Policy Evaluation for Reinforcement Learning from Human Feedback: A Sample Complexity Analysis

Zihao Li
Princeton University

Xiang Ji
Princeton University

Minshuo Chen
Princeton University

Mengdi Wang
Princeton University

Abstract

A recently popular approach to solving reinforcement learning is with data from human preferences. In fact, human preference data are now used with classic reinforcement learning algorithms such as actor-critic methods, which involve evaluating an intermediate policy over a reward learned from human preference data with distribution shift, known as off-policy evaluation (OPE). Such algorithm includes (i) learning reward function from human preference dataset, and (ii) learning expected cumulative reward of a target policy. Despite the huge empirical success, existing OPE methods with preference data often lack theoretical understandings and rely heavily on heuristics. In this paper, we study the sample efficiency of OPE with human preference and establish a statistical guarantee for it. Specifically, we approach OPE with learning the value function by fitted-Q-evaluation with a deep neural network. By appropriately selecting the size of a ReLU network, we show that one can leverage any low-dimensional manifold structure in the Markov decision process and obtain a sample-efficient estimator without suffering from the curse of high data ambient dimensionality. Under the assumption of high reward smoothness, our results *almost align with the classical OPE results with observable reward data*. To the best of our knowledge, this is the first result that establishes a *provably efficient* guarantee for off-policy evaluation with RLHF.

1 Introduction

Reinforcement Learning with Human Feedback (RLHF) is a popular approach to solve reinforcement learning problems using data based on human feedback. In recent years, RLHF has achieved significant success in large language models, clinical trials, auto-driving, robotics, etc. (Ouyang et al., 2022; Gao et al., 2022; Glaese et al., 2022; Hussein et al., 2017; Jain et al., 2013; Kupcsik et al., 2018; Menick et al., 2022; Nakano et al., 2021; Novoseller et al., 2020; Chakraborty et al., 2024). In conventional reinforcement learning, the goal of the learner is to find the optimal policy using data containing reward observations; in RLHF, the learner does not have direct access to the reward signal but instead can only receive human preferences over actions or trajectories, which are assumed to be made in line with the actual reward. RLHF approaches are popular because it is believed that humans are accurate at comparing and ranking a small number of items based on their preference, and such preferences can be modeled with classic human choice models such as the Plackett-Luce (PL) ranking model (Bradley & Terry, 1952).

One popular way to do RLHF is by first constructing a reward function based on human preference data and then using this reward in a classic reinforcement learning algorithm (Christiano et al., 2017; Ibarz et al., 2018). A prominent example is RLHF with actor-critic methods, which are commonly used in robotics applications (Lee et al., 2021b; Liu et al., 2022; Liang et al., 2022). In such methods, the actor is responsible for learning and improving the policy of the agent so that the agent's policy can converge to the optimal policy; meanwhile, the critic evaluates the value function of the current policy in an off-policy manner in order to guide the actor to improve the current policy in the following iteration. In such an algorithm, the intermediate policy that the critic needs to evaluate could be arbitrary and non-optimal, and the data it uses during the evaluation are generated by the policy from previous steps, which is different from the one to be evaluated. Such evaluation is called off-policy, and it

is crucial to study its error in order to understand the algorithm's convergence.

In fact, off-policy evaluation (OPE) can be formulated more generally beyond the actor-critic setting and is itself an important problem in reinforcement learning. When on-policy Monte Carlo rollout is time-consuming, costly, or risky, OPE becomes necessary and has therefore found many applications such as healthcare (Murphy et al., 2001), recommendation system (Chapelle & Li, 2011), education (Mandel et al., 2014), robotics (Levine et al., 2020) and dialogue system (Jiang et al., 2021). Due to the mismatch between the visitation distributions of the behavior and target policies, evaluation in the off-policy setting is entirely different from the on-policy one and can be much more challenging and delicate. Most existing studies of OPE focus on the standard setting and rely on direct reward observations in the dataset. In contrast, OPE with preference data has been barely studied.

There has been a recent line of work that theoretically studies RLHF in the offline setting. For learning the optimal policy, Zhu et al. (2023) proposed the pessimistic MLE algorithm for stepwise preferences and analyzed its suboptimality in the linear function approximation case. Zhan et al. (2023) studied the same problem under trajectory-wise preferences in the general function approximation setting. Li et al. (2023) studied offline RLHF under the dynamic choice model. However, these works all focus on the policy learning problem, and there has yet been any result that studies RLHF in the unrealizable setting, where the function approximation cannot perfectly express the reward and value function and has to incur an approximation error. In this paper, we study the estimation error of OPE with human preference data. To stay close to the practice, we consider the function approximation with deep neural networks. We seek to delineate how things like the distribution mismatch and network size can affect the final estimation error. Our challenges are three-folded: (i) The reward itself is not directly observable and needs to be learned from the human preference data; (ii) our reward learning is subject to error from distribution mismatch and unrealizable function approximation. This error is propagated to the later policy evaluation and needs to be carefully controlled. (iii) In many applications, we need to handle a high-dimensional state-action space, in which deep neural networks might need a lot of samples to operate.

Our results. In this work, we address these challenges by analyzing the Fitted Q-Evaluation with Human Preferences algorithm, which is based on Maximum Likelihood Estimation and Fitted Q-evaluation. Specifically, our contributions are as follows:

- We provide an upper bound for the estimation error of OPE with human preferences and show how distribution shift, intrinsic dimension, underlying model class and sample size of the human preference data can affect the OPE error. Our analysis shows that the error from the reward learning part using human preferences propagates only in a moderate way and does not bottleneck the final estimation error.

- We consider function approximation with deep neural networks for nonparametric reward and Q-functions. In this setting, we prove that by choosing hyperparameters appropriately, deep ReLU network can adapt to the intrinsic low-dimension structure of the preference data and keep the estimation error mostly independent from the the ambient dimension.

Notably, our result matches the existing error bounds for standard OPE and demonstrates OPE with human preferences can be as sample-efficient, even in the absence of an observable reward. This provides a theoretical support for the actor-critic methods with RLHF in practice (Lee et al., 2021b; Liang et al., 2022), which essentially solve OPE with human preference data. To the best of our knowledge, this paper offer the first theoretical guarantee for off-policy evaluation with human preference.

1.1 Related Work

RLHF for policy learning. In recent years RLHF and inverse reinforcement learning (IRL) has been widely applied to robotics, recommendation system, and large language model (Ouyang et al., 2022; Lindner et al., 2022; Menick et al., 2022; Jaques et al., 2020; Lee et al., 2021b; Nakano et al., 2021; Chandrasekaran & Mainampati, 2021). Policy optimization applied to train text generation systems with pre-specified rewards Cho et al. (2018) and ranking systems Zhang et al. (2023). To model human choice behavior, (Shah et al., 2015; Ouyang et al., 2022; Saha & Krishnamurthy, 2022) learn reward from pairwise comparison and ranking. Pacchiano et al. (2021) study pairwise comparison with function approximation in pairwise comparison. To learn the optimal policy, most of them implement actor-critic-type algorithms such as PPO or SAC. From a theoretical view, Zhu et al. (2023); Zhan et al. (2020) study various cases of preference-based comparison in trajectory bandit problems with linear function approximation and establish a theoretical guarantee for their algorithm. Li et al. (2023) study RLHF under human feedback setting. However, none of these theoretical works have studied policy evaluation or a deep neural network setting.

Off-policy evaluation. Except for the FQE studied in this paper, there have been various methods to do off-policy evaluation. One popular type is using importance sampling to reweigh samples by the distribution shift ratio (Precup, 2000), but importance sampling suffers from large variance, especially when the sampling policy is unknown, which causes an error with an exponential term in the length of the horizon in the worst case. To address this issue, some variants with reduced variance such as marginal importance sampling (MIS) (Xie et al., 2019) and doubly robust estimation (Jiang & Li, 2016; Thomas & Brunskill, 2016; Uehara et al., 2020) have been developed. For the tabular setting with complete data coverage, (Yin & Wang, 2020) show that MIS is an asymptotically efficient OPE estimator, which matches the Cramer-Rao lower bound in Jiang & Li (2016). Duan & Wang (2020) studies OPE with linear function approximation. Zhang et al. (2022) studies OPE with differentiable function approximation. Moreover, a line of work Zhan et al. (2022); Lee et al. (2021a); Nachum et al. (2019); Nachum & Dai (2020) focuses on policy evaluation using MIS and linear programming, without function approximation.

Deep neural network in reinforcement learning. In recent years, deep learning has enjoyed tremendous empirical success in RL applications(Mnih et al., 2013; Silver et al., 2016; Gu et al., 2016; Mao et al., 2016). In RLHF, deep neural networks are also crucial when the state-action space is immense or continuous, such as Ouyang et al. (2022); Stiennon et al. (2020); Wu et al. (2021). Specifically, the popular GPT 3.5 is fine-tuned by first learning reward function from the human feedback dataset, and then implementing PPO to learn the optimal policy. Recently, there have been some theoretical studies of deep RL. Fan et al. (2020) focuses on the online policy learning problem and studies DQN with feed-forward ReLU network; Yin et al. (2022) studies offline policy learning with realizable, general differentiable function approximation. Notably, a recent study Ji et al. (2022) provides an analysis of the estimation error of nonparametric FQE using a feed-forward CNN network. However, all these works study deep RL with an observable reward given by the environment, which lies in the standard RL regime. To the best of our knowledge, there has not been any result on the theoretical guarantee one can achieve with deep RL when the data are human preferences.

Notation For a scalar $a > 0$, $\lceil a \rceil$ denotes the ceiling function, which gives the smallest integer that is no less than a; $\lfloor a \rfloor$ denotes the floor function, which gives the largest integer that is no larger than a. For any scalars a and b, $a \vee b$ denotes $\max(a, b)$ and $a \wedge b$ denotes $\min(a, b)$. For a vector or a matrix, $\| \cdot \|_0$ de-

notes the number of nonzero entries and $\|\cdot\|_\infty$ denotes the maximum magnitude of entries. Given a function $f : \mathbb{R}^D \to \mathbb{R}$ and a multi-index $s = [s_1, \cdots, s_D]^\top$, $\partial^s f$ denotes $\frac{\partial^{|s|} f}{\partial x_1^{s_1} \cdots \partial x_D^{s_D}}$. $\|f\|_{L^p}$ denote the L^p norm of function f. For an arbitrary space $\mathcal{X}$, we use $\Delta(\mathcal{X})$ to denote the set of all probability distribution on $\mathcal{X}$. For two vectors $x, y \in \mathbb{R}^d$, we denote $x \cdot y = \sum_i^d x_i y_i$ as the inner product of x, y. We denote the set of all probability measures on $\mathcal{X}$ as $\Delta(\mathcal{X})$. We use $[n]$ to represent the set of integers from 0 to $n-1$. For every set $\mathcal{M} \subset \mathcal{X}$ for metric space $\mathcal{X}$, we define its ϵ-covering number with respect to norm $\|\cdot\|$ by $N(\mathcal{M}, \|\cdot\|, \epsilon)$. We adopt the convention $0/0 = 0$. Given distributions p and q, if p is absolutely continuous with respect to q, the Pearson χ^2-divergence is defined as $\chi^2(p, q) := \mathbb{E}_q \left[\left(\frac{dp}{dq} - 1 \right)^2 \right]$.

2 Preliminaries

2.1 Markov Decision Process

We define a finite-horizon MDP model $M = (\mathcal{S}, \mathcal{A}, H, \{P_h\}_{h \in [H]}, \{r_h\}_{h \in [H]})$, H is the horizon length, in each step $h \in [H]$, the agent starts from state s_h in the state space $\mathcal{S}$, chooses an action $a_h \in \mathcal{A}$ with probability $\pi_h(a_h \mid s_h)$, receives a reward of $r_h(s_h, a_h)$ and transits to the next state s' with probability $P_h(s' \mid s_h, a_h)$. Here $\mathcal{A}$ is a finite action set with $|\mathcal{A}|$ actions and $P_h(\cdot|s_h, a_h) \in \Delta(s_h, a_h)$ is the transition kernel condition on state action pair (s, a). For convenience we assume that $r_h(s, a) \in [0, 1]$ for all $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$. Without loss of generality, we assume that the initial state of each episode s_0 is fixed. Note that this will not add difficulty to our analysis. For any policy $\pi = \{\pi_h\}_{h \in [H]}$ the state value function is

$$V_h^\pi(s) = \mathbb{E}_\pi \left[\sum_{t=h}^H r_t(s_t, a_t) \mid s_h = s \right],$$

and the action value function is

$$Q_h^\pi(s, a) = \mathbb{E}_\pi \left[\sum_{t=h}^H r_t(s_t, a_t) \mid s_h = s, a_h = a \right],$$

here the expectation $\mathbb{E}_\pi$ is taken with respect to the randomness of the trajectory induced by π, i.e. is obtained by taking action $a_t \sim \pi_t(\cdot \mid s_t)$ and observing $s_{t+1} \sim P_h(\cdot \mid s_t, a_t)$. For any function $f : \mathcal{S} \to \mathbb{R}$, we define the transition operator:

$$\mathcal{P}_h^\pi f(s, a)$$
$$= \mathbb{E}[f(s_{h+1}, a_{h+1}) \pi(a_{h+1} \mid s_{h+1}) \mid s_h = s, a_h = a].$$

We also denote the Bellman operator at time h under policy π as $\mathcal{T}_h^\pi$:

$$\mathcal{T}_h^\pi f(s, a) := r_h(s, a) + \mathcal{P}_h^\pi f(s, a).$$

For a given policy $\pi = \{\pi_h\}_{h \in [H]}$, we use $q_h^\pi(s, a)$ to denote the state-acton distribution induced by π, i.e. the probability of visiting (s, a) on step h when the policy is π.

2.2 Riemannian Manifold

Real-world data often exhibit low-dimensional geometric structures and can be viewed as samples near a low-dimensional manifold. In this section, we briefly introduce the concept of Riemannian manifolds. Let $\mathcal{M}$ be a d-dimensional Riemannian manifold isometrically embedded in $\mathbb{R}^D$. A chart for $\mathcal{M}$ is a pair (U, ϕ) such that $U \subset \mathcal{M}$ is open and $\phi : U \to \mathbb{R}^d$ is a homeomorphism, i.e., ϕ is a bijection, its inverse and itself are continuous. Two charts (U, ϕ) and (V, ψ) are called $\mathcal{C}^k$ compatible if and only if

$$\phi \circ \psi^{-1} : \psi(U \cap V) \to \phi(U \cap V)$$
$$\text{and} \quad \psi \circ \phi^{-1} : \phi(U \cap V) \to \psi(U \cap V)$$

are both $\mathcal{C}^k$ functions (k-th order continuously differentiable). A $\mathcal{C}^k$ atlas of $\mathcal{M}$ is a collection of $\mathcal{C}^k$ compatible charts $\{(U_i, \phi_i)\}$ such that $\bigcup_i U_i = \mathcal{M}$. An atlas of $\mathcal{M}$ contains an open cover of $\mathcal{M}$ and mappings from each open cover to $\mathbb{R}^d$. We call a manifold *smooth* if it has a $\mathcal{C}^\infty$ atlas.

We also define the reach (Niyogi et al., 2006; Federer, 1959) of a manifold to characterize the curvature of $\mathcal{M}$.

Definition 2.1 (Reach, Definition 2.1 in Aamari et al. (2019)). *The medial axis of $\mathcal{M}$ is defined as $\overline{\mathcal{T}}(\mathcal{M})$, which is the closure of*

$$\mathcal{T}(\mathcal{M}) = \left\{ x \in \mathbb{R}^D \mid \exists x_1 \neq x_2 \in \mathcal{M} \text{ such that} \right.$$
$$\left. \|x - x_1\|_2 = \|x - x_2\|_2 = \inf_{y \in \mathcal{M}} \|x - y\|_2 \right\}.$$

The reach ω of $\mathcal{M}$ is the minimum distance between $\mathcal{M}$ and $\overline{\mathcal{T}}(\mathcal{M})$, i.e.

$$\omega = \inf_{x \in \overline{\mathcal{T}}(\mathcal{M}), y \in \mathcal{M}} \|x - y\|_2.$$

Roughly speaking, reach measures how fast a manifold "bends" in the embedded Euclidean space. A manifold with a large reach "bends" relatively slowly. On the contrary, a small ω signifies more complicated local geometric structures, which are possibly hard to fully capture.

2.3 Hölder Functions on Manifolds

To capture the smoothness of the Bellman operator and the human preference reward function, we define a Hölder function on a smooth manifold $\mathcal{M}$.

Definition 2.2 (Hölder function). *Let $\mathcal{M}$ be a d-dimensional compact Riemannian manifold isometrically embedded in $\mathbb{R}^D$. Let $\{(U_i, P_i)\}_{i \in \mathcal{A}}$ be an atlas of $\mathcal{M}$ where the P_i 's are orthogonal projections onto tangent spaces. For a positive integer k and $l \in (0, 1]$ such that $k + l = \alpha$, a function $f : \mathcal{M} \mapsto \mathbb{R}$ is α-Hölder continuous if for each chart (U_i, P_i) in the atlas, we have*

(i) $f \circ P_i^{-1} \in C^s$ with $\left| D^{\mathbf{k}} \left(f \circ P_i^{-1} \right) \right| \leq 1$ for any $|\mathbf{k}| \leq k, x \in U_i$;

(ii) for any $|\mathbf{k}| = k$ and $x_1, x_2 \in U_i$,

$$\left| D^{\mathbf{k}} \left(f \circ P_i^{-1} \right) \right|_{P_i(x_1)} - D^{\mathbf{k}} \left(f \circ P_i^{-1} \right) \right|_{P_i(x_2)} |$$
$$\leq \| P_i(x_1) - P_i(x_2) \|_2^l.$$

Moreover, we denote the collection of α-Hölder functions on $\mathcal{M}$ as $\mathcal{H}^\alpha(\mathcal{M})$. For a multivariate function, we say it is α-Hölder if it is α-Hölder in every dimension.

Definition 2.2 requires that all s-th order derivatives of $f \circ P_i^{-1}$ are Hölder continuous. We recover the standard Hölder class on an Euclidean space if P_i is the identity mapping.

2.4 ReLU Neural Network Class

We consider a rectified linear unit (ReLU) neural network as our function class. Given an input $x \in \mathbb{R}^D$, an L-layer ReLU neural network computes an output as

$$f(x) = W_L \cdot \text{ReLU}(W_{L-1} \cdots \text{ReLU}(W_1 x + b_1) \cdots + b_{L-1}) + b_L, \quad (1)$$

where $W_1, \ldots, W_L$ and $b_1, \ldots, b_L$ are weight matrices and vectors of proper sizes. Here $\text{ReLU}(x) = \max\{x, 0\}$.

We denote $\mathcal{E}(R, \tau, L, p, I)$ as a class of neural networks with bounded weight parameters and bounded output:

$$\mathcal{E}(R, \tau, L, p, I) = \left\{ f \mid f(x) \text{ in the form of equation 1} \right.$$

with L-layers and width bounded by p, and satisfies $\|f\|_\infty \leq R, \|W_i\|_{\infty,\infty} \leq \tau, \|b_i\|_\infty \leq \tau$ for $i = 1, \ldots, L$,

$$\left. \sum_{i=1}^L \|W_i\|_0 + \|b_i\|_0 \leq I \right\}. \quad (2)$$

3 Preference-Based Fitted Q-Evaluation

We consider the *preference-based* nonparametric off-policy evaluation (OPE) problem in a finite-horizon

time-inhomogeneous MDP. The transition kernel $\{P_h\}_{h \in [H]}$ and reward function $\{r_h\}_{h \in [H]}$ are both unknown to us. Specifically, our objective is to evaluate the value of π from a fixed initial distribution ξ over horizon H, given by

$$v^\pi := \mathbb{E}\bigg[\sum_{h=1}^{H} r_h(s_h, a_h) \mid s_1 \sim \xi\bigg],$$

where $a_h \sim \pi(\cdot|s_h)$ and $s_{h+1} \sim P_h(s_{h+1}|s_h, a_h)$. We are given two datasets for every timestep $h \in [H]$. To learn the transition information in the MDP, we are given a dataset containing samples of state transitions $\mathcal{D}_h := \{(s_{h,k}, a_{h,k}, s'_{h,k})\}_{k=1}^{K}$, which is i.i.d. sampled by an unknown behavior policy π_0. We also assume a given dataset containing human preference, $\mathcal{D}_h^{HF} := \{(\tilde{s}_{h,k}^{(1)}, \tilde{a}_{h,k}^{(1)}), (\tilde{s}_{h,k}^{(2)}, \tilde{a}_{h,k}^{(2)}), y_{h,k}\}_{k=1}^{K_{HF}}$, in which for each $h \in [H]$, $\{(\tilde{s}_{h,k}^{(1)}, \tilde{a}_{h,k}^{(1)}), (\tilde{s}_{h,k}^{(2)}, \tilde{a}_{h,k}^{(2)})\}_{k=1}^{K_{HF}}$ are i.i.d. samples generated from a fixed distribution η_h, which could come from a pretrained model (Ouyang et al., 2022) or a reward-free exploration algorithm (Lee et al., 2021b), $y_{h,k}$ is the human preference made at episode k that will be clarified in the following. We assume that human preference among state-action pairs is generated under the following model.

Assumption 3.1 (Human preference model)**.** *For every $h \in [H]$, given arbitrary M state-action pairs $\{s_h^{(m)}, a_h^{(m)}\}_{m \in [M]}$, we assume that the human prefers y_h according to the following model:*

$$\mathbb{P}\big(y_h = (s_h^{(i)}, a_h^{(i)}) \mid (s_h^{(m)}, a_h^{(m)})_{m \in [M]}\big) \propto \exp(r_h(s_h^{(i)}, a_h^{(i)})),$$
$$\tag{3}$$

where $i \in [M]$.

Such assumption has been widely adopted to model human choice behavior in RL with human preference, e.g. Sharma et al. (2017); Wulfmeier et al. (2016); Ziebart et al. (2008); Zhou et al. (2020); Ouyang et al. (2022); Zhu et al. (2023). Note that under Assumption 3.1, shifting the reward function by a constant for all (s, a) simultaneously doesn't affect human preference probability, thus reward function is not identifiable. For identifiability of the reward, without loss of generality, we also need to make the following assumption.

Assumption 3.2 (Reward identifiability)**.** *We assume that there exists a fixed state-action pair (s_h^f, a_h^f) for every $h \in [H]$, such that $r_h(s^f, a^f) = 0$. Moreover, we assume that the human agent can prefer (s^f, a^f) for all $(h, k) \in [H] \times [K]$.*

With Assumption 3.2, we further assume that $y_{h,k}$, the human preference at step k, is chosen among $\{(\tilde{s}_{h,k}^{(1)}, \tilde{a}_{h,k}^{(1)}), (\tilde{s}_{h,k}^{(2)}, \tilde{a}_{h,k}^{(2)}), (s^f, a^f)\}$, generated independently according to Assumption 3.1 with $M = 3$. Assumption 3.2 eliminates arbitrary constant shifting in

the reward so that we can be determined by a unique reward function from Assumption 3.1. Such assumption is akin to empirical works such as Christiano et al. (2017); Stiennon et al. (2020), in which (s^f, a^f) is interpreted as human rater making an error, and also widely accepted by choice models literature, in which a fixed option is assumed to be of zero-utility, e.g. Adusumilli & Eckardt (2019); Buchholz et al. (2021). We are now ready to present our algorithm.

Algorithm 1 FQE with human preferences

Require: Datasets $\{\mathcal{D}_h^{HF}\}_{h \in [H]}, \{\mathcal{D}_h\}_{h \in [H]}$, neural network classes $\mathcal{F}(R, \tau, L, p, I)$ and $\mathcal{G}(\tilde{R}, \tilde{\tau}, \tilde{L}, \tilde{p}, \tilde{I})$ defined by equation 2, target policy $\pi = \{\pi_h\}_{h \in [H]}$

1: **for** step $h = H, \ldots, 1$ **do**
2:     Learn $\bar{r}_h$ with MLE :

$$r_h^l = \arg\max_{g \in \mathcal{G}_h} \sum_{k=1}^{K} \bigg\{ f(y_{h,k})$$
$$- \log\bigg(\exp(f(s^f, a^f) + \sum_{i=1}^{2} \exp\big(f(s_{h,k}^{(i)}, a_{h,k}^{(i)}))\big)\bigg) \bigg\}.$$

3:     Normalize the reward across states and set $\widehat{r}_h(s, a) = r_h^l(s, a) - r_h^l(s^f, a^f)$.
4:     Update $\widehat{Q}_h^\pi = \widehat{\mathcal{T}}_h^\pi \widehat{Q}_{h+1}^\pi$ by equation 4.
5: **end for**
6: Compute policy value with the estimated Q-function:

$$\widehat{v}^\pi = \int_{\mathcal{X}} \widehat{Q}_1^\pi(s, a)\xi(s)\pi_1(a \mid s)d(s, a).$$

7: **Output:** $\widehat{v}^\pi$

The pseudocode for our algorithm is presented in Algorithm 1. For every $h \in [H]$, our algorithm consists of two stages: (i) We learn the reward by a *maximum likelihood estimation* method. Specifically, we use a deep ReLU function class $\mathcal{G}(\tilde{R}, \tilde{\tau}, \tilde{L}, \tilde{p}, \tilde{I})$ to do the function approximation for all $h \in [H]$. By Assumption 3.2, we align the estimated reward across states by Step 3 to ensure that $\widehat{r}_h(s^f, a^f) = 0$. Our estimation for reward aligns with the line of RLHF empirical studies including Ouyang et al. (2022); Lee et al. (2021b); Stiennon et al. (2020); Wu et al. (2021), in which a reward function is learned from a human preference dataset assuming an underlying softmax human policy; (ii) To estimate v^π, we use *neural FQE* to estimate Q_h^π in a backward, recursive fashion, with $\mathcal{F}(R, \tau, L, p, I)$ as the neural network class for all $h \in [H]$. $\widehat{Q}_h^\pi$, our estimate at step h, is set to be

$\widehat{T}_h^\pi \widehat{Q}_{h+1}^\pi$, whose update rule is based on the Bellman equation:

$$
\widehat{\mathcal{T}}_h^\pi \widehat{Q}_{h+1}^\pi = \arg\min_{f \in \mathcal{F}} \sum_{k=1}^K \Bigg(f(s_{h,k}, a_{h,k}) - \widehat{r}_h(s_{h,k}, a_{h,k})
$$
$$
- \int_{\mathcal{A}} \widehat{Q}_{h+1}^\pi(s'_{h,k}, a) \pi_h(a \mid s'_{h,k}) da \Bigg)^2,
\tag{4}
$$

where $\widehat{\mathcal{T}}_h^\pi$ is an intermediate estimate of the Bellman operator $\mathcal{T}_h^\pi$, $\widehat{Q}_{h+1}^\pi$ is an intermediate estimate of Q_{h+1}^π.

4   Main Results

In this section, we prove an upper bound on the estimation error of v^π by Algorithm 1. First, let us state two assumptions on the MDP.

Assumption 4.1 (Low-dimensional state-action space). *The state-action space $\mathcal{X}$ is a d-dimensional closed Riemannian manifold, which is embedded isometrically in Euclidean space R^D. We also assume that there exists constant $B > 0$ such that $\|x\|_\infty \leq B$ for any $x \in \mathcal{X}$. The reach of $\mathcal{X}$ is $\omega > 0$.*

Assumption 4.1 characterizes the low intrinsic dimension of the MDP embedded in high dimensions. We consider the case that the "intrinsic dimension" of $\mathcal{X}$ is $d \ll D$. Such a setting is common in practice because the representation or features people use are often redundant compared to the latent structures of the problem. For instance, images of a dynamical system are widely believed to admit such low-dimensional latent structures Gong et al. (2019); Hinton & Salakhutdinov (2006); Pope et al. (2021); Osher et al. (2017). Common pre-trained language models can also have a low intrinsic dimension as shown in recent empirical studies (Aghajanyan et al., 2020; Hu et al., 2021).

Assumption 4.2 (Bellman completeness). *Under target policy π, for any time step h and any $f \in \mathcal{F}$, we have $\mathcal{P}_h^\pi f \in \mathcal{H}^\alpha(\mathcal{M})$, where $0 < \alpha \leq +\infty$.*

Assumption 4.2 is about the closure of a function class under the Bellman operator, which is widely adopted in RL literature Ji et al. (2022); Chen & Jiang (2019b); Yang et al. (2020); Du et al. (2020). In fact, some classic MDP settings also have this property implicitly, e.g., linear MDP (Jin et al., 2020) or RKHS MDP (Yang et al., 2020). Such an assumption is necessary for the Bellman residual: without such an assumption, Wang et al. (2020) provides a lower bound that the sample complexity is exponential in the horizon for guaranteeing a constant error in OPE.

Assumption 4.3. *We assume that the human preference reward function can be expressed as $r_h(s, a) =$*

$f_h(\psi_h(s, a))$ *for every $h \in [H]$, where $\psi_h : \mathcal{X} \to [0, B']^{\tilde{d}}$ and $f : [0, B']^{\tilde{d}} \to [0, 1]$. Furthermore, $\psi_h \in \mathcal{H}^{\tilde{\alpha}}(\mathcal{X})$ and $f \in \mathcal{H}^{\tilde{\alpha}}([0, B']^{\tilde{d}})$, where $0 < \tilde{\alpha} \leq +\infty$.*
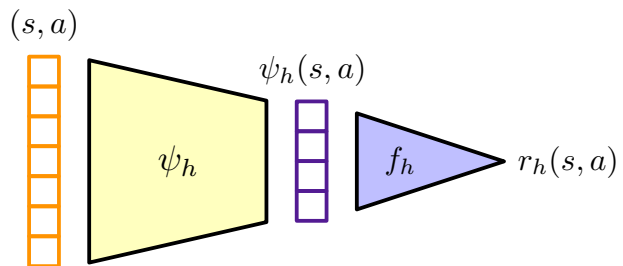

Figure 1: Reward function with low-dim feature.

In Assumption 4.3, we assume that the reward function admits an underlying feature ψ_h, which is sufficiently smooth and maps the input to a space with lower dimension $\tilde{d}$. Empirically, many studies have shown that learning human preferences requires only a smaller dataset and architecture, which implies that such preference rewards admit a low dimension latent feature, see e.g. Christiano et al. (2017); Wang et al. (2022); Stiennon et al. (2020); Chandrasekaran & Mainampati (2021). We summarize our results for reward learning in Theorem 4.4:

Theorem 4.4 (Reward convergence rate). *With Assumption 4.3, for every $0 < \epsilon_0 < 1$, if the ReLU class $\mathcal{G}_h(\tilde{R}, \tilde{\tau}, \tilde{L}, \tilde{p}, \tilde{I})$ satisfies*

$$
\tilde{L} = O\left(\log \frac{1}{\epsilon_0} \right), \tilde{p} = O\left(\epsilon_0^{-\frac{2\tilde{d}}{\tilde{\alpha}}} \cdot \tilde{d} \right), \tilde{R} = 1,
$$
$$
\tilde{I} = O\left(\epsilon_0^{-\frac{2\tilde{d}}{\tilde{\alpha}}} \cdot \tilde{d} \cdot \log \frac{1}{\epsilon_0} \right), \tilde{\tau} = \max\{B, 1, \sqrt{\tilde{d}}, \omega^2\},
$$

then we have

$$
\mathbb{E}\left[\int_{\mathcal{X}} \left(\widehat{r}_h(s, a) - r_h(s, a) \right)^2 d\eta_h(s, a) \right]
$$
$$
\leq O\left(\epsilon_0^2 + \epsilon_0^{-2\tilde{d}/\tilde{\alpha}} \cdot \frac{\log K_{HF}}{K_{HF}} \cdot \log^3 \frac{2}{\epsilon_0} \right)
\tag{5}
$$

for all $h \in [H]$. Here the expectation $\mathbb{E}[\cdot]$ is taken over the data-generating process of $\mathcal{D}^{HF}$. Recall that η_h is the sampling distribution of (s, a) for the human preference data.

Proof. See Appendix B for details. $\square$

Theorem 4.4 upper bounds the mean square error given by neural MLE. Specifically, the estimation error consists of two terms: (i) the complexity of function class $\mathcal{G}(\tilde{R}, \tilde{\tau}, \tilde{L}, \tilde{p}, \tilde{I})$, measured by its covering number, and (ii) model approximation error ϵ_0. Such a scenario is the classic bias-variance tradeoff, which

needs to be carefully balanced by selecting the appropriate network size. Specifically, by setting $\epsilon_0 = O\big(K_{HF}^{-\frac{\tilde{\alpha}}{2\tilde{\alpha}+2\tilde{d}}}\big)$, we have the mean square error bounded by $O\big(K_{HF}^{-\frac{2\tilde{\alpha}}{2\tilde{\alpha}+2\tilde{d}}} \cdot \log^3 K_{HF}\big)$. To depict the distribution shift of the dataset, we introduce the following metric:

Definition 4.5 (Restricted χ^2-divergence). *For two distribution q_1, q_2 defined on a sample space $\mathcal{X}$, and a function class $\mathcal{Q}$ of functions that maps $\mathcal{X}$ to $\mathbb{R}$, we define their restricted χ^2 divergence $\chi^2_{\mathcal{Q}}(q_1, q_2)$ as*

$$\chi^2_{\mathcal{Q}}(q_1, q_2) = \sup_{f \in \mathcal{Q}} \frac{\mathbb{E}_{q_1}[f]^2}{\mathbb{E}_{q_2}[f^2]} - 1.$$

We are now ready to present our main theorem, which provides a statistical upper bound for the estimation error in off-policy evaluation.

Theorem 4.6. *Suppose Assumption 4.1 and 4.2 hold. By Algorithm 1, we have*

$$\mathbb{E}[|\widehat{v}^\pi - v^\pi|] \leq O\Bigg(H\kappa_1 \cdot \left(K^{-\frac{\alpha}{2\alpha+d}} + \sqrt{\frac{D}{K}}\right) \cdot \log^{3/2} K$$

$$+ H(\kappa_1 + \kappa_2) \cdot K_{HF}^{-\frac{\tilde{\alpha}}{2\tilde{\alpha}+2\tilde{d}}} \cdot \log K^{3/2}\Bigg)$$

if ReLU network class $\mathcal{G}_h(\tilde{R}, \tilde{\tau}, \tilde{L}, \tilde{p}, \tilde{I})$ satisfies

$$\tilde{L} = O\left(\frac{\tilde{\alpha}}{2\tilde{\alpha}+\tilde{d}} \cdot \log K_{HF}\right), \tilde{p} = O\left(K_{HF}^{\frac{\tilde{d}}{2\tilde{\alpha}+\tilde{d}}} \cdot \tilde{d}\right),$$

$$\tilde{I} = O\left(\frac{\alpha}{2\alpha+d} \cdot K_{HF}^{\frac{\tilde{d}}{2\tilde{\alpha}+\tilde{d}}} \cdot \tilde{d} \cdot \log K_{HF}\right),$$

$$\tilde{\tau} = \max\{B, 1, \sqrt{\tilde{d}}, \omega^2\}, \tilde{R} = 1,$$

and $\mathcal{F}_h(R, \tau, L, p, I)$ satisfies

$$L = O\left(\frac{\alpha}{2\alpha+d} \cdot \log K\right), p = O\left(K^{\frac{d}{2\alpha+d}}\right),$$

$$I = O\left(\frac{\alpha}{2\alpha+d} \cdot K^{\frac{d}{2\alpha+d}} \cdot \log K\right),$$

$$\tau = \max\{B, 1, \sqrt{d}, \omega^2\}, R = H.$$

The distributional mismatch of the transition dataset is captured by $\kappa_1 = \sum_{h=1}^{H} \sqrt{1 + \chi^2_{\mathcal{Q}}(q_h^\pi, q_h^{\pi_0})}$, and the distributional mismatch of the human preference dataset is captured by $\kappa_2 = \sum_{h=1}^{H} \sqrt{1 + \chi^2_{\mathcal{Q}}(q_h^\pi, \eta_h)}$, in which q_h^π and $q_h^{\pi_0}$ are the visitation distributions of π and π_0 at step h respectively, η_h is the sampling distribution of $\{(s_h^{(i)}, a_h^{(i)})\}_{i=1}^{2}$ and $\mathcal{Q}$ is the Minkowski sum between the ReLU function class $\mathcal{F}(R, \tau, L, p, I)$ and the Hölder function class $\mathcal{H}^\alpha(\mathcal{X})$. $O(\cdot)$ hides factors depending on $\log D, \log H, B, B', \omega$ and surface area of $\mathcal{X}$.

Proof. See Appendix A for details. $\square$

Theorem 4.6 explains the success of fine-tuning processes in RLHF such as Ouyang et al. (2022); Stiennon et al. (2020), in which a compact dataset with human feedback and a relatively small model is adequate for acquiring a precise reward function. Particularly, the reward learning error contribute a $O(H(\kappa_1 + \kappa_2) \cdot K_{HF}^{-\frac{\tilde{\alpha}}{2\tilde{\alpha}+2\tilde{d}}} \cdot \log K_{HF}^{3/2})$ term in the total estimation error of v^π. When $\tilde{d} \ll d$ and $\tilde{\alpha} \gg \alpha$, i.e. the reward function has a smaller intrinsic dimension and higher smoothness, the error inherited from reward learning is much smaller than the error caused by learning the underlying MDP. We further discuss the implication of Theorem 4.6 in details:

Provable benefits of human feedback. Designing a satisfying reward function is hard in reinforcement learning. When learning from human feedback, we have no access to a given reward function as in previous off-policy evaluation works (Ji et al., 2022; Duan & Wang, 2020; Uehara et al., 2022). Compared to the guarantee standard neural FQE such as Ji et al. (2022), in which the rewards are explicitly given by the environment, Algorithm 1 learns reward function from human feedback by neural MLE, which results in an additional $\tilde{O}\big(K_{HF}^{-\frac{\tilde{\alpha}}{2\tilde{\alpha}+2\tilde{d}}}\big)$ term. Such a term does not affect our error bound as long as $K_{HF} \asymp K^{\frac{\alpha/\tilde{\alpha}}{(2\alpha+d)/(2\tilde{\alpha}+2\tilde{d})}}$. We would like to highlight that when the human-preference reward has an intrinsic smooth and low dimensional structure, i.e. $\tilde{\alpha} \gg \alpha$ and $\tilde{d} \ll d$, we can train the reward with a $\mathcal{D}^{HF}$ with size $K_{HF} \ll K$. This explains why a small dataset is sufficient for fine-tuning a large model. Under such a circumstance, our evaluation error is $\tilde{O}(K^{-\frac{\alpha}{d+2\alpha}})$, which matches the standard results in doing FQE with an observable reward, such as Ji et al. (2022). Although OPE with human preference contains less information as the reward is unobservable in the sampled dataset, Theorem 4.6 shows that it is as efficient as standard OPE with observable reward function.

(II) Adaptation to intrinsic dimension. Note that our estimation error is mainly influenced by the intrinsic dimensionality d and $\tilde{d}$, instead of the representation dimensionality D. Consequently, the bound of estimation error for our method is much smaller than the error in methods that ignore the problem's intrinsic dimension, such as Nguyen-Tang et al. (2022).

(III) Charaterizing distribution shift. The term κ_1 and κ_2 in our result quantify the divergence between the visitation distribution of the target policy and the sampling policy of human preference data and transition information, using a restricted χ^2-divergence. It's

worth noting that this restricted χ^2-divergence is always less than or equal to the widely used absolute density ratio (Nguyen-Tang et al., 2022; Chen & Jiang, 2019a; Xie & Jiang, 2020), and often, it can be significantly smaller. This occurs because the probability measures q_h^π and $q_h^{\pi_0}$ may vary considerably in small regions of the sample space, while their integration of a smooth function in $\mathcal{Q}$ across the entire sample space can be close. The absolute density ratio captures the former variation, whereas the restricted χ^2-divergence captures the latter. We also highlight that although the error in the reward learning part propagates to the FQE part, the distribution mismatch in human preference only exhibits an additive relation with the distribution mismatch of the transition dataset $\mathcal{D}$ without amplification.

5 Proof Sketch

In this section we briefly sketch the proofs of efficiencies of Algorithm 1. For detailed proof, see Appendix A and B.

Proof outline for Theorem 4.4. In the reward estimation stage, by implementing MLE on human preference probability $\rho(\cdot \mid s, a^{(0)}, a^{(1)}, a^{default})$, we claim that with high probability, the model error under the distribution of the sampling policy can be bounded by the approximation error of neural network class $\mathcal{G}_h$ and its complexity. Specifically, we prove that

$$
\mathbb{E}\bigg[\mathbb{E}_{\eta_h}\big[\|\rho_{l,h}(\cdot \mid (s_h^{(0)}, a_h^{(0)}), (s_h^{(1)}, a_h^{(1)}))
$$
$$
- \rho_{b,h}(\cdot \mid (s_h^{(0)}, a_h^{(0)}), (s_h^{(1)}, a_h^{(1)}))\|_1^2\big]\bigg]
$$
$$
\leq O\bigg(\frac{1}{K_{HF}}\log\big(H \cdot N(\mathcal{G}_h, \|\cdot\|_\infty, 1/K_{HF})\big) + \epsilon\bigg),
$$

where ρ_h^l is the preference distribution learned by MLE, ϵ is the approximation error of neural function class $\mathcal{G}_h$. Notably, our method takes model misspecification into account, which is novel in MLE error analysis. By the definition of preference in equation 3 and the condition that $r_h(s^f, a^f) = 0$, we can prove that

$$
\mathbb{E}\bigg[\int_{\mathcal{X}} |r_h(s,a) - \widehat{r}_h(s,a)|^2 d\eta_h(s,a)\bigg]
$$
$$
\leq O\bigg(\frac{1}{K_{HF}}\log\big(H \cdot N(\mathcal{G}_h, \|\cdot\|_\infty, 1/K_{HF})\big) + \epsilon\bigg).
$$

Finally, by the approximation ability of neural networks, we can give an upper bound for the function class complexity in terms of $O(\epsilon^{-d/\alpha} \cdot \log 1/\epsilon)$, while maintaining a given approximation error. Set $\epsilon = \epsilon_0^2$ and we conclude the proof.

Proof outline for Theorem 4.6. Theorem 4.6 is proved by an inductive fashion. In particular, with

notation

$$
E_h := \mathbb{E}\bigg[\int_{\mathcal{X}} \big|\mathcal{T}_h^\pi Q_{h+1}^\pi(s,a) - \widehat{\mathcal{T}}_h^\pi \widehat{Q}_{h+1}^\pi(s,a)\big| dq_h^\pi(s,a)\bigg],
$$
$$
\bar{T}_h f(s,a) = \widehat{r}_h(s,a) + \mathcal{P}_h f(s,a),
$$

by Cauchy-Schwarz inequality we can prove that

$$
E_h \leq E_{h+1}
$$
$$
+ \sqrt{1 + \chi_{\mathcal{Q}}^2(q_h^\pi, q_h^{\pi_0})}
$$
$$
\cdot \mathbb{E}\bigg[\int_{\mathcal{X}} \big(\widehat{\mathcal{T}}_h^\pi \widehat{Q}_{h+1}^\pi(s,a) - \bar{\mathcal{T}}_h^\pi \widehat{Q}_{h+1}^\pi(s,a)\big)^2 dq_h^{\pi_0}(s,a)\bigg]^{1/2}
$$
$$
+ \sqrt{1 + \chi_{\mathcal{Q}}^2(q_h^\pi, \eta_h)}
$$
$$
\cdot \mathbb{E}\bigg[\int_{\mathcal{X}} \big(r_{h+1}(s,a) - \widehat{r}_{h+1}(s,a)\big)^2 d\eta_h(s,a)\bigg]^{1/2}
$$

which consists of a distribution shift term, an L^2 norm between difference of $\widehat{\mathcal{T}}_h$ and $\bar{\mathcal{T}}_h$ projected on the estimate of Q_{h+1}, and an estimation error for human preference reward function. We further decomposed the right-hand side into the generalization and approximation error in neural network regression, and the error inherited in reward estimation. Note that we face the trade-off between the size of neural net size and its approximation limit. We recommend the readers to Appendix B for details. By choosing the neural network class size appropriately, we prove that

$$
\mathbb{E}\bigg[\int_{\mathcal{X}} \big(\mathcal{T}_h^\pi \widehat{Q}_{h+1}^\pi(s,a) - \widehat{\mathcal{T}}_h^\pi \widehat{Q}_{h+1}^\pi(s,a)\big)^2 dq_h^{\pi_0}(s,a)\bigg]
$$
$$
\lesssim H\bigg(K^{-\frac{2\alpha}{2\alpha+d}} + K_{HF}^{-\frac{2\tilde{\alpha}}{2\tilde{\alpha}+2d}} + \frac{D}{K}\bigg).
$$

Sum everything together and we conclude the proof.

6 Conclusion

This paper studies nonparametric off-policy evaluation in reinforcement learning with human preferences. We consider a two-stage algorithm: (i) We use a fully connected ReLU network to learn the reward function from human choices by neural maximum likelihood estimation, and (ii) we leverage the learned reward to approximate Q-functions by minimizing a Bellman mean square error, with another ReLU neural network as the function approximator. Our theory proves that when state-action space exhibits low-dimensional structures and the underlying reward function enjoys a smooth representation, FQE with human preference data achieves a $\tilde{O}(HK^{-\frac{\alpha}{2\alpha+d}})$ finite-sample estimation error, which matches the standard FQE case in which the reward is observable.

References

Eddie Aamari, Jisu Kim, Frédéric Chazal, Bertrand Michel, Alessandro Rinaldo, and Larry Wasserman. Estimating the reach of a manifold. 2019.

Karun Adusumilli and Dita Eckardt. Temporal-difference estimation of dynamic discrete choice models. *arXiv preprint arXiv:1912.09509*, 2019.

Armen Aghajanyan, Luke Zettlemoyer, and Sonal Gupta. Intrinsic dimensionality explains the effectiveness of language model fine-tuning, 2020.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39:324, 1952.

Nicholas Buchholz, Matthew Shum, and Haiqing Xu. Semiparametric estimation of dynamic discrete choice models. *Journal of Econometrics*, 223(2): 312–327, 2021.

Souradip Chakraborty, Amrit Bedi, Alec Koppel, Huazheng Wang, Dinesh Manocha, Mengdi Wang, and Furong Huang. Parl: A unified framework for policy alignment in reinforcement learning. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Balasubramaniyan Chandrasekaran and Manasa Mainampati. A human in the loop based robotic system by using soft actor critic with discrete actions. In *2021 4th International Conference on Mechatronics, Robotics and Automation (ICMRA)*, pp. 19–25. IEEE, 2021.

Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. *Advances in neural information processing systems*, 24, 2011.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pp. 1042–1051. PMLR, 2019a.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning, 2019b.

Minshuo Chen, Haoming Jiang, Wenjing Liao, and Tuo Zhao. Nonparametric regression on low-dimensional manifolds using deep relu networks: Function approximation and statistical recovery. *Information and Inference: A Journal of the IMA*, 11 (4):1203–1253, 2022.

Woon Sang Cho, Pengchuan Zhang, Yizhe Zhang, Xiujun Li, Michel Galley, Chris Brockett, Mengdi Wang, and Jianfeng Gao. Towards coherent and cohesive long-form text generation. *arXiv preprint arXiv:1811.00511*, 2018. URL https://ar5iv.org/abs/1811.00511.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Simon S. Du, Sham M. Kakade, Ruosong Wang, and Lin F. Yang. Is a good representation sufficient for sample efficient reinforcement learning?, 2020.

Yaqi Duan and Mengdi Wang. Minimax-optimal off-policy evaluation with linear function approximation, 2020.

Jianqing Fan, Zhaoran Wang, Yuchen Xie, and Zhuoran Yang. A theoretical analysis of deep q-learning, 2020.

Herbert Federer. Curvature measures. *Transactions of the American Mathematical Society*, 93(3):418–491, 1959.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization, 2022.

Amelia Glaese, Nat McAleese, Maja Trebacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*, 2022.

Sixue Gong, Vishnu Naresh Boddeti, and Anil K. Jain. On the intrinsic dimensionality of image representations, 2019.

Shixiang Gu, Timothy Lillicrap, Ilya Sutskever, and Sergey Levine. Continuous deep q-learning with model-based acceleration. In *International conference on machine learning*, pp. 2829–2838. PMLR, 2016.

Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Ahmed Hussein, Mohamed Medhat Gaber, Eyad Elyan, and Chrisina Jayne. Imitation learning: A survey of learning methods. *ACM Computing Surveys (CSUR)*, 50(2):1–35, 2017.

Borja Ibarz, Jan Leike, Tobias Pohlen, Geoffrey Irving, Shane Legg, and Dario Amodei. Reward learning from human preferences and demonstrations in atari. *Advances in neural information processing systems*, 31, 2018.

Ashesh Jain, Brian Wojcik, Thorsten Joachims, and Ashutosh Saxena. Learning trajectory preferences

for manipulators via iterative improvement. *Advances in neural information processing systems*, 26, 2013.

Natasha Jaques, Judy Hanwen Shen, Asma Ghandeharioun, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Shane Gu, and Rosalind Picard. Human-centric dialog training via offline reinforcement learning. *arXiv preprint arXiv:2010.05848*, 2020.

Xiang Ji, Minshuo Chen, Mengdi Wang, and Tuo Zhao. Sample complexity of nonparametric off-policy evaluation on low-dimensional manifolds using deep networks. *arXiv preprint arXiv:2206.02887*, 2022.

Haoming Jiang, Bo Dai, Mengjiao Yang, Tuo Zhao, and Wei Wei. Towards automatic evaluation of dialog systems: A model-free off-policy evaluation approach. *arXiv preprint arXiv:2102.10242*, 2021.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 652–661. PMLR, 2016.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pp. 2137–2143. PMLR, 2020.

Andras Kupcsik, David Hsu, and Wee Sun Lee. Learning dynamic robot-to-human object handover from human feedback. *Robotics Research: Volume 1*, pp. 161–176, 2018.

Jongmin Lee, Wonseok Jeon, Byungjun Lee, Joelle Pineau, and Kee-Eung Kim. Optidice: Offline policy optimization via stationary distribution correction estimation. In *International Conference on Machine Learning*, pp. 6120–6130. PMLR, 2021a.

Kimin Lee, Laura Smith, and Pieter Abbeel. Pebble: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pretraining. *arXiv preprint arXiv:2106.05091*, 2021b.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Zihao Li, Zhuoran Yang, and Mengdi Wang. Reinforcement learning with human feedback: Learning dynamic choices via pessimism, 2023.

Xinran Liang, Katherine Shu, Kimin Lee, and Pieter Abbeel. Reward uncertainty for exploration in preference-based reinforcement learning. *arXiv preprint arXiv:2205.12401*, 2022.

David Lindner, Sebastian Tschiatschek, Katja Hofmann, and Andreas Krause. Interactively learning preference constraints in linear bandits, 2022.

Runze Liu, Fengshuo Bai, Yali Du, and Yaodong Yang. Meta-reward-net: Implicitly differentiable reward learning for preference-based reinforcement learning. *Advances in Neural Information Processing Systems*, 35:22270–22284, 2022.

Travis Mandel, Yun-En Liu, Sergey Levine, Emma Brunskill, and Zoran Popovic. Offline policy evaluation across representations with applications to educational games. In *AAMAS*, volume 1077, 2014.

Hongzi Mao, Mohammad Alizadeh, Ishai Menache, and Srikanth Kandula. Resource management with deep reinforcement learning. In *Proceedings of the 15th ACM workshop on hot topics in networks*, pp. 50–56, 2016.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, et al. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*, 2022.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

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

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

Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. *Advances in neural information processing systems*, 32, 2019.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

Thanh Nguyen-Tang, Sunil Gupta, Hung Tran-The, and Svetha Venkatesh. Sample complexity of offline reinforcement learning with deep relu networks, 2022.

Partha Niyogi, Stephen Smale, and Shmuel Weinberger. Finding the homology of submanifolds with high confidence from random samples. 2006.

Ellen Novoseller, Yibing Wei, Yanan Sui, Yisong Yue, and Joel Burdick. Dueling posterior sampling for preference-based reinforcement learning. In *Conference on Uncertainty in Artificial Intelligence*, pp. 1029–1038. PMLR, 2020.

Stanley Osher, Zuoqiang Shi, and Wei Zhu. Low dimensional manifold model for image processing. *SIAM Journal on Imaging Sciences*, 10(4):1669–1690, 2017.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Aldo Pacchiano, Aadirupa Saha, and Jonathan Lee. Dueling rl: reinforcement learning with trajectory preferences. *arXiv preprint arXiv:2111.04850*, 2021.

Phillip Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning, 2021.

Doina Precup. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, pp. 80, 2000.

Aadirupa Saha and Akshay Krishnamurthy. Efficient and optimal algorithms for contextual dueling bandits under realizability. In *International Conference on Algorithmic Learning Theory*, pp. 968–994. PMLR, 2022.

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

Nihar Shah, Sivaraman Balakrishnan, Joseph Bradley, Abhay Parekh, Kannan Ramchandran, and Martin Wainwright. Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. In *Artificial intelligence and statistics*, pp. 856–865. PMLR, 2015.

Mohit Sharma, Kris M Kitani, and Joachim Groeger. Inverse reinforcement learning with conditional choice probabilities. *arXiv preprint arXiv:1709.07597*, 2017.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 2139–2148. PMLR, 2016.

Masatoshi Uehara, Masahiro Kato, and Shota Yasui. Off-policy evaluation and learning for external validity under a covariate shift. *Advances in Neural Information Processing Systems*, 33:49–61, 2020.

Masatoshi Uehara, Chengchun Shi, and Nathan Kallus. A review of off-policy evaluation in reinforcement learning. *arXiv preprint arXiv:2212.06355*, 2022.

Ruosong Wang, Dean P. Foster, and Sham M. Kakade. What are the statistical limits of offline rl with linear function approximation?, 2020.

Xiaofei Wang, Kimin Lee, Kourosh Hakhamaneshi, Pieter Abbeel, and Michael Laskin. Skill preferences: Learning to extract and execute robotic skills from human feedback. In *Conference on Robot Learning*, pp. 1259–1268. PMLR, 2022.

Jeff Wu, Long Ouyang, Daniel M Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. Recursively summarizing books with human feedback. *arXiv preprint arXiv:2109.10862*, 2021.

Markus Wulfmeier, Peter Ondruska, and Ingmar Posner. Maximum entropy deep inverse reinforcement learning, 2016.

Tengyang Xie and Nan Jiang. Q* approximation schemes for batch reinforcement learning: A theoretical comparison, 2020.

Tengyang Xie, Yifei Ma, and Yu-Xiang Wang. Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. *Advances in Neural Information Processing Systems*, 32, 2019.

Zhuoran Yang, Chi Jin, Zhaoran Wang, Mengdi Wang, and Michael I Jordan. On function approximation in reinforcement learning: Optimism in the face of large state spaces. *arXiv preprint arXiv:2011.04622*, 2020.

Ming Yin and Yu-Xiang Wang. Asymptotically efficient off-policy evaluation for tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 3948–3958. PMLR, 2020.

Ming Yin, Mengdi Wang, and Yu-Xiang Wang. Offline reinforcement learning with differentiable function approximation is provably efficient. *arXiv preprint arXiv:2210.00750*, 2022.

Huixin Zhan, Feng Tao, and Yongcan Cao. Human-guided robot behavior learning: A gan-assisted preference-based reinforcement learning approach, 2020.

Wenhao Zhan, Baihe Huang, Audrey Huang, Nan Jiang, and Jason Lee. Offline reinforcement learning with realizability and single-policy concentrability. In *Conference on Learning Theory*, pp. 2730–2775. PMLR, 2022.

Wenhao Zhan, Masatoshi Uehara, Nathan Kallus, Jason D. Lee, and Wen Sun. Provable offline reinforcement learning with human feedback, 2023.

Ruiqi Zhang, Xuezhou Zhang, Chengzhuo Ni, and Mengdi Wang. Off-policy fitted q-evaluation with differentiable function approximators: Z-estimation and inference theory. *arXiv preprint arXiv:2202.04970*, 2022.

Zeyu Zhang, Yi Su, Hui Yuan, Yiran Wu, Rishab Balasubramanian, Qingyun Wu, Huazheng Wang, and Mengdi Wang. Unified off-policy learning to rank: a reinforcement learning perspective. *arXiv preprint arXiv:2306.07528*, 2023. URL `https://ar5iv.org/abs/2306.07528`.

Yang Zhou, Rui Fu, and Chang Wang. Learning the car-following behavior of drivers using maximum entropy deep inverse reinforcement learning. *Journal of advanced transportation*, 2020:1–13, 2020.

Banghua Zhu, Jiantao Jiao, and Michael I. Jordan. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons, 2023.

Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pp. 1433–1438. Chicago, IL, USA, 2008.

A Proof for Theorem 4.4

Proof. In this proof, we denote the human labeling policy with $\rho_{b,h}(\cdot \mid (\tilde{s}_h^{(1)}, \tilde{a}_h^{(1)}), (\tilde{s}_h^{(2)}, \tilde{a}_h^{(2)}))$. Recall that the preference for step h and episode k is chosen among $\{(\tilde{s}_{h,k}^{(1)}, \tilde{a}_{h,k}^{(1)}), (\tilde{s}_{h,k}^{(2)}, \tilde{a}_{h,k}^{(2)}), (s^f, a^f)\}$. In this proof, we use the notation of $\Upsilon_h^k := \{(\tilde{s}_{h,k}^{(1)}, \tilde{a}_{h,k}^{(1)}), (\tilde{s}_{h,k}^{(2)}, \tilde{a}_{h,k}^{(2)}), (s^f, a^f)\}$. Recall that by Assumption 3.1 and 3.2, Υ_h^k is sampled by distribution η_h for every $k \in [K]$, and the human preference $y_{h,k} = (s,a)$ is generated according to

$$\rho_{b,h}(s,a \mid \Upsilon_h^k) = \frac{\exp(r_h(s,a))}{\exp(r_h(\tilde{s}_{h,k}^{(1)}, \tilde{a}_{h,k}^{(1)})) + \exp(r_h(\tilde{s}_{h,k}^{(2)}, \tilde{a}_{h,k}^{(2)})) + \exp(r_h(s^f, a^f))},$$

here the state-action pair $(s,a) \in \Upsilon_h^k$. For all $h \in [H]$, we define a set $\Upsilon_h := \{(s_h^{(1)}, a_h^{(1)}), (s_h^{(2)}, a_h^{(2)}), (s^f, a^f)\}$, where $(s_h^{(1)}, a_h^{(1)})$ and $(s_h^{(2)}, a_h^{(2)})$ are fixed state-action pairs. Then Υ_h^k can be regarded as the realization of Υ under distribution η_h. We also define a vector-value function class defined on Υ_h, which is parametrized by function $r \in \mathcal{G}$:

$$\Pi_h(\Upsilon_h) = \left\{ \rho \in \mathbb{R}^3 \mid \rho_r(s,a \mid \Upsilon_h) = \exp(r(s,a)) / \sum_{(s',a') \in \Upsilon} \exp(r(s',a')) \text{ for some } r \in \mathcal{G} \right\}.$$

Let $\mathcal{N}_{[]}(\Pi_h, \|\cdot\|_\infty, 1/n)$ be the smallest $1/n$-upper bracketing of Π_h. And $|\mathcal{N}_{[]}(\Pi_h, \|\cdot\|_\infty, 1/n)| = N_{[]}(\Pi_h, \|\cdot\|_\infty, 1/n)$, where $N_{[]}(\Pi_h, \|\cdot\|_\infty, 1/n)$ is the bracketing number of Π_h. For a given neural network $\mathcal{G}$ and a reward function r, we denote

$$\epsilon_{misspec} := \min_{r' \in \mathcal{G}} \|r' - r\|_\infty, \qquad r_{h,o} = \operatorname{argmin}_{r' \in \mathcal{G}} \|r' - r\|_\infty.$$

to measure the approximation ability of $\mathcal{G}$ for the r. Our proof for Theorem 4.4 lies in three steps:

(i) First, we prove that with MLE, we have

$$\mathbb{E}_{(s_h^{(i)}, a_h^{(i)}) \sim \eta_h}\left[\|\rho(\cdot \mid \Upsilon_h) - \rho_h^l(\cdot \mid \Upsilon_h)\|_1^2 \right] \leq \mathcal{O}\left(\frac{\log\left(H \cdot N_{[]}(\Pi_h, \|\cdot\|_\infty, 1/K)/\delta\right)}{K_{HF}} \right) + \epsilon_{misspec} \qquad (6)$$

with probability at least $1-\delta$. Here $\mathbb{E}_{(s_h^{(i)}, a_h^{(i)}) \sim \eta_h}[\cdot]$ means that the expectation is taken under the randomness of $(s_h^{(i)}, a_h^{(i)}) \sim \eta_h$, $i \in \{1,2\}$, $\rho_{l,h}$ is the choice probability learned by MLE in step h, and the L^1 norm is taken over all possible state action pairs in Υ_h. By invoking Lemma C.6 we get the expectation version.

(ii) Second, we prove that with Assumption 3.1 and 3.1, we can convert the previous TV-bound for human choice probability into an estimation error for the reward function $\widehat{r}_h(s,a)$, with an error bound of

$$\mathbb{E}\left[\int_{\mathcal{X}} |r_h(s,a) - (r_h^l(s,a) - r_h^l(s^f, a^f))|^2 d\eta_h(s,a) \right]$$
$$\leq O\left(\frac{1}{K_{HF}} \log\left(H \cdot N(\mathcal{G}, \|\cdot\|_\infty, 1/K_{HF})\right) + \epsilon_{misspec} \right).$$

Recall that r_h^l is the reward function we learned directly from MLE. Note that such a bound measures the tradeoff between approximation and generalization: for the network to have better approximation ability, i.e. a smaller $\epsilon_{misspec}$, a larger neural network is necessary, which increases the complexity of the neural network that is measured in term $\log\left(H \cdot N(\mathcal{G}, \|\cdot\|_\infty, 1/K_{HF})\right)$.

(iii) Third, we prove that we can handle the tradeoff between approximation and generalization by choosing an appropriate neural network have equation 5 hold. Specifically, under Assumption 4.3, we can obtain a fast convergence rate for our estimation $\widehat{r}_h$.

Step (i). For notation simplicity, we denote $\rho_{r_{h,o}}$ by $\rho_{o,h}$ and $\rho_{r_{h,l}}$ by $\rho_{l,h}$. By MLE procedure, we have

$$\frac{1}{K_{HF}} \sum_{k=1}^{K_{HF}} \log\left(\frac{\rho_{l,h}(y_{h,k} \mid \Upsilon_h^k)}{\rho_{b,h}(y_{h,k} \mid \Upsilon_h^k)}\right) = \frac{1}{K_{HF}} \sum_{k=1}^{K_{HF}} \log\left(\frac{\rho_{l,h}(y_{h,k} \mid \Upsilon_h^k)}{\rho_{o,h}(y_{h,k} \mid \Upsilon_h^k)}\right) + \frac{1}{K_{HF}} \sum_{k=1}^{K_{HF}} \log\left(\frac{\rho_{o,h}(y_{h,k} \mid \Upsilon_h^k)}{\rho_{b,h}(y_{h,k} \mid \Upsilon_h^k)}\right)$$

$$\geq \frac{1}{K_{HF}} \sum_{k=1}^{K_{HF}} \log(\rho_{b,h}(y_{h,k} \mid \Upsilon_h^k)) - \frac{1}{K_{HF}} \sum_{k=1}^{K_{HF}} \log(\rho_{o,h}(y_{h,k} \mid \Upsilon_h^k))$$

$$= \frac{1}{K_{HF}} \sum_{k=1}^{K_{HF}} \left\{ r_h(s_{h,k}, a_{h,k}) - r_{h,o}(s_{h,k}, a_{h,k}) \right.$$

$$\left. + \log\left(\sum_{(s,a)\in\Upsilon_h} \exp(r_{h,o}(s,a))\right) - \log\left(\sum_{(s,a)\in\Upsilon_h} \exp(r_h(s,a))\right) \right\}$$

$$\geq -4\epsilon_{misspec},$$

here the first inequality comes from $\widehat{r}_h^l \in \mathcal{G}$ and $\rho_{l,h} = \arg\max_{\rho_r, r\in\mathcal{G}} \frac{1}{K} \sum_{k=1}^K \log\left(\rho_h(y_h^k \mid s_h^k)\right)$, the second inequality comes from the fact that $\log(\sum_{i=1}^3 \exp(x_i))$ being 1-Lipschitz for every component. Meanwhile, by Markov's inequality and Boole's inequality, it holds with probability at least $1-\delta$ that for all $\bar{\rho} \in \mathcal{N}_{[]}(\Pi, \|\cdot\|_\infty, 1/n)$, we have

$$\sum_{k=1}^{K_{HF}} \frac{1}{2} \log\left(\frac{\bar{\rho}(y_h^k \mid \Upsilon_h^k)}{\rho_{b,h}(y_h^k \mid \Upsilon_h^k)}\right) \leq K_{HF} \log\left(\mathbb{E}_{(s_h^{(i)}, a_h^{(i)})\sim\eta_h, y_h\sim\rho_{b,h}}\left[\exp\left(\frac{1}{2}\log\left(\frac{\bar{\rho}(y_h \mid \Upsilon)}{\rho_{b,h}(y_h \mid \Upsilon)}\right)\right)\right]\right)$$

$$+ \log\left(\frac{N_{[]}(\Pi_h, \|\cdot\|_\infty, 1/K_{HF})}{\delta}\right),$$

specify $\bar{\rho}$ to be the upper bracket of $\rho_{l,h}$, we have

$$-2K_{HF}\cdot\epsilon_{misspec} \leq K_{HF} \log\left(\mathbb{E}_{(s_h^{(i)}, a_h^{(i)})\sim\eta_h, y_h\sim\rho_{b,h}}\left[\exp\left(\frac{1}{2}\log\left(\frac{\bar{\rho}(y_h \mid \Upsilon_h)}{\rho_{b,h}(y_h \mid \Upsilon_h)}\right)\right)\right]\right)$$

$$+ \log\left(\frac{N_{[]}(\Pi_h, \|\cdot\|_\infty, 1/K_{HF})}{\delta}\right)$$

$$\leq K_{HF}\cdot\log\left(\mathbb{E}_{(s_h^{(i)}, a_h^{(i)})\sim\eta_h, y_h\sim\rho_{b,h}}\left[\sqrt{\frac{\bar{\rho}(y_h \mid \Upsilon_h)}{\rho_{b,h}(y_h \mid \Upsilon_h)}}\right]\right) + \log\left(\frac{N_{[]}(\Pi_h, \|\cdot\|_\infty, 1/K_{HF})}{\delta}\right)$$

$$= K_{HF}\cdot\log\left(\mathbb{E}_{(s_h^{(i)}, a_h^{(i)})\sim\eta_h}\left[\sum_{y_h\in\Upsilon_h} \sqrt{\bar{\rho}(y_h \mid \Upsilon_h)\cdot\rho_{b,h}(y_h \mid \Upsilon_h)}\right]\right)$$

$$+ \log\left(\frac{N_{[]}(\Pi_h, \|\cdot\|_\infty, 1/K_{HF})}{\delta}\right),$$

Utilizing the $\log x \leq x - 1$, we have

$$1 - \mathbb{E}_{(s_h^{(i)}, a_h^{(i)})\sim\eta_h}\left[\sum_{y_h\in\Upsilon_h} \sqrt{\bar{\rho}(y_h \mid \Upsilon_h)\cdot\rho_{b,h}(y_h \mid \Upsilon_h)}\right] \leq \frac{1}{K_{HF}} \log\left(\frac{N_{[]}(\Pi_h, \|\cdot\|_\infty, 1/K_{HF})}{\delta}\right) + \epsilon_{misspec}.$$

Therefore we can bound the Hellinger distance between $\rho_{b,h}$ and $\bar{\rho}$,

$$h(\bar{\rho}, \rho_{b,h}) = \mathbb{E}_{(s_h^{(i)}, a_h^{(i)})\sim\eta_h}\left[\sum_{y_h\in\Upsilon_h} \left(\bar{\rho}(a \mid s_h)^{1/2} - \rho_{b,h}(a \mid s_h)^{1/2}\right)^2\right] \tag{7}$$

$$\leq 2\mathbb{E}_{(s_h^{(i)}, a_h^{(i)})\sim\eta_h}\left[1 - \sum_{y_h\in\Upsilon_h} \sqrt{\bar{\rho}(a \mid s_h)\cdot\rho_{b,h}(a \mid s_h)}\right] + \frac{1}{K_{HF}} \tag{8}$$

$$\leq \frac{2}{K_{HF}} \log\left(\frac{N_{[]}(\Pi, \|\cdot\|_\infty, 1/K_{HF})}{\delta}\right) + \frac{1}{K_{HF}} + \epsilon_{misspec}, \tag{9}$$

here the second inequality comes from the fact that $\bar{\rho}$ is a upper bracketing of Π. Moreover, it is easy to verify that

$$\mathbb{E}_{(s_h^{(i)},a_h^{(i)})\sim\eta_h}\Big[\sum_{y_h\in\Upsilon_h}((\bar{\rho}(a\mid s_h)^{1/2}+\rho_{b,h}(a\mid s_h)^{1/2}))^2\Big] \leq 2\mathbb{E}_{(s_h^{(i)},a_h^{(i)})\sim\eta_h}\Big[\sum_{y_h\in\Upsilon_h}(\bar{\rho}(a\mid s_h)+\rho_{b,h}(a\mid s_h))\Big] \quad (10)$$

$$\leq \frac{2}{K_{HF}}+4, \quad (11)$$

where the second inequality comes from the fact that $\bar{\pi}$ is the $1/n$-upper bracket of a probability distribution. Combining the equation 7 and equation 10, by Cauchy-Schwarz inequality, we have

$$\mathbb{E}_{(s_h^{(i)},a_h^{(i)})\sim\eta_h}\big[\|\bar{\rho}(\cdot\mid\Upsilon_h)-\rho_{b,h}(\cdot\mid\Upsilon_h)\|_1^2\big]$$
$$\leq \frac{15}{K_{HF}}\cdot\log\left(\frac{N_{[]}(\Pi,\|\cdot\|_\infty,1/K_{HF})}{\delta}\right)+\epsilon_{misspec},$$

here the L_1 norm is taken under all $(s,a)\in\Upsilon_h$. Meanwhile,

$$\|\bar{\rho}(\cdot\mid\Upsilon_h)-\rho_{b,h}(\cdot\mid\Upsilon_h)\|_1^2-\|\rho_{l,h}(\cdot\mid\Upsilon_h)-\rho_{b,h}(\cdot\mid\Upsilon_h)\|_1^2 \leq \left(4+\frac{1}{K_{HF}}\right)\cdot\frac{1}{K_{HF}},$$

therefore we have

$$\mathbb{E}_{(s_h^{(i)},a_h^{(i)})\sim\eta_h}\big[\|\rho_{l,h}(\cdot\mid\Upsilon_h)-\rho_{b,h}(\cdot\mid\Upsilon_h)\|_1^2\big] \leq \frac{20}{K_{HF}}\cdot\log\left(\frac{N_{[]}(\Pi_h,\|\cdot\|_\infty,1/K_{HF})}{\delta}\right)+\epsilon_{misspec}.$$

By Lemma Next, we bound $N_{[]}(\Pi_h,\|\cdot\|_\infty,1/K)$ by $N(\mathcal{G},\|\cdot\|_\infty,1/4K)$. For all $h\in[H]$, recall the definition

$$\Pi_h(\Upsilon_h)=\Big\{\rho_r(a\mid s)=\exp(r(s,a))/\sum_{a'\in\Upsilon_h}\exp(r(s,a'))\text{ for some }r\in\mathcal{F}_h\Big\}.$$

It is easy to check that
$$|\rho_r(a\mid s)-\rho_{r'}(a\mid s)|\leq 2\cdot\|r-r'\|_\infty,\forall(s,a)\in\mathcal{S}\times\mathcal{A}.$$

Recall that $N(\mathcal{F}_h,\|\cdot\|_\infty,1/K)$ is the covering number for model class $\mathcal{F}_h$. We have

$$N_{[]}(\Pi_h,\|\cdot\|_\infty,1/K_{HF})\leq N(\mathcal{G}_h,\|\cdot\|_\infty,1/4K_{HF}) \quad (12)$$

always hold for all $h\in[H]$ by Lemma C.2. Thus we conclude the proof in **Step (i)**.

Step (ii). Note that by Lemma C.6, we can turn into a bound in expectation:

$$\mathbb{E}\Big[\mathbb{E}_{(s_h^{(i)},a_h^{(i)})\sim\eta_h}\big[\|\rho_{l,h}(\cdot\mid\Upsilon_h)-\rho_{b,h}(\cdot\mid\Upsilon_h)\|_1^2\big]\Big] \quad (13)$$

$$\leq O\left(\frac{1}{K_{HF}}\log\left(H\cdot N(\mathcal{G}_h,\|\cdot\|_\infty,1/K_{HF})\right)+\epsilon_{misspec}\right),$$

where the expectation is taken with respect to the data-generating process. Recall that

$$\rho_{l,h}((s,a)\mid\Upsilon_h)=\frac{\exp(r_h^l(s,a))}{\sum_{a\in\Upsilon_h}\exp(r_h^l(s,a))},$$
$$\rho_{b,h}((s,a)\mid\Upsilon_h)=\frac{\exp(r_h(s,a))}{\sum_{a\in\Upsilon_h}\exp(r_h(s,a))},$$

with $r_h(s^f,a^f)=0$. By basic algebra, we have

$$\sum_{(s,a)\in\Upsilon_h}|r_h(s,a)-(r_h^l(s,a)-r^l(s^f,a^f)|^2 = \sum_{(s,a)\in\Upsilon_h}\left|\log\left(\frac{\rho_{l,h}(s,a\mid\Upsilon_h)}{\rho_{b,h}(s,a\mid\Upsilon_h)}\right)-\log\left(\frac{\rho_{l,h}(s^f,a^f\mid\Upsilon_h)}{\rho_{b,h}(s^f,a^f\mid\Upsilon_h)}\right)\right|^2$$
$$\leq 20\cdot\|\rho_{l,h}(\cdot\mid\Upsilon_h)-\rho_{b,h}(\cdot\mid\Upsilon_h)\|_1^2$$

for any arbitrary Υ_h. Combine this with equation 13, we have

$$\mathbb{E}\left[\int_{\mathcal{X}} |r_h(s,a) - (r_h^l(s,a) - r_h^l(s^f,a^f))|^2 d\eta_h(s,a)\right] \leq O\left(\frac{1}{K_{HF}} \log\left(H \cdot N(\mathcal{G}_h, \|\cdot\|_\infty, 1/K_{HF})\right) + \epsilon_{misspec}\right).$$

Recall that in Algorithm 1, we construct $\widehat{r}_h(s,a) = r_h^l(s,a) - r_h^l(s^f,a^f)$. Therefore we have

$$\mathbb{E}\left[\int_{\mathcal{X}} |r_h(s,a) - \widehat{r}_h(s,a)|^2 d\eta_h(s,a)\right] \leq O\left(\frac{1}{K_{HF}} \log\left(H \cdot N(\mathcal{G}_h, \|\cdot\|_\infty, 1/K_{HF})\right) + \epsilon_{misspec}\right). \tag{14}$$

Step (iii). Note that the bound in 14 consists of two terms: the approximation error of the neural network $\mathcal{G}_h$, and the model class complexity of $\mathcal{G}_h$. With Assumption 4.3, we connect the equation 14 with the selection of a proper neural network class. Recall that $f \in \mathcal{H}^{\alpha_{HF}}([0,1]_{HF}^d)$, and $\psi \in \mathcal{H}^{\alpha_{HF}}$. We consider $\widehat{f}(\widehat{\psi}(x))$, with $\widehat{\phi} : \mathcal{M} \to [0,1]^{d_{HF}}$. Then the approximation error in the norm L^∞ has the following decomposition:

$$|\widehat{f} \circ \widehat{\psi}(x) - f \circ \psi(x)| \leq |\widehat{f} \circ \widehat{\psi}(x) - \widehat{f} \circ \psi(x)| + |\widehat{f} \circ \psi(x) - f \circ \psi(x)| \tag{15}$$

$$\lesssim \|\widehat{f} - f\|_\infty + \max_{i \in [d^{HF}]} \|\widehat{\psi}_i - \psi_i\|_\infty, \tag{16}$$

therefore to bound the approximation error of $r = f \circ \psi$, we only need to approximate ψ and f respectively. By Lemma 2 in Nguyen-Tang et al. (2022), for every function $h \in \mathcal{H}^{\alpha_{HF}}$ and bounded in $[0,R]$ and a neural network class $\mathcal{E}(L,p,I,\tau,R)$, if

$$L = O\left(\log \frac{1}{\epsilon}\right), p = O\left(\epsilon^{-\frac{d_{HF}}{\alpha_{HF}}}\right), I = O\left(\epsilon^{-\frac{d_{HF}}{\alpha_{HF}}} \log \frac{1}{\epsilon}\right),$$
$$\tau = \max\left\{B, 1, \sqrt{d}, \omega^2\right\}, R = 1,$$

then we have

$$\min_{f \in \mathcal{E}} \|f - h\|_\infty \leq \epsilon.$$

therefore, by equation 15, to guarantee $\min_{g \in \mathcal{G}_h(\tilde{R},\tilde{\tau},\tilde{L},\tilde{p},\tilde{I})} \|g - r\|_\infty \leq O(\epsilon)$, we only need to guarantee that $\|f - \widehat{f}\|_\infty \leq \epsilon$ and $\|\psi_h - \widehat{\psi}_h^i\|_\infty \leq \epsilon$ respectively. It suffices to choose

$$\tilde{L} = O\left(\log \frac{1}{\epsilon}\right), \tilde{p} = O\left(\epsilon^{-\frac{d_{HF}}{\alpha_{HF}}} \cdot d_{HF}\right), \tilde{I} = O\left(\epsilon^{-\frac{d_{HF}}{\alpha_{HF}}} \cdot d_{HF} \cdot \log \frac{1}{\epsilon}\right),$$
$$\tilde{\tau} = \max\{B, 1, \sqrt{d_{HF}}, \omega^2\}, \tilde{R} = 1. \tag{17}$$

Note that by Lemma C.3, this implies a log-covering number of

$$\log N(H \cdot \mathcal{G}_h, \|\cdot\|_\infty, 1/K_{HF}) \leq O\left(\log \left(\frac{2L^2(pB+2)\kappa^L p^{L+1}}{1/K_{HF}}\right)^{K_{HF}}\right) \tag{18}$$

$$\leq O\left(\epsilon^{-d_{HF}/\alpha_{HF}} \cdot \log^3 \frac{2}{\epsilon} \cdot \frac{\log K_{HF}}{K_{HF}}\right).$$

Recall that in **Step (ii)** we have

$$\mathbb{E}\left[\int_{\mathcal{X}} (r_h(s,a) - \widehat{r}_h(s,a))^2 d\eta_h(s,a)\right] \leq O\left(\epsilon + \log N(H \cdot \mathcal{G}_h, \|\cdot\|_\infty, 1/K)\right).$$

Set $\epsilon = \epsilon_{misspec}$ in equation 17 and equation 18, then we have

$$\mathbb{E}\left[\int_{\mathcal{X}} (r_h(s,a) - \widehat{r}_h(s,a))^2 dq_h^\pi\right] \leq O\left(\epsilon_{misspec} + \epsilon_{misspec}^{-d_{HF}/\alpha_{HF}} \cdot \log^3 \frac{2}{\epsilon_{misspec}} \cdot \frac{\log K_{HF}}{K_{HF}}\right)$$

for all $\mathcal{G}_h(\tilde{R}, \tilde{\tau}, \tilde{L}, \tilde{p}, \tilde{I})$ with $h \in [H]$. Finally, set $\epsilon_{misspec} = \epsilon_0^2$, and we conclude the proof of Lemma 4.4. $\qquad\square$

B   Proof For Theorem 4.6

Proof. In Theorem 4.6, we aim to bound the L_1 distance between our estimate v^π and $\widehat{v}^\pi$. We decompose error recursively for a summation of H-in step errors and then change the probability measure. The key is to find a upper bound for $\mathbb{E}_{(s,a)\sim q_h^\pi(s,a)}\left[\left|\widehat{\mathcal{T}}_h^\pi \widehat{Q}_{h+1}^\pi - \mathcal{T}_h^\pi Q_{h+1}^\pi\right|\right]$. First, by definition of $\hat{v}^\pi$ and Bellman equation, we have

$$\mathbb{E}\left|v^\pi - \widehat{v}^\pi\right| = \mathbb{E}\left[\left|\int_{\mathcal{X}} (Q_1^\pi - \widehat{Q}_1^\pi)(s,a) dq_1^\pi(s,a)\right|\right]$$
$$\leq \mathbb{E}\left[\int_{\mathcal{X}} |Q_1^\pi - \widehat{Q}_1^\pi|(s,a) dq_1^\pi(s,a)\right],$$

here the expectation is taken with respect to the data generating process, i.e. $(s,a) \sim q_1^\pi$. The first equality comes from the Bellman equation, and the second inequality comes from Jensen's inequality. Now we analyze how the estimation error would propagate with the horizon length H, and prove our result by mathematic induction. For the expected error for $\widehat{Q}_h$, we have

$$\mathbb{E}\left[\int_{\mathcal{X}} |Q_h^\pi(s,a) - \widehat{Q}_h^\pi(s,a)|\right]$$
$$= \mathbb{E}\left[\int_{\mathcal{X}} |\mathcal{T}_h^\pi Q_{h+1}^\pi(s,a) - \widehat{\mathcal{T}}_h^\pi \widehat{Q}_{h+1}^\pi(s,a)| dq_h^\pi(s,a)\right]$$
$$\leq \underbrace{\mathbb{E}\left[\int_{\mathcal{X}} |\mathcal{T}_h^\pi Q_{h+1}^\pi(s,a) - \mathcal{T}_h^\pi \widehat{Q}_{h+1}^\pi(s,a)| dq_h^\pi(s,a)\right]}_{\text{(i)}}$$
$$+ \underbrace{\mathbb{E}\left[\int_{\mathcal{X}} |\mathcal{T}_h^\pi \widehat{Q}_{h+1}^\pi(s,a) - \widehat{\mathcal{T}}_h^\pi \widehat{Q}_{h+1}^\pi(s,a)| dq_h^\pi(s,a)\right]}_{\text{(ii)}},$$

To bound (i), note that $\mathcal{T}f(s,a) - \mathcal{T}g(s,a) = \mathbb{E}\left[f(s',a') - g(s',a') \mid s,a\right]$, where (s',a') is the succeeding state-action pair after (s,a). Therefore by Jensen's inequality, we have

$$\text{(i)} \leq \mathbb{E}\left[\int_{\mathcal{X}} \mathbb{E}\left[|Q_{h+1}(s',a') - \widehat{Q}_{h+1}(s',a')| \mid s,a\right] dq_h^\pi(s,a)\right]$$
$$\leq \mathbb{E}\left[\int_{\mathcal{X}} |Q_{h+1}(s,a) - \widehat{Q}_{h+1}(s,a)| dq_{h+1}^\pi(s,a)\right].$$

Next, we bound (ii) by the distribution shift captured by $\chi_\mathcal{Q}^2$ distance. For notation simplicity, we define the surrogate Bellman operator:
$$\bar{\mathcal{T}}_h^\pi \widehat{Q}_{h+1}^\pi := \widehat{r}_h + \mathcal{P}_h^\pi \widehat{Q}_{h+1}^\pi,$$

and we have

$$\text{(ii)} \leq \mathbb{E}\left[\int_{\mathcal{X}} |\widehat{\mathcal{T}}_h^\pi \widehat{Q}_{h+1}^\pi(s,a) - \bar{\mathcal{T}}_h^\pi \widehat{Q}_{h+1}^\pi(s,a)| dq_h^\pi(s,a)\right] + \mathbb{E}\left[\int_{\mathcal{X}} |\mathcal{T}_h^\pi \widehat{Q}_{h+1}^\pi(s,a) - \bar{\mathcal{T}}_h^\pi \widehat{Q}_{h+1}^\pi(s,a)| dq_h^\pi(s,a)\right]$$
$$= \underbrace{\mathbb{E}\left[\int_{\mathcal{X}} |\widehat{\mathcal{T}}_h^\pi \widehat{Q}_{h+1}^\pi(s,a) - \bar{\mathcal{T}}_h^\pi \widehat{Q}_{h+1}^\pi(s,a)| dq_h^\pi(s,a)\right]}_{\text{(iii)}} + \underbrace{\mathbb{E}\left[\int_{\mathcal{X}} |\widehat{r}_h(s,a) - r_h(s,a)| dq_h^\pi(s,a)\right]}_{\text{(iv)}}.$$

We now aim to bound (iii) and (iv) by the distribution shift of $\mathcal{D}$ and $\mathcal{D}^{HF}$ with respect to π, respectively. First, notice that

$$\text{(iv)} \leq \mathbb{E}\left[\mathbb{E}\left[\sqrt{\int_{\mathcal{X}} \left(r_h(s,a) - \widehat{r}_h(s,a)\right)^2 d\eta_h(s,a)} \cdot \sqrt{1 + \chi_\mathcal{Q}^2(q_h^\pi, \eta_h)} \mid \mathcal{D}_{h+1}^{HF}, \ldots, \mathcal{D}_H^{HF}\right]\right]$$

$$\leq \sqrt{1 + \chi^2_{\mathcal{Q}}(q^\pi_h, \eta_h)} \cdot \sqrt{\mathbb{E}\left[\mathbb{E}\left[\int_{\mathcal{X}} \left(r_h(s,a) - \widehat{r}_h(s,a)\right)^2 d\eta_h(s,a) \mid \mathcal{D}_{h+1}, \ldots, \mathcal{D}_H\right]\right]}$$

$$= \sqrt{1 + \chi^2_{\mathcal{Q}}(q^\pi_h, \eta_h)} \cdot \sqrt{\mathbb{E}\left[\int_{\mathcal{X}} \left(r_h(s,a) - \widehat{r}_h(s,a)\right)^2 d\eta_h(s,a)\right]},$$

By Theorem 4.4, we have

$$\mathbb{E}\left[\int_{\mathcal{X}} \left(r_h(s,a) - \widehat{r}_h(s,a)\right)^2 d\eta_h(s,a)\right] \leq O\left(\epsilon_0^2 + \epsilon_0^{-2\tilde{d}/\tilde{\alpha}} \cdot \frac{\log K_{HF}}{K_{HF}} \cdot \log^3 \frac{2}{\epsilon_0}\right)$$

for all $\epsilon_0 \in (0,1)$ and $\mathcal{G}$ defined in Theorem 4.4. Therefore we have

$$\text{(iv)} \leq \sqrt{1 + \chi^2_{\mathcal{Q}}(q^\pi_h, \eta_h)} \cdot O\left(\epsilon_0^2 + \epsilon_0^{-2\tilde{d}/\tilde{\alpha}} \cdot \frac{\log K_{HF}}{K_{HF}} \cdot \log^3 \frac{2}{\epsilon_0}\right).$$

For (iii), we have

$$\text{(iii)} \leq \mathbb{E}\left[\mathbb{E}\left[\sqrt{\int_{\mathcal{X}} \left(\bar{\mathcal{T}}^\pi_h \widehat{Q}^\pi_{h+1}(s,a) - \widehat{\mathcal{T}}^\pi_h \widehat{Q}^\pi_{h+1}(s,a)\right)^2 dq^{\pi_0}_h(s,a)} \cdot \sqrt{1 + \chi^2_{\mathcal{Q}}(q^\pi_h, q^{\pi_0}_h)} \mid \mathcal{D}_{h+1}, \ldots, \mathcal{D}_H\right]\right]$$

$$\leq \sqrt{1 + \chi^2_{\mathcal{Q}}(q^\pi_h, q^{\pi_0}_h)} \cdot \sqrt{\mathbb{E}\left[\mathbb{E}\left[\int_{\mathcal{X}} \left(\bar{\mathcal{T}}^\pi_h \widehat{Q}^\pi_{h+1}(s,a) - \widehat{\mathcal{T}}^\pi_h \widehat{Q}^\pi_{h+1}(s,a)\right)^2 dq^{\pi_0}_h(s,a) \mid \mathcal{D}_{h+1}, \ldots, \mathcal{D}_H\right]\right]}$$

$$= \sqrt{1 + \chi^2_{\mathcal{Q}}(q^\pi_h, q^{\pi_0}_h)} \cdot \sqrt{\mathbb{E}\left[\int_{\mathcal{A}} \left(\bar{\mathcal{T}}^\pi_h \widehat{Q}^\pi_{h+1}(s,a) - \widehat{\mathcal{T}}^\pi_h \widehat{Q}^\pi_{h+1}(s,a)\right)^2 dq^{\pi_0}_h(s,a)\right]}. \tag{19}$$

The first inequality comes from Lemma C.5, and the third equality comes from the definition of conditional expectation. By equation 19, we bound the estimation error under the distribution induced by π by the distribution shift between π_0 and π. Now we are operating on the distribution induced by π_0, and there's no distribution shift between distribution induced by π_0 and the data generating distribution.

Now we have the following decomposition:

$$\mathbb{E}\left[\int_{\mathcal{X}} \left(\bar{\mathcal{T}}^\pi_h \widehat{Q}^\pi_{h+1}(s,a) - \widehat{\mathcal{T}}^\pi_h \widehat{Q}^\pi_{h+1}(s,a)\right)^2 dq^{\pi_0}_h(s,a)\right]$$

$$= \mathbb{E}\left[\frac{2}{K} \sum_{k=1}^K \left(\bar{\mathcal{T}}^\pi_h \widehat{Q}^\pi_{h+1}(s_{h,k}, a_{h,k}) - \widehat{\mathcal{T}}^\pi_h \widehat{Q}^\pi_{h+1}(s_{h,k}, a_{h,k})\right)^2\right]$$

$$+ \mathbb{E}\left[\int_{\mathcal{X}} \left(\bar{\mathcal{T}}^\pi_h \widehat{Q}^\pi_{h+1}(s,a) - \widehat{\mathcal{T}}^\pi_h \widehat{Q}^\pi_{h+1}(s,a)\right)^2 dq^{\pi_0}_h(s,a)\right]$$

$$- \mathbb{E}\left[\frac{2}{K} \sum_{k=1}^K \left(\bar{\mathcal{T}}^\pi_h \widehat{Q}^\pi_{h+1}(s_{h,k}, a_{h,k}) - \widehat{\mathcal{T}}^\pi_h \widehat{Q}^\pi_{h+1}(s_{h,k}, a_{h,k})\right)^2\right]$$

$$= \epsilon_{approx} + \epsilon_{general},$$

where we define the empirical approximation error and generalization error respectively, i.e.

$$\epsilon_{approx} := \mathbb{E}\left[\frac{2}{K} \sum_{k=1}^K \left(\bar{\mathcal{T}}^\pi_h \widehat{Q}^\pi_{h+1}(s_{h,k}, a_{h,k}) - \widehat{\mathcal{T}}^\pi_h \widehat{Q}^\pi_{h+1}(s_{h,k}, a_{h,k})\right)^2\right],$$

$$\epsilon_{general} := \mathbb{E}\left[\int_{\mathcal{X}} \left(\bar{\mathcal{T}}^\pi_h \widehat{Q}^\pi_{h+1}(s,a) - \widehat{\mathcal{T}}^\pi_h \widehat{Q}^\pi_{h+1}(s,a)\right)^2 dq^{\pi_0}_h(s,a)\right]$$

$$- \mathbb{E}\left[\frac{2}{K} \sum_{k=1}^K \left(\bar{\mathcal{T}}^\pi_h \widehat{Q}^\pi_{h+1}(s_{h,k}, a_{h,k}) - \widehat{\mathcal{T}}^\pi_h \widehat{Q}^\pi_{h+1}(s_{h,k}, a_{h,k})\right)^2\right].$$

Here both expectation are taken with respect to the dataset generating process. By Lemma C.6, we have

$$\epsilon_{general} \le \frac{104R^2}{3K} \log \mathcal{N}\left(\delta/4R, \mathcal{F}(R,\kappa,L,p,K), \|\cdot\|_\infty\right) + \left(4 + \frac{1}{2R}\right)\delta.$$

We now only need to bound the empirical approximation error ϵ_{approx}. Lemma C.4 allows us to have we have

$$
\epsilon_{approx} \le 4 \inf_{f \in \mathcal{F}(R,\|,L,p,I)} \int_\mathcal{X} \left(f(s,a) - \bar{\mathcal{T}}_h^\pi \widehat{Q}_h^\pi(s,a)\right)^2 dq_h^{\pi_0}(s,a)
$$
$$
+ 48\sigma^2 \frac{\log \mathcal{N}\left(\delta, \mathcal{F}(R,\kappa,L,p,I), \|\cdot\|_\infty\right) + 2}{K}
$$
$$
+ \left(8\sqrt{6}\sqrt{\frac{\log \mathcal{N}\left(\delta, \mathcal{F}(R,\kappa,L,p,I), \|\cdot\|_\infty\right) + 2}{K}} + 8\right)\sigma\delta, \tag{20}
$$

where σ is the variance of the sample point $\widehat{r}_h(s_{h,k}, a_{h,k}) + \int_\mathcal{A} \widehat{Q}_h(s_{h,k}, a)\pi(a \mid s_{h,k})da$ with mean $\bar{\mathcal{T}}_h\widehat{Q}_h(s_h, a_h)$, strictly upper bounded by HR. Note that equation 20 is controlled by two terms: the L^2 approximation ability of $\mathcal{F}_h$ under distribution q_h^π, and the logarithm of function class covering number, which represents the function class complexity. We now only need to estimate the term $\inf_{f \in \mathcal{F}(R,\|,L,p,I)} \int_\mathcal{X} \left(f(s,a) - \bar{\mathcal{T}}_h^\pi \widehat{Q}_h^\pi(s,a)\right)^2 dq_h^{\pi_0}(s,a)$. Here the only problem is $\widehat{r}_h$ is generated by a neural network and is therefore not Holder in general, which makes it untractable. Nevertheless, by Theorem 4.4, the difference between $\hat{r}_h$ and the Hölder function r_h can be bounded, and we have

$$
\inf_{f \in \mathcal{F}(R,\tau,L,p,I)} \int_\mathcal{X} \left(f(s,a) - \bar{\mathcal{T}}_h^\pi \widehat{Q}_h^\pi(s,a)\right)^2 dq_h^{\pi_0}(s,a)
$$
$$
\le 2\Bigg\{ \inf_{f \in \mathcal{F}(R,\tau,L,p,I)} \int_\mathcal{X} \left(f(s,a) - r_h(s,a) - \mathcal{P}_h^\pi \widehat{Q}_h^\pi(s,a)\right)^2 dq_h^{\pi_0}(s,a)
$$
$$
+ \int_\mathcal{X} \left(r_h(s,a) - \widehat{r}_h(s,a)\right)^2 dq_h^{\pi_0}(s,a) \Bigg\}
$$
$$
\le 2\Bigg\{ \inf_{f \in \mathcal{F}(R,\tau,L,p,I)} \int_\mathcal{X} \left(f(s,a) - r_h(s,a) - \mathcal{P}_h^\pi \widehat{Q}_h^\pi(s,a)\right)^2 dq_h^{\pi_0}(s,a)
$$
$$
+ O\left(\epsilon_0^2 + \epsilon_0^{-2d_{HF}/\alpha_{HF}} \cdot \log^3 \frac{2}{\epsilon_0} \cdot \frac{\log K_{HF}}{K_{HF}}\right) \Bigg\}, \tag{21}
$$

Plugging this back to ϵ_{approx} and combining with ϵ_{approx}, we have

$$
\text{(ii)} \le \Bigg\{ \inf_{f \in \mathcal{F}(R,\tau,L,p,I)} \int_\mathcal{X} \left(f(s,a) - r_h(s,a) - \mathcal{P}_h^\pi \widehat{Q}_h^\pi(s,a)\right)^2 dq_h^{\pi_0}(s,a)
$$
$$
+ O\left(\epsilon_0^2 + \epsilon_0^{-2d_{HF}/\alpha_{HF}} \cdot \log^3 \frac{2}{\epsilon_0} \cdot \frac{\log K_{HF}}{K_{HF}}\right)
$$
$$
+ \frac{R^2}{K} \log \mathcal{N}\left(\delta/4R, \mathcal{F}(R,\kappa,L,p,K), \|\cdot\|_\infty\right) + \left(4 + \frac{1}{2R}\right)\delta
$$
$$
+ \sigma^2 \frac{\log \mathcal{N}\left(\delta, \mathcal{F}(R,\kappa,L,p,I), \|\cdot\|_\infty\right)}{K}
$$
$$
+ \sigma\delta \cdot \sqrt{\frac{\log \mathcal{N}\left(\delta, \mathcal{F}(R,\kappa,L,p,I), \|\cdot\|_\infty\right)}{K}} \Bigg\}^{1/2} \cdot \sqrt{1 + \chi_\mathcal{Q}^2(q_h^\pi, q_h^{\pi_0})}
$$
$$
+ O\left(\epsilon_0^2 + \epsilon_0^{-2d_{HF}/\alpha_{HF}} \cdot \log^3 \frac{2}{\epsilon_0} \cdot \frac{\log K_{HF}}{K_{HF}}\right) \cdot \sqrt{1 + \chi_\mathcal{Q}^2(q_h^\pi, \eta_h)}.
$$

Combining everything together, we have

$$
\mathbb{E}\left[\int_\mathcal{X} |Q_h^\pi(s,a) - \widehat{Q}_h^\pi(s,a)| dq_h(s,a)\right]
$$

$$\lesssim \mathbb{E}\left[\int_{\mathcal{X}}|Q_{h+1}^{\pi}(s,a) - \widehat{Q}_{h+1}^{\pi}(s,a)|dq_{h+1}(s,a)\right]$$

$$+\left\{\inf_{f\in\mathcal{F}(R,\tau,L,p,I)}\int_{\mathcal{X}}\left(f(s,a) - r_h(s,a) - \mathcal{P}_h^{\pi}\widehat{Q}_h^{\pi}(s,a)\right)^2 dq_h^{\pi_0}(s,a)\right.$$

$$+O\left(\epsilon_0^2 + \epsilon_0^{-2d_{HF}/\alpha_{HF}}\cdot\log^3\frac{2}{\epsilon_0}\cdot\frac{\log K_{HF}}{K_{HF}}\right)$$

$$+\frac{R^2}{K}\log\mathcal{N}\left(\delta/4R, \mathcal{F}(R,\tau,L,p,K), \|\cdot\|_\infty\right) + \left(4 + \frac{1}{2R}\right)\delta$$

$$+\sigma\delta\cdot\sqrt{\frac{\log\mathcal{N}\left(\delta, \mathcal{F}(R,\tau,L,p,I), \|\cdot\|_\infty\right)}{K}}$$

$$+\sigma^2\frac{\log\mathcal{N}\left(\delta, \mathcal{F}(R,\tau,L,p,I), \|\cdot\|_\infty\right)}{K}\right\}^{1/2}\cdot\sqrt{1 + \chi_{\mathcal{Q}}^2(q_h^{\pi}, q_h^{\pi_0})}$$

$$+O\left(\epsilon_0^2 + \epsilon_0^{-2d_{HF}/\alpha_{HF}}\cdot\log^3\frac{2}{\epsilon_0}\cdot\frac{\log K_{HF}}{K_{HF}}\right)\cdot\sqrt{1 + \chi_{\mathcal{Q}}^2(q_h^{\pi}, \eta_h)}. \tag{22}$$

To balance the bias-variance tradeoff, we only need to initialize the size of function class $\mathcal{F}_h$ and $\mathcal{G}_h$ deliberately. By Lemma 2 in Nguyen-Tang et al. (2022) and Assumption 4.1, in order to yield a function f such that $\|f - \bar{\mathcal{T}}_h^{\pi}\widehat{Q}_h^{\pi}\|_\infty \le \epsilon$, we need

$$L = O\left(\log\frac{1}{\epsilon}\right), p = O\left(\epsilon^{-\frac{d}{\alpha}}\right), I = O\left(\epsilon^{-\frac{d}{\alpha}}\cdot\log\frac{1}{\epsilon}\right),$$

$$\tau = \max\{B, H, \sqrt{d}, \omega^2\}, R = H,$$

recall that d is the intrinsic dimension of (s,a), B is the upper bound of (s,a) in L_∞, ω is the reach of Riemannian manifold. $O(\cdot)$ hides factors of $\log D, \alpha, d$ and surface area of $\mathcal{X}$.

By Lemma C.3, for a ReLU neural network $\mathcal{H}$, we have

$$\log\mathcal{N}(\delta, \mathcal{H}(L,p,I,\tau,R)) = \log\left(\frac{2L^2(pB+2)\tau^L p^{L+1}}{\delta}\right)^I$$

$$\lesssim \epsilon^{-\frac{d}{\alpha}}\log^3\frac{1}{\epsilon}\log\frac{1}{\delta},$$

here $\mathcal{H}\in\{\mathcal{F},\mathcal{G}\}$, we hide the constant depending on $\log B, \omega$ and $\log\log n$. Plugging it in, we have

$$\mathbb{E}\left[\int_{\mathcal{X}}|Q_h^{\pi}(s,a) - \widehat{Q}_h^{\pi}(s,a)|dq_h(s,a)\right]$$

$$\le \mathbb{E}\left[\int_{\mathcal{X}}|Q_{h+1}^{\pi}(s,a) - \widehat{Q}_{h+1}^{\pi}(s,a)|dq_{h+1}(s,a)\right]$$

$$+\left\{O\left(\epsilon_0^2 + \epsilon_0^{-2d_{HF}/\alpha_{HF}}\cdot\log^3\frac{2}{\epsilon_0}\cdot\frac{\log K_{HF}}{K_{HF}}\right)\right.$$

$$+\frac{H^2}{K}\left((\epsilon^{-d/\alpha} + D)\log^3\frac{1}{\epsilon}\log\frac{1}{\delta}\right) + \left(4 + \frac{1}{2H}\right)\delta$$

$$+H\delta\cdot\sqrt{\frac{\epsilon^{-d/\alpha} + \log^3\frac{1}{\epsilon}\log\frac{1}{\delta}}{K}}\right\}^{1/2}\cdot\sqrt{1 + \chi_{\mathcal{Q}}^2(q_h^{\pi}, q_h^{\pi_0})}$$

$$+O\left(\epsilon_0^2 + \epsilon_0^{-2d_{HF}/\alpha_{HF}}\cdot\log^3\frac{2}{\epsilon_0}\cdot\frac{\log K_{HF}}{K_{HF}}\right)\cdot\sqrt{1 + \chi_{\mathcal{Q}}^2(q_h^{\pi}, \eta_h)}.$$

Now we need to balance the tradeoff between model complexity and model approximation ability. We choose ϵ, ϵ_0 to satisfy $\epsilon^2 = \frac{1}{K}\epsilon^{-d/\alpha}$, $\epsilon_0 = \frac{1}{K_{HF}}\epsilon_0^{-2d/\alpha}$, which gives $\epsilon = K^{-\frac{\alpha}{2\alpha+d}}$ and $\epsilon_0 = K^{-\frac{\alpha_{HF}}{2\alpha_{HF}+2d_{HF}}}$. It suffices to pick $\delta = 1/K$. And we conclude that

$$\mathbb{E}\left[\int_{\mathcal{X}}|Q_h^{\pi}(s,a) - \widehat{Q}_h^{\pi}(s,a)|dq_h(s,a)\right]$$

$$\leq \mathbb{E}\left[\int_{\mathcal{X}} |Q_{h+1}^{\pi}(s,a) - \widehat{Q}_{h+1}^{\pi}(s,a)| dq_{h+1}(s,a)\right]$$

$$+ c \cdot \sqrt{1 + \chi_{\mathcal{Q}}^2(q_h^{\pi}, q_h^{\pi_0})} H\left(K^{-\frac{\alpha}{2\alpha+d}} + K_{HF}^{-\frac{\alpha_{HF}}{2\alpha_{HF}+2d_{HF}}} + \sqrt{\frac{D}{K}}\right) \log^{3/2} K$$

$$+ c' \cdot \sqrt{1 + \chi_{\mathcal{Q}}^2(q_h^{\pi}, \eta_h)} \cdot K_{HF}^{-\frac{\alpha_{HF}}{2\alpha_{HF}+2d_{HF}}},$$

where constant c depends on $\log D, \log K_{HF}, d, \alpha, p, q, B, \omega$ and surface area of $\mathcal{X}$. Note that by Lemma 2 in Nguyen-Tang et al. (2022), it suffices to choose $\mathcal{F}(R, \tau, L, p, I)$ such that

$$L = O\left(\frac{\alpha}{2\alpha+d} \cdot \log K\right), p = O\left(K^{\frac{d}{2\alpha+d}}\right), I = O\left(\frac{\alpha}{2\alpha+d} \cdot K^{\frac{d}{2\alpha+d}} \cdot \log K\right),$$

$$\tau = \max\{B, 1, \sqrt{d}, \omega^2\}, R = H.$$

By our selection of $\epsilon_0 = K^{-\frac{\alpha_{HF}}{2\alpha_{HF}+2d_{HF}}}$, by Theorem 4.4, it suffices to choose $\mathcal{G}(\tilde{R}, \tilde{\tau}, \tilde{L}, \tilde{p}, \tilde{I})$ such that

$$\tilde{L} = O\left(\frac{\alpha_{HF}}{2\alpha_{HF}+d_{HF}} \cdot \log K_{HF}\right), \tilde{p} = O\left(K_{HF}^{\frac{d_{HF}}{2\alpha_{HF}+d_{HF}}} \cdot d_{HF}\right), \tilde{I} = O\left(\frac{\alpha}{2\alpha+d} \cdot K_{HF}^{\frac{d_{HF}}{2\alpha_{HF}+d_{HF}}} \cdot d_{HF} \cdot \log K_{HF}\right),$$

$$\tilde{\tau} = \max\{B, 1, \sqrt{d_{HF}}, \omega^2\}, \tilde{R} = 1.$$

By simple induction, we can prove that

$$\mathbb{E}\left[\int_{\mathcal{X}} |Q_1^{\pi}(s,a) - \widehat{Q}_1^{\pi}(s,a)| dq_h(s,a)\right] \leq \sum_{h=1}^{H} \sqrt{1 + \chi_{\mathcal{Q}}^2(q_h^{\pi}, q_h^{\pi_0})} \cdot H\left(K^{-\frac{\alpha}{2\alpha+d}} + \sqrt{\frac{D}{K}}\right) \log^{3/2} K$$

$$+ \sum_{h=1}^{H} \left(\sqrt{1 + \chi_{\mathcal{Q}}^2(q_h^{\pi}, \eta_h)} + \sqrt{1 + \chi_{\mathcal{Q}}^2(q_h^{\pi}, q_h^{\pi_0})}\right) \cdot K_{HF}^{-\frac{d_{HF}}{2\alpha_{HF}+d_{HF}}} \cdot \log K^{3/2}.$$

Taking expectation and we conclude the proof.

$\square$

C Auxiliary Lemma

Lemma C.1. *For a positive random variable X and constant $c_1, c_2 > 0$, if with probability at least $1 - \delta$ that $X \leq (c_1 + \log(1/\delta)) \cdot c_2$ for all $\delta \in [0, 1]$, then we have $\mathbb{E}[X] \leq (1 + c_1) \cdot c_2$.*

Proof. By condition, we have

$$\mathbb{P}\left(\frac{X}{c_2} - c_1 > t \right) \leq e^{-t},$$

invoking $\mathbb{E}[X] = \int_{t \geq 0} \mathbb{P}(X > t) d\mu(t)$, and we have

$$\mathbb{E}\left[\frac{X}{c_2} - c_1 \right] \leq \int_{t \geq 0} e^{-t} dt = 1,$$

and we conclude the proof. $\qquad\square$

Lemma C.2. *Consider a class $\mathcal{F}$ of functions $m_\theta : \theta \in \Theta$ indexed by a parameter θ in an arbitrary index set Θ with a metric d. Suppose that the dependence on θ is Lipschitz in the sense that*

$$|m_{\theta_1}(x) - m_{\theta_2}(x)| \leq d(\theta_1, \theta_2) F(x)$$

for some function $F : \mathcal{X} \to \mathbb{R}$, for every $\theta_1, \theta_2 \in \Theta$ and $x \in \mathcal{X}$. Then, for any norm $\| \cdot \|$, the bracketing numbers of this class are bounded by the covering numbers:

$$N_{[]}(\mathcal{F}, \| \cdot \|, 2\epsilon \|F\|) \leq N(\Theta, d, \epsilon).$$

Proof. See Lemma 2.14 in Sen (2018) for details. $\qquad\square$

Lemma C.3 (Lemma 7 in Chen et al. (2022))**.** *Given $\delta > 0$, the δ-covering number of the neural network class $\mathcal{F}(R, \kappa, L, p, K)$ satisfies*

$$\mathcal{N}\left(\delta, \mathcal{F}(R, \kappa, L, p, K), \| \cdot \|_\infty \right) \leq \left(\frac{2L^2(pB + 2)\tau^L p^{L+1}}{\delta} \right)^I.$$

Lemma C.4 (Lemma 5 in Chen et al. (2022))**.** *Fix the neural network class $\mathcal{F}(M, L, J, I, \tau_1, \tau_2, V)$. Let $\mathcal{X}$ be a d-dimensional compact Riemannian manifold that satisfies Assumption 4.1. We are given a function f_0. We are also given samples $S_n = \{(x_i, y_i)\}_{i=1}^n$, where x_i are i.i.d. sampled from a distribution $\mathcal{P}_x$ on $\mathcal{X}$ and $y_i = f_0(x_i) + \zeta_i$. ζ_i's are i.i.d. sub-Gaussian random noise with variance σ^2, uncorrelated with x_i's. If we compute an estimator*

$$\widehat{f}_n = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (f(x_i) - y_i)^2$$

with the neural network class $\mathcal{F} = \mathcal{F}(M, L, J, I, \tau_1, \tau_2, V)$. For any constant $\delta \in (0, 2V)$, we have

$$\mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n (\widehat{f}_n(x) - f_0(x))^2 \right] \leq 2 \inf_{f \in \mathcal{F}(M, L, J, I, \tau_1, \tau_2, V)} \int_{\mathcal{X}} (f(x) - f_0(x))^2 \, d\mathcal{P}_x(x)$$

$$+ 24\sigma^2 \frac{\log \mathcal{N}(\delta, \mathcal{F}(M, L, J, I, \tau_1, \tau_2, V), \| \cdot \|_\infty) + 2}{n}$$

$$+ \left(4\sqrt{6} \sqrt{\frac{\log \mathcal{N}(\delta, \mathcal{F}(M, L, J, I, \tau_1, \tau_2, V), \| \cdot \|_\infty) + 2}{n}} + 4 \right) \sigma\delta,$$

where $\mathcal{N}(\delta, \mathcal{F}(M, L, J, I, \tau_1, \tau_2, V), \| \cdot \|_\infty)$ denotes the δ-covering number of $\mathcal{F}(M, L, J, I, \tau_1, \tau_2, V)$ with respect to the ℓ_∞ norm, i.e., there exists a discretization of $\mathcal{F}(M, L, J, I, \tau_1, \tau_2, V)$ into $\mathcal{N}(\delta, \mathcal{F}(M, L, J, I, \tau_1, \tau_2, V), \| \cdot \|_\infty)$ distinct elements, such that for any $f \in \mathcal{F}$, there is $\bar{f}$ in the discretization satisfying $\|\bar{f} - f\|_\infty \leq \epsilon$.

Lemma C.5. *Given a function class $\mathcal{Q}$ that contains functions mapping from $\mathcal{X}$ to $\mathbb{R}$ and two probability distributions q_1 and q_2 supported on $\mathcal{X}$, for any $g \in \mathcal{Q}$,*

$$\mathbb{E}_{x \sim q_1}[g(x)] \leq \sqrt{\mathbb{E}_{x \sim q_2}\left[g^2(x)\right]\left(1 + \chi_{\mathcal{Q}}^2\left(q_1, q_2\right)\right)}.$$

Proof.

$$\mathbb{E}_{x \sim q_1}[g(x)] = \sqrt{\mathbb{E}_{x \sim q_2}\left[g^2(x)\right] \frac{\mathbb{E}_{x \sim q_1}[g(x)]^2}{\mathbb{E}_{x \sim q_2}\left[g^2(x)\right]}}$$

$$\leq \sqrt{\mathbb{E}_{x \sim q_2}\left[g^2(x)\right] \sup_{f \in \mathcal{Q}} \frac{\mathbb{E}_{x \sim q_1}[f(x)]^2}{\mathbb{E}_{x \sim q_2}\left[f^2(x)\right]}}$$

$$= \sqrt{\mathbb{E}_{x \sim q_2}\left[g^2(x)\right]\left(1 + \chi_{\mathcal{Q}}^2\left(q_1, q_2\right)\right)},$$

where the last step is by the definition of $\chi_{\mathcal{Q}}^2\left(q_1, q_2\right) := \sup_{f \in \mathcal{Q}} \frac{\mathbb{E}_{q_1}[f]^2}{\mathbb{E}_{q_2}[f^2]} - 1$. $\square$

Lemma C.6 (Lemma 6 in Chen et al. (2022)). *For any constant $\delta \in (0, 2R)$, the generalization error $\epsilon_{general}$ satisfies*

$$\epsilon_{general} \leq \frac{104R^2}{3K} \log \mathcal{N}\left(\delta/4R, \mathcal{F}(R, \kappa, L, p, K), \|\cdot\|_\infty\right) + \left(4 + \frac{1}{2R}\right)\delta.$$

Checklist

1. For all models and algorithms presented, check if you include:

   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes/No/Not Applicable] **Yes.**

   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes/No/Not Applicable] **Yes.**

   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes/No/Not Applicable] **N/A.**

2. For any theoretical claim, check if you include:

   (a) Statements of the full set of assumptions of all theoretical results. [Yes/No/Not Applicable] **Yes.**

   (b) Complete proofs of all theoretical results. [Yes/No/Not Applicable] **Yes.**

   (c) Clear explanations of any assumptions. [Yes/No/Not Applicable] **Yes.**

3. For all figures and tables that present empirical results, check if you include:

   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes/No/Not Applicable] **N/A.**

   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes/No/Not Applicable] **N/A.**

   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes/No/Not Applicable] **Yes.**

   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes/No/Not Applicable] **N/A.**

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [Yes/No/Not Applicable] **N/A.**

   (b) The license information of the assets, if applicable. [Yes/No/Not Applicable] **N/A.**

   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes/No/Not Applicable] **N/A.**

   (d) Information about consent from data providers/curators. [Yes/No/Not Applicable] **N/A.**

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes/No/Not Applicable] **N/A.**

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Yes/No/Not Applicable] **N/A.**

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Yes/No/Not Applicable] **N/A.**

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Yes/No/Not Applicable] **N/A.**