Abstract

In this paper, we investigate the problem of offline reinforcement learning with human feedback where feedback is available in the form of preference between trajectory pairs rather than explicit rewards. Our proposed algorithm consists of two main steps: (1) estimate the implicit reward using Maximum Likelihood Estimation (MLE) with general function approximation from offline data and (2) solve a distributionally robust planning problem over a confidence set around the MLE. We consider the general reward setting where the reward can be defined over the whole trajectory and provide a novel guarantee that allows us to learn any target policy with a polynomial number of samples, as long as the target policy is covered by the offline data. This guarantee is the first of its kind with general function approximation. To measure the coverage of the target policy, we introduce a new single-policy concentrability coefficient, which can be upper bounded by the per-trajectory concentrability coefficient. We also establish lower bounds that highlight the necessity of such concentrability and the difference from standard RL, where state-action-wise rewards are directly observed. We further extend and analyze our algorithm when the feedback is given over action pairs.

1 Introduction

In standard reinforcement learning (RL) setting, the agent learns to maximize an observed numerical reward signal. However, finding appropriate numerical rewards can often be challenging in practice, and getting rewards right significantly impacts the effectiveness of RL algorithms (Wirth et al., 2017). To address this challenge, preference-based RL with human feedback (RLHF) has emerged as a promising alternative (Christiano et al., 2017). In RLHF, the agent does not receive a numerical reward signal, but rather feedback from a human expert in the form of preferences for a state-action trajectory in given pairs of trajectories. RLHF has gained considerable attention across multiple application domains, including games (MacGlashan et al., 2017; Christiano et al., 2017; Warnell et al., 2018), large language models (Ziegler et al., 2019; Stiennon et al., 2020; Wu et al., 2021; Nakano et al., 2021; Ouyang et al., 2022; Glaese et al., 2022; Bai et al., 2022; Ramamurthy et al., 2022; Liu et al., 2023), and robot learning (Brown et al., 2019; Shin et al., 2023).

In this work, we focus on the problem of offline RLHF, where the learning process relies exclusively on pre-collected offline data without active interaction with the environment. Offline RL has gained significant attention in various applications where conducting real-time online experiments may be costly. In the context of RLHF, an offline setting is particularly relevant due to the high cost and latency associated with obtaining human feedback. One of the key challenges in offline RL is the limited coverage of available offline data. Since coverage of the entire state-action space is rarely feasible in practice (Chen and Jiang, 2019a), recent empirical and theoretical approaches to offline RL leverage pessimism so as to rely only on the coverage of one comparator policy (possibly the optimal one), i.e., the so-called partial coverage condition (Yu et al., 2020; Kidambi et al., 2020; Rashidinejad et al., 2021a; Li et al., 2022a; Shi et al., 2022; Yin and Wang, 2021; Xie et al., 2021; Uehara and Sun, 2021; Zhan et al., 2022a). In the context of RLHF, it is also crucial to develop algorithms that work under the partial coverage condition.

Despite its significance, there are very few algorithms specifically designed for offline RLHF with strong statistical guarantees. In this work, we provide such algorithms and guarantees when preferences depend on unknown reward functions over trajectories. Notably, we consider general reward functions that can be defined over the whole trajectory rather than just state-action pairs. This is consistent with many practical settings in natural language processing. For instance, all benchmarks presented in RL4LM (Ramamurthy et al., 2022) use metrics defined over the entire trajectories.

1 Department of Electrical and Computer Engineering, Princeton University, Princeton, USA.
2 Department of Computer Science, Cornell University, Ithaca, USA.
3 School of Operations Research and Information Engineering, Cornell University, Ithaca, USA.
Correspondence to: Wen Sun <ws455@cornell.edu>.

Interactive Learning with Implicit Human Feedback Workshop at ICML 2023, Honolulu, Hawaii, USA. Copyright 2023 by the author(s).
Our main contributions can be summarized as follows:

- We propose a simple algorithm with general function approximation that consists of two main steps: (1) estimate the implicit reward using Maximum Likelihood Estimation (MLE) with general function approximation from offline data and (2) solve a distributionally robust planning problem over a confidence set around the MLE.
- We prove that our algorithm can effectively compete with a target policy as long as the offline data cover the target policy. Our analysis leverages a newly defined concentrability coefficient which is tailored to RLHF. As the concentrability coefficient differs from that in the standard RL setting where state-action-wise rewards are directly observed, we establish lower bounds that highlight the necessity of our partial coverage condition.
- We extend the algorithm to the setting where the transition kernel is unknown, where we not only construct confidence sets for the reward function but also for the system dynamics. Notably, even though the reward can be trajectory-wise, we only need to estimate the per-step transition dynamics to ensure efficient learning.
- We further extend our results to the action-based comparison model, where preferences are defined over individual actions instead of entire trajectories based on the advantage function of the optimal policy (Ramachandran and Amir, 2007; Zhu et al., 2023). In comparison to the case of the trajectory-wise comparison model, we can establish a partial coverage guarantee using a concentrability coefficient on pairs of state-action pairs rather than trajectories. In this scenario, our sample complexity only scales with a bound on the advantage function, which can be much smaller than a bound on per-trajectory rewards as shown in Ross et al. (2011); Agarwal et al. (2019).

2 Related Work

Reinforcement learning from human feedback. The closest work to ours is (Zhu et al., 2023), which also studies offline RLHF, but their algorithm and analysis are restricted to linear models. Our algorithm and analysis extend to general function approximation. Indeed, general classes such as neural networks are commonly employed in practice (Christiano et al., 2017; Abdelkareem et al., 2022). In the special case of linear rewards and preferences over trajectories, while our algorithms differ, our guarantees recover theirs. So, our guarantees are more general; see Remark 1. Moreover, they only consider the setting where the transition kernel is known, while our work can also handle unknown transitions. Finally, in the case of action-based preferences, Zhu et al. (2023) cannot provide guarantees with partial coverage, even under their restriction to linear models. We demonstrate how to achieve meaningful guarantees under partial coverage and a soft margin (Assumption 6).

Wirth et al. (2017) provide a survey of Preference-based RL (PbRL). PbRL has received considerable attention in theoretical RL (Yue et al., 2012; Novoseller et al., 2020; Xu et al., 2020; Pacchiano et al., 2021; Chen et al., 2022) but the focus is largely on offline PbRL. To the best of our knowledge, Zhu et al. (2023) is the only previous work to provide theoretical guarantees for offline PbRL.

Offline RL. In offline RL, one of the most critical challenges is addressing the issue of insufficient coverage in the offline data. It is well-known that naive methods are unable to learn the optimal policy in such scenarios (Rashidinejad et al., 2021b). To tackle this problem, numerous algorithms have been proposed with theoretical guarantees (Liu et al., 2020; Kumar et al., 2020; Jin et al., 2021; Rashidinejad et al., 2021b; Uehara and Sun, 2021; Li et al., 2022b; Shi et al., 2022; Jin et al., 2020; Xie et al., 2021; Zhan et al., 2022a).

The most relevant work is (Uehara and Sun, 2021), which focuses on offline model-based RL with general function approximation. However, their methods cannot be directly applied to RLHF since per-step rewards are not observable in our setting. Furthermore, even in the standard RL setting, the construction of confidence intervals differs between our approach and theirs. Another related paper is Cheng et al. (2022), which considers the general offline pessimistic RL framework in the standard setting and also subtracts a reference term in their algorithm, similar to ours. However, our motivations for such reference terms are quite different from theirs. Additional detailed comparisons are given in Section 4.1 and Remark 3.

3 Preliminaries

We first introduce our offline RLHF setting with general function approximation.

Markov decision processes. We consider an episodic time-inhomogeneous Markov Decision Process (MDP) denoted by \( \mathcal{M} \), which consists of a state space \( \mathcal{S} \), an action space \( \mathcal{A} \), an initial state distribution \( P_0 \in \Delta_S \), and a horizon \( H \in \mathbb{N}^+ \). At each step \( h \in [H - 1] \), we use \( P_h : \mathcal{S} \times \mathcal{A} \to \Delta_S \) to denote the ground truth transitions. The ground truth reward function for the entire trajectory is denoted by \( r^* : \mathcal{T} \to [0, r_{\text{max}}] \), where \( \mathcal{T} = (\mathcal{S} \times \mathcal{A})^H \) represents the set of all possible trajectories. Note that \( r^* \) is a trajectory-wise reward, which is more general than state-action-wise rewards commonly considered in standard RL, which is the special case where for some \( \{ r^*_{h\text{max}}^{h+1} \}_{h=1}^H \) we have \( r^*(\tau) = \sum_{h=1}^H r^*_h(s_h, a_h) \) for a trajectory \( \tau = (s_1, a_1, \cdots, s_H, a_H) \).

A history-dependent policy \( \pi := \{ \pi_h \}_{h=1}^H \) is characterized by \( \pi_h : (\mathcal{S} \times \mathcal{A})^{h-1} \times \mathcal{S} \to \Delta_{\mathcal{A}} \), specifying the probability of selecting actions for the agent at each step \( h \in [H] \) based on the entire history. We denote the set of all such history-
We focus on the problem of offline RLHF in this work. Assumption 1. The optimal policy, denoted \( \pi^* \), is the policy that maximizes the expected reward with respect to the true reward \( r^* \) and system dynamics \( P^* \), i.e., \( \pi^* := \arg \max_{\pi \in \Pi_{\text{his}}} J(\pi; r^*, P^*) \). As the true reward function \( r^* \) is dependent on the entire trajectory, the optimal policy \( \pi^* \) is generally history-dependent. Thus, designing offline RLHF algorithms that can learn history-dependent policies is crucial.

For any policy \( \pi \), we can define its state-action visitation measure as follows: \( d^\pi_h(s, a) = \mathbb{P}^s \pi^* P^* (s_h = s, a_h = a), \forall h \in [H] \), where \( \mathbb{P}^s \pi^* P^* (\cdot) \) denotes the distribution of the trajectory when executing policy \( \pi \) in \( P^* \). We will also use \( d^\pi (\tau) \) to denote \( \mathbb{P}^s \pi^* P^* (\tau) \) for the whole trajectory \( \tau \).

A policy is Markovian if at each step it depends solely on the current state. When the reward is state-action-wise and the policy is Markovian, we can define the associated V- and Q-functions as \( V_h^\pi(s) = \mathbb{E}_\pi \left[ \sum_{t=1}^H r_t^\pi(s_t, a_t) | s_h = s, \forall h \in [H] \right] \), \( Q_h^\pi(s, a) = \mathbb{E}_\pi \left[ \sum_{t=1}^H r_t^\pi(s_t, a_t) | s_h = s, a_h = a, \forall h \in [H] \right] \). It is well-known that when the reward is state-action-wise, the optimal policy \( \pi^* \) is both Markovian and deterministic. Furthermore, we have \( V_h^\pi(s) = \sup_\pi V_h^\pi(s) \) and \( Q_h^\pi(s, a) = \sup_\pi Q_h^\pi(s, a) \) for all \( h \in [H] \). For brevity, we will use \( V^\pi \) and \( Q^\pi \) to represent the optimal state-value function and Q-function, respectively.

The advantage function of the optimal policy, denoted by \( A^* \), is defined to be \( A^*_h(s, a) = Q^*_h(s, a) - V^*_h(s) \) for all \( h \in [H] \), \( s \in S, a \in A \).

Offline reinforcement learning with human feedback. We focus on the problem of offline RLHF in this work. Specifically, in the trajectory-based pairwise comparison setting, we are provided with an offline dataset \( \mathcal{D} = \{(\tau^{n,0}, \tau^{n,1}, o^n)\}_{n=1}^N \), where \( \tau^{n,0} = \{s_h^{n,0}, a_h^{n,0}\}_{h=1}^H \) and \( \tau^{n,1} = \{s_h^{n,1}, a_h^{n,1}\}_{h=1}^H \) are i.i.d. sampled from the distributions \( \mu_0 \) and \( \mu_1 \), respectively, and \( o^n \in \{0, 1\} \) indicates preference for \( \tau^{n,1} \) over \( \tau^{n,0} \). We assume it satisfies the following preference model:

**Assumption 1 (Preference-based model).** Given a pair of trajectories \( (\tau^0, \tau^1) \), \( o \in \{0, 1\} \) satisfies
\[
P(o = 1 | \tau_0, \tau_1) = P(\tau_1 \text{ is preferred over } \tau_0 | \tau_0, \tau_1)
= \Phi(r^*(\tau_1) - r^*(\tau_0))
\]
where \( \Phi : \mathbb{R} \to [0, 1] \) is a monotonically increasing link function.

A commonly used link function is the sigmoid function \( \sigma(x) = 1/(1 + \exp(-x)) \), leading to the Bradley-Terry-Luce (BTL) model (Christiano et al., 2017).

The objective of offline RLHF is to learn a high-quality policy \( \hat{\pi} \in \Pi_{\text{his}}, \) i.e., with \( J(\pi_{\text{tar}}; r^*, P^*) - J(\hat{\pi}; r^*, P^*) \leq \epsilon \) where \( \pi_{\text{tar}} \) is a target policy we want to compete with (potentially \( \pi^* \)).

**General function approximation.** In our paper, we estimate the reward \( r^* \) with general function approximation. We introduce a function class \( G_r \), such as linear functions or neural networks, to approximate the true reward. For each \( r \in G_r \) and trajectory pair \( (\tau^0, \tau^1) \), we denote the induced preference model with respect to \( r \) as \( P_r(o | \tau^0, \tau^1) \), defined as
\[
P_r(o = 1 | \tau^0, \tau^1) = \Phi(r(\tau^1) - r(\tau^0)).
\]

We use bracketing numbers to measure the complexity of \( \{P_r : r \in G_r\} \).

**Definition 1 (ε-bracketing number of preferences).** We say \((g^1, g^2)\) is an \( \epsilon \)-bracket if \( g^1(\cdot | \tau^0, \tau^1) \leq g^2(\cdot | \tau^0, \tau^1) \) and \( \|g^1(\cdot | \tau^0, \tau^1) - g^2(\cdot | \tau^0, \tau^1)\| \leq \epsilon \) for all trajectory-pairs \((\tau^0, \tau^1)\). The \( \epsilon \)-bracketing number of \( G_r \), denoted by \( N_{\epsilon G_r} \), is the minimal number of \( \epsilon \)-brackets \((g^n, g^n)_{n=1}^N \) needed so that for any \( r \in G_r \) there is a bracket \( i \in [N] \) containing it, meaning \( g^1(\cdot | \tau^0, \tau^1) \leq P_r(\cdot | \tau^0, \tau^1) \leq g^2(\cdot | \tau^0, \tau^1) \) for all trajectory-pairs \((\tau^0, \tau^1)\).

The \( \epsilon \)-bracket number is widely used in statistics (van de Geer, 2000) to study MLE and related M-estimates. One example for which we can bound the \( \epsilon \)-bracket number is linear rewards under the BTL model (Pacchiano et al., 2021; Zhu et al., 2023).

**Proposition 1.** Suppose \( \|\phi(\tau)\|_2 \leq R \forall \tau \in \mathcal{T}, G_r \subseteq \{r : (\phi(\tau), \theta) : \|\theta\|_2 \leq B\} \) for some featureization \( \phi : \mathcal{T} \to \mathbb{R}^d \) and \( B > 0 \), and the link function is \( \Phi(\cdot) = \sigma(\cdot) \). Then for any \( \epsilon \leq 1 \), \( \log N_{\epsilon G_r} \leq O(d \log \frac{BR^2}{\epsilon}) \).

The proof is deferred to Appendix A. To handle unknown transitions, we similarly use function classes \( \{G_{P_h}\}_{h=0}^{H-1} \) to approximate the transition probabilities \( \{P_h^*\}_{h=0}^{H-1} \). Similarly, we use \( \log N_{\epsilon G_{P_h}} \) to denote the \( \epsilon \)-bracket number of \( G_{P_h} \). The formal definition is deferred to Appendix D.

## 4 Trajectory-Based Pairwise-Comparison with Known Transition

In this section, we present our algorithm and analyze the sample complexity for the trajectory-based pairwise-comparison setting when the ground truth transition \( P^* \) is known. In Sections 5 and 6, we will further explore the unknown transition setting and the action-based comparison setting.
4.1 Algorithm

Our proposed algorithm, FREEHAND described in Algorithm 1, consists of the following two steps.

Confidence set construction via MLE. We construct a confidence set for the ground truth reward from the implicit preference feedback. We achieve this by selecting reward models that nearly maximize the log-likelihood of observed data up to a slackness parameter $\zeta$. We will show that the result, $R(D)$, approximates the following confidence set:

$$ R'(D) := \{ r \in \mathcal{G}_r : \mathbb{E}_{\tau_0 \sim \mu_0, \tau_1 \sim \mu_1} [ | r(\tau_1) - r(\tau_0) | ] \\ - | r^*(\tau_1) - r^*(\tau_0) |^2 \leq \zeta \} $$

for a certain $\zeta$. Here the distance between $r$ and $r^*$ is measured using the total variation distance (i.e., $\ell_1$ norm) of $r(\tau_1) - r(\tau_0)$ and $r^*(\tau_1) - r^*(\tau_0)$ over the offline data.

Distributionally robust policy optimization. After constructing the confidence set, we search for the policy that maximizes the policy value under the least favorable reward model, the $r \in R(D)$ minimizing the policy value $J(\pi; r, P^*)$ minus $\mathbb{E}_{\tau \sim \mu_{\text{ref}}} [ r(\tau) ]$, where $\mu_{\text{ref}}$ is an arbitrary known reference trajectory distribution. It is generally recommended to set $\mu_{\text{ref}}$ to $\mu_1$, as we will explain later, possibly a sample-average approximation thereof based on $\{ \tau^{1,1}, \ldots, \tau^{N,1} \}$. By selecting the least favorable reward model instead of the MLE solution $\hat{\tau}$, we penalize policies that are not well-covered by the offline data. The need for a reference policy arises because the approximated confidence set measures the uncertainty for reward difference between two trajectories ($r(\tau_1) - r(\tau_0)$), but it cannot measure the uncertainty of the reward of a single trajectory.

In the following, we compare our algorithm to existing works. (Zhu et al., 2023) consider a pessimistic offline RL algorithm for RLHF specialized to the linear reward class setting, while our FREEHAND can handle general function approximation. Specifically, they construct the confidence set using the feature-covariance-rotated $\ell_2$-ball around the MLE $\hat{\theta}$, where $\hat{\tau}(\tau) = \langle \phi(\tau), \hat{\theta} \rangle$. In contrast, our confidence set is obtained directly from the log-likelihood objective and is generic. Uehara and Sun (2021) proposes a model-based pessimistic offline RL algorithm when we have access to rewards. The confidence set construction correspondingly differs significantly. Cheng et al. (2022) considers a general offline pessimistic RL framework. In their policy optimization step, they also subtract the value of a reference policy. This similarity is superficial, however, as the motivations are different. We subtract the value because we can only measure the difference between rewards of any two trajectories, while their motivation is to obtain a certain robustness result (their proposition 3).

Algorithm 1 FREEHAND: offline Reinforcement Learning with Human feedback

1: Input: offline dataset $D$, slackness parameter $\zeta$, reference distribution $\mu_{\text{ref}}$, true transition $P^*$
2: MLE: compute $\hat{r} = \arg \max_{r \in \mathcal{G}_r} \sum_{n=1}^{N} \log P_r(o = o^n | \tau^{n,1}, \tau^{n,0})$
3: Confidence set construction: construct $R(D) = \{ r \in \mathcal{G}_r : \sum_{n=1}^{N} \log P_r(o = o^n | \tau^{n,0}, \tau^{n,1}) \geq \sum_{n=1}^{N} \log P_{\hat{r}}(o = o^n | \tau^{n,0}, \tau^{n,1}) - \zeta \}$
4: Distributionally robust planning: return $\hat{\pi} = \arg \max_{\pi \in \Pi_{\text{ref}}} \min_{r \in R(D)} (J(\pi; r, P^*) - \mathbb{E}_{\tau \sim \mu_{\text{ref}}} [ r(\tau) ])$. 

4.2 Analysis

To analyze the sample complexity of FREEHAND, we first quantify the discrepancy between the offline data $D$ and the distribution induced by the target policy $\pi_{\text{tar}}$.

Definition 2 (Concentrability coefficient w.r.t. a reward class $\mathcal{G}_r$, a target policy $\pi_{\text{tar}}$, and a reference policy $\mu_{\text{ref}}$ is defined as

$$ C_r(\mathcal{G}_r, \pi_{\text{tar}}, \mu_{\text{ref}}) := \max \left\{ 0, \sup_{r \in \mathcal{G}_r} \mathbb{E}_{\tau \sim \pi_{\text{tar}}, \tau \sim \mu_{\text{ref}}} [ r^*(\tau^0) - r^*(\tau^1) - r(\tau^0) + r(\tau^1) ] \right\}. $$

Note, when we choose $\mu_{\text{ref}} = \mu_1$, by Jensen’s inequality, the value of $C_r(\mathcal{G}_r, \pi_{\text{tar}}, \mu_1)$ can always be upper bounded by the per-trajectory concentration coefficient: $C_r(\mathcal{G}_r, \pi_{\text{tar}}, \mu_1) \leq \sqrt{C_{\text{tr}}}$ for any $\mathcal{G}_r$, where $C_{\text{tr}} := \max_{\tau \in \mathcal{T}} \mathbb{E}_{\tau \sim \mu_{\text{ref}}} [ r^*(\tau^0) - r^*(\tau^1) - r(\tau^0) + r(\tau^1) ]$. Moreover, while $C_r(\mathcal{G}_r, \pi_{\text{tar}}, \mu_1)$ becomes $\sqrt{C_{\text{tr}}}$ in the worst case (e.g., when $\mathcal{G}_r$ is the set of all functions mapping from $\mathcal{T}$ to $\mathbb{R}$), it can generally be much smaller. For example, when using linear models, it is a relative condition number, as explained in Appendix C. Finally, when $\mu_{\text{ref}} = d_{\text{tr}}$, our coefficient becomes 0. This implies that $C_r(\mathcal{G}_r, \pi_{\text{tar}}, \mu_1)$ could be small when $\pi_{\text{tar}}$ and $\mu_{\text{ref}}$ are close. While the concept of concentrability coefficient has been used in offline RL with explicit reward feedback (Chen and Jiang, 2019b; Song et al., 2022), this property is unique when the feedback is in the form of preferences.

In our following PAC analysis, we further assume the reward class $\mathcal{G}_r$ is realizable and bounded.

Assumption 2 (Realizability). We have $r^* \in \mathcal{G}_r$.

Assumption 3 (Boundedness). We have $0 \leq r(\tau) \leq r_{\text{max}}$ for all $r \in \mathcal{G}_r$ and $\tau \in \mathcal{T}$.
Theorem 1. For any $\delta \in (0, 1]$, let $\zeta = c_{\text{MLE}} \log(N_{G_r}(1/N)/\delta)$ where $c_{\text{MLE}} > 0$ is a universal constant, then under Assumption 1, 2, and 3, with probability $1 - \delta$, we have

$$J(\pi_{\text{tar}}; r^*, P^*) - J(\hat{\pi}; r^*, P^*) \leq \sqrt{cC_r^2(G_r, \pi_{\text{tar}}, \mu_{\text{ref}})\kappa^2 \log(N_{G_r}(1/N)/\delta)} \cdot \frac{1}{\epsilon^2},$$

where $c > 0$ is a universal constant and $\kappa = (\inf_{x \in [-r_{\max}, r_{\max}]} \Phi(x))^{-1}$.

Theorem 1 indicates that FREEHAND can learn an $\epsilon$-optimal policy compared to $\pi_{\text{tar}}$ with a sample complexity of

$$N = \tilde{O}\left(\frac{C_r^2(G_r, \pi_{\text{tar}}, \mu_{\text{ref}})\kappa^2 \log(N_{G_r}(1/N)/\delta)}{\epsilon^2}\right).$$

Next we provide a detailed explanation of this sample complexity. Firstly, $C_r(G_r, \pi_{\text{tar}}, \mu_{\text{ref}})$ represents the extent to which the dataset $\mathcal{D}$ covers the target policy $\pi_{\text{tar}}$. In our theorem, to obtain a non-vacuous PAC guarantee, we only require the dataset $\mathcal{D}$ to cover the target policy $\pi_{\text{tar}}$ (i.e., $C_r(G_r, \pi_{\text{tar}}, \mu_{\text{ref}}) < \infty$). The distributionally robust optimization step plays a crucial role in obtaining this guarantee under partial coverage. In particular, invoking the above-mentioned third property of $C_r(G_r, \pi_{\text{tar}}, \mu_{\text{ref}})$, when setting $\pi_{\text{ref}} = \mu_{\text{ref}}$, (2) is reduced to

$$J(\mu_{\text{ref}}; r^*, P^*) \leq J(\hat{\pi}; r^*, P^*)$$

This encourages us to choose $\mu_{\text{ref}} = \mu_1$ (or $\mu_0$) as it will allow us to ensure our performance is at least larger than the performance associated with the offline data.

Secondly, $\log(N_{G_r}(1/N))$ measures the complexity of the function class $G_r$. For example, when using linear models, it takes $\tilde{O}(d)$. We refer the reader to van de Geer (2000) for bracketing number computations for more general classes.

Thirdly, $\kappa$ represents the non-linearity of the link function $\Phi$, which determines the difficulty of estimating the reward from human preferences. This dependence on $\kappa$ is present in the existing literature of RLHF, both in online settings (Pacchiano et al., 2021; Chen et al., 2022) and offline settings (Zhu et al., 2023).

Remark 1 (Comparison to Zhu et al. (2023)). By specializing our result to the linear model, we recover the result in Zhu et al. (2023). Specifically, the bracketing number is calculated in Proposition 1, and $C_r(G_r, \pi_{\text{tar}}, \mu_{\text{ref}})$ is reduced to a relative condition number. The details are deferred to Appendix C.

Remark 2. In practice, to compute $E_{r \sim \mu_1}[r(\tau)]$ in the second step, we can use the sample average, with an additional cost of $\sqrt{\log(1/\delta)}/N$ in the suboptimality bound in Eq. (2).

4.3 Discussion of the Concentrability Coefficient

In the worst-case scenario (i.e., $G_r$ is the set of all functions mapping from $\mathcal{F}$ to $\mathbb{R}$), the value of $C_r(G_r, \pi_{\text{tar}}, \mu_1)$ is reduced to the per-trajectory concentrability coefficient $C_{\text{tr}}$. The per-trajectory concentrability coefficient is generally larger than the per-step concentrability coefficient $C_{\text{st}}$ commonly used in the general offline RL literature. Specifically, $C_{\text{st}}$ is defined as

$$C_{\text{st}} := \max_{s, a, h} \frac{d^2_{\text{tr}}(s, a)}{\mu_{0, h}(s, a)},$$

where $\mu_{0, h}(s, a)$ represents the marginal distribution at step $h$. In this section, we show the dependence on the per-trajectory concentrability coefficient is necessary for our offline RLHF context. This is intuitively because our RLHF setting involves reward functions defined over trajectories, reflecting the fact that human feedback is also trajectory-based.

In the next proposition, we first show that the per-trajectory concentrability coefficient $C_{\text{tr}}$ can be exponentially larger than the per-step concentrability coefficient $C_{\text{st}}$.

Proposition 2. For any $S \geq 1, A \geq 2, H \geq 1, C \geq 1$, there exists an MDP $\mathcal{M}$ with horizon $H$, a policy $\pi_{\text{tar}}$ and a data distribution $\mu_0$ such that $|S| = S, |A| = A$ and $C_{\text{tr}} = C$ while $C_{\text{st}} = C^H$.

Proposition 2 indicates that $C_{\text{tr}}$ can be significantly larger than $C_{\text{st}}$. A natural question arises as to whether we can obtain suboptimality guarantees using $C_{\text{st}}$. Unfortunately, the following lower bounds reveal that the suboptimality can scale with $C_{\text{st}}^{H-1}$ in the worst case:

Theorem 2. Set $\pi_{\text{tar}} = \pi^*$. Then, for any $C > 1$ and $H \geq 1$, there exists a dataset distribution $\mu_1$ such that we have

$$\inf_{\hat{\pi}} \sup_{(M, \mu_0) \in \Theta_{\text{st}}(C)} \mathbb{E}_{\mathcal{D}}[J(\pi^*; r^*, P^*) - J(\hat{\pi}; r^*, P^*)] \geq \min\left\{C - 1, \sqrt{\frac{\max\{C, 2\}^{H-1}(C - 1)}{N}}\right\},$$

where $\hat{\pi}$ is any measurable function of the data $\mathcal{D}$ (and knows the information of $\mu_1$). $\Theta_{\text{st}}(C)$ is the set of all MDP, offline distribution $(M, \mu_0)$ such that $C_{\text{st}} \leq C$. Note $\mathbb{E}_{\mathcal{D}}$ is taken with respect to the randomness in $\mathcal{D}$.

In addition, with similar hard instances constructed in Theorem 2, we can show that scaling with $C_{\text{tr}}$ is necessary in our setting:

Theorem 3. Set $\pi_{\text{tar}} = \pi^*$. Then, for any $C > 1$ and $H \geq 1$, there exists a dataset distribution $\mu_1$ such that we
have
\[
\inf \sup_{\pi} \mathbb{E}_D [J(\pi^*; r^*, P^*) - J(\tilde{\pi}; r^*, P^*)] \geq \min \left\{ C - 1, \sqrt{\frac{C - 1}{N}} \right\},
\]
where \( \tilde{\pi} \) is any measurable function of the data \( D \) (and knows the information of \( \mu_1 \)). \( \Theta_{tr} \) is the set of all MDP, offline distribution \((\mathcal{M}, \mu_0)\) such that \( C_{tr} \leq C \). Note \( \mathbb{E}_D \) is taken with respect to the randomness in \( D \).

Note that when \( \mu_1 \) is known, we can set \( \mu_{ref} = \mu_1 \) in Algorithm 1 and then \( C_\tau(\mathcal{G}, \pi_{tar}, \mu_1) \leq \sqrt{C_{tr}} \), which implies the sample complexity in Theorem 1 indeed nearly matches this lower bound with respect to \( C_{tr} \) and \( N \) when \( N \) is sufficiently large.

In summary, Theorem 2 and Theorem 3 imply that the scaling with the per-trajectory concentrability coefficient is essential in the trajectory-based pairwise-comparison setting, and it cannot be relaxed to the per-step concentrability without additional assumptions, such as on the reward structure.

## 5 Trajectory-Based Comparison with Unknown Transition

We extend the setting presented in Section 4 to the scenario where the transition function \( P^\star \) is unknown. The algorithm is described in Algorithm 2. Compared to Algorithm 1, we simply added a similar step to handle unknown transitions. Hereafter, we use the convention \( P_0(h|s,a) := P_0(h) \).

Our sample complexity will depend on the following additional concentration coefficient:

**Definition 3** (Concentrability coefficient for the transition). The concentrability coefficient w.r.t. transition classes \( \{\mathcal{G}_{\text{h}}\} \) and a target policy \( \pi_{\text{tar}} \) is defined as
\[
C_\tau(\{\mathcal{G}_{\text{h}}\}, \pi_{\text{tar}}) := \max_{h:0 \leq h \leq H-1} \sup_{\mathcal{G}_{\text{h}}} \mathbb{E}_{h,0 \leq h \leq H-1} [P_h(h|s,a) - P_h(h|s,a)]
\]
\[
\sqrt{\mathbb{E}_{h,0 \leq h \leq H-1} [P_h(h|s,a) - P_h(h|s,a)]^2}.
\]

Note this is always upper-bounded by the density-ratio-based concentrability coefficient, \( C_P(\{\mathcal{G}_{\text{h}}\}, \pi_{\text{tar}}) \leq \sup_{(s, a, h) \in S \times A \times [H]} d_{\text{tr}}^2(s, a, h) \).

We also assume the transition classes \( \{\mathcal{G}_{\text{h}}\}^{H-1}_{h=0} \) are realizable:

**Assumption 4** (Realizability). Suppose that we have \( P^\star_h \in \mathcal{G}_{\text{h}} \) for all \( h \) where \( 0 \leq h \leq H - 1 \). In addition, any choice \( P_h \in \mathcal{G}_{\text{h}} \) for \( 0 \leq h \leq H - 1 \) are valid transition distributions.

**Algorithm 2** FREEHAND–transition

<table>
<thead>
<tr>
<th>Input</th>
<th>offline dataset ( D ), slackness parameter ( \zeta, \zeta_{\text{ref}} ), reference distribution ( \mu_{\text{ref}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLE for reward</td>
<td>compute ( \hat{r} = \arg\max_{r \in \mathcal{G}} \sum_{n=1}^N \log P_r(o = o^n</td>
</tr>
<tr>
<td>MLE for transition</td>
<td>compute ( \hat{P}<em>h = \arg\max</em>{h \in \mathcal{G}} \sum_{n=1}^N \sum_{i=0}^1 \log P_h(s_{h+1}^n</td>
</tr>
<tr>
<td>Confidence set construction</td>
<td>for ( 0 \leq h \leq H - 1 ), construct ( R(D) = \left{ r \in \mathcal{G} : \sum_{n=1}^N \log P_r(o = o^n</td>
</tr>
<tr>
<td>Distributionally robust planning</td>
<td>return ( \tilde{\pi} = \arg\max_{\pi \in \Pi_{\text{h}}} \min_{r \in R(D), r \in \mathcal{G}} J(\pi^*; r, {P_{h \leq h}(H-1)} - \mathbb{E}<em>{\tau \sim \mu</em>{\text{ref}}} [r]) )</td>
</tr>
</tbody>
</table>

Then with the above assumptions, we have the following theorem to characterize the sample complexity when the transition is unknown:

**Theorem 4.** For any \( \delta \in (0, 1] \), let \( \zeta = c_{\text{MLE}} \log(N_{\mathcal{G}}(1/N)/\delta), \zeta_{\text{ref}} = c_{\text{p}} \log(N_{\mathcal{G}}(1/N)/\delta) \), where \( c_{\text{MLE}}, c_{\text{p}} > 0 \) are universal constants, then under Assumption 1,2,3 and 4, we have
\[
J(\pi_{\text{tar}}; r^*, P^*) - J(\tilde{\pi}; r^*, P^*) \leq \sqrt{c_{\text{MLE}}^2(\mathcal{G}, \pi_{\text{tar}}, \mu_{\text{ref}}) N \log(N_{\mathcal{G}}(1/N)/\delta)}
\]
\[
+ H \max \sqrt{c_{\text{p}}^2(\mathcal{G}, \pi_{\text{tar}}) \log(N_{\mathcal{G}}(1/N)/\delta)}
\]
where \( c > 0 \) and \( \kappa \) are the same as Theorem 1 and \( N_{\mathcal{G}} := \max_{0 \leq h \leq H-1} N_{\mathcal{G}} \).

Compared to Theorem 1, we introduce an additional term in our guarantee to account for the unknown transitions. Once again, our result demonstrates that the learned policy can achieve performance comparable to any target policy \( \pi_{\text{tar}} \) covered by the offline data, i.e., \( C_\tau(\mathcal{G}, \pi_{\text{tar}}, \mu_{\text{ref}}) < \infty, C_P(\{\mathcal{G}_{\text{h}}\}, \pi_{\text{tar}}) < \infty \).

**Remark 3** (Comparison to Uehara and Sun (2021)). Like us, Uehara and Sun (2021) proposed a model-based RL algorithm that works under partial coverage, but in the standard RL setting and with a known state-action-wise reward function. In addition to the difference in settings, which is the primary difference, our approach moreover differs from their approach because while they construct
6 Action-Based Comparison

Next, we turn our attention to the action-based comparison setting (Ramachandran and Amir, 2007; Zhu et al., 2023), where human evaluators provide preferences between pairs of actions instead of pairs of trajectories. In this section, we assume that the reward function \( r^* \) is state-action-wise:

\[
r^*(\tau) = \sum_{h=1}^{H} r_h^*(s_h, a_h) \text{ for } \tau = (s_1, a_1, \ldots, s_H, a_H).
\]

And, we consider a preference model based on the action-based comparison model.

**Assumption 4 (Action-based comparison model).** Given a pair of actions \( a_h^0, a_h^1 \) and state \( s_h, \) \( o \in \{0, 1\} \) satisfies

\[
P(o_h = 1 \mid s_h, a_h^0, a_h^1) = \Phi(Q_h^*(s_h, a_h^1) - Q_h^*(s_h, a_h^0)).
\]

Here, the aforementioned distribution can be equivalently expressed as

\[
P(o_h = 1 \mid s_h, a_h^0, a_h^1) = \Phi(A_h^*(s_h, a_h^0) - A_h^*(s_h, a_h^1)),
\]

where \( A_h \) denotes the optimal advantage function. Consequently, we introduce general function classes \( \mathcal{G}_h \) to estimate the optimal advantage function \( A_h^* \). In addition, for each \( A_h \in \mathcal{G}_h \) and \( (s, a^0, a^1) \in S \times \mathcal{A} \times \mathcal{A} \), we use \( P_h(\cdot \mid s, a, a') \) to represent the human preference model with respect to \( A_h \), defined as

\[
P_h(o = 1 \mid s, a, a') := \Phi(A_h(s, a') - A_h(s, a)).
\]

We again use the \( \epsilon \)-bracket number of such advantage function classes to quantify their complexity, denoted as \( \mathcal{N}_{\mathcal{G}_h} \).

The full formal definition is provided in Appendix D.

6.1 Algorithm

Our algorithm comprises two steps. In the first step, our objective is to estimate the optimal advantage function using MLE. In the second step, we determine the policy by selecting the action with the highest advantage value based on the learned advantage function.

Algorithm 3 FREEHAND-action

1. **Input:** offline dataset \( D \).
2. **MLE:** compute \( A_h = \arg\max_{A_h \in \mathcal{G}_h} \sum_{n=1}^{N} \log P_h(o = n | s_h^n, a_h^n, a_h^{n+1}) \), \( \forall h \in [H] \).
3. **Greedy policy:** return \( \hat{\pi}_h(s) = \arg\max_{a \in \mathcal{A}} \hat{A}_h(s, a) \)

policy. Our analysis depends on the following assumption on the margin of \( Q^* \):

**Assumption 6 (Soft margin).** There exists \( \alpha_0 \in \mathbb{R}^+, \beta \in (0, \infty) \) such that for all \( a \in \mathcal{A}, h \in [H], \alpha > 0 \), we have

\[
E_{(s,a^0) \sim d_h^*} = (Q_h^*(s_h, a_h^*) - Q_h^*(s_h, a_h^0)) \leq (\alpha/\alpha_0)^{\beta}.
\]

The soft margin is widely used in the literature on classification, decision making, and RL (Audibert and Tsybakov, 2007; Perchet and Rigollet, 2013; Luedtke and Chambaz, 2020; Hu et al., 2021; 2022; Uehara et al., 2023). Note, when the optimal Q function satisfies a gap (as in Simchowitz and Jamieson, 2019; Wu et al., 2022), the soft margin assumption holds with \( \beta = \infty \).

Next, we introduce the concentrability coefficient for the action-based comparison setting, which is defined as follows.

**Definition 4 (Concentrability coefficient for action-based comparison).**

\[
C_{act} := \sup_{h \in [H], A_h \in \mathcal{G}_h} \frac{E_{(s,a^0) \sim d_h^*, a^1 \sim \text{Unif}(|s|)}[l(A_h(s, a^0, a^1))]}{E_{s \sim \mu_{0,h}, a^0 \sim \mu_{0,h}(\cdot | s), a^1 \sim \mu_{1,h}(\cdot | s)}[l(A_h(s, a^0, a^1))]},
\]

where \( l(A_h(s, a^0, a^1)) := |A_h^*(s, a^0) - A_h^*(s, a^1)|^2 \) and \( \text{Unif}(|s|) \) is the uniform policy over \( \mathcal{A} \).

We observe that

\[
C_{act} \leq \left( \sup_{h \in [H], s \in S} \frac{\pi_h^*(a^0 | s)}{d_h^*(s)} \right) \cdot \left( \sup_{h \in [H], s \in S, a \in \mathcal{A}} \frac{\pi_h^*(a^0 | s)}{\mu_{0,h}(a^0 | s)} \right) \cdot \left( \frac{1}{|\mathcal{A}|} \sup_{h \in [H], s \in S, a \in \mathcal{A}} \frac{1}{\mu_{1,h}(a^1 | s)} \right).
\]

Based on this bound, we can consider simple sufficient conditions for \( C_{act} \) to be finite. Firstly, regarding the first term, it is sufficient for the dataset distribution \( \mu_h \) to cover the states visited by the optimal policy \( \pi^* \), denoted as \( d_h^* \). Regarding the second term, we require \( \mu_{0,h} \) to cover \( \pi_h^* \). Additionally, the third term can be upper bounded when \( \mu_{1,h} \) can cover the whole action space. This is mild because \( \forall s \in \ldots \)
With the aforementioned assumptions, we can establish the
/ \cdot (1 / \epsilon)^2. As mentioned earlier, the quantity C_{act} represents the extent to which the
distribution induced by the optimal policy is covered by the
offline data. Therefore, there is no need for a potentially
stringent condition that requires the offline data to cover the
total state space like Zhu et al. (2023).

Furthermore, our guarantee is designed to overcome the
limitations of existing approaches. In Theorem 1, our upper-
bound is influenced by the parameter $\kappa$. When using a
common sigmoid link function, this parameter scales with
$\Theta(\exp(r_{max}))$. As a result, in dense reward settings where
$r_{max}$ scales with $H$, this scaling factor may lead to an explicit
dependence of $\Theta(\exp(H))$. Similar observations have been
made in previous works (Zhu et al., 2023; Pacchiano et al., 2021; Chen et al., 2022). However, even if $r_{max}$ scales with $H$, it is known that the $\ell_{\infty}$-norm of the advantage function, denoted as $b_{\max}$, can take much smaller values (Ross et al., 2011; Agarwal et al., 2019). Hence, we can avoid the explicit dependence on $\Theta(\exp(H))$.

7 Conclusions

We propose the first algorithm for RLHF with preferences
over trajectories with general function approximation and
under partial coverage. We establish lower bounds that explain the differences between our RLHF model and standard
RL with direct reward feedback. Moreover, we extend our
algorithm to unknown transitions and to preference feedback
over actions, while maintaining strong guarantees under partial coverage.

References


\S; \mu_h(s) > 0 \text{ is not controllable to the learner; but } \forall (s, a) \in \mathcal{S} \times \mathcal{A}; \mu_{1,h}(a | s) > 0 \text{ is controllable to the learner in the data-collection process. To summarize, } C_{act} < \infty \text{ primarily requires partial coverage over the state space with respect to the optimal policy, which is preferable in practical applications where } \mathcal{S} \text{ can be very large. Additionally, we introduce several assumptions on the function classes similar to those in Section 4.}

Assumption 7. For all $h \in [H]$, we have $A^*_h \in \mathcal{G}_{A_h}$.

Assumption 8. For all $h \in [H]$ and $A_h \in \mathcal{G}_{A_h}$, we have $|A_h(s, a)| \leq b_{\max}$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$.

With the aforementioned assumptions, we can establish the sample complexity of FREEHAND-action.

Theorem 5. Under Assumption 5,6,7 and 8, we have with probability at least $1 - \delta$ that

$$J(\pi^*; r^*, P^*) - J(\tilde{\pi}; r^*, P^*) \leq cH|A| \left(\frac{2}{|A|} \beta \right) \frac{1}{\alpha_0} \left(\frac{1}{\delta} \right)^{2/\beta} \left(\frac{\kappa^2 C_{act} \log(HN_{\mathcal{G}_{A_h}}(1/N)/\delta)}{N}\right)^{2/\beta},$$

where $N_{\mathcal{G}_{A_h}} := \max_{h \in [H]} \inf_{\pi \in [-b_{\max}, b_{\max}] \mathcal{G}_{A_h}} \Psi(\pi)$.

Theorem 5 suggests that FREEHAND-action can learn a near-optimal policy as long as $C_{act}$ takes a finite value under a soft margin. Specifically, when a hard margin is imposed (i.e., $\beta = \infty$), FREEHAND-action can learn an $\epsilon$-optimal policy with a sample complexity of $N = \mathcal{O}(1/\epsilon^2)$, which is faster than a typical rate $\mathcal{O}(1/\epsilon^2)$. As mentioned earlier, the quantity $C_{act}$ represents the extent to which the distribution induced by the optimal policy is covered by the offline data. Therefore, there is no need for a potentially stringent condition that requires the offline data to cover the entire state space like Zhu et al. (2023).


A Proof of Proposition 1

Let \( \mathcal{F} \) denote the function class \( \{ f_r : f_r(r^0, \tau^1) = P_r(o = 1|r^0, \tau^1), r \in G_r \} \). Let \( \mathcal{I}_\mathcal{F}(\epsilon) \) denote the \( \epsilon \)-bracket number with respect to \( \ell_\infty \)-norm, i.e., the minimum integer \( M \) such that there exist \( M \) functions \( \{ f^1, \ldots, f^M \} \) such that for each \( f_r \in \mathcal{F} \), we have \( \sup_{r^0, \tau^1} |f_r(r^0, \tau^1) - f^i(r^0, \tau^1)| \leq \epsilon \) for some \( i \in [M] \). Then we know there exists a set of function \( \mathcal{F} \) with \( |\mathcal{F}| = \mathcal{I}_\mathcal{F}(\epsilon/4) \) such that for each \( f_r \in \mathcal{F} \), there exists \( \mathcal{F} \) satisfying
\[
\sup_{r^0, \tau^1} |f_r(r^0, \tau^1) - \mathcal{F}(r^0, \tau^1)| \leq \epsilon/4.
\]

Now we construct a bracket \( (g^1, g^2) \) defined as follows:
\[
g^1(o = 1|r^0, \tau^1) = \mathcal{F}(r^0, \tau^1) - \epsilon/4, \\
g^1(o = 0|r^0, \tau^1) = 1 - \mathcal{F}(r^0, \tau^1) - \epsilon/4, \\
g^2(o = 1|r^0, \tau^1) = \mathcal{F}(r^0, \tau^1) + \epsilon/4, \\
g^2(o = 0|r^0, \tau^1) = 1 - \mathcal{F}(r^0, \tau^1) + \epsilon/4.
\]

Then clearly we have \( g^1(o = 1|\tau^0, \tau^1) \leq P_r(o = 1|\tau^0, \tau^1) \leq g^2(o = 1|\tau^0, \tau^1) \) and \( \|g^1(o = 1|\tau^0, \tau^1) - g^2(o = 1|\tau^0, \tau^1)\|_1 \leq \epsilon \). This implies that \( \mathcal{N}_\mathcal{F}(\epsilon, \mathcal{I}_\mathcal{F}(\epsilon/4)) \).

Now we only need to bound \( \mathcal{I}_\mathcal{F}(\epsilon/4) \). Consider \( \theta \) and \( \theta' \) with \( \|\theta - \theta'\|_2 \leq \epsilon_1 \) and let \( r(r') \) denote the reward \( \langle \phi, \theta \rangle \) \( \langle \phi, \theta' \rangle \). Then we know for all \( \tau \),
\[
|r(\tau) - r'(\tau)| \leq R \epsilon_1.
\]

Fix the trajectory pair \( (\tau^0, \tau^1) \). Without loss of generality, we assume \( \exp(r'(r^0)) + \exp(r'(r^1)) \leq \exp(r'(r^0)) + \exp(r'(r^1)) \). Then we have
\[
\exp(r(r^0)) + \exp(r(r^1)) \leq \exp(r'(r^0)) + \exp(r'(r^1)) \leq \exp(R \epsilon_1) \left( \exp(r(r^0)) + \exp(r(r^1)) \right).
\]

On the other hand, we have
\[
|f_r(r^0, \tau^1) - f_r(r^0, \tau^1)| = \frac{1}{\left( \exp(r'(r^0)) + \exp(r'(r^1)) \right)} \\ \cdot \left( \frac{1}{\exp(r(r^0)) + \exp(r(r^1))} \right) \\ \cdot \left( \exp(r(r^1)) \left( \exp(r'(r^0)) + \exp(r'(r^1)) \right) - \exp(r'(r^1)) \left( \exp(r(r^0)) + \exp(r(r^1)) \right) \right).
\]

Therefore, if \( \exp(r(r^1)) \left( \exp(r'(r^0)) + \exp(r'(r^1)) \right) - \exp(r'(r^1)) \left( \exp(r(r^0)) + \exp(r(r^1)) \right) \geq 0 \), then we have
\[
\left( \exp(r(r^1)) \left( \exp(r'(r^0)) + \exp(r'(r^1)) \right) - \exp(r'(r^1)) \left( \exp(r(r^0)) + \exp(r(r^1)) \right) \right) \leq \exp(R \epsilon_1) \exp(r(r^1)) \left( \exp(r(r^0)) + \exp(r(r^1)) \right) - \exp(-R \epsilon_1) \exp(r(r^1)) \left( \exp(r(r^0)) + \exp(r(r^1)) \right)
\]
\[
= (\exp(R \epsilon_1) - \exp(-R \epsilon_1)) \exp(r(r^1)) \left( \exp(r(r^0)) + \exp(r(r^1)) \right).
\]

Otherwise, we have
\[
\exp(r(r^1)) \left( \exp(r'(r^0)) + \exp(r'(r^1)) \right) - \exp(r'(r^1)) \left( \exp(r(r^0)) + \exp(r(r^1)) \right) \leq \exp(R \epsilon_1) \exp(r(r^1)) \left( \exp(r(r^0)) + \exp(r(r^1)) \right) - \exp(-R \epsilon_1) \exp(r(r^1)) \left( \exp(r(r^0)) + \exp(r(r^1)) \right)
\]
\[
= (\exp(R \epsilon_1) - 1) \exp(r(r^1)) \left( \exp(r(r^0)) + \exp(r(r^1)) \right).
\]

Therefore we have
\[
|f_r(r^0, \tau^1) - f_r(r^0, \tau^1)| \leq \frac{\exp(R \epsilon_1) - \exp(-R \epsilon_1)) \exp(r(r^1)) \left( \exp(r(r^0)) + \exp(r(r^1)) \right)}{\left( \exp(r(r^0)) + \exp(r(r^1)) \right)} \left( \exp(r(r^0)) + \exp(r(r^1)) \right)
\]
\[
\leq \exp(2R \epsilon_1) - 1.
\]

This implies that for any \( \epsilon \leq 1 \),
\[
\log \mathcal{I}_\mathcal{F}(\epsilon/4) \leq \log \mathcal{I}_{\mathcal{F},B} \left( \frac{2 \ln 2}{R} \frac{\epsilon}{\epsilon} \right) \leq O \left( d \log \frac{BR}{\epsilon} \right),
\]
where \( \mathcal{I}_{\mathcal{F},B}(\cdot) \) is the covering number of a \( d \)-dimensional ball centered at the origin with radius \( B \) with respect to \( \ell_2 \)-norm and the last step is from (Wainwright, 2019). This concludes our proof.

B Proof of Theorem 1

The proof of Theorem 1 consists of two steps, deriving the guarantee of MLE and analyzing the performance of pessimistic offline RL.

Step 1: MLE guarantee. We first need to show that the confidence set \( \mathcal{R}(D) \) contains the true reward \( r^* \) with high probability. This can be proved via the following lemma which characterizes the guarantee of MLE:
Lemma 1 (Performance of MLE). Fix any δ ∈ (0, 1]. Then with probability at least 1 − δ/2 we have that for all reward function r ∈ G_r,
\[ \sum_{n=1}^{N} \log \left( \frac{P_r(o^n|τ_n,0, τ_n,1)}{P_r(o^n|τ_n,0, τ_n,1)} \right) \leq c_{MLE} \log(N_{G_r}(1/N)/δ), \]
where c_{MLE} > 0 is a universal constant.

We defer the proof to Appendix B.1. Denote the event in Assumption 2. Where
variation (TV) distance between
and ensure for all \( r \) theorem between

We know that under event \( E_1 \), we have
\[ \sum_{n=1}^{N} \log P_r(o^n|τ_n,0, τ_n,1) \geq \sum_{n=1}^{N} \log P_r(o^n|τ_n,0, τ_n,1) - c_{MLE} \log(N_{G_r}(1/N)/δ), \]
which implies that \( r^* \in R(D) \) since we know \( r^* \in G_r \) from Assumption 2.

Nevertheless, the confidence set \( R(D) \) is constructed via loglikelihood and we indeed prefer a bound on the total variation (TV) distance between \( P_r \) and \( P_\tau \), where \( r \in R(D) \) to facilitate our subsequent analysis. We can obtain such a bound as shown in the following lemma from the literature (Liu et al., 2022)[Proposition 14],(Zhan et al., 2022b)[Lemma 9]:

Lemma 2. With probability at least 1 − δ/2, we have for all reward function r ∈ G_r that
\[ \mathbb{E}_{τ_0, τ_1} \left[ \left\| P_r(·|τ_0, τ_1) - P_\tau(·|τ_0, τ_1) \right\|^2 \right] \leq \frac{c_{TV}}{N} \left( \sum_{n=1}^{N} \log \left( \frac{P_r(o^n|τ_n,0, τ_n,1)}{P_r(o^n|τ_n,0, τ_n,1)} \right) + \log(N_{G_r}(1/N)/δ) \right), \]
where \( c_{TV} > 0 \) is a universal constant.

Denote the event in Lemma 2 by \( E_2 \) and then we know \( \mathbb{P}(E_2) \geq 1 − δ/2 \). Then from Lemma 1 and Lemma 2 we know that under event \( E_1 \cap E_2 \), we have for all \( r \in R(D) \):
\[ \mathbb{E}_{τ_0, τ_1} \left[ \left\| P_r(·|τ_0, τ_1) - P_\tau(·|τ_0, τ_1) \right\|^2 \right] \leq \frac{c}{N} \log(N_{G_r}(1/N)/δ), \]
(4)
where \( c > 0 \) is a universal constant.

Then under Assumption 3, we can apply the mean value theorem between \( r^*(τ_1) - r^*(τ_0) \) and \( r(τ_1) - r(τ_0) \) to (4) and ensure for all \( r \in R(D) \) that
\[ \mathbb{E}_{τ_0, τ_1} \left[ \left\| (r^*(τ_1) - r^*(τ_0)) - (r(τ_1) - r(τ_0)) \right\|^2 \right] \leq \frac{c_k^2 \log(N_{G_r}(1/N)/δ)}{N}, \]
where \( k := \frac{1}{\int_{[−r_{max}, r_{max}]_\mathbb{R}} P(τ) \, dτ} \) measures the non-linearity of the link function \( Ψ \).

Step 2: Pessimistic offline RL. Let \( r^{inf} \) denote \( \argmin_{τ \in R(D)} J(π; r, P^*) - \mathbb{E}_{τ \sim \mu_{ref}}[r(τ)] \). Then we can bound the suboptimality of \( π \) as follows:
\[ J(π_{tar}; r^*, P^*) - J(π; r^*, P^*) = (J(π_{tar}; r^*, P^*) - \mathbb{E}_{τ_\sim \mu_{ref}}[r^*(τ)]) - (J(π; r^*, P^*) - \mathbb{E}_{τ_\sim \mu_{ref}}[r^*(τ)]) \leq \left( J(π_{tar}; r^*, P^*) - \mathbb{E}_{τ_\sim \mu_{ref}}[r^*(τ)] \right) - \left( J(π; r^*, P^*) - \mathbb{E}_{τ_\sim \mu_{ref}}[r^*(τ)] \right) \leq \left( J(π_{tar}; r^*, P^*) - \mathbb{E}_{τ_\sim \mu_{ref}}[r^*(τ)] \right) - \left( J(π; r^*, P^*) - \mathbb{E}_{τ_\sim \mu_{ref}}[r^*(τ)] \right) \leq C_r(G_r, π_{tar}, μ_{ref}) \cdot \sqrt{\frac{c_{TV}^2(G_r, π_{tar}, μ_{ref}) r^2 \log(N_{G_r}(1/N)/δ)}{N}}, \]
where the second step is due to \( π \) = \( \argmax_{τ \in R(D)} \min_{τ \in R(D)} J(π; r, P^*) - \mathbb{E}_{τ_\sim \mu_{ref}}[r(τ)] \), the third step is due to \( r^{inf} \) = \( \argmin_{τ \in R(D)} J(π; r, P^*) - \mathbb{E}_{τ_\sim \mu_{ref}}[r(τ)] \), the fifth step comes from the definition of \( C_r(G_r, π_{tar}, μ_{ref}) \) (Definition 2) and the last step leverages (5). This concludes our proof.

B.1 Proof of Lemma 1

The proof largely follows (Zhan et al., 2022b). Suppose \( F \) is a 1/N-bracket of \( G_r \) with \( |F| = N_{G_r}(1/N) \) and we denote the set of all right brackets in \( F \) by \( \hat{F} \), i.e., \( \hat{F} := \{ f : \exists f', \text{ such that } [f', f] \in F \} \). Then fix any \( f \in \hat{F} \), we have:
\[ \mathbb{E} \left[ \exp \left( \sum_{n=1}^{N} \log \left( \frac{f(o^n|τ_n,0, τ_n,1)}{P_{\tau}(o^n|τ_n,0, τ_n,1)} \right) \right) \right] \leq \prod_{n=1}^{N} \mathbb{E} \left[ \exp \left( \log \left( \frac{f(o^n|τ_n,0, τ_n,1)}{P_{\tau}(o^n|τ_n,0, τ_n,1)} \right) \right) \right] = \prod_{n=1}^{N} \mathbb{E} \left[ \frac{f(o^n|τ_n,0, τ_n,1)}{P_{\tau}(o^n|τ_n,0, τ_n,1)} \right]. \]
\[ F = \prod_{n=1}^{N} \mathbb{E} \left[ \sum_o f(o | \tau^{n,0}, \tau^{n,1}) \right] \leq \left( 1 + \frac{1}{N} \right)^N \leq e, \]

where the first step is due to each sample in \( D \) is i.i.d., the third step uses Tower property and the fourth step is from the fact that \( F \) is a minimum 1/N-bracket.

Then by Markov’s inequality we have for any \( \delta \in (0, 1) \),

\[ P \left( \sum_{n=1}^{N} \log \left( \frac{f(o^n | \tau^{n,0}, \tau^{n,1})}{P_r(o^n | \tau^{n,0}, \tau^{n,1})} \right) > \log(1/\delta) \right) \leq \mathbb{E} \left[ \exp \left( \sum_{n=1}^{N} \log \left( \frac{f(o^n | \tau^{n,0}, \tau^{n,1})}{P_r(o^n | \tau^{n,0}, \tau^{n,1})} \right) \right) \right] \cdot \exp \left[ - \log(1/\delta) \right] \leq e \delta. \]

By union bound, we have for all \( f \in F \),

\[ P \left( \sum_{n=1}^{N} \log \left( \frac{f(o^n | \tau^{n,0}, \tau^{n,1})}{P_r(o^n | \tau^{n,0}, \tau^{n,1})} \right) > c_{\text{MLE}} \log(N_{G_r}(1/N)/\delta) \right) \leq \delta/2, \]

where \( c_{\text{MLE}} > 0 \) is a universal constant.

Therefore from the definition of 1/N-bracket net, we know for all \( r \in G_r \), there exists \( f \in F \) such that \( P_r(o | \tau^{0}, \tau^{1}) \leq f(o | \tau^{0}, \tau^{1}) \) for any trajectories \((\tau^{0}, \tau^{1})\). This implies that for all \( r \in G_r \),

\[ P \left( \sum_{n=1}^{N} \log \left( \frac{P_r(o^n | \tau^{n,0}, \tau^{n,1})}{f(o^n | \tau^{n,0}, \tau^{n,1})} \right) > c_{\text{MLE}} \log(N_{G_r}(1/N)/\delta) \right) \leq \delta/2. \]

This concludes our proof.

C  Comparison with (Zhu et al., 2023)

(Zhu et al., 2023) considers the linear reward setting under BTL model and they can achieve the following sample complexity:

\[ N = \mathcal{O} \left( \frac{C_{\text{lin}}^2 \exp(4BR) d \log(1/\delta)}{\epsilon^2} \right), \]

where \( R \) and \( B \) are the norm bounds on the feature vectors \( \phi \) and parameter \( \theta \) (defined in Proposition 1). The concentrability coefficient \( C_{\text{lin}} \) is defined as

\[ C_{\text{lin}} := \|E_{r^0 \sim \pi_{\text{tar}}, r^1 \sim \mu_{\text{ref}}} [\phi(r^0) - \phi(r^1)] \|_{\Sigma_D^{-1}}, \]

and \( \Sigma_D \) is the empirical covariance matrix of the dataset \( \frac{1}{N} \sum_{n=1}^{N} (\phi(\tau^{n,0}) - \phi(\tau^{n,1}))(\phi(\tau^{n,0}) - \phi(\tau^{n,1}))^\top \).

Note that all the analysis and proofs in this paper still hold when we define the concentrability coefficient as

\[ C_{\text{r}}(G_r, \pi_{\text{tar}}, \mu_{\text{ref}}) := \max \left\{ 0, \sup_{r \in G_r} \frac{E_{r^0 \sim \pi_{\text{tar}}, r^1 \sim \mu_{\text{ref}}} [r^*(\tau^0) - r^*(\tau^1) - (r^0) + r(\tau^1)]}{\sqrt{\frac{1}{N} \sum_{n=1}^{N} |r^*(\tau^{n,0}) - r^*(\tau^{n,1}) - r(\tau^{n,0}) + r(\tau^{n,1})|^2}} \right\}. \]

Then when specializing the result in Theorem 1 to the linear reward setting under BTL model with this version of concentrability coefficient, the sample complexity is

\[ N = O \left( (C_{\text{r}}(G_r, \pi_{\text{tar}}, \mu_{\text{ref}}))^2 \exp(2r_{\max} d \log(BR/\delta)) \right). \]

We know that \( BR \geq r_{\max} \). In addition, note that in this case, we have \( C_{\text{lin}} \geq 0 \) and for any \( r \in G_r \),

\[ \left| E_{r^0 \sim \pi_{\text{tar}}, r^1 \sim \mu_{\text{ref}}} [r^*(\tau^0) - r^*(\tau^1) - (r^0) + r(\tau^1)] \right| \leq \left| E_{r^0 \sim \pi_{\text{tar}}, r^1 \sim \mu_{\text{ref}}} [\phi(r^0) - \phi(r^1)] \right| \|_{\Sigma_D^{-1}} \cdot \| \theta^* - \theta \|_{\Sigma_D} \]

where \( \| r^*(\tau) = \langle \phi(\tau), \theta^* \rangle \) and \( r(\tau) = \langle \phi(\tau), \theta \rangle \). Therefore we have

\[ C_{\text{r}}(G_r, \pi_{\text{tar}}, \mu_{\text{ref}}) \leq C_{\text{lin}}. \]

This implies that Theorem 1 can recover the sample complexity for linear reward setting under BTL model in (Zhu et al., 2023) with only some additional log factors.

D  Omitted Details

In this section we supplement the definition of bracket number for the transition class and advantage function class.

Definition 5 (c-bracket number of transition probability classes). Suppose \( f^1, f^2 \) is a function with \( f^1(\cdot \mid s, a), f^2(\cdot \mid s, a) \in \mathbb{R}^{\mid S} \) for all \( (s, a) \in S \times A \). Then we say \( f^1, f^2 \) is a c-bracket if \( f^1(\cdot \mid s, a) \leq f^2(\cdot \mid s, a) \) and \( \| f^1(\cdot \mid s, a) - f^2(\cdot \mid s, a) \|_1 \leq c \) for all \( (s, a) \). The c-bracket number of a transition probability class \( G_p_h \) where \( h \in [H - 1] \) is the minimum integer \( N \) satisfying that there exist \( N \) c-brackets \( (f^{n,1}, f^{n,2}) \) such that for any function \( p_h \in G_p_h \) there is a bracket \( (f^{i,1}, f^{i,2}) \) where \( i \in [N] \) containing it, i.e., \( f^{i,1}(\cdot \mid s, a) \leq p_h(\cdot \mid s, a) \leq f^{i,2}(\cdot \mid s, a) \) for all \( (s, a) \).

Definition 6 (c-bracket number of initial state distribution classes). Suppose \( f^1, f^2 \in \mathbb{R}^{\mid S} \). Then we say \( f^1, f^2 \) is a c-bracket if \( f^1 \leq f^2 \) and \( \| f^1 - f^2 \|_1 \leq c \). The
We consider the case where $\pi$ is a Markovian policy. Then we define $\pi_{b,h}(\cdot|s)$ for every $h \in [H], s \in S$ as follows:

$$
\pi_{b,h}(a|s) = \begin{cases} 
\pi_{\text{tar},h}(a|s)/C, & \text{if } a = a^1, \\
\pi_{\text{tar},h}(a|s) + (1 - 1/C)\pi_{\text{tar},h}(a^1|s), & \text{if } a = a^2, \\
\pi_{\text{tar},h}(a|s), & \text{otherwise},
\end{cases}
$$

and we select $\mu_0 = d^{(\pi_{\text{tar},h})}_{1}$. Then since $d_h^{(s)}(s)$ is fixed for any policy $\pi$, we know

$$
C^{st} = \sup_{s,a,h} \pi_{\text{tar},h}(a|s)/\pi_{b,h}(a|s) = C.
$$

On the other hand, we have

$$
C^{tr} = \sup_{a_1,H,s,a,h} \prod_{h \in [H]} \frac{\pi_{\text{tar},h}(a_h|s_h)}{\pi_{b,h}(a_h|s_h)}
\leq \sup_{s,a,h} \prod_{h \in [H]} \frac{\pi_{\text{tar},h}(a^1|s_h)}{\pi_{b,h}(a^1|s_h)} = C^H.
$$

This concludes our proof.

### E.2 Proof of Theorem 2

The proof is inspired by the hard instances in (Rashidinejad et al., 2021b). We consider the case $C \geq 2$ and $1 < C < 2$ respectively.

**Case 1:** $C \geq 2$. Consider the case where there is only one state $s$ and two actions $a^1, a^2$. Set the dataset distribution $\mu_0 = \mu_1$ where

$$
\mu_{0,h}(s,a^1) = \frac{1}{C}, \quad \mu_{0,h}(s,a^2) = 1 - \frac{1}{C}, \quad \forall h \in [H],
$$

and

$$
\mu_{0}(\tau) = \prod_{h=1}^{H} \mu_{0,h}(s_h,a_h) \text{ for all trajectory } \tau = (s_1, a_1, \cdots, s_H, a_H).
$$

We consider two different reward function $r^1$ and $r^2$:

$$
r^1(\tau) = \begin{cases} 
\frac{3}{2} + x, & \text{if all the actions in } \tau \text{ are } a^1, \\
\frac{1}{2}, & \text{otherwise},
\end{cases}
$$

$$
r^2(\tau) = \begin{cases} 
\frac{3}{2} - x, & \text{if all the actions in } \tau \text{ are } a^1, \\
\frac{1}{2}, & \text{otherwise},
\end{cases}
$$

Here $0 < x < \frac{1}{2}$ is a quantity we will specify later and we denote the special trajectory where all the actions are $a^1$ by $\tau^*$. Then we have two MDPs, $M_1$ and $M_2$ whose reward functions are $r^1$ and $r^2$ respectively. It can be easily verified that $(M_1, \mu_0) \in \Theta_{st}(C)$, $(M_2, \mu_0) \in \Theta_{st}(C)$.

Further, let $L(\pi; M)$ denote the suboptimality of policy $\pi$ in $M$, then we have for all policies $\pi$,

$$
L(\pi; M_1) + L(\pi; M_2) \geq x.
$$

Now we can apply Fano’s inequality, which leads to the following inequality

$$
\inf \sup \mathbb{E}_D[L(\pi, M)] \geq \frac{x}{2} - N \cdot \text{KL}
\left(\mu_0 \otimes \mu_1 \otimes P_{r_1} \| \mu_0 \otimes \mu_1 \otimes P_{r_2}\right)
\left(\mu_0 \otimes \mu_1 \otimes P_{r_1} \| \mu_0 \otimes \mu_1 \otimes P_{r_2}\right).
$$

Now we only need to bound $\text{KL}
\left(\mu_0 \otimes \mu_1 \otimes P_{r_1} \| \mu_0 \otimes \mu_1 \otimes P_{r_2}\right)$, which can be computed as follows:

$$
\text{KL}
\left(\mu_0 \otimes \mu_1 \otimes P_{r_1} \| \mu_0 \otimes \mu_1 \otimes P_{r_2}\right)
\leq 2 \sum_{\tau^{0}=\tau^{*}, \tau^{*} \neq \tau^{*}} \mu_0(\tau^{0})\mu_1(\tau^{*})\text{KL} \left(\text{Bern}(\sigma(x)) \| \text{Bern}(\sigma(-x))\right)
\leq 2 \exp(1/2)x^2.
$$


Then by letting \( x = \min \left\{ \frac{1}{2}, \sqrt{\frac{CN}{2 \exp(1/2)N}} \right\} \), we have
\[
\inf \sup_{\pi, M \in \{M_1, M_2\}} \mathbb{E}_D[L(\pi, M)] 
\geq \frac{x}{2} = \min \left\{ \frac{1}{4}, \sqrt{\frac{C^H}{8 \exp(1/2)N}} \right\}.
\]

**Case 2:** \( 1 < C < 2 \). Consider the case where there are two one states \( s^1, s^2 \) and two actions \( a^1, a^2 \). We suppose the initial state distribution of \( P_0^* \) is fixed as \( P_0^*(s^1) = C - 1 \) and \( P_0^*(s^2) = 2 - C \). In addition, the state will stay the same throughout the whole episode. Then we can set the dataset distribution \( \mu_0 = \mu_1 \) where
\[
\mu_0(\tau) = \begin{cases} 
\frac{2(C-1)}{C}, & \text{if the state of } \tau \text{ is } s^1, \\
\frac{C-2}{C}, & \text{if the state of } \tau \text{ is } s^2, \\
0, & \text{if the state of } \tau \text{ is } s^2 \text{ and the actions are all } a^1, \\
\frac{2-C}{C}, & \text{if the state of } \tau \text{ is } s^2 \text{ and the actions contain } a^2. 
\end{cases}
\]

Then we know
\[
\mu_{0,b}(s^1, a^1) = \mu_{0,b}(s^1, a^2) = \frac{C-1}{C},
\]
\[
\mu_{0,b}(s^2, a^1) = \frac{2-C}{C}, \quad \mu_{0,b}(s^2, a^2) = 0, \quad \forall h \in [H].
\]

We consider two different reward function \( r^1 \) and \( r^2 \):
\[
r^1(\tau) = \begin{cases} 
\frac{1}{2} + x, & \text{if the state of } \tau \text{ is } s^1 \text{ and all the actions in } \tau \text{ are } a^1, \\
\frac{1}{2}, & \text{otherwise.}
\end{cases}
\]
\[
r^2(\tau) = \begin{cases} 
\frac{1}{2} - x, & \text{if the state of } \tau \text{ is } s^1 \text{ and all the actions in } \tau \text{ are } a^1, \\
\frac{1}{2}, & \text{otherwise.}
\end{cases}
\]

Here \( 0 < x < \frac{1}{2} \) is a quantity we will specify later and we denote the special trajectory where the state is \( s^1 \) and all the actions are \( a^1 \) by \( \tau^* \). Then we have two MDPs, \( M_1 \) and \( M_2 \) whose reward functions are \( r^1 \) and \( r^2 \) respectively. It can be easily verified that \((M_1, \mu_0) \in \Theta_{st}(C), (M_2, \mu_0) \in \Theta_{st}(C)\).

In addition, we have for all policies \( \pi \),
\[
L(\pi; M_1) + L(\pi; M_2) \geq (C-1)x.
\]

Therefore by Fano’s inequality, we have
\[
\inf \sup_{\pi, M \in \{M_1, M_2\}} \mathbb{E}_D[L(\pi, M)] 
\geq \frac{(C-1)x}{2} \left( 2 - N \cdot KL\left( \mu_0 \otimes \mu_1 \otimes P_{r^1} || \mu_0 \otimes \mu_1 \otimes P_{r^2} \right) \right),
\]
where the KL divergence can be computed as follows:
\[
KL\left( \mu_0 \otimes \mu_1 \otimes P_{r^1} || \mu_0 \otimes \mu_1 \otimes P_{r^2} \right)
= 2 \sum_{\tau^0 = \tau^*, r^* \neq \tau^*} \mu_0(\tau^0) \mu_1(\tau^1) KL(\text{Bern}(\sigma(x)) || \text{Bern}(\sigma(-x)))
\leq 4(C-1) \exp(1/2)x^2 \frac{2HC}{2HC}. \]

Then by letting \( x = \min \left\{ \frac{1}{2}, \sqrt{\frac{2\sigma C}{4 \exp(1/2)(C-1)N}} \right\} \), we have
\[
\inf \sup_{\pi, M \in \{M_1, M_2\}} \mathbb{E}_D[L(\pi, M)] 
\geq \frac{(C-1)x}{2} = \min \left\{ \frac{C-1}{4}, \sqrt{\frac{2HC(C-1)}{16 \exp(1/2)N}} \right\}.
\]

In conclusion, we have for any \( C > 1 \) and \( H \geq 1 \),
\[
\inf \sup_{\pi, (M, \mu_0) \in \Theta_{st}(C)} \mathbb{E}_D[J(\pi^*, r^*, P^*) - J(\tilde{\pi}, r^*, P^*)] \geq \min \left\{ C-1, \sqrt{\frac{(\max\{C,2\})H^{-1}(C-1)}{N}} \right\}.
\]

### E.3 Proof of Theorem 3

The proof is quite similar to the proof of Theorem 2. We consider the case \( C \geq 2 \) and \( 1 < C < 2 \) respectively.

**Case 1: \( C \geq 2 \).** Consider the case where there is only one state \( s \) and two actions \( a^1, a^2 \). Set the dataset distribution \( \mu_0 = \mu_1 \) where
\[
\mu_0(\tau^*) = \frac{1}{C}, \quad \mu_0(\tau^1) = 1 - \frac{1}{C},
\]
where \( \tau^* \) is the trajectory where the actions are all \( a^1 \) and \( \tau^1 \) is the trajectory where the actions are all \( a^2 \).

We consider two different reward function \( r^1 \) and \( r^2 \):
\[
r^1(\tau) = \begin{cases} 
\frac{1}{2} + x, & \text{if all the actions in } \tau \text{ are } a^1, \\
\frac{1}{2}, & \text{otherwise.}
\end{cases}
\]
\[
r^2(\tau) = \begin{cases} 
\frac{1}{2} - x, & \text{if all the actions in } \tau \text{ are } a^1, \\
\frac{1}{2}, & \text{otherwise.}
\end{cases}
\]

Here \( 0 < x < \frac{1}{2} \) is a quantity we will specify later. Then we have two MDPs, \( M_1 \) and \( M_2 \) whose reward functions are \( r^1 \) and \( r^2 \) respectively. It can be easily verified that \((M_1, \mu_0) \in \Theta_{st}(C), (M_2, \mu_0) \in \Theta_{st}(C)\).
Further, let $L(\pi; M)$ denote the suboptimality of policy $\pi$ in $M$, then we have for all policies $\pi$,

$$L(\pi; M_1) + L(\pi; M_2) \geq x.$$ 

Now we can apply Fano’s inequality, which leads to the following inequality

$$\inf_{\pi} \sup_{M \in \{M_1, M_2\}} \mathbb{E}_D[L(\pi, M)] 
\geq \frac{x}{2} \left( 2 - N \cdot \text{KL} \left( \mu_0 \otimes \mu_1 \otimes P_{\tau_r} \| \mu_0 \otimes \mu_1 \otimes P_{\tau} \right) \right).$$

Now we only need to bound $\text{KL} \left( \mu_0 \otimes \mu_1 \otimes P_{\tau_r} \| \mu_0 \otimes \mu_1 \otimes P_{\tau} \right)$, which can be computed as follows:

$$\text{KL} \left( \mu_0 \otimes \mu_1 \otimes P_{\tau_r} \| \mu_0 \otimes \mu_1 \otimes P_{\tau} \right) = 2 \sum_{\tau^0 = \tau^*, \tau^1 = \tau} \mu_0(\tau^0) \mu_1(\tau^1) \cdot \text{KL}(\sigma(x) \| \sigma(-x)) \leq 2 \exp(1/2) x^2 \cdot \frac{C}{2}.$$ 

Then by letting $x = \min \left\{ \frac{1}{2}, \sqrt{\frac{C}{2 \exp(1/2) N}} \right\}$, we have

$$\inf_{\pi} \sup_{M \in \{M_1, M_2\}} \mathbb{E}_D[L(\pi, M)] \geq \frac{x}{2} = \min \left\{ \frac{1}{4}, \sqrt{\frac{C}{8 \exp(1/2) N}} \right\}.$$ 

**Case 2:** $1 < C < 2$. Consider the case where there are two one states $s^1, s^2$ and two actions $a^1, a^2$. We suppose the initial state distribution of $P_{\tau_r}$ is fixed as $P_{\tau_r}(s^1) = C - 1$ and $P_{\tau_r}(s^2) = 2 - C$. In addition, the state will stay the same throughout the whole episode. Then we can set the dataset distribution $\mu_0 = \mu_1$ where

$$\mu_0(\tau) = \begin{cases} 
\frac{2(1-C)}{C}, & \text{if the state of } \tau \text{ is } s^1 \\
\frac{2-C}{C}, & \text{if the state of } \tau \text{ is } s^2 \\
0, & \text{otherwise}
\end{cases}$$

Let $\tau^*$ be the trajectory where the state is $s^1$ and the actions are all $a^2$.

We further consider two different reward functions $r^1$ and $r^2$:

$$r^1(\tau) = \begin{cases} 
\frac{1}{2} + x, & \text{if the state is } s^1 \text{ and all the actions in } \tau \text{ are } a^1, \\
\frac{1}{2}, & \text{otherwise}.
\end{cases}$$

Here $0 < x < \frac{1}{4}$ is a quantity we will specify later. Then we have two MDPs, $M_1$ and $M_2$ whose reward functions are $r^1$ and $r^2$ respectively. It can be easily verified that $(M_1, \mu_0) \in \Theta_{\tau_r}(C), (M_2, \mu_0) \in \Theta_{\tau_r}(C)$.

In addition, we have for all policies $\pi$,

$$L(\pi; M_1) + L(\pi; M_2) \geq (C - 1)x.$$ 

Therefore by Fano’s inequality, we have

$$\inf_{\pi} \sup_{M \in \{M_1, M_2\}} \mathbb{E}_D[L(\pi, M)] \geq \frac{(C - 1)x}{2} \left( 2 - N \cdot \text{KL} \left( \mu_0 \otimes \mu_1 \otimes P_{\tau_r} \| \mu_0 \otimes \mu_1 \otimes P_{\tau} \right) \right),$$

where the KL divergence can be computed as follows:

$$\text{KL} \left( \mu_0 \otimes \mu_1 \otimes P_{\tau_r} \| \mu_0 \otimes \mu_1 \otimes P_{\tau} \right) = 2 \sum_{\tau^0 = \tau^*, \tau^1 \neq \tau^*} \mu_0(\tau^0) \mu_1(\tau^1) \cdot \text{KL}(\sigma(x) \| \sigma(-x)) \leq 2(C - 1) \exp(1/2) x \frac{C}{2}.$$ 

Then by letting $x = \min \left\{ \frac{1}{2}, \sqrt{\frac{C}{2 \exp(1/2)(C - 1) N}} \right\}$, we have

$$\inf_{\pi} \sup_{M \in \{M_1, M_2\}} \mathbb{E}_D[L(\pi, M)] \geq \frac{(C - 1)x}{2} = \min \left\{ \frac{C - 1}{4}, \sqrt{\frac{(C - 1)}{8 \exp(1/2) N}} \right\}.$$ 

In conclusion, we have for any $C > 1$ and $H \geq 1$,

$$\inf_{\pi} \sup_{(M, \mu_0) \in \Theta_{\tau_r}(C)} \mathbb{E}_D[J(\pi^*; r^*, P^*) - J(\hat{\pi}; r^*, P^*)] \geq \min \left\{ C - 1, \sqrt{\frac{C - 1}{N}} \right\}.$$ 

**F Proof of Theorem 4**

The proof still consists of two steps, deriving the guarantee of MLE and analyzing the performance of pessimistic offline RL.

**Step 1: MLE guarantee.** Note that Lemma 1 and Lemma 2 still applies here. Let $E_1$ and $E_2$ denote the event in Lemma 1 and Lemma 2 respectively. Following almost the same arguments, we have the following guarantee for the estimation of the system dynamics:
Lemma 3. Under Assumption 4, with probability at least $1 - \delta/2$, the following event holds true:

1. $P_h^* \in \mathcal{P}_h(D)$, $P_0^* \in \mathcal{P}_{\text{tra}}(D)$, $\forall h \in [H - 1]$,

2. $E(s_h, a_h)_{\sim \mu_0, h} \left[ \left\| P_h(\cdot | s, a) - P_h^*(\cdot | s, a) \right\|_1^2 \right]$
   $\leq \frac{c \log(HN_{P_h}(1/N)/\delta)}{N}, \forall h \in [H - 1], P_h \in \mathcal{P}_h(D)$,

3. $E_{s \sim \nu_{0, 1}} \left[ \left\| P_h(s) - P_0^*(s) \right\|_1^2 \right]$
   $\leq \frac{c \log(HN_{P_h}(1/N)/\delta)}{N}, \forall P_0 \in \mathcal{P}_0(D)$.

The proof is omitted here. Let $\mathcal{E}_3$ denote the event in Lemma 3.

Step 2: Pessimistic offline RL. We first introduce the following lemma which suggests that under event $\mathcal{E}_3$, we can evaluate the expected cumulative reward of $\pi_{\text{tra}}$ with respect to any reward function $r \in \mathcal{G}_r$ via the system dynamics $P_h \in \mathcal{P}_h(D)$:

Lemma 4. Suppose Assumption 3 is true. Then under $\mathcal{E}_3$, we have for all reward function $r \in \mathcal{G}_r$ and $P = \{P_h \}_{h=1}^{H}$ where $P_h \in \mathcal{P}_h(D)$ that

$$J(\pi_{\text{tra}}; r, P^*) - J(\pi_{\text{tra}}; r, P)$$

$\leq H \max_{r_{\text{tar}}} \sqrt{\frac{cT_2^2(\mathcal{G}_{P_h}, \pi_{\text{tra}}) \log(HN_P(1/N)/\delta)}{N}},$

where $N_P = \max_{0 \leq h \leq H - 1} \{N_{P_h} \}$.

The proof is deferred to Appendix F.1.

Let $(r_{\pi}^{\text{inf}}, P_{\pi}^{\text{inf}})$ denote argmin$_{r \in \mathcal{G}_r, \pi \in \Pi_0} \mathbb{E}_{r \sim \mu}(r(0))$. Then under event $\mathcal{E}_3$, we can bound the suboptimality of $\tilde{\pi}$ as follows:

$$J(\pi_{\text{tra}}; r^*, P^*) - J(\pi_{\text{tra}}; r^*, P^*)$$

$= (J(\pi_{\text{tra}}; r^*, P^*) - \mathbb{E}_{r_{\pi_{\text{tra}}} \sim \mu}(r_{\pi_{\text{tra}}}(r)))$

$- (J(\pi_{\text{tra}}; r^*, P^*) - \mathbb{E}_{r_{\pi_{\text{tra}}} \sim \mu}(r_{\pi_{\text{tra}}}(r)))$

$= \left( (J(\pi_{\text{tra}}; r^*, P^*) - \mathbb{E}_{r_{\pi_{\text{tra}}} \sim \mu}(r_{\pi_{\text{tra}}}(r))) \right)$

$+ \left( (J(\pi_{\text{tra}}; r^*, P^*) - \mathbb{E}_{r_{\pi_{\text{tra}}} \sim \mu}(r_{\pi_{\text{tra}}}(r))) \right)$

$+ \left( (J(\pi_{\text{tra}}; r^*, P^*) - \mathbb{E}_{r_{\pi_{\text{tra}}} \sim \mu}(r_{\pi_{\text{tra}}}(r))) \right)$. 

where the third and fourth step are due to the definition of $\tilde{\pi}, (r_{\pi}^{\text{inf}}, P_{\pi}^{\text{inf}})$ and (1) in Lemma 3. The last step comes from Lemma 4 and the proof of Theorem 1. This concludes our proof.

F.1 Proof of Lemma 4

Let $P_h$ be the system dynamics $(P_0^*, \{P_t^* \}_{t=1}^{H-1})$ for all $0 \leq h \leq H - 1$. Then we have

$$J(\pi_{\text{tra}}; r, P^*) - J(\pi_{\text{tra}}; r, P)$$

$= \sum_{h=1}^{H-1} (J(\pi_{\text{tra}}; r, P^*) - J(\pi_{\text{tra}}; r, P^{h-1}))$

$+ (J(\pi_{\text{tra}}; r, P^*) - J(\pi_{\text{tra}}; r, P))$. 

For any $h \in [H - 1]$, we have

$$J(\pi_{\text{tra}}; r, P^h) - J(\pi_{\text{tra}}; r, P^{h-1})$$

$= \mathbb{E}_{s_1, a_1, \cdots, s_h, a_h \sim (\pi_{\text{tra}}; P^*)} \left[ \sum_{s_{h+1}} P_h(s_{h+1} | s_h, a_h) \mathbb{E}(\pi_{\text{tra}}, P) \right.$

$\left. [r(\pi) | s_1, a_1, \cdots, s_{h+1}] - \sum_{s_{h+1}} P_h(s_{h+1} | s_h, a_h) \mathbb{E}(\pi_{\text{tra}}, P) \right]$. 

$\mathbb{E}(\pi_{\text{tra}}, P)$
= \mathbb{E}(s_1, a_1, \ldots, s_h, a_h) \sim (\tau, \rho) \left[ \sum_{s_{h+1}} (P_h(s_{h+1}|s_h, a_h) - P_h(s_{h+1}|s_h, a_h)) \right] \\
\leq r_{\max} \mathbb{E}(s_1, a_1, \ldots, s_h, a_h) \sim (\tau, \rho) \left[ \left\| P_h(s_{h+1}|s_h, a_h) - P_h(s_{h+1}|s_h, a_h) \right\|_1 \right] \\
\leq r_{\max} \sqrt{cC^2_h(\tau) \log(HN^2_{\Phi}(1/N)\delta)}.

Therefore we conclude that

\displaystyle J(\pi_{\text{tar}}; r, P^*) - J(\pi_{\text{tar}}; r, P) \leq H r_{\max} \sqrt{cC^2_h(\tau) \log(HN^2_{\Phi}(1/N)\delta)}.$

**G Proof of Theorem 5**

We first derive the guarantee of MLE for estimating $A^*$. Similar to Lemma 1 and Lemma 2, we have the following lemma in the action-based comparison setting:

**Lemma 5.** Under Assumption 7, with probability at least $1 - \delta$, the following event holds true: $\forall h \in [H],$

\[
\mathbb{E}_{s \sim \mu_h, a^0 \sim \mu_0, h(\cdot|s), a^1 \sim \mu_1, h(\cdot|s)} \left[ \left\| P_h(\cdot|s, a^0, a^1) - P_h(\cdot|s, a^0, a^1) \right\|_1 \right] \leq c \log(HN_{\Phi_h}(1/N)\delta).\]

The proof is omitted here. Let $\mathcal{E}_4$ denote the event in Lemma 5. Then under Assumption 8, we can apply the mean value theorem and obtain that under $\mathcal{E}_4$, we have for all $h \in [H]$ that

\[
\mathbb{E}_{s \sim \mu_h, a^0 \sim \mu_0, h(\cdot|s), a^1 \sim \mu_1, h(\cdot|s)} \left[ \left| A_h^*(s, a^0) - A_h^*(s, a^1) \right| - \hat{A}_h(s, a^0) + \hat{A}_h(s, a^1) \right| \leq \frac{c \kappa^2 \log(HN_{\Phi_h}(1/N)\delta)}{N}, \forall h \in [H].
\]

Recall that $\kappa = \frac{\min_{x \in [-r_{\max}, r_{\max}] \Phi(x)}}{N}$. On the other hand, note that we have the following performance lemma:

**Lemma 6.** For any deterministic Markovian policies $\pi$ and $\pi'$, we have

\[
J(\pi; r, P^*) - J(\pi'; r, P^*) = \sum_{h=1}^{H} \mathbb{E}_{s \sim \tilde{d}_h^*} \left[ Q_h^*(s, \pi(s)) - Q_h^*(s, \pi'(s)) \right]
\]

The proof is deferred to Appendix G.1. The rest of the proof largely follows Uehara et al. (2023). Under the event $\mathcal{E}_4$, we can bound the suboptimality of $\hat{\pi}$ as follows:

\[
J(\pi^*; r, P^*) - J(\hat{\pi}; r, P^*) \leq \sum_{h=1}^{H} \mathbb{E}_{s \sim d_h^*} \left[ \sum_{a \in A} \mathbbm{1}(A_h(s, a) \geq \hat{A}_h(s, \pi_h^*(s))) - \mathbbm{1}(Q_h^*(s, \pi_h^*(s)) < Q_h^*(s, \pi_h^*(s))) \right]
\]

where the first step comes from Lemma 6 and the second step is due to the definition of $\hat{\pi}$. Then for any $\alpha > 0$, we have

\[
\mathbb{E}_{s \sim d_h^*} \left[ \sum_{a \in A} \mathbbm{1}(A_h(s, a) \geq \hat{A}_h(s, \pi_h^*(s))) - \mathbbm{1}(Q_h^*(s, \pi_h^*(s)) < Q_h^*(s, \pi_h^*(s))) \right] \leq \sum_{a \in A} \mathbbm{1}
\]

By Assumption 6, we have

\[
\mathbb{E}_{s \sim d_h^*} \left[ \sum_{a \in A} \mathbbm{1}(Q_h^*(s, \pi_h^*(s)) > Q_h^*(s, \alpha) \geq Q_h^*(s, \pi_h^*(s)) - \alpha) \right]
\]

For the second term, we have

\[
\mathbb{E}_{s \sim d_h^*} \left[ \sum_{a \in A} \mathbbm{1}(Q_h^*(s, \pi_h^*(s)) - Q_h^*(s, \alpha) \right]
\]
\[ -\hat{A}_h(s, \pi^*_h(s)) + \hat{A}_h(s, a) \geq \alpha \] 

\[ = \frac{1}{\alpha^2} E_{s \sim d^*_h} \left[ \sum_{a \in A} \alpha^2 (A^*_h(s, \pi^*_h(s)) - A^*_h(s, a) - A_h(s, \pi^*_h(s)) + A_h(s, a)) \right] \]

\[ \leq \frac{1}{\alpha^2} E_{s \sim d^*_h} \left[ \sum_{a \in A} |A^*_h(s, \pi^*_h(s)) - A^*_h(s, a) - A_h(s, \pi^*_h(s)) + A_h(s, a)|^2 \right] \]

\[ \leq \frac{c |A| C_{act} \kappa^2 \log(HN_{\gamma_h^*}(1/N) / \delta)}{\alpha^2 N}, \]

where the last step comes from the definition of \( C_{act} \) and (6).

Therefore by picking appropriate \( \alpha \), we have with probability at least \( 1 - \delta \) that

\[ J(\pi^*; r^*, P^*) - J(\hat{\pi}; r^*, P^*) \leq cH |A| \left( \frac{2}{\beta} \right)^{\beta + 2} \left( \frac{1}{\alpha_0} \right)^{\frac{2}{\beta + 2}} \]

\[ \cdot \left( \kappa^2 C_{act} \log(HN_{\gamma_h^*}(1/N) / \delta) / N \right)^{\frac{1}{\beta + 2}}. \]

### G.1 Proof of Lemma 6

For any two policies \( \pi \) and \( \pi' \), we have that

\[ J(\pi'; r^*, P^*) - J(\pi; r^*, P^*) \]

\[ = E_{\pi'} [r_1^*(s_1, a_1) + V^*_2(s_2) - V^*_1(s_1)] \]

\[ = E_{\pi'} [V^*_2(s_2) - (V^*_1(s_1) - r_1^*(s_1, a_1))] \]

\[ = E_{\pi'} [V^*_2(s_2) - V^*_2(s_2)] + E_{\pi'} [Q^*_1(s_1, a_1) - V^*_1(s_1)] \]

\[ = E_{\pi'} [V^*_2(s_2) - V^*_2(s_2)] + E_{\pi'} [\{Q^*_1(s_1, \cdot), \pi'_1(\cdot | s_1) - \pi_1(\cdot | s_1)\}] \]

\[ = \cdots = \sum_{h=1}^H E_{\pi'} [\{Q^*_h(s_h, \cdot), \pi'_h(\cdot | s) - \pi_h(\cdot | s)\}]. \]

This concludes our proof.