning*, pp. 26517–26547. PMLR, 2022b.

A. Offline CVaR Evaluation

We consider estimating the CVaR of Z^π with $d = 1$. Given a threshold $\tau \in (0, 1)$, the CVaR_τ of Z^π is defined as (assuming finite-horizon MDPs):

$$\text{CVaR}_\tau(Z^\pi) := \max_{b \in [0, H]} \left(b - \frac{1}{\tau} \mathbb{E}_{z \sim Z^\pi} \max\{b - z, 0\} \right).$$

CVaR intuitively measures the expected value of the random variable belonging to the tail part of the distribution and is often used as a risk-sensitive measure. The following lemma shows that $\text{CVaR}_\tau(Z^\pi)$ is Lipschitz continuous with respect to metric d_{tv} and the Lipschitz constant is $2H/\tau$.

Lemma A.1. *Let $f, f' \in \Delta([0, H])$ be two densities. Then we have*

$$\text{CVaR}_\tau(f) - \text{CVaR}_\tau(f') \leq \frac{2H}{\tau} \cdot d_{tv}(f, f').$$

Proof. Let $f, f' \in \Delta([0, H])$ denote two densities. Then we have

$$\begin{aligned} & \text{CVaR}_\tau(f) - \text{CVaR}_\tau(f') \\ &= \max_{b \in [0, H]} \left(b - \frac{1}{\tau} \mathbb{E}_{z \sim f} \max\{b - z, 0\} \right) - \max_{b \in [0, H]} \left(b - \frac{1}{\tau} \mathbb{E}_{z \sim f'} \max\{b - z, 0\} \right) \\ &\leq \left(b_0 - \frac{1}{\tau} \mathbb{E}_{z \sim f} \max\{b_0 - z, 0\} \right) - \left(b_0 - \frac{1}{\tau} \mathbb{E}_{z \sim f'} \max\{b_0 - z, 0\} \right) \\ &= \frac{1}{\tau} \left(\mathbb{E}_{z \sim f'} \max\{b_0 - z, 0\} - \mathbb{E}_{z \sim f} \max\{b_0 - z, 0\} \right) \\ &= \frac{1}{\tau} \int_{[0, H]} (f'(z) - f(z)) \max\{b_0 - z, 0\} dz \\ &\leq \frac{H}{\tau} \int_{[0, H]} |f'(z) - f(z)| dz \\ &\leq \frac{2H}{\tau} d_{tv}(f, f') \end{aligned}$$

where the first inequality holds by picking $b_0 = \arg \max_{b \in [0, H]} \left(b - \frac{1}{\tau} \mathbb{E}_{z \sim f} \max\{b - z, 0\} \right)$. □

Thus using our bound from Corollary 4.7, we get:

$$\left| \text{CVaR}_\tau(Z^\pi) - \text{CVaR}_\tau(\hat{f}) \right| \leq \frac{4C^{1/2}H^{2.5}}{\tau} \sqrt{\frac{\log(\max_h |\mathcal{F}|_h / \delta)}{n}},$$

with probability at least $1 - \delta$.

B. Examples

In this section, we discuss two examples: one is tabular MDPs, and the other is Linear Quadratic Regulators. For simplicity of presentation, we focus on scalar rewards and finite horizon.

B.1. Tabular MDPs

We consider tabular MDP (i.e., $|\mathcal{X}|$ and $|\mathcal{A}|$ are finite) with continuous known reward distributions. Specifically, we consider the sparse reward case where we only have a reward at the last time step H and have zero rewards at time step $h < H$. For each (x, a) , Denote $r_H(x, a) \in \Delta([0, 1])$.

Note that in this setup, via induction, it is easy to verify that for any h, x, a , $Z_h^\pi(\cdot | x, a)$ is a mixture of the distributions $\{r_H(x, a) : x \in \mathcal{X}, a \in \mathcal{A}\}$, i.e., for any h, x, a , there exists a probability weight vector $w \in \Delta(|\mathcal{X}||\mathcal{A}|)$, such that

$Z_h^\pi(\cdot|x, a) = \sum_{x', a' \in \mathcal{X} \times \mathcal{A}} w(x', a') r_H(\cdot|x', a')$. Note that the parameters $w(x, a)$ are unknown due to the unknown transition operator P , and need to be learned. Thus, in this case, we can design function class \mathcal{F}_h as follows:

$$\mathcal{F}_h = \left\{ f(\cdot|x, a) = \sum_{x', a' \in \mathcal{X} \times \mathcal{A}} w_{x,a}(x', a') r_H(\cdot|x', a') : \right. \\ \left. \{w_{x,a} \in \Delta(|\mathcal{X}||\mathcal{A}|)\}_{x,a \in \mathcal{X} \times \mathcal{A}} \right\}.$$

It is not hard to verify that $\{\mathcal{F}_h\}_{h=1}^H$ does satisfy the Bellman complete condition. The log of the bracket number of \mathcal{F}_h is polynomial with respect to $|\mathcal{X}||\mathcal{A}|$.

Lemma B.1. *In the above example, the complexity of \mathcal{F}_h is bounded: $\log N_{[]}(\epsilon, \mathcal{F}_h, \|\cdot\|_\infty) \leq O(|\mathcal{X}|^2|\mathcal{A}|^2 \log(r_\infty|\mathcal{X}||\mathcal{A}|/\epsilon))$ where $r_\infty := \|r_H\|_\infty$.*

Thus Algorithm 1 is capable of finding an accurate estimator of Z^π with sample complexity scaling polynomially with respect to the size of the state and action spaces and horizon.

B.2. Linear Quadratic Regulator

The second example is LQR. We have $\mathcal{X} \subset \mathbb{R}^{d_x}, \mathcal{A} \subset \mathbb{R}^{d_a}$.

$$x_{h+1} = Ax_h + Ba_h, \\ r(x_h, a_h) = -(x_h^\top Qx_h + a_h^\top Ra_h) + \varepsilon$$

where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. Since the optimal policy for LQR is a linear policy, we consider evaluating a linear policy $\pi(x) := Kx$ where $K \in \mathbb{R}^{d_a \times d_x}$. For this linear policy, $Z_h^\pi(\cdot|x, a)$ is a Gaussian distribution, i.e., $Z_h^\pi(\cdot|x, a) = \mathcal{N}(\mu_h(x, a), \sigma_h(x, a))$, where $\mu_h(x, a)$ and $\sigma_h(x, a)$ has closed form solutions.

Lemma B.2. *For LQR defined above, $\mu_h(x, a)$ and $\sigma_h(x, a)$ has the following closed form solutions*

$$\mu_h(x, a) = -(Ax + Ba)^\top U_{h+1}(Ax + Ba) \\ - x^\top Qx - a^\top Ra, \\ \sigma_h^2(x, a) = (H - h + 1)\sigma^2$$

where we denote $U_h = \sum_{i=h}^H ((A + BK)^{i-h-1})^\top (Q + K^\top RK)(A + BK)^{i-h-1}$.

Thus our function class \mathcal{F}_h can be designed as follows:

$$\mathcal{F}_h = \left\{ f(\cdot|x, a) = \mathcal{N}(\cdot | x^\top M_1x + a^\top M_2x + a^\top M_3a, \right. \\ \left. (H - h + 1)\sigma^2), \forall M_1, M_2, M_3 \right\}$$

We can show that this function class satisfies Bellman completeness. Furthermore, here, we can refine C in Assumption 4.1 to a relative condition number following the derivation in Uehara & Sun (2021). More specifically, C is $\sup_{w \neq 0, h} \frac{w^\top \mathbb{E}_{a^\pi} [\phi(x, a) \phi^\top(x, a)] w}{w^\top \mathbb{E}_\rho [\phi(x, a) \phi^\top(x, a)] w}$ where $\phi(x, a) = (x^\top, a^\top)^\top \otimes (x^\top, a^\top)^\top$ is a quadratic feature and \otimes is the Kronecker product. Under some regularity assumption (i.e., the norms of M_1, M_2, M_3 are bounded, which is the case when the dynamical system induced by the linear policy is stable), this function class has bounded statistical complexity.

Lemma B.3. *We assume there exist parameters m_x, m_a, m_1, m_2, m_3 for which $\|x\|_2 \leq m_x$ for all $x \in \mathcal{X}$ and $\|a\|_2 \leq m_a$ for all $a \in \mathcal{A}$, and $\|M_i\|_F \leq m_i$ for $i = 1, 2, 3$. Then we have*

$$\log N_{[]}(\epsilon, \mathcal{F}_h, \|\cdot\|_\infty) \leq \text{Poly} \left(d_x, d_a, \log \frac{m_x m_a m_1 m_2 m_3}{\epsilon \sigma} \right).$$

It is unclear if quantile regression TD or categorical TD can achieve meaningful guarantees on LQR since the function classes used by them do not satisfy Bellman completeness (i.e., given any conditional density $f(\cdot|x, a)$, $\mathcal{T}^\pi f$ will not be discrete since here $r(x, a)$ is continuous). Even for regular RL, without Bellman completeness, in the off-policy setting, it is possible that TD-based algorithms may diverge, and TD fixed point solutions can be arbitrarily bad.

C. Supporting Lemmas

C.1. Maximum Likelihood Estimation

In this section, we adapt the theoretical results of MLE (Agarwal et al., 2020b) to more general versions. We will follow the notation in Appendix E of Agarwal et al. (2020b) and restate the setting here for completeness.

We consider a sequential conditional probability estimation problem. Let \mathcal{X} and \mathcal{Y} denote the instance space and the target space, respectively. We are given a function class $\mathcal{F} : (\mathcal{X} \times \mathcal{Y}) \rightarrow \mathbb{R}$ with which we want to model the true conditional distribution f^* . To this end, we are given a dataset $D := \{(x_i, y_i)\}_{i=1}^n$, where $x_i \sim \mathcal{D}_i$ and $y_i \sim p(\cdot | x_i) = f^*(x, \cdot)$.

We only assume that there exists f_i^* for each $i \in [n]$ such that $\mathbb{E}_{x \sim \mathcal{D}_i} d_{tv}(f_i^*(x), f^*(x)) = 0$. Note that this assumption only considers x on the support of \mathcal{D}_i and is thus weaker than saying $f^* \in \mathcal{F}$.

For the data generating process, we assume the data distribution \mathcal{D}_i is history-dependent, i.e., it can depend on the previous samples: $x_1, y_1, \dots, x_{i-1}, y_{i-1}$.

Let $\mathcal{D}' = \{(x'_i, y'_i)\}_{i=1}^n$ denote the tangent sequence which is generated by $x'_i \sim \mathcal{D}_i$ and $y'_i \sim p(\cdot | x'_i)$. The tangent sequence is independent when conditioned on \mathcal{D} .

Lemma C.1 (Adapted version of Lemma 25 (Agarwal et al., 2020b)). *Let $f_1 \in \mathcal{X} \mapsto \Delta(\mathcal{Y})$ be a conditional probability density and $f_2 \in \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}_{\geq 0}$ (satisfying $\int_{\mathcal{Y}} f_2(x, y) dy \leq s$ for all $x \in \mathcal{X}$). Let $\mathcal{D} \in \Delta(\mathcal{X})$ be any distribution. Then, we have*

$$\mathbb{E}_{x \sim \mathcal{D}} \left(\int_{\mathcal{Y}} |f_1(x, y) - f_2(x, y)| dy \right)^2 \leq (2 + 2s) \left((s - 1) - 2 \log \mathbb{E}_{x \sim \mathcal{D}, y \sim f_1(x, \cdot)} \exp \left(-\frac{1}{2} \log (f_1(x, y)/f_2(x, y)) \right) \right).$$

Proof of Lemma C.1. First, we have

$$\begin{aligned} & \mathbb{E}_{x \sim \mathcal{D}} \left(\int_{\mathcal{Y}} |f_1(x, y) - f_2(x, y)| dy \right)^2 = \mathbb{E}_{x \sim \mathcal{D}} \left(\int_{\mathcal{Y}} \left| \sqrt{f_1(x, y)} - \sqrt{f_2(x, y)} \right| \left(\sqrt{f_1(x, y)} + \sqrt{f_2(x, y)} \right) dy \right)^2 \\ & \leq \mathbb{E}_{x \sim \mathcal{D}} \int_{\mathcal{Y}} \left(\sqrt{f_1(x, y)} - \sqrt{f_2(x, y)} \right)^2 dy \cdot \int_{\mathcal{Y}} \left(\sqrt{f_1(x, y)} + \sqrt{f_2(x, y)} \right)^2 dy \\ & = \mathbb{E}_{x \sim \mathcal{D}} \int_{\mathcal{Y}} \left(\sqrt{f_1(x, y)} - \sqrt{f_2(x, y)} \right)^2 dy \cdot 2 \int_{\mathcal{Y}} (f_1(x, y) + f_2(x, y)) dy - \int_{\mathcal{Y}} \left(\sqrt{f_1(x, y)} - \sqrt{f_2(x, y)} \right)^2 dy \\ & = \mathbb{E}_{x \sim \mathcal{D}} \int_{\mathcal{Y}} \left(\sqrt{f_1(x, y)} - \sqrt{f_2(x, y)} \right)^2 dy \cdot 2 \int_{\mathcal{Y}} (f_1(x, y) + f_2(x, y)) dy \\ & \leq \underbrace{\mathbb{E}_{x \sim \mathcal{D}} \int_{\mathcal{Y}} \left(\sqrt{f_1(x, y)} - \sqrt{f_2(x, y)} \right)^2 dy}_{(*)} \cdot (2 + 2s). \end{aligned}$$

where the first inequality holds for Cauchy–Schwarz inequality. For (*), we have

$$\begin{aligned} (*) & = \mathbb{E}_{x \sim \mathcal{D}} \int_{\mathcal{Y}} \left(\sqrt{f_1(x, y)} - \sqrt{f_2(x, y)} \right)^2 dy \leq (s - 1) + 2 - 2 \mathbb{E}_{x \sim \mathcal{D}} \int_{\mathcal{Y}} \sqrt{f_1(x, y)f_2(x, y)} dy \\ & = (s - 1) + 2 \left(1 - \mathbb{E}_{x \sim \mathcal{D}} \int_{\mathcal{Y}} \sqrt{f_1(x, y)f_2(x, y)} dy \right) \leq (s - 1) - 2 \log \left(\mathbb{E}_{x \sim \mathcal{D}} \int_{\mathcal{Y}} \sqrt{f_1(x, y)f_2(x, y)} dy \right) \\ & \leq (s - 1) - 2 \log \mathbb{E}_{x \sim \mathcal{D}, y \sim f_1(x, \cdot)} \sqrt{f_2(x, y)/f_1(x, y)} \\ & = (s - 1) - 2 \log \mathbb{E}_{x \sim \mathcal{D}, y \sim f_1(x, \cdot)} \exp \left(-\frac{1}{2} \log (f_1(x, y)/f_2(x, y)) \right) \end{aligned}$$

where the second inequality holds because $1 - x \leq -\log x$. □

Lemma C.2 (Adapted version of Theorem 21 (Agarwal et al., 2020b)). *Fix $\delta \in (0, 1)$. Let $N_{\square}(\epsilon, \mathcal{F}, \|\cdot\|_{\infty})$ denote the*

ϵ -bracketing number of \mathcal{F} w.r.t. $\|\cdot\|_\infty$. Then for any estimator \hat{f} that depends on D , with probability at least $1 - \delta$, we have

$$\sum_{i=1}^n \mathbb{E}_{x \sim \mathcal{D}_i} d_{tv}^2 \left(\hat{f}(x, \cdot), f^*(x, \cdot) \right) \leq \frac{3n\epsilon^2 |\mathcal{Y}|^2}{2} + 2n\epsilon |\mathcal{Y}| + (4 + 2\epsilon |\mathcal{Y}|) \left(\frac{1}{2} \sum_{i=1}^n \log (f^*(x_i, y_i) / \hat{f}(x_i, y_i)) + \log N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|_\infty) + \log(1/\delta) \right)$$

where $|\mathcal{Y}|$ denotes $\int_{\mathcal{Y}} dy$.

Proof of Lemma C.2. We take an ϵ -bracket of \mathcal{F} , $\{[l_i, u_i] : i = 1, 2, \dots\}$, and denote $\tilde{\mathcal{F}} = \{u_i : i = 1, 2, \dots\}$. Pick $\tilde{f} \in \tilde{\mathcal{F}}$ satisfying $\hat{f} \leq \tilde{f}$, so \tilde{f} also depends on D . Applying Lemma 24 of (Agarwal et al., 2020b) to function class $\tilde{\mathcal{F}}$ and estimator \hat{f} and using Chernoff method, we have

$$\underbrace{-\log \mathbb{E}_{D'} \exp(L(\tilde{f}(D), D'))}_{(i)} \leq \underbrace{-L(\tilde{f}(D), D) + \log N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|_\infty) + \log(1/\delta)}_{(ii)}. \quad (3)$$

holds with probability at least $1 - \delta$. We set $L(f, D) = \sum_{i=1}^n -1/2 \log(f^*(x_i, y_i) / f(x_i, y_i))$. Then the right hand side of (3) is

$$\begin{aligned} (ii) &= \frac{1}{2} \sum_{i=1}^n \log(f^*(x_i, y_i) / \tilde{f}(x_i, y_i)) + \log N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|_\infty) + \log(1/\delta) \\ &\leq \frac{1}{2} \sum_{i=1}^n \log(f^*(x_i, y_i) / \hat{f}(x_i, y_i)) + \log N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|_\infty) + \log(1/\delta). \end{aligned}$$

On the other hand, by the definition of total variation distance and the fact that $a^2 \leq 2b^2 + 2c^2$ whenever $0 \leq a \leq b + c$, we have

$$\begin{aligned} \sum_{i=1}^n \mathbb{E}_{x \sim \mathcal{D}_i} d_{tv}^2 \left(\hat{f}(x, \cdot), f^*(x, \cdot) \right) &= \frac{1}{4} \sum_{i=1}^n \mathbb{E}_{x \sim \mathcal{D}_i} \left(\int_{\mathcal{Y}} |\hat{f}(x, y) - f^*(x, y)| dy \right)^2 \\ &\leq \frac{1}{2} \sum_{i=1}^n \mathbb{E}_{x \sim \mathcal{D}_i} \underbrace{\left(\int_{\mathcal{Y}} |\hat{f}(x, y) - \tilde{f}(x, y)| dy \right)^2}_{(iii)} + \frac{1}{2} \sum_{i=1}^n \mathbb{E}_{x \sim \mathcal{D}_i} \underbrace{\left(\int_{\mathcal{Y}} |\tilde{f}(x, y) - f^*(x, y)| dy \right)^2}_{(iv)}. \end{aligned}$$

For (iii), by the definition of \tilde{f} , we have (iii) $\leq n\epsilon^2 |\mathcal{Y}|^2$. For (iv), we apply Lemma C.1 with $f_1 = f^*$ and $f_2 = \tilde{f}$ (thus $s = 1 + \epsilon |\mathcal{Y}|$) and get

$$\begin{aligned} (iv) &= 2n\epsilon |\mathcal{Y}| (2 + \epsilon |\mathcal{Y}|) - \sum_{i=1}^n (8 + 4\epsilon |\mathcal{Y}|) \left(\log \mathbb{E}_{x, y \sim f^*(x, \cdot)} \exp \left(-\frac{1}{2} \log \left(f^*(x, y) / \tilde{f}(x, y) \right) \right) \right) \\ &= 2n\epsilon |\mathcal{Y}| (2 + \epsilon |\mathcal{Y}|) - \sum_{i=1}^n (8 + 4\epsilon |\mathcal{Y}|) \left(\log \mathbb{E}_{x, y \sim \mathcal{D}_i} \exp \left(-\frac{1}{2} \log \left(f^*(x, y) / \tilde{f}(x, y) \right) \right) \right) \\ &= 2n\epsilon |\mathcal{Y}| (2 + \epsilon |\mathcal{Y}|) - (8 + 4\epsilon |\mathcal{Y}|) \log \mathbb{E}_{x, y \sim \mathcal{D}'} \left[\exp \left(\sum_{i=1}^n -\frac{1}{2} \log \left(f^*(x, y) / \tilde{f}(x, y) \right) \right) \middle| D \right] \\ &= 4n\epsilon |\mathcal{Y}| + 2n\epsilon^2 |\mathcal{Y}|^2 + (8 + 4\epsilon |\mathcal{Y}|) \cdot (i). \end{aligned}$$

By plugging (iii) and (iv) back we get

$$\sum_{i=1}^n \mathbb{E}_{x \sim \mathcal{D}_i} d_{tv}^2 \left(\hat{f}(x, \cdot), f^*(x, \cdot) \right) \leq 2n\epsilon |\mathcal{Y}| + \frac{3}{2} n\epsilon^2 |\mathcal{Y}|^2 + (4 + 2\epsilon |\mathcal{Y}|) \cdot (i).$$

Notice that (i) \leq (ii), so we complete the proof by plugging (ii) into the above. \square

Lemma C.3. Fixed $\delta \in (0, 1)$. Let \hat{f} denote the maximum likelihood estimator,

$$\hat{f} = \arg \max_{f \in \mathcal{F}} \sum_{i=1}^n \log f(x_i, y_i).$$

Then according to different assumptions on the size of \mathcal{F} , we have the following two conclusions:

(1) If $|\mathcal{F}| < \infty$, we have

$$\sum_{i=1}^n \mathbb{E}_{x \sim \mathcal{D}_i} d_{tv}^2(\hat{f}(x, \cdot), f^*(x, \cdot)) \leq 4 \log |\mathcal{F}| / \delta \quad (4)$$

with probability at least $1 - \delta$.

(2) For general \mathcal{F} , we have

$$\sum_{i=1}^n \mathbb{E}_{x \sim \mathcal{D}_i} d_{tv}^2(\hat{f}(x, \cdot), f^*(x, \cdot)) \leq 10 \log N_{[]}((n|\mathcal{Y}|)^{-1}, \mathcal{F}, \|\cdot\|_\infty) / \delta \quad (5)$$

with probability at least $1 - \delta$.

Proof of Lemma C.3. By Lemma C.2, we have

$$\begin{aligned} \sum_{i=1}^n \mathbb{E}_{x \sim \mathcal{D}_i} d_{tv}^2(\hat{f}(x, \cdot), f^*(x, \cdot)) \leq \\ \frac{3n\epsilon^2|\mathcal{Y}|^2}{2} + 2n\epsilon|\mathcal{Y}| + (4 + 2\epsilon|\mathcal{Y}|) \left(\underbrace{\frac{1}{2} \sum_{i=1}^n \log(f^*(x_i, y_i) / \hat{f}(x_i, y_i))}_{(\diamond)} + \log N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|_\infty) + \log(1/\delta) \right) \end{aligned} \quad (6)$$

with probability at least $1 - \delta$. Since \hat{f} is the maximum likelihood estimator and there exists f_i^* that agrees with f^* on the support of \mathcal{D}_i , we have

$$\log(f^*(x_i, y_i) / \hat{f}(x_i, y_i)) = \log(f_i^*(x_i, y_i) / \hat{f}(x_i, y_i)) \leq 0$$

and thus $(\diamond) \leq 0$. When $|\mathcal{F}| < \infty$, we can set $\epsilon = 0$, and then (6) exactly becomes (4). For general \mathcal{F} , we set $\epsilon = (n|\mathcal{Y}|)^{-1}$ and then get

$$\begin{aligned} \sum_{i=1}^n \mathbb{E}_{x \sim \mathcal{D}_i} d_{tv}^2(\hat{f}(x, \cdot), f^*(x, \cdot)) &\leq \frac{3}{2n} + 2 + \left(4 + \frac{2}{n}\right) \log N_{[]}((n|\mathcal{Y}|)^{-1}, \mathcal{F}, \|\cdot\|_\infty) / \delta \\ &\leq 4 + 6 \log N_{[]}((n|\mathcal{Y}|)^{-1}, \mathcal{F}, \|\cdot\|_\infty) / \delta \leq 10 \log N_{[]}((n|\mathcal{Y}|)^{-1}, \mathcal{F}, \|\cdot\|_\infty) / \delta, \end{aligned}$$

which is exactly (5). □

C.2. Total Variation Distance and Wasserstein Distance

The following lemma states that the total variation distance is equal to the optimal coupling in a sense. The proof can be found in [Levin & Peres \(2017\)](#) (Proposition 4.7).

Lemma C.4. Let f_1 and f_2 be two probability distributions on \mathcal{X} . Then

$$d_{tv}(f_1, f_2) = \inf_{c \in \mathcal{C}} \Pr(x \neq y)$$

where \mathcal{C} is the set of all couplings of f_1 and f_2 .

The following lemma shows the dual representation of the Wasserstein distance. The proof can be found in Villani (2021) (Theorem 1.3) and Villani et al. (2009) (Theorem 5.10).

Lemma C.5 (Kantorovich duality). *Let $f_1, f_2 \in \Delta(\mathcal{X})$ where \mathcal{X} is a Polish space (e.g., Euclidean space). It can be shown that, for any $1 \leq p < \infty$,*

$$d_{w,p}^p(f_1, f_2) = \sup_{\psi, \phi} \int \psi(x) f_1(x) dx - \int \phi(x) f_2(x) dx \quad \text{s.t.} \quad \psi(x) - \phi(y) \leq \|x - y\|^p, \quad \forall x, y \in \mathcal{X}.$$

Lemma C.6. *Let f_1 and f_2 be two distributions on a bounded set \mathcal{X} . Then*

$$d_{w,p}^p(f_1, f_2) \leq \text{diam}^p(\mathcal{X}) \cdot d_{tv}(f_1, f_2)$$

where $\text{diam}(\mathcal{X}) = \sup_{x,y \in \mathcal{X}} \|x - y\|$ is the diameter of \mathcal{X} .

Proof. By definition, we have

$$\begin{aligned} d_{w,p}^p(f_1, f_2) &= \inf_{c \in \mathcal{C}} \mathbb{E}_{x,y \sim c} \|x - y\|^p = \inf_{c \in \mathcal{C}} \mathbb{E}_{x,y \sim c} [\mathbb{1}[x \neq y] \cdot \|x - y\|^p] \\ &\leq \text{diam}^p(\mathcal{X}) \cdot \inf_{c \in \mathcal{C}} \mathbb{E}_{x,y \sim c} \mathbb{1}[x \neq y] = \text{diam}^p(\mathcal{X}) \cdot d_{tv}(f_1, f_2) \end{aligned}$$

where by \mathcal{C} we denote the set of all couplings of f_1 and f_2 , and the last equality holds because of Lemma C.4. \square

Corollary C.7. *Let f_1 and f_2 be two distributions on $[0, m]^d$. Then*

$$d_{w,p}^p(f_1, f_2) \leq (m\sqrt{d})^p \cdot d_{tv}(f_1, f_2).$$

Since the total variation distance is at most one, we have the following.

Corollary C.8. *Let f_1 and f_2 be two distributions on $[0, m]^d$. Then*

$$d_{w,p}^p(f_1, f_2) \leq (m\sqrt{d})^p.$$

D. Missing Proofs in Section 4

D.1. Proof of Theorem 4.2

Proof. Note that for all $h \in [H]$, we have

$$\begin{aligned} &\mathbb{E}_{x,a \sim d_h^\pi} d_{tv} \left([\mathcal{T}^\pi \hat{f}_{h+1}](x, a), [Z_{h+1}^\pi](x, a) \right) \\ &= \frac{1}{2} \mathbb{E}_{x,a \sim d_h^\pi} \sup_{g: \|g\|_\infty \leq 1} \left| \mathbb{E}_{\substack{x' \sim P(x,a) \\ a' \sim \pi(x') \\ r \sim r(x,a)}} \left(\mathbb{E}_{y \sim \hat{f}_{h+1}(\cdot|x',a')} g(r+y) - \mathbb{E}_{y \sim Z_{h+1}^\pi(\cdot|x',a')} g(r+y) \right) \right| \\ &\leq \frac{1}{2} \mathbb{E}_{\substack{x,a \sim d_h^\pi \\ x' \sim P(x,a) \\ a' \sim \pi(\cdot|x) \\ r \sim r(x,a)}} \sup_{g: \|g\|_\infty \leq 1} \left| \mathbb{E}_{y \sim \hat{f}_{h+1}(\cdot|x',a')} g(r+y) - \mathbb{E}_{y \sim Z_{h+1}^\pi(\cdot|x',a')} g(r+y) \right| \\ &= \frac{1}{2} \mathbb{E}_{\substack{x,a \sim d_h^\pi \\ x' \sim P(x,a) \\ a' \sim \pi(\cdot|x)}} \sup_{g: \|g\|_\infty \leq 1} \left| \mathbb{E}_{y \sim \hat{f}_{h+1}(\cdot|x',a')} g(y) - \mathbb{E}_{y \sim Z_{h+1}^\pi(\cdot|x',a')} g(y) \right| \\ &= \mathbb{E}_{x',a' \sim d_{h+1}^\pi} d_{tv} \left(\hat{f}_{h+1}(x', a'), Z_{h+1}^\pi(x', a') \right). \end{aligned}$$

Here the inequality holds for Jensen's inequality. The second equality holds since the randomness of r lies outside the supremum, so we can consider r as a constant within the supremum, allowing us to set $\tilde{g}(y) = g(r + y)$ for which we have $\|\tilde{g}\|_\infty \leq 1$ thus removing the additive term r . Hence, by triangle inequality, we have

$$\begin{aligned} & \mathbb{E}_{x,a \sim d_h^\pi} d_{tv} \left(\hat{f}_h(x, a), Z_h^\pi(x, a) \right) = \mathbb{E}_{x,a \sim d_h^\pi} d_{tv} \left(\hat{f}_h(x, a), [\mathcal{T}^\pi Z_{h+1}^\pi](x, a) \right) \\ & \leq \underbrace{\mathbb{E}_{x,a \sim d_h^\pi} d_{tv} \left(\hat{f}_h(x, a), [\mathcal{T}^\pi \hat{f}_{h+1}](x, a) \right)}_{(i)} + \underbrace{\mathbb{E}_{x,a \sim d_h^\pi} d_{tv} \left([\mathcal{T}^\pi \hat{f}_{h+1}](x, a), [\mathcal{T}^\pi Z_{h+1}^\pi](x, a) \right)}_{(ii)}. \end{aligned}$$

By Assumption 4.1 and Jensen's inequality, we have (i) $\leq \sqrt{C} \zeta_h$ because

$$\begin{aligned} & \mathbb{E}_{x,a \sim d_h^\pi} d_{tv} \left(\hat{f}_h(x, a), [\mathcal{T}^\pi \hat{f}_{h+1}](x, a) \right) \\ & \leq \left\{ \mathbb{E}_{x,a \sim d_h^\pi} d_{tv}^2 \left(\hat{f}_h(x, a), [\mathcal{T}^\pi \hat{f}_{h+1}](x, a) \right) \right\}^{1/2} \leq \sqrt{C} \zeta_h. \end{aligned}$$

And by the above derivation we have (ii) $\leq \mathbb{E}_{x,a \sim d_{h+1}^\pi} d_{tv} \left(\hat{f}_{h+1}(x, a), Z_{h+1}^\pi(x, a) \right)$. Hence,

$$\mathbb{E}_{x,a \sim d_h^\pi} d_{tv} \left(\hat{f}_h(x, a), Z_h^\pi(x, a) \right) \leq \sqrt{C} \zeta_h + \mathbb{E}_{x,a \sim d_{h+1}^\pi} d_{tv} \left(\hat{f}_{h+1}(x, a), Z_{h+1}^\pi(x, a) \right).$$

Summing over $h = 1, \dots, H$ on both sides, we get

$$\mathbb{E}_{x,a \sim d_1^\pi} d_{tv} \left(\hat{f}_1(x, a), Z_1^\pi(x, a) \right) \leq \sqrt{C} \sum_{h=1}^H \zeta_h + \mathbb{E}_{x,a \sim d_{H+1}^\pi} d_{tv} \left(\hat{f}_{H+1}(x, a), Z_{H+1}^\pi(x, a) \right) = \sqrt{C} \sum_{h=1}^H \zeta_h.$$

where the equality holds since $\hat{f}_{H+1} = Z_{H+1}^\pi = 0$ by definition. Now we complete the proof by noticing the following

$$\begin{aligned} d_{tv} \left(\hat{f}, Z^\pi \right) &= \frac{1}{2} \sup_{g: \|g\|_\infty \leq 1} \left| \mathbb{E}_{x,a \sim d_1^\pi} \left(\mathbb{E}_{y \sim \hat{f}_1(\cdot|x,a)} g(y) - \mathbb{E}_{y \sim Z^\pi(\cdot|x,a)} g(y) \right) \right| \\ &\leq \frac{1}{2} \mathbb{E}_{x,a \sim d_1^\pi} \sup_{g: \|g\|_\infty \leq 1} \left| \mathbb{E}_{y \sim \hat{f}_1(\cdot|x,a)} g(y) - \mathbb{E}_{y \sim Z^\pi(\cdot|x,a)} g(y) \right| = \mathbb{E}_{x,a \sim d_1^\pi} d_{tv} \left(\hat{f}_1(x, a), Z_1^\pi(x, a) \right). \end{aligned}$$

□

D.2. Proof of Lemma 4.4

Proof. Observing Algorithm 1, when $h = H$, we are estimating the conditional distribution Z_H^π via MLE. Under Assumption 4.3 which implies that there exists a function $g \in \mathcal{F}_H$ that agrees with Z_H^π on the support of ρ , we can apply Lemma C.3, which leads to

$$\mathbb{E}_{x,a \sim \rho} d_{tv}^2 \left(\hat{f}_H(x, a), Z_H^\pi(x, a) \right) \leq \frac{4H}{n} \log(|\mathcal{F}_H|/\delta)$$

with probability at least $1 - \delta$. When $h < H$, we are estimating the conditional distribution $\mathcal{T}^\pi \hat{f}_{h+1}$ via MLE. Also note that thanks to the random data split, we have \hat{f}_{h+1} being independent of the dataset \mathcal{D}_h (\hat{f}_{h+1} only depends on datasets $\mathcal{D}_{h+1}, \dots, \mathcal{D}_H$). Therefore, under Assumption 4.3 which implies that there exists a function $g \in \mathcal{F}_h$ that agrees with $\mathcal{T}^\pi \hat{f}_{h+1}$ on the support of ρ , we can apply Lemma C.3, which leads to

$$\mathbb{E}_{x,a \sim \rho} d_{tv}^2 \left(\hat{f}_h(x, a), [\mathcal{T}^\pi \hat{f}_{h+1}](x, a) \right) \leq \frac{4H}{n} \log(|\mathcal{F}_H|/\delta)$$

with probability at least $1 - \delta$. We complete the proof by taking the union bound for $h \in [H]$.

□

D.3. Proof of Lemma 4.6

Proof. The proof is similar to Lemma 4.4. Observing Algorithm 1, when $h = H$, we are basically estimating the conditional distribution Z_H^π via MLE. Hence, under Assumption 4.3 which implies that there exists a function $g \in \mathcal{F}_H$ that agrees with Z_H^π on the support of ρ , we can apply Lemma C.3, which leads to

$$\mathbb{E}_{x,a \sim \rho} d_{tv}^2 \left(\hat{f}_H(x, a), Z_H^\pi(x, a) \right) \leq \frac{10H}{n} \log \left(N_{[]} \left((nH^d)^{-1}, \mathcal{F}_H, \|\cdot\|_\infty \right) / \delta \right)$$

with probability at least $1 - \delta$. When $h < H$, we are estimating the conditional distribution $\mathcal{T}^\pi \hat{f}_{h+1}$ via MLE. Therefore, under Assumption 4.3 which implies that there exists a function $g \in \mathcal{F}_h$ that agrees with $\mathcal{T}^\pi \hat{f}_{h+1}$ on the support of ρ , we can apply Lemma C.3, which leads to

$$\mathbb{E}_{x,a \sim \rho} d_{tv}^2 \left(\hat{f}_h(x, a), [\mathcal{T}^\pi \hat{f}_{h+1}](x, a) \right) \leq \frac{10H}{n} \log \left(N_{[]} \left((nH^d)^{-1}, \mathcal{F}_h, \|\cdot\|_\infty \right) / \delta \right)$$

with probability at least $1 - \delta$. We complete the proof by taking the union bound for $h \in [H]$. \square

D.4. Proof of Lemma 4.9

Proof. First, it deserves to verify that the “metric” $(\mathbb{E}_{x,a \sim d^\pi} d_{w,p}^{2p})^{1/(2p)}$ we are using satisfies the triangle inequality and is thus indeed a metric. To this end, we note that, for any three densities $f_1, f_2, f_3 : \mathcal{X} \times \mathcal{A} \mapsto \Delta([0, (1 - \gamma)^{-1}]^d)$, the following holds since $d_{w,p}$ is a metric,

$$\left(\mathbb{E}_{x,a \sim d^\pi} d_{w,p}^{2p}(f_1(x, a), f_2(x, a)) \right)^{\frac{1}{2p}} \leq \left(\mathbb{E}_{x,a \sim d^\pi} \left(d_{w,p}(f_1(x, a), f_3(x, a)) + d_{w,p}(f_3(x, a), f_2(x, a)) \right)^{2p} \right)^{\frac{1}{2p}}.$$

Then by Minkowski inequality, the above

$$\leq \left(\mathbb{E}_{x,a \sim d^\pi} d_{w,p}^{2p}(f_1(x, a), f_3(x, a)) \right)^{\frac{1}{2p}} + \left(\mathbb{E}_{x,a \sim d^\pi} d_{w,p}^{2p}(f_3(x, a), f_2(x, a)) \right)^{\frac{1}{2p}},$$

for which we conclude triangle inequality for $(\mathbb{E}_{x,a \sim d^\pi} d_{w,p}^{2p})^{1/(2p)}$. Since other axioms of metrics are trivial to verify, we conclude that it is indeed a metric. Hence, we can safely proceed.

To establish the contractive property, we start with the following lemma, which shows that the distributional Bellman operator is roughly “ γ -contractive” in a sense but with distribution shifts.

Lemma D.1. *For any $f, f' \in \mathcal{F}$, $x \in \mathcal{X}$ and $a \in \mathcal{A}$, we have*

$$d_{w,p}^p([\mathcal{T}^\pi f](x, a), [\mathcal{T}^\pi f'](x, a)) \leq \mathbb{E}_{x' \sim P(x, a), a' \sim \pi(x')} \gamma^p d_{w,p}^p(f(x', a'), f'(x', a')).$$

Proof of Lemma D.1. By the dual form of Wasserstein distance (Lemma C.5), we have

$$\begin{aligned} d_{w,p}^p([\mathcal{T}^\pi f](x, a), [\mathcal{T}^\pi f'](x, a)) &= \sup_{(\psi, \phi) \in \Gamma} \mathbb{E}_{z \sim [\mathcal{T}^\pi f](x, a)} \psi(z) - \mathbb{E}_{z \sim [\mathcal{T}^\pi f'](x, a)} \phi(z) \\ &= \sup_{(\psi, \phi) \in \Gamma} \mathbb{E}_{\substack{x' \sim P(x, a) \\ a' \sim \pi(x') \\ r \sim r(x, a)}} \left(\mathbb{E}_{y \sim f(x', a')} \psi(r + \gamma y) - \mathbb{E}_{y \sim f'(x', a')} \phi(r + \gamma y) \right) \\ &\leq \mathbb{E}_{\substack{x' \sim P(x, a) \\ a' \sim \pi(x') \\ r \sim r(x, a)}} \underbrace{\sup_{(\psi, \phi) \in \Gamma} \left(\mathbb{E}_{y \sim f(x', a')} \psi(r + \gamma y) - \mathbb{E}_{y \sim f'(x', a')} \phi(r + \gamma y) \right)}_{(*)} \end{aligned} \quad (7)$$

where $\Gamma = \{(\psi, \phi) : \psi(x) - \phi(y) \leq \|x - y\|^p\}$. The second equality holds by the definition of Bellman operator.

Regarding (*), for any $(\psi, \phi) \in \Gamma$, we define $\tilde{\psi}(y) = \psi(r + \gamma y)/\gamma^p$ and $\tilde{\phi}(y) = \phi(r + \gamma y)/\gamma^p$. Then, we have

$$(*) = \gamma^p \sup_{(\psi, \phi) \in \Gamma} \left(\mathbb{E}_{y \sim f(x', a')} \tilde{\psi}(y) - \mathbb{E}_{y \sim f'(x', a')} \tilde{\phi}(y) \right).$$

We note that, for any x, y ,

$$\tilde{\psi}(x) - \tilde{\phi}(y) = \frac{\psi(r + \gamma x) - \phi(r + \gamma y)}{\gamma^p} \leq \frac{\|(r + \gamma x) - (r + \gamma y)\|^p}{\gamma^p} = \|x - y\|^p.$$

Here the inequality holds since $(\psi, \phi) \in \Gamma$. Hence, $(\tilde{\psi}, \tilde{\phi}) \in \Gamma$ as well. In other words, for any given ψ and ϕ , their correspondences $\tilde{\psi}$ and $\tilde{\phi}$ are also in Γ . Thus we can take the supremum directly over the latter, which leads to

$$(*) \leq \gamma^p \sup_{(\tilde{\psi}, \tilde{\phi}) \in \Gamma} \left(\mathbb{E}_{y \sim f(x', a')} \tilde{\psi}(y) - \mathbb{E}_{y \sim f'(x', a')} \tilde{\phi}(y) \right) = \gamma^p d_{w,p}^p(f(x', a'), f'(x', a'))$$

where the equality holds due to the dual form of Wasserstein distance (Lemma C.5) again. Then we plug the above into (7) and get

$$d_{w,p}^p([\mathcal{T}^\pi f](x, a), [\mathcal{T}^\pi f'](x, a)) \leq \mathbb{E}_{x' \sim P(x, a), a' \sim \pi(x')} \gamma^p d_{w,p}^p(f(x', a'), f'(x', a')).$$

where we have removed the randomness of $r \sim r(x, a)$ originally appeared in (7) since the term inside the expectation is now completely independent of r . \square

By Lemma D.1, we have

$$\begin{aligned} & \left(\mathbb{E}_{x, a \sim d^\pi} d_{w,p}^{2p}([\mathcal{T}^\pi f](x, a), [\mathcal{T}^\pi f'](x, a)) \right)^{\frac{1}{2p}} = \left(\mathbb{E}_{x, a \sim d^\pi} \left(d_{w,p}^p([\mathcal{T}^\pi f](x, a), [\mathcal{T}^\pi f'](x, a)) \right)^2 \right)^{\frac{1}{2p}} \\ & \leq \gamma \cdot \left(\mathbb{E}_{x, a \sim d^\pi} \left(\mathbb{E}_{x' \sim P(x, a), a' \sim \pi(x')} d_{w,p}^p(f(x', a'), f'(x', a')) \right)^2 \right)^{\frac{1}{2p}} \\ & \leq \gamma \cdot \left(\underbrace{\mathbb{E}_{\substack{x, a \sim d^\pi \\ x' \sim P(x, a), a' \sim \pi(x')}} d_{w,p}^{2p}(f(x', a'), f'(x', a'))}_{(\dagger)} \right)^{\frac{1}{2p}} \end{aligned}$$

where the last inequality holds because of Jensen's inequality. Since $d^\pi(x, a) = \gamma \mathbb{E}_{\tilde{x}, \tilde{a} \sim d^\pi} P(x|\tilde{x}, \tilde{a})\pi(a|x) + (1 - \gamma)\mu(x)\pi(x|a)$, we have $\mathbb{E}_{\tilde{x}, \tilde{a} \sim d^\pi} P(x|\tilde{x}, \tilde{a})\pi(a|x) \leq \gamma^{-1} d^\pi(x, a)$. Therefore,

$$(\dagger) \leq \gamma^{-1} \mathbb{E}_{x, a \sim d^\pi} d_{w,p}^{2p}(f(x, a), f'(x, a)).$$

Hence, we conclude that

$$\begin{aligned} & \left(\mathbb{E}_{x, a \sim d^\pi} d_{w,p}^{2p}([\mathcal{T}^\pi f](x, a), [\mathcal{T}^\pi f'](x, a)) \right)^{\frac{1}{2p}} \\ & \leq \gamma \cdot \left(\gamma^{-1} \mathbb{E}_{x, a \sim d^\pi} d_{w,p}^{2p}(f(x, a), f'(x, a)) \right)^{\frac{1}{2p}} \\ & = \gamma^{1 - \frac{1}{2p}} \cdot \left(\mathbb{E}_{x, a \sim d^\pi} d_{w,p}^{2p}(f(x, a), f'(x, a)) \right)^{\frac{1}{2p}}. \end{aligned}$$

\square

D.5. Proof of Theorem 4.11

Proof. We will prove the following theorem which is more general.

Theorem D.2. *Under Assumption 4.10, suppose we have a sequence of functions $\hat{f}_1, \dots, \hat{f}_T : \mathcal{X} \times \mathcal{A} \mapsto \Delta([0, (1-\gamma)^{-1}]^d)$ and a sequence of values $\zeta_1, \dots, \zeta_T \in \mathbb{R}$ such that*

$$\left(\mathbb{E}_{x, a \sim \rho} d_{w,p}^{2p} \left(\hat{f}_t(x, a), [\mathcal{T}^\pi \hat{f}_{t-1}](x, a) \right) \right)^{\frac{1}{2p}} \leq \zeta_t$$

holds for all $t \in [T]$. Let our estimator $\hat{f} := \mathbb{E}_{x \sim \mu, a \sim \pi(x)} \hat{f}_T(x, a)$. Then we have, for all $p \geq 1$,

$$d_{w,p} \left(\hat{f}, Z^\pi \right) \leq \left(\frac{C}{1-\gamma} \right)^{\frac{1}{2p}} \sum_{t=1}^T \gamma^{(T-t)(1-\frac{1}{2p})} \cdot \zeta_t + \frac{\sqrt{d} \cdot \gamma^{T(1-\frac{1}{2p})}}{(1-\gamma)^{1+\frac{1}{2p}}}. \quad (8)$$

Proof of Theorem D.2. Recall that we defined the conditional distributions $\bar{Z}^\pi(x, a) \in \Delta([0, (1-\gamma)^{-1}]^d)$ which is the distribution of the return under policy π starting with state action (x, a) . It is easy to see that $Z^\pi = \mathbb{E}_{x \sim \mu, a \sim \pi(x)} [\bar{Z}^\pi(x, a)]$. We start with the following.

$$\begin{aligned} & \left(\mathbb{E}_{x, a \sim d^\pi} d_{w,p}^{2p} \left(\hat{f}_t(x, a), \bar{Z}^\pi(x, a) \right) \right)^{\frac{1}{2p}} \\ & \leq \left(\mathbb{E}_{x, a \sim d^\pi} d_{w,p}^{2p} \left(\hat{f}_t(x, a), [\mathcal{T}^\pi \hat{f}_{t-1}](x, a) \right) \right)^{\frac{1}{2p}} + \left(\mathbb{E}_{x, a \sim d^\pi} d_{w,p}^{2p} \left([\mathcal{T}^\pi \hat{f}_{t-1}](x, a), \bar{Z}^\pi(x, a) \right) \right)^{\frac{1}{2p}} \\ & \leq C^{\frac{1}{2p}} \left(\mathbb{E}_{x, a \sim \rho} d_{w,p}^{2p} \left(\hat{f}_t(x, a), [\mathcal{T}^\pi \hat{f}_{t-1}](x, a) \right) \right)^{\frac{1}{2p}} + \left(\mathbb{E}_{x, a \sim d^\pi} d_{w,p}^{2p} \left([\mathcal{T}^\pi \hat{f}_{t-1}](x, a), [\mathcal{T}^\pi \bar{Z}^\pi](x, a) \right) \right)^{\frac{1}{2p}} \\ & \leq C^{\frac{1}{2p}} \zeta_t + \gamma^{1-\frac{1}{2p}} \left(\mathbb{E}_{x, a \sim d^\pi} d_{w,p}^{2p} \left(\hat{f}_{t-1}(x, a), \bar{Z}^\pi(x, a) \right) \right)^{\frac{1}{2p}} \end{aligned}$$

where the first inequality is due to triangle inequality (proved in Appendix D.4), the second inequality holds because of the coverage assumption (Assumption 4.10), and the last inequality holds due to the contractive property of the distributional Bellman operator (Lemma 4.9). Unrolling the recursion of t , we arrive at

$$\begin{aligned} & \left(\mathbb{E}_{x, a \sim d^\pi} d_{w,p}^{2p} \left(\hat{f}_T(x, a), \bar{Z}^\pi(x, a) \right) \right)^{\frac{1}{2p}} \\ & \leq \sum_{t=1}^T \gamma^{(T-t)(1-\frac{1}{2p})} C^{\frac{1}{2p}} \zeta_t + \gamma^{T(1-\frac{1}{2p})} \left(\mathbb{E}_{x, a \sim d^\pi} d_{w,p}^{2p} \left(\hat{f}_0(x, a), \bar{Z}^\pi(x, a) \right) \right)^{\frac{1}{2p}} \\ & \leq \sum_{t=1}^T \gamma^{(T-t)(1-\frac{1}{2p})} C^{\frac{1}{2p}} \zeta_t + \gamma^{T(1-\frac{1}{2p})} \cdot \frac{\sqrt{d}}{1-\gamma} \end{aligned} \quad (9)$$

where the last inequality is due to Corollary C.8 which shows that

$$d_{w,p}(\hat{f}_0(x, a), \bar{Z}^\pi(x, a)) \leq \text{diam}([0, (1-\gamma)^{-1}]^d) \leq \frac{\sqrt{d}}{(1-\gamma)}.$$

Since $d^\pi(x, a) = \gamma \mathbb{E}_{\tilde{x}, \tilde{a} \sim d^\pi} P(x|\tilde{x}, \tilde{a})\pi(a|x) + (1-\gamma)\mu(x)\pi(x|a)$, we have $\mu(x)\pi(x|a) \leq (1-\gamma)^{-1}d^\pi(x, a)$ and thus

$$\begin{aligned} & \left(\mathbb{E}_{x \sim \mu, a \sim \pi(x)} d_{w,p}^{2p} \left(\hat{f}_T(x, a), \bar{Z}^\pi(x, a) \right) \right)^{\frac{1}{2p}} \leq \left((1-\gamma)^{-1} \mathbb{E}_{x, a \sim d^\pi} d_{w,p}^{2p} \left(\hat{f}_T(x, a), \bar{Z}^\pi(x, a) \right) \right)^{\frac{1}{2p}} \\ & = (1-\gamma)^{-\frac{1}{2p}} \left(\mathbb{E}_{x, a \sim d^\pi} d_{w,p}^{2p} \left(\hat{f}_T(x, a), \bar{Z}^\pi(x, a) \right) \right)^{\frac{1}{2p}} \leq (1-\gamma)^{-\frac{1}{2p}} \left(\sum_{t=1}^T \gamma^{(T-t)(1-\frac{1}{2p})} C^{\frac{1}{2p}} \zeta_t + \gamma^{T(1-\frac{1}{2p})} \cdot \frac{\sqrt{d}}{1-\gamma} \right) \end{aligned} \quad (10)$$

where the last inequality is for (9).

Applying the dual representation of Wasserstein distance (Lemma C.5) to $d_{w,p}^p(\hat{f}, Z^\pi)$, we have

$$\begin{aligned}
 d_{w,p}^p(\hat{f}, Z^\pi) &= d_{w,p}^p\left(\mathbb{E}_{x \sim \mu, a \sim \pi(x)} \hat{f}_T(x, a), \mathbb{E}_{x \sim \mu, a \sim \pi(x)} \bar{Z}^\pi(x, a)\right) \\
 &= \sup_{\psi, \phi \in \Gamma} \mathbb{E}_{x \sim \mu, a \sim \pi(x)} \left(\mathbb{E}_{z \sim \hat{f}_T(x, a)} \psi(z) - \mathbb{E}_{z \sim \bar{Z}^\pi(x, a)} \phi(z) \right) \\
 &\leq \mathbb{E}_{x \sim \mu, a \sim \pi(x)} \sup_{\psi, \phi \in \Gamma} \left(\mathbb{E}_{z \sim \hat{f}_T(x, a)} \psi(z) - \mathbb{E}_{z \sim \bar{Z}^\pi(x, a)} \phi(z) \right) \\
 &= \mathbb{E}_{x \sim \mu, a \sim \pi(x)} d_{w,p}^p(\hat{f}_T(x, a), \bar{Z}^\pi(x, a)) \\
 &\leq \left(\mathbb{E}_{x \sim \mu, a \sim \pi(x)} d_{w,p}^{2p}(\hat{f}_T(x, a), \bar{Z}^\pi(x, a)) \right)^{\frac{1}{2}}. \tag{11}
 \end{aligned}$$

where $\Gamma = \{(\psi, \phi) : \psi(x) - \phi(y) \leq \|x - y\|^p\}$. By chaining (10) and (11) we complete the proof. \square

By assuming there exists a common upper bound ζ (i.e., $\zeta_t \leq \zeta, \forall t$), we can further simplify (8) by noticing the following. First, since the sum of geometric series is bounded in the following sense

$$\sum_{t=1}^T \gamma^{(T-t)(1-\frac{1}{2p})} \leq \frac{1}{1 - \gamma^{(1-\frac{1}{2p})}},$$

we can get

$$d_{w,p}(\hat{f}, Z^\pi) \leq \left(\frac{C}{1-\gamma} \right)^{\frac{1}{2p}} \frac{\zeta}{(1-\gamma^{1-\frac{1}{2p}})} + \frac{\sqrt{d} \cdot \gamma^{T(1-\frac{1}{2p})}}{(1-\gamma)^{1+\frac{1}{2p}}}.$$

Second, we note that the right-hand side above attains the maximum when $p = 1$. Therefore

$$\begin{aligned}
 d_{w,p}(\hat{f}, Z^\pi) &\leq \left(\frac{C}{1-\gamma} \right)^{\frac{1}{2p}} \cdot \frac{\zeta}{1-\gamma^{\frac{1}{2}}} + \frac{\sqrt{d} \cdot \gamma^{\frac{T}{2}}}{(1-\gamma)^{\frac{3}{2}}} \\
 &\leq \frac{2C^{\frac{1}{2p}}}{(1-\gamma)^{\frac{3}{2}}} \cdot \zeta + \frac{\sqrt{d} \cdot \gamma^{\frac{T}{2}}}{(1-\gamma)^{\frac{3}{2}}}
 \end{aligned}$$

where the last inequality holds since $1 - \gamma^{1/2} \geq (1 - \gamma)/2$. \square

D.6. Proof of Lemma 4.13

Proof. We only show the proof for finite function class since the proof for infinite class is essentially the same.

For Algorithm 2, we are iteratively estimating the conditional distribution $\mathcal{T}^\pi \hat{f}_{t-1}$. Note that thanks to the random data split, we have \hat{f}_{t-1} being independent of the dataset \mathcal{D}_t (\hat{f}_{t-1} only depends on datasets $\mathcal{D}_1, \dots, \mathcal{D}_{t-1}$). Therefore, under Assumption 4.12 which implies that there exists a function $g \in \mathcal{F}$ that agrees with $\mathcal{T}^\pi \hat{f}_{t-1}$ on the support of ρ , we can apply Lemma C.3, which leads to

$$\mathbb{E}_{x, a \sim \rho} d_{tv}^2(\hat{f}_t(x, a), [\mathcal{T}^\pi \hat{f}_{t-1}](x, a)) \leq \frac{4T}{n} \log(|\mathcal{F}|T/\delta)$$

with probability at least $1 - \delta$. Here we have taken the union bound for $t \in [T]$. For the result of Wasserstein distance, we apply Corollary C.7 and get

$$\mathbb{E}_{x, a \sim \rho} d_{w,p}^{2p}(\hat{f}_t(x, a), [\mathcal{T}^\pi \hat{f}_{t-1}](x, a)) \leq \left(\frac{\sqrt{d}}{1-\gamma} \right)^{2p} \mathbb{E}_{x, a \sim \rho} d_{tv}^2(\hat{f}_t(x, a), [\mathcal{T}^\pi \hat{f}_{t-1}](x, a)).$$

\square

D.7. Proof of Corollary 4.14

Proof. We only prove for the finite function class ($|\mathcal{F}| < \infty$) since the proof for the infinite function class is quite similar. We start with Theorem 4.11, plug in the result of Lemma 4.13, and get

$$\begin{aligned} d_{w,p}(\hat{f}, Z^\pi) &\leq \frac{2C^{\frac{1}{2p}}}{(1-\gamma)^{\frac{3}{2}}} \cdot \frac{\sqrt{d}}{1-\gamma} \cdot \left(\frac{4T}{n} \log(|\mathcal{F}|T/\delta)\right)^{\frac{1}{2p}} + \frac{\sqrt{d} \cdot \gamma^{\frac{T}{2}}}{(1-\gamma)^{\frac{3}{2}}} \\ &= \frac{\sqrt{d}}{(1-\gamma)^{\frac{3}{2}}} \left(\frac{2C^{\frac{1}{2p}}}{1-\gamma} \cdot \left(\frac{4T}{n} \log(|\mathcal{F}|T/\delta)\right)^{\frac{1}{2p}} + \gamma^{\frac{T}{2}} \right) \end{aligned} \quad (12)$$

We choose

$$T = \frac{\log\left(C^{\frac{1}{2p}} \cdot \iota^{\frac{1}{2p}} \cdot (1-\gamma)^{-1} \cdot n^{-\frac{1}{2p}}\right)}{\log\left(\gamma^{\frac{1}{2}}\right)} \quad \text{where } \iota = \log(|\mathcal{F}|/\delta),$$

which leads to

$$\gamma^{\frac{T}{2}} = \frac{C^{\frac{1}{2p}} \cdot \iota^{\frac{1}{2p}} \cdot n^{-\frac{1}{2p}}}{1-\gamma}.$$

Thus, the second additive term of (12) will be smaller than the first one. Hence, we conclude that

$$d_{w,p}(\hat{f}, Z^\pi) \leq 2 \cdot \frac{\sqrt{d}}{(1-\gamma)^{\frac{3}{2}}} \cdot \frac{2C^{\frac{1}{2p}}}{1-\gamma} \cdot \left(\frac{4T}{n} \log(|\mathcal{F}|T/\delta)\right)^{\frac{1}{2p}} \leq \tilde{O}\left(\frac{\sqrt{d}(C \log(|\mathcal{F}|T/\delta))^{\frac{1}{2p}}}{(1-\gamma)^{\frac{5}{2}} \cdot n^{\frac{1}{2p}}}\right).$$

□

D.8. Proof of Lemma B.1

Proof. The bracketing number of the probability simplex $\Delta(|\mathcal{X}||\mathcal{A}|)$ is bounded by $N_{[]}(\epsilon, \Delta(|\mathcal{X}||\mathcal{A}|), \|\cdot\|_\infty) \leq (c/\epsilon)^{|\mathcal{X}||\mathcal{A}|}$ where c is a constant. Hence, we have $N_{[]}(\epsilon, (\Delta(|\mathcal{X}||\mathcal{A}|))^{\otimes |\mathcal{X}||\mathcal{A}|}, \|\cdot\|_\infty) \leq (c/\epsilon)^{|\mathcal{X}|^2|\mathcal{A}|^2}$.

Let $\tilde{\Delta}$ denote an ϵ -bracket of $(\Delta(|\mathcal{X}||\mathcal{A}|))^{\otimes |\mathcal{X}||\mathcal{A}|}$. Then we can construct a bracket of \mathcal{F}_h as follows

$$\tilde{\mathcal{F}}_h = \left\{ \left[\underline{f}, \bar{f} \right] : \underline{f}(x, a) = \sum_{x', a'} \underline{w}_{x,a}(x', a') r_H(x', a'), \bar{f}(x, a) = \sum_{x', a'} \bar{w}_{x,a}(x', a') r_H(x', a'), \forall [\underline{w}, \bar{w}] \in \tilde{\Delta} \right\}.$$

We claim that $\tilde{\mathcal{F}}_h$ is a $\epsilon r_\infty |\mathcal{X}||\mathcal{A}|$ -bracket of \mathcal{F}_h . To see this, we have

$$\left\| \bar{f} - \underline{f} \right\|_\infty \leq \sum_{x', a'} |\underline{w}_{x,a}(x', a') - \bar{w}_{x,a}(x', a')| r_H(x', a') \leq \epsilon \sum_{x', a'} r_H(x', a') \leq \epsilon r_\infty |\mathcal{X}||\mathcal{A}|.$$

Therefore, we conclude $N_{[]}(\epsilon r_\infty |\mathcal{X}||\mathcal{A}|, \mathcal{F}_h, \|\cdot\|_\infty) \leq |\tilde{\Delta}| \leq (c/\epsilon)^{|\mathcal{X}|^2|\mathcal{A}|^2}$. By substitution we arrive at $N_{[]}(\epsilon, \mathcal{F}_h, \|\cdot\|_\infty) \leq (c r_\infty |\mathcal{X}||\mathcal{A}|/\epsilon)^{|\mathcal{X}|^2|\mathcal{A}|^2}$. Then we complete the proof by taking a logarithm. □

D.9. Proof of Lemma B.2

$$\begin{aligned}
 \mu_h(x, a) &= \sum_{i=h}^H -(x_i^\top Q x_i + a_i^\top R a_i) = -x^\top Q x - a^\top R a - \sum_{i=h+1}^H -(x_i^\top Q x_i + a_i^\top R a_i) \\
 &= -x^\top Q x - a^\top R a - \sum_{i=h+1}^H (x_i^\top Q x_i + x_i^\top K^\top R K x_i) \\
 &= -x^\top Q x - a^\top R a - \sum_{i=h+1}^H x_i^\top (Q + K^\top R K) x_i \\
 &= -x^\top Q x - a^\top R a - \sum_{i=h+1}^H ((A + BK)^{i-h-1} (Ax + Ba))^\top (Q + K^\top R K) ((A + BK)^{i-h-1} (Ax + Ba)) \\
 &= -x^\top Q x - a^\top R a - (Ax + Ba)^\top \left(\sum_{i=h+1}^H ((A + BK)^{i-h-1})^\top (Q + K^\top R K) (A + BK)^{i-h-1} \right) (Ax + Ba).
 \end{aligned}$$

D.10. Proof of Lemma B.3

Lemma D.3. For any $x, a, b \in \mathbb{R}$, we have $\exp(-(x-a)^2) - \exp(-(x-b)^2) \leq \sqrt{2/e} \cdot |a-b|$.

Proof of Lemma D.3. When $a \geq b$, it is equivalent to $\exp(-(x-a)^2) - \exp(-(x-b)^2) \leq \sqrt{2/e} \cdot (a-b)$. Thus it suffices to show that $g(x, a) := \exp(-(x-a)^2) - \sqrt{2/e} \cdot a$ is non-increasing in a . We take the first derivative with respect to a and then get

$$\frac{\partial}{\partial a} g(x, a) = 2(x-a) \exp(-(x-a)^2) - \sqrt{\frac{2}{e}} \leq 0$$

since it is easy to verify that $\max_x |x \exp(-x^2)| \leq 1/\sqrt{2e}$. This completes the proof for $a \geq b$.

When $a < b$, it suffices to show that $h(x, a) := \exp(-(x-a)^2) + \sqrt{2/e} \cdot a$ is non-decreasing in a . We take the first derivative with respect to a and then get

$$\frac{\partial}{\partial a} h(x, a) = 2(x-a) \exp(-(x-a)^2) + \sqrt{\frac{2}{e}} \geq 0.$$

Thus we are done. \square

Lemma D.4. For any $\mu_1, \mu_2 \in \mathbb{R}$, it holds that $\max_x \mathcal{N}(x | \mu_1, \sigma^2) - \mathcal{N}(x | \mu_2, \sigma^2) \leq \frac{1}{\sigma^2 \sqrt{2\pi e}} \cdot |\mu_1 - \mu_2|$.

Proof of Lemma D.4.

$$\begin{aligned}
 \mathcal{N}(x | \mu_1, \sigma^2) - \mathcal{N}(x | \mu_2, \sigma^2) &= \frac{1}{\sigma \sqrt{2\pi}} \left(\exp\left(-\frac{1}{2} \left(\frac{x - \mu_1}{\sigma}\right)^2\right) - \exp\left(-\frac{1}{2} \left(\frac{x - \mu_2}{\sigma}\right)^2\right) \right) \\
 &\leq \frac{1}{\sigma \sqrt{2\pi}} \cdot \sqrt{\frac{2}{e}} \cdot \left| \frac{\mu_1}{\sigma \sqrt{2}} - \frac{\mu_2}{\sigma \sqrt{2}} \right| = \frac{1}{\sigma^2 \sqrt{2\pi e}} |\mu_1 - \mu_2|
 \end{aligned}$$

where the inequality holds for Lemma D.3. \square

Lemma D.5. For LQR, let \mathcal{M}_i ($i = 1, 2, 3$) denotes the set of possible matrices of M_i . We assume that, there exists parameters m_x and m_a for which $\|x\|_2 \leq m_x$ and $\|a\|_2 \leq m_a$ for all $x \in \mathcal{X}$ and $a \in \mathcal{A}$. Then we have

$$N_{\square}(\epsilon, \mathcal{F}_h, \|\cdot\|_\infty) \leq \prod_{i=1,2,3} N\left(\frac{\epsilon \sigma^2 (H-h+1) \sqrt{2\pi e}}{2(m_x^2 + m_x m_a + m_a^2)}, \mathcal{M}_i, \|\cdot\|_F\right).$$

Here $N_{\square}(\cdot)$ and $N(\cdot)$ denote the bracketing number and covering number, respectively.

Proof of Lemma D.5. We denote by $\widetilde{\mathcal{M}}_1$, $\widetilde{\mathcal{M}}_2$, and $\widetilde{\mathcal{M}}_3$ the ϵ -covers of \mathcal{M}_1 , \mathcal{M}_2 , and \mathcal{M}_3 , respectively. We construct the following function class

$$\widetilde{\mathcal{F}}_h = \left\{ \tilde{f}(\cdot|x, a) = \mathcal{N}(\cdot \mid x^\top \widetilde{M}_1 x + a^\top \widetilde{M}_2 x + a^\top \widetilde{M}_3 a, (H-h+1)\sigma^2), \forall \widetilde{M}_1 \in \widetilde{\mathcal{M}}_1, \widetilde{M}_2 \in \widetilde{\mathcal{M}}_2, \widetilde{M}_3 \in \widetilde{\mathcal{M}}_3 \right\}.$$

We claim that $\widetilde{\mathcal{F}}_h$ is a cover of \mathcal{F}_h . To see this, note that for any $f \in \mathcal{F}_h$, there exists $\tilde{f} \in \widetilde{\mathcal{F}}_h$ ($i = 1, 2, 3$) for which $\|M_i - \widetilde{M}_i\|_F \leq \epsilon$, and thus

$$\begin{aligned} \|\tilde{f} - f\|_\infty &= \max_{x,a,z} \left| \mathcal{N}(z \mid x^\top M_1 x + a^\top M_2 x + a^\top M_3 a, (H-h+1)\sigma^2) \right. \\ &\quad \left. - \mathcal{N}(z \mid x^\top \widetilde{M}_1 x + a^\top \widetilde{M}_2 x + a^\top \widetilde{M}_3 a, (H-h+1)\sigma^2) \right| \\ &\leq \frac{1}{(H-h+1)\sigma^2 \sqrt{2\pi e}} \underbrace{\left| x^\top (M_1 - \widetilde{M}_1) x + a^\top (M_2 - \widetilde{M}_2) x + a^\top (M_3 - \widetilde{M}_3) a \right|}_{(\heartsuit)}. \end{aligned}$$

where the last inequality holds for Lemma D.4. For (\heartsuit) , we have

$$(\heartsuit) \leq \|x\|_2 \|M_1 - \widetilde{M}_1\|_F \|x\|_2 + \|a\|_2 \|M_2 - \widetilde{M}_2\|_F \|x\|_2 + \|a\|_2 \|M_3 - \widetilde{M}_3\|_F \|a\|_2 \leq \epsilon(m_x^2 + m_x m_a + m_a^2).$$

Hence, we have

$$\|\tilde{f} - f\|_\infty \leq \epsilon \cdot \frac{m_x^2 + m_x m_a + m_a^2}{(H-h+1)\sigma^2 \sqrt{2\pi e}}.$$

This implies

$$N\left(\frac{\epsilon(m_x^2 + m_x m_a + m_a^2)}{(H-h+1)\sigma^2 \sqrt{2\pi e}}, \mathcal{F}_h, \|\cdot\|_\infty\right) \leq N(\epsilon, \mathcal{M}_1, \|\cdot\|_F) \cdot N(\epsilon, \mathcal{M}_2, \|\cdot\|_F) \cdot N(\epsilon, \mathcal{M}_3, \|\cdot\|_F).$$

We note that $N_{\square}(2\epsilon, \mathcal{F}_h, \|\cdot\|_\infty) \leq N(\epsilon, \mathcal{F}_h, \|\cdot\|_\infty)$. Hence we complete the proof. \square

Proof of Lemma B.3. Let $\mathcal{M}_i = \{M : \|M\|_F \leq m_i\}$ ($i = 1, 2, 3$) denote the set of possible matrices M_i . Then we have $N(\epsilon, \mathcal{M}_1, \|\cdot\|_F) \leq (3m_1/\epsilon)^{d_x \times d_x}$, $N(\epsilon, \mathcal{M}_2, \|\cdot\|_F) \leq (3m_2/\epsilon)^{d_x \times d_a}$, and $N(\epsilon, \mathcal{M}_3, \|\cdot\|_F) \leq (3m_3/\epsilon)^{d_a \times d_a}$. By Lemma D.5, we have that

$$\begin{aligned} &N_{\square}(\epsilon, \mathcal{F}_h, \|\cdot\|_\infty) \\ &\leq \left(\frac{6m_1(m_x^2 + m_x m_a + m_a^2)}{\epsilon\sigma^2(H-h+1)\sqrt{2\pi e}} \right)^{d_x \times d_x} \left(\frac{6m_2(m_x^2 + m_x m_a + m_a^2)}{\epsilon\sigma^2(H-h+1)\sqrt{2\pi e}} \right)^{d_x \times d_a} \left(\frac{6m_3(m_x^2 + m_x m_a + m_a^2)}{\epsilon\sigma^2(H-h+1)\sqrt{2\pi e}} \right)^{d_a \times d_a} \\ &\leq \left(\frac{6m_1(m_x^2 + m_x m_a + m_a^2)}{\epsilon\sigma^2\sqrt{2\pi e}} \right)^{d_x \times d_x} \left(\frac{6m_2(m_x^2 + m_x m_a + m_a^2)}{\epsilon\sigma^2\sqrt{2\pi e}} \right)^{d_x \times d_a} \left(\frac{6m_3(m_x^2 + m_x m_a + m_a^2)}{\epsilon\sigma^2\sqrt{2\pi e}} \right)^{d_a \times d_a}. \end{aligned}$$

Taking a logarithm on both sides, we get

$$\begin{aligned} &\log N_{\square}(\epsilon, \mathcal{F}_h, \|\cdot\|_\infty) \\ &\leq O\left(d_x^2 \log \frac{m_1(m_x^2 + m_x m_a + m_a^2)}{\epsilon\sigma^2} + d_x d_a \log \frac{m_2(m_x^2 + m_x m_a + m_a^2)}{\epsilon\sigma^2} + d_a^2 \log \frac{m_3(m_x^2 + m_x m_a + m_a^2)}{\epsilon\sigma^2} \right). \end{aligned}$$

\square

E. Experiment Details

We release our code at <https://github.com/ziqian2000/Fitted-Likelihood-Estimation>.

E.1. Implementation Details of Combination Lock Environment

We first clarify our implementation of the combination lock environment.

Reward. We denote r^+ and r^- as the random reward for latent state $w_H = 0$ and $w_H = 1$, respectively. For the one-dimensional case, they are sampled from Gaussian distributions: $r^+ \sim \mathcal{N}(1, 0.1^2)$ and $r^- \sim \mathcal{N}(-1, 0.1^2)$. For the second experiment with two-dimensional reward, they are defined as

$$r^+ = x + \frac{2x}{\|x\|_2} \quad \text{where} \quad x \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.05 & 0 \\ 0 & 0.05 \end{bmatrix}\right), \quad r^- \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.05 & 0 \\ 0 & 0.05 \end{bmatrix}\right).$$

Visually, most samples of r^+ appear in a ring centered at the origin with a radius of 2.

State. The state is constructed by three components, that is, state $x = (x_1, x_2, x_3)^\top$ for which x_1 is the one-hot encoding of latent state, x_2 is the one-hot encoding of time step h , and x_3 is a vector of Gaussian noise sampled independently from $\mathcal{N}(0, 0.1^2)$.

The optimal action a_h^* is chosen to be 0 for all $h \in [H]$ for simplicity. We list other environment hyperparameters in Table 3 for reference.

Table 3. Hyperparameters for the combination lock environment. The two columns denote the respective hyperparameters employed in one-dimensional and two-dimensional experiments.

| | 1-DIMENSIONAL | 2-DIMENSIONAL |
|---------------------|---------------|---------------|
| HORIZON | 20 | 10 |
| NUMBER OF ACTIONS | 2 | 2 |
| DIMENSION OF STATES | 30 | 30 |

E.2. Implementation Details of Algorithms

All algorithms, with the exception of Diff-FLE, is implemented by a neural network consisting of two layers, each with 32 neurons, connected by the ReLU activation functions. Diff-FLE employs a three-layered neural network, each layer containing 256 neurons, connected by the ReLU functions. Some shared hyperparameters are listed in Table 4.

Table 4. Shared hyperparameters. Note that the size of the dataset is written as a product, which is determined by the way we generate the offline data: the first number means the number of samples generated for each latent state and each time step, the second number means the number of time steps (i.e., horizon), and the third number means the size of the latent space.

| | 1-DIMENSIONAL | 2-DIMENSIONAL |
|-----------------|----------------------------|----------------------------|
| SIZE OF DATASET | $10000 \times 20 \times 2$ | $10000 \times 10 \times 2$ |
| BATCH SIZE | 500 | 500 |

Categorical Algorithm. We present the implementation of the two-dimensional version of the categorical algorithm, which is not presented in the prior work (Bellemare et al., 2017). As a reminder, for the one-dimensional counterpart, for each atom of the next state, we first calculate its target position, then distribute the probability of that atom based on the distance of the target position to the closest two atoms. In the two-dimensional case, we discretize on each dimension, resulting in a grid-shaped discretization. Therefore, the probability of the atoms of the next state will be distributed based on the distance to the *four* closest atoms (generally, it will be distributed to 2^n atoms in the n -dimensional case). The other implementation details are the same as the one-dimensional case. The list of hyperparameters can be found in the Table 5.

Quantile Algorithm. We followed the implementation of Dabney et al. (2018). The list of hyperparameters can be found in the Table 6.

Diff-FLE. Our implementation is based on DDPM (Ho et al., 2020). However, our neural network is much simpler than theirs, as mentioned above. The list of hyperparameters can be found in the Table 7.

Table 5. Hyperparameters for the categorical algorithm.

| | 1-DIMENSIONAL | 2-DIMENSIONAL |
|----------------------|---------------|--------------------|
| NUMBER OF ATOMS | 100 | 30^2 |
| LEARNING RATE | 10^{-2} | 3×10^{-2} |
| NUMBER OF ITERATIONS | 200 | 100 |
| DISCRETIZED RANGE | $[-1.5, 1.5]$ | $[-4, 4]^2$ |

Table 6. Hyperparameters for quantile Algorithm.

| | 1-DIMENSIONAL |
|----------------------|---------------|
| NUMBER OF QUANTILES | 100 |
| LEARNING RATE | 10^{-3} |
| NUMBER OF ITERATIONS | 1000 |

GMM-FLE. For the training of GMM-FLE, we applied gradient ascent on the log-likelihood. While many classic approaches (e.g., the Expectation-Maximization (EM) algorithm) exist, we found no significant performance gap between gradient ascent and EM in our trials on both one-dimensional and two-dimensional data. Therefore, we opted for the gradient ascent, which matches our theory better. The list of hyperparameters is listed in Table 8.

E.3. Full Experiment Results

Table 9 is the full version of Table 1, and Table 10 is the full version of Table 2.

Table 7. Hyperparameters for Diff-FLE.

| | 1-DIMENSIONAL | 2-DIMENSIONAL |
|----------------------------|---------------|---------------|
| STEPS OF DIFFUSION PROCESS | 200 | 200 |
| STARTING VARIANCE | 10^{-3} | 10^{-3} |
| FINAL VARIANCE | 0.1 | 0.1 |
| VARIANCE INCREASING | LINEAR | LINEAR |
| LEARNING RATE | 10^{-3} | 10^{-3} |
| NUMBER OF ITERATIONS | 5000 | 15000 |

Table 8. Hyperparameters for GMM-FLE.

| | 1-DIMENSIONAL | 2-DIMENSIONAL |
|---------------------------------|---------------|--------------------|
| NUMBER OF GAUSSIAN DISTRIBUTION | 10 | 10 |
| LEARNING RATE | 10^{-4} | 2×10^{-4} |
| NUMBER OF ITERATIONS | 20000 | 10000 |

Table 9. Full version of Table 1.

| h | CATE ALG | QUAN ALG | DIFF-FLE | GMM-FLE |
|-----|-------------------|-------------------|-------------------|-------------------|
| 1 | 0.071 ± 0.015 | 0.603 ± 0.011 | 0.292 ± 0.073 | 0.039 ± 0.004 |
| 2 | 0.067 ± 0.012 | 0.609 ± 0.014 | 0.305 ± 0.055 | 0.041 ± 0.005 |
| 3 | 0.068 ± 0.013 | 0.612 ± 0.017 | 0.305 ± 0.079 | 0.039 ± 0.009 |
| 4 | 0.073 ± 0.013 | 0.593 ± 0.015 | 0.288 ± 0.073 | 0.038 ± 0.003 |
| 5 | 0.074 ± 0.015 | 0.602 ± 0.009 | 0.285 ± 0.054 | 0.036 ± 0.009 |
| 6 | 0.077 ± 0.011 | 0.612 ± 0.010 | 0.268 ± 0.040 | 0.030 ± 0.008 |
| 7 | 0.080 ± 0.014 | 0.602 ± 0.014 | 0.290 ± 0.066 | 0.034 ± 0.004 |
| 8 | 0.080 ± 0.016 | 0.584 ± 0.018 | 0.273 ± 0.039 | 0.039 ± 0.013 |
| 9 | 0.081 ± 0.019 | 0.529 ± 0.028 | 0.247 ± 0.034 | 0.048 ± 0.010 |
| 10 | 0.079 ± 0.017 | 0.494 ± 0.018 | 0.234 ± 0.043 | 0.044 ± 0.012 |
| 11 | 0.080 ± 0.016 | 0.514 ± 0.018 | 0.244 ± 0.038 | 0.039 ± 0.012 |
| 12 | 0.089 ± 0.009 | 0.518 ± 0.013 | 0.232 ± 0.015 | 0.032 ± 0.007 |
| 13 | 0.089 ± 0.011 | 0.481 ± 0.016 | 0.219 ± 0.027 | 0.029 ± 0.016 |
| 14 | 0.081 ± 0.015 | 0.416 ± 0.026 | 0.221 ± 0.021 | 0.033 ± 0.012 |
| 15 | 0.083 ± 0.015 | 0.330 ± 0.028 | 0.178 ± 0.033 | 0.026 ± 0.015 |
| 16 | 0.081 ± 0.009 | 0.283 ± 0.017 | 0.170 ± 0.045 | 0.027 ± 0.013 |
| 17 | 0.082 ± 0.008 | 0.252 ± 0.008 | 0.167 ± 0.037 | 0.034 ± 0.013 |
| 18 | 0.070 ± 0.010 | 0.217 ± 0.012 | 0.133 ± 0.019 | 0.023 ± 0.008 |
| 19 | 0.078 ± 0.011 | 0.167 ± 0.019 | 0.109 ± 0.031 | 0.018 ± 0.008 |
| 20 | 0.077 ± 0.014 | 0.076 ± 0.009 | 0.067 ± 0.024 | 0.013 ± 0.005 |

Table 10. Full version of Table 2.

| h | CATE ALG | DIFF-FLE | GMM-FLE |
|-----|-------------------|-------------------|-------------------|
| 1 | 0.483 ± 0.003 | 0.357 ± 0.031 | 0.438 ± 0.008 |
| 2 | 0.483 ± 0.003 | 0.344 ± 0.030 | 0.424 ± 0.048 |
| 3 | 0.480 ± 0.003 | 0.339 ± 0.023 | 0.450 ± 0.042 |
| 4 | 0.469 ± 0.002 | 0.327 ± 0.019 | 0.478 ± 0.048 |
| 5 | 0.466 ± 0.001 | 0.310 ± 0.019 | 0.493 ± 0.050 |
| 6 | 0.466 ± 0.001 | 0.289 ± 0.031 | 0.491 ± 0.061 |
| 7 | 0.470 ± 0.003 | 0.256 ± 0.032 | 0.510 ± 0.080 |
| 8 | 0.465 ± 0.002 | 0.234 ± 0.023 | 0.505 ± 0.099 |
| 9 | 0.453 ± 0.001 | 0.207 ± 0.014 | 0.502 ± 0.094 |
| 10 | 0.446 ± 0.002 | 0.143 ± 0.011 | 0.376 ± 0.101 |