
Outcome-Based Online Reinforcement Learning: Algorithms and Fundamental Limits

Anonymous Author(s)

Affiliation

Address

email

Abstract

Reinforcement learning with outcome-based feedback faces a fundamental challenge: when rewards are only observed at trajectory endpoints, how do we assign credit to the right actions? This paper provides the first comprehensive analysis of this problem in online RL with general function approximation. We develop a provably sample-efficient algorithm achieving $\tilde{O}(C_{\text{cov}}H^3/\varepsilon^2)$ sample complexity, where C_{cov} is the coverability coefficient of the underlying MDP. By leveraging general function approximation, our approach works effectively in large or infinite state spaces where tabular methods fail, requiring only that value functions and reward functions can be represented by appropriate function classes. Our results also characterize when outcome-based feedback is statistically separated from per-step rewards, revealing an unavoidable exponential separation for certain MDPs. For deterministic MDPs, we show how to eliminate the completeness assumption, dramatically simplifying the algorithm. We further extend our approach to preference-based feedback settings, proving that equivalent statistical efficiency can be achieved even under more limited information. Together, these results constitute a theoretical foundation for understanding the statistical properties of outcome-based reinforcement learning.

1 Introduction

Reinforcement learning with outcome-based feedback is a fundamental paradigm where agents receive rewards only at the end of complete trajectories rather than at individual steps. This feedback model naturally arises in many applications, from large language model training (Ouyang et al., 2022; Bai et al., 2022; Jaech et al., 2024), where human preferences are provided for entire outputs rather than individual tokens, to clinical trials, where patient outcomes are only observable after a complete treatment regimen. Despite the prevalence of such settings, the statistical implications of outcome-based feedback for online exploration remain poorly understood.

In traditional reinforcement learning (Sutton et al., 1998), agents observe rewards immediately after each action, providing a granular signal that directly links actions to their consequences. In contrast, outcome-based feedback presents a fundamental challenge: when rewards are only observed at the trajectory level, determining which specific actions contributed to the final outcome becomes significantly more difficult. This credit assignment problem is particularly acute in sequential decision-making tasks with long horizons, where many different action combinations could lead to the observed outcome.

While recent work (Jia et al., 2025) has shown that outcome-based feedback is sufficient for offline reinforcement learning under certain conditions, the feasibility of efficient online exploration with only trajectory-level feedback remains an open question. Online learning—where an agent actively

explores to gather new data—is essential for adaptive systems that must learn in dynamic environments without pre-collected datasets. This leads to our central question:

When is online exploration with outcome-based reward statistically tractable?

This question has been studied in the setting where the reward function is assumed to be well-structured (Efroni et al., 2021; Pacchiano et al., 2021; Chatterji et al., 2021; Cassel et al., 2024; Lancewicki and Mansour, 2025), with a primary focus on the *linear* reward functions. Similar reliance on the well-behaved reward structure also appears in the recent work on Reinforcement Learning from Human Feedback (RLHF) (Chen et al., 2022b,a; Wu and Sun, 2023; Wang et al., 2023), where only *preference* feedback is available. However, well-behaved reward structure is dedicated and might fail to capture many real-world scenarios with general function approximation. In this paper, we address this question by providing a comprehensive theoretical analysis of outcome-based online reinforcement learning with general function approximation. We investigate when efficient exploration is possible with only trajectory-level feedback and characterize the fundamental statistical limits of learning in this setting. Our main results are as follows:

- (1) We present a model-free algorithm for outcome-based online RL with general function approximation (Algorithm 1) that relies solely on trajectory-level reward feedback rather than per-step feedback. Our algorithm achieves a complexity bound of $\tilde{O}(C_{\text{cov}} H^3 / \varepsilon^2)$ under standard realizability and completeness assumptions, where C_{cov} is the coverability coefficient that measures an intrinsic complexity of the underlying MDP. This bound applies in the general function approximation setting where state spaces may be large or infinite, requiring only that value functions can be represented by an appropriate function class with bounded statistical complexity.
- (2) For the special case of deterministic MDPs, we present a simpler algorithm based on Bellman residual minimization (Algorithm 2) that achieves similar theoretical guarantees with improved computational efficiency.
- (3) As extension, we generalize our approach to preference-based reinforcement learning (Appendix F.1), where feedback comes in the form of binary preferences between trajectory pairs under the Bradley-Terry-Luce model. This extension bridges the gap to practical reinforcement learning from human feedback (RLHF) scenarios, where even outcome reward feedback is rare.
- (4) We also identify a fundamental separation between outcome-based and per-step feedback (Section 5). Specifically, there exists a MDP with known transition dynamics and horizon $H = 2$, and the reward being a d -dimensional generalized linear function, while in this problem $e^{\Omega(d)}$ samples are necessary to learn a near-optimal policy with only outcome reward. However, such a problem is known to be *easy* with per-step reward feedback, in the sense that existing algorithms can return an ε -optimal within $\tilde{O}(d^2 / \varepsilon^2)$ rounds with per-step feedback. This separation demonstrates that delicate analysis based on well-behaved reward structure can fail catastrophically when only outcome reward feedback is available.

Our results provide a theoretical foundation for understanding when outcome-based exploration is tractable and when it presents insurmountable statistical barriers. By characterizing these fundamental limits, we offer guidance for the development of efficient algorithms for learning from trajectory-level feedback in online settings and highlight the precise conditions under which outcome-based feedback is statistically equivalent to per-step feedback.

2 Preliminaries

Markov Decision Process. An MDP M is specified by a tuple $(\mathcal{S}, \mathcal{A}, \mathbb{T}, \rho, R, H)$, with state space \mathcal{S} , action space \mathcal{A} , horizon H , transition kernel $\mathbb{T} = (\mathbb{T}_h : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S}))_{h=1}^{H-1}$, ρ is initial state distribution $\rho \in \Delta(\mathcal{S})$, and the mean reward function $R = (R_h : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1])_{h=1}^H$. At the start of each episode, the environment randomly draws an initial state $s_1 \sim \rho$, and then at each step $h \in [H]$, after the agent takes action a_h , the environment generates the next state $s_{h+1} \sim \mathbb{T}(\cdot | s_h, a_h)$. The episode terminates immediately after a_H is taken, and, for notational simplicity, we denote s_{H+1} to be the deterministic terminal state. We denote $\tau = (s_1, a_1, \dots, s_H, a_H)$ to be the trajectory, and throughout this paper we always assume the reward function is normalized, i.e., $R(\tau) := \sum_{h=1}^H R_h(s_h, a_h) \in [0, 1]$ almost surely.

87 In addition to the states, the learner may also observe the reward feedback after the episode terminates.
 88 In the *process reward* feedback setting, the learner receives a random reward vector $(r_1, \dots, r_H) \in$
 89 $[0, 1]^H$ such that $\mathbb{E}[r_h|\tau] = R_h(s_h, a_h)$ for each $h \in [H]$. In the *outcome reward* setting, the learner
 90 only receives a single reward value $r \in [H]$ such that $\mathbb{E}[r|\tau] = \sum_{h=1}^H R_h(s_h, a_h)$.

91 **Policies, value functions, and the Bellman operator.** A (randomized) policy π is specified as
 92 $\{\pi_h : \mathcal{S} \rightarrow \Delta(\mathcal{A})\}$, and it induces a distribution \mathbb{P}^π of trajectory $\tau = (s_1, a_1, \dots, s_H, a_H)$ by $s_1 \sim \rho$,
 93 and for each $h \in [H]$, $a_h \sim \pi_h(s_h)$, $s_{h+1} \sim \mathbb{T}_h(s_h, a_h)$. We let $\mathbb{E}^\pi[\cdot]$ to be the corresponding
 94 expectation.

95 The expected cumulative reward of a policy π is given by $J(\pi) := \mathbb{E}^\pi \left[\sum_{h=1}^H R_h(s_h, a_h) \right]$. The
 96 value function and Q -function of π is defined as

$$V_h^\pi(s) := \mathbb{E}^\pi \left[\sum_{\ell=h}^H R_\ell(s_\ell, a_\ell) \middle| s_h = s \right], \quad Q_h^\pi(s, a) := \mathbb{E}^\pi \left[\sum_{\ell=h}^H R_\ell(s_\ell, a_\ell) \middle| s_h = s, a_h = a \right].$$

97 Let π^* denote an optimal policy (i.e., $\pi^* \in \operatorname{argmax}_\pi J(\pi)$), and let V^* and Q^* be the corresponding
 98 value function and Q -function. It is well-known that (V^*, Q^*) satisfies the following Bellman
 99 equation for each $s \in \mathcal{S}, a \in \mathcal{A}, h \in [H]$:

$$V_h^*(s) = \max_{a \in \mathcal{A}} Q_h^*(s, a), \quad Q_h^*(s, a) = R_h(s, a) + \mathbb{E}_{s' \sim \mathbb{T}_h(\cdot|s, a)} V_{h+1}^*(s'), \quad (1)$$

100 with the convention that $V_{H+1}^* = 0$. Therefore, we define the Bellman operator \mathcal{T}_h as follows: for
 101 any $f : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$, $\mathcal{T}_h f$ is defined as

$$[\mathcal{T}_h f](s, a) := R_h(s, a) + \mathbb{E}_{s' \sim \mathbb{T}_h(\cdot|s, a)} \max_{a' \in \mathcal{A}} f(s', a'). \quad (2)$$

102 Then, it is straightforward to verify that the Bellman equation reduces to $Q_h^* = \mathcal{T}_h Q_{h+1}^*$ for $h \in [H]$.

103 **Complexity measure of the MDP.** Coverability is a natural notion for measuring the difficulty of
 104 learning in the underlying MDP (Xie et al., 2022).

Definition 1 (Coverability). *For a given MDP M and a policy class Π , the coverability C_{cov} is defined as*

$$C_{\text{cov}}(\Pi; M) := \min_{\mu_1, \dots, \mu_H \in \Delta(\mathcal{S} \times \mathcal{A})} \max_{h \in [H], \pi \in \Pi} \left\| \frac{d_h^\pi}{\mu_h} \right\|_\infty,$$

105 where $\left\| \frac{d_h^\pi}{\mu_h} \right\|_\infty := \max_{s \in \mathcal{S}, a \in \mathcal{A}} \frac{d_h^\pi(s, a)}{\mu_h(s, a)}$.

106 The coverability coefficient of an MDP is an inherent measure of the diversity of the state-action
 107 distributions. Our main upper bounds scale with the coverability of the underlying MDP M^* , and in
 108 this case we abbreviate $C_{\text{cov}}(\Pi) := C_{\text{cov}}(\Pi; M^*)$ for succinctness.

109 **Function approximation.** In this paper, we work with (model-free) function approximation, where
 110 the learner have access to a *value function class* $\mathcal{F} = \mathcal{F}_1 \times \dots \times \mathcal{F}_H$ and a *reward function class*
 111 $\mathcal{R} = \mathcal{R}_1 \times \dots \times \mathcal{R}_H$ with each $\mathcal{F}_h, \mathcal{R}_h \subseteq (\mathcal{S} \times \mathcal{A} \rightarrow [0, 1])$.

112 The function class \mathcal{F} and \mathcal{R} consist of candidate functions to approximate Q^* and the ground-truth
 113 reward function R^* .¹ In the literature of RL with general function approximation, it is typically
 114 assumed that the function classes are *realizable*, i.e., $Q^* \in \mathcal{F}$ and $R^* \in \mathcal{R}$. In this paper, we adopt
 115 the following relaxed realizability condition with a fixed approximation error $\varepsilon_{\text{app}} \geq 0$.

116 **Assumption 1** (Realizability). *There exists $Q^\sharp \in \mathcal{F}$ and $R^\sharp \in \mathcal{R}$ such that $\max_{h \in [H]} \|Q_h^\sharp - Q_h^*\|_\infty \leq$
 117 $\varepsilon_{\text{app}}, \max_{h \in [H]} \|R_h^\sharp - R_h^*\|_\infty \leq \varepsilon_{\text{app}}$.*

118 For each value function $f \in \mathcal{F}$, it induces a greedy policy π_f given by $\pi_{f, h}(s) := \operatorname{argmax}_{a \in \mathcal{A}} f(s, a)$.
 119 Therefore, the value function class \mathcal{F} induces a policy class $\Pi_{\mathcal{F}} := \{\pi_f : f \in \mathcal{F}\}$, and we take our
 120 policy class $\Pi = \Pi_{\mathcal{F}}$ for the remaining part of this paper.

121 The complexity of the function class is measured by the covering number.

¹In the following, we always write R^* for the true reward function to avoid confusion.

Definition 2 (Covering number). For a function class $\mathcal{H} \subseteq (\mathcal{X} \rightarrow \mathbb{R})$ and parameter $\alpha \geq 0$, an α -covering of \mathcal{H} (with respect to the sup norm) is a subset $\mathcal{H}' \subseteq \mathcal{H}$ such that for any $f \in \mathcal{H}$, there exists $f' \in \mathcal{H}'$ with $\sup_{x \in \mathcal{X}} |f(x) - f'(x)| \leq \alpha$. We define the α -covering number of \mathcal{H} as $N(\mathcal{H}, \alpha) := \min\{|\mathcal{H}'| : \mathcal{H}' \text{ is a } \alpha\text{-covering of } \mathcal{H}\}$.

Bellman completeness. A reward function $R = (R_1, \dots, R_H) \in \mathcal{R}$ induces a Bellman operator as

$$[\mathcal{T}_{R,h}f](s, a) := R_h(s, a) + \mathbb{E}_{s' \sim \mathbb{T}_h(\cdot|s,a)} \max_{a' \in \mathcal{A}} f_{h+1}(s', a'), \quad \forall f = (f_1, \dots, f_H) \in \mathcal{F},$$

where we also adopt the notation $f_{H+1} = 0$ for any $f \in \mathcal{F}$. Most literature on RL with general function approximation also makes use of a richer comparator function class $\mathcal{G} = \mathcal{G}_1 \times \dots \times \mathcal{G}_H$ that satisfies the following *Bellman completeness* (Jin et al., 2021a; Xie et al., 2021, 2022, etc.).

Assumption 2 (Bellman completeness). For each $h \in [H]$, $\mathcal{F}_h \subseteq \mathcal{G}_h$. For any $f \in \mathcal{F}$ and $R \in \mathcal{R}$, it holds $\inf_{g_h \in \mathcal{G}_h} \|\mathcal{T}_{R,h}f_{h+1} - g_h\|_\infty \leq \varepsilon_{\text{app}}$ for $h \in [H]$.

Miscellaneous notation. For any $p \in [0, 1]$, we define $\text{Bern}(p)$ to be the Bernoulli distribution with $\mathbb{P}(X = 1) = p$. For functions f and g , we use $f = O(g)$ to denote that there exists a universal constant C such that $f \leq C \cdot g$.

3 Sample-Efficient Online RL with Outcome Reward

In this section, we present a model-free RL algorithm with outcome reward, which achieves sample complexity guarantee scaling with the coverability coefficient and the log-covering number of the function classes.

3.1 Main Result

We present [Algorithm 1](#), which is based on the principle of optimism. For simplicity of presentation, we assume that the initial state s_1 is fixed.

The crux of the proposed algorithm is a new method for performing Fitted-Q Iteration with only outcome reward, in contrast to most existing RL algorithms (with general function approximation) that make use the process reward (r_1, \dots, r_H) to fit the Q-function for each step (Du et al., 2021; Jin et al., 2021a, etc.). A natural first idea is to fit, given a dataset $\mathcal{D} = \{(\tau, r)\}$ consisting of previously observed (trajectory, outcome reward) pairs, a reward model from the reward function class \mathcal{R} by optimizing the following reward model loss:

$$\mathcal{L}_{\mathcal{D}}^{\text{RM}}(R) := \sum_{(\tau, r) \in \mathcal{D}} \left(\sum_{h=1}^H R_h(s_h, a_h) - r \right)^2. \quad (3)$$

As discussed below in [Remark 1](#), directly relying on this estimated reward model can lead to bad performance. Instead, we jointly address consistency of the value functions and reward models, as detailed below.

For any proxy reward model $R \in \mathcal{R}$ and a value function $f \in \mathcal{F}$, we define the Bellman error at step $h \in [H]$ as

$$\mathcal{E}_{\mathcal{D},h}(f_h, f_{h+1}; R) := \sum_{(\tau, r) \in \mathcal{D}} \left(f_h(s_h, a_h) - R_h(s_h, a_h) - \max_{a'} f_{h+1}(s_{h+1}, a') \right)^2, \quad (4)$$

a measure of violation of the Bellman equation (1) with the proxy reward model R . Then, we introduce the Bellman loss defined as

$$\mathcal{L}_{\mathcal{D}}^{\text{BE}}(f; R) := \sum_{h=1}^H \mathcal{E}_{\mathcal{D},h}(f_h, f_{h+1}; R) - \inf_{g \in \mathcal{G}} \sum_{h=1}^H \mathcal{E}_{\mathcal{D},h}(g_h, f_{h+1}; R), \quad (5)$$

where we subtract the infimum of $g \in \mathcal{G}$ over the helper function class \mathcal{G} , a common approach to overcoming the double-sampling problem (Zanette et al., 2020; Jin et al., 2021a).

Algorithm 1 Outcome-Based Exploration with Optimism

input: Q-function class \mathcal{F} , reward function class \mathcal{R} , comparator class \mathcal{G} , parameter $\lambda > 0$, reference policy π_{ref} .

initialize: $\mathcal{D} \leftarrow \emptyset$.

1: **for** $t = 1, 2, \dots, T$ **do**

2: Compute the optimistic estimates:

$$(f^{(t)}, R^{(t)}) = \max_{f \in \mathcal{F}, R \in \mathcal{R}} \lambda f_1(s_1) - \mathcal{L}_{\mathcal{D}}^{\text{BE}}(f; R) - \mathcal{L}_{\mathcal{D}}^{\text{RM}}(R).$$

3: Select policy $\pi^{(t)} \leftarrow \pi_{f^{(t)}}$.

4: **for** $h = 1, 2, \dots, H$ **do**

5: Execute $\pi^{(t)} \circ_h \pi_{\text{ref}}$ for one episode and obtain $(\tau^{(t,h)}, r^{(t,h)})$

6: Update dataset: $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\tau^{(t,h)}, r^{(t,h)})\}$.

7: **end for**

8: **end for**

9: Output $\hat{\pi} = \text{Unif}(\pi^{(1:T)})$.

158 **Algorithm.** First fix an arbitrary policy π_{ref} (can be the policy which takes an arbitrary action a_0 at
159 al states). The proposed algorithm takes in a value function class \mathcal{F} , a reward function class \mathcal{R} and a
160 comparator function class \mathcal{G} , and performs the following two steps for each iteration $t = 1, 2, \dots, T$:

161 1. (Optimism) Compute optimistic estimates of (Q^*, R^*) through solving the following joint maxi-
162 mization problem with the dataset \mathcal{D} consisting of all previously observed (trajectory, outcome
163 reward) pairs:

$$(f^{(t)}, R^{(t)}) = \max_{f \in \mathcal{F}, R \in \mathcal{R}} \lambda f_1(s_1) - \mathcal{L}_{\mathcal{D}}^{\text{BE}}(f; R) - \mathcal{L}_{\mathcal{D}}^{\text{RM}}(R), \quad (6)$$

164 where for any $f \in \mathcal{F}$ we denote $f_1(s_1) := \max_{a \in \mathcal{A}} f_1(s_1, a)$ to be the value of f at the initial
165 state. Therefore, the optimization problem (6) enforces optimism by balancing the estimated
166 value $f_1(s_1)$ and the estimation error $\mathcal{L}_{\mathcal{D}}^{\text{BE}}(f; R) + \mathcal{L}_{\mathcal{D}}^{\text{RM}}(R)$ though a hyper-parameter $\lambda \geq 0$.

167 2. (Data collection) Based on the optimism estimate $f^{(t)}$, the algorithm selects $\pi^{(t)} := \pi_{f^{(t)}}$. To
168 collect data, the algorithm then executes the exploration policies $\pi \circ_h \pi_{\text{ref}}$ for each $h \in [H]$,
169 where for any policy π and π_{ref} , we let $\pi \circ_h \pi_{\text{ref}}$ be the policy that executes π for the first h steps,
170 and then executes π_{ref} starting at the $(h + 1)$ -th step.

171 **Theoretical analysis.** For Algorithm 1, we provide the following sample complexity guarantee,
172 which scales with the coverability $C_{\text{cov}} = C_{\text{cov}}(\Pi_{\mathcal{F}})$, where $\Pi_{\mathcal{F}} = \{\pi_f : f \in \mathcal{F}\}$ is the policy class
173 induced by \mathcal{F} . To simplify the presentation, we denote $\log N_T := \inf_{\alpha \geq 0} (\log N(\alpha) + T\alpha)$, where
174 $N(\alpha)$ is defined as

$$N(\alpha) := \max_{h \in [H]} \{N(\mathcal{F}_h, \alpha), N(\mathcal{R}_h, \alpha), N(\mathcal{G}_h, \alpha)\}.$$

175 With the function classes being parametric, it is clear that $\log N_T \leq O(d \log(T))$.

176 **Theorem 1.** Let $\delta \in (0, 1)$. Suppose that Assumption 1 and Assumption 2 hold, and the parameters
177 are chosen as

$$\lambda = c_0 \max \left\{ \frac{H \log(N_T H^2 / \delta)}{\varepsilon}, TH \varepsilon_{\text{app}} \right\}, \quad T \geq c_1 \frac{C_{\text{cov}} H^2 \log(T)}{\varepsilon^2} \cdot \log(N_T H^2 / \delta), \quad (7)$$

178 where $c_0, c_1 > 0$ are absolute constants. Then with probability at least $1 - \delta$, the output policy $\hat{\pi}$ of
179 Algorithm 1 satisfies $V^*(s_1) - V^{\hat{\pi}}(s_1) \leq \varepsilon + O(C_{\text{cov}} H^2 \log(T) \cdot \varepsilon_{\text{app}})$.

180 The proof of Theorem 1 is deferred to Appendix C. Particularly, we note that when the function
181 classes satisfy $\log N_T \leq \tilde{O}(d)$ and $\varepsilon_{\text{app}} = 0$, Algorithm 1 outputs an ε -optimal policy with sample
182 complexity

$$TH \leq \tilde{O}\left(\frac{C_{\text{cov}} d H^3}{\varepsilon^2}\right).$$

Notably, the coverability C_{cov} measures the inherent complexity of the underlying MDP M^* (Xie et al., 2022) and it is independent of the reward function class. As our result only depends on the coverability C_{cov} and the statistical complexity of the function classes, it does *not* rely on the structure of reward functions, while previous works assume the reward functions are either linear (Efroni et al., 2021; Cassel et al., 2024) or admit low trajectory eluder dimension (Chen et al., 2022b,a).

Remark 1. In Algorithm 1, the reward functions $R^{(t)}$ and Q -functions $f^{(t)}$ are jointly optimized (see Eq. (6)). A natural question is whether these can be optimized separately—i.e., first learning a fitted reward model and then applying optimism to the Q -functions based on the learned reward model. We show that this decoupled approach can lead to failures: due to reward model mismatch, the algorithm may become ‘trapped’ in regions where the exploratory policy fails to gather informative data. As a result, the sample complexity can become infinite in the worst case. See Section F.1 in the appendix for details.

3.2 A Simpler Algorithm for Deterministic MDPs

A disadvantage of Algorithm 1 is that it requires solving a max-min optimization problem (6), as the Bellman loss $\mathcal{L}_{\mathcal{D}}^{\text{BE}}$ involves a minimization problem over \mathcal{G} . While such computationally inefficient optimization problems are the common subroutines of existing function approximation RL algorithms (Jin et al., 2021a; Foster et al., 2021, 2022; Chen et al., 2022a, etc.), it turns out that Algorithm 1 can be significantly simplified when the transition dynamics in underlying MDP are deterministic.

Assumption 3. The transition kernel \mathbb{T} is deterministic, i.e., for any $h \in [H]$ and $s_h \in \mathcal{S}, a_h \in \mathcal{A}$, there is a unique state $s_{h+1} \in \mathcal{S}$ such that $\mathbb{T}_h(s_{h+1} \mid s_h, a_h) = 1$.

Note that in this setting, the initial state s_1 and the outcome reward r can still be random. This setting is also referred to as *Deterministic Contextual MDP* in Xie et al. (2024).

Value difference as reward model. A key observation is that, when the underlying MDP M^* is deterministic, the Bellman equation (1) trivially reduces to the following equality

$$Q_h^*(s_h, a_h) = R_h^*(s_h, a_h) + V_{h+1}^*(s_{h+1}),$$

which holds almost surely. Hence, for any trajectory τ , it holds that

$$R^*(\tau) = \sum_{h=1}^H R_h^*(s_h, a_h) = \sum_{h=1}^H [Q_h^*(s_h, a_h) - V_{h+1}^*(s_{h+1})].$$

Therefore, any value function $f \in \mathcal{F}$ induces an outcome reward model $R^f : (\mathcal{S} \times \mathcal{A})^H \rightarrow \mathbb{R}$ defined as

$$R^f(\tau) = \sum_{h=1}^H [f_h(s_h, a_h) - f_{h+1}(s_{h+1})],$$

where we adopt the notation $f_h(s) := \max_{a \in \mathcal{A}} f_h(s, a)$ for $h \in [H]$. This observation motivates the following Bellman Residual loss:

$$\mathcal{L}_{\mathcal{D}}^{\text{BR}}(f) := \sum_{(\tau, r) \in \mathcal{D}} \left(\sum_{h=1}^H [f_h(s_h, a_h) - f_{h+1}(s_{h+1})] - r \right)^2, \quad (8)$$

where $\mathcal{D} = \{(\tau, r)\}$ is any dataset consisting of (trajectory, outcome reward) pairs.

Bellman Residual Minimization (BRM) with Optimism. For deterministic MDP, we propose Algorithm 2 as a simplification of our main algorithm. Similar to Algorithm 1, the proposed algorithm takes in the value function class \mathcal{F} and alternates between the following two steps for each round $t = 1, 2, \dots, T$:

1. (Optimism) Compute optimistic estimates of Q^* through solving the following maximization problem with the dataset \mathcal{D} consisting of all previously observed (trajectory, outcome reward) pairs:

$$f^{(t)} = \max_{f \in \mathcal{F}, R \in \mathcal{R}} \lambda f_1(s_1) - \mathcal{L}_{\mathcal{D}}^{\text{BR}}(f), \quad (9)$$

Algorithm 2 Outcome-Based Exploration with Optimism for Deterministic MDP

input: Function class \mathcal{F} , parameter $\lambda > 0$.

initialize: $\mathcal{D} \leftarrow \emptyset$, initial estimate $f^{(1)} \in \mathcal{F}$.

1: **for** $t = 1, 2, \dots, T$ **do**

2: Receive $s^{(t)}$ and compute the optimistic estimates:

$$f^{(t)} = \max_{f \in \mathcal{F}} \lambda f_1(s_1^{(t)}) - \mathcal{L}_{\mathcal{D}}^{\text{BR}}(f).$$

3: Select policy $\pi^{(t)} \leftarrow \pi_{f^{(t)}}$.

4: Execute $\pi^{(t)}$ to obtain a trajectory $\tau^{(t)} = (s_1^{(t)}, a_1^{(t)}, \dots, s_H^{(t)}, a_H^{(t)})$ with outcome reward $r^{(t)}$.

5: Update dataset: $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\tau^{(t)}, r^{(t)})\}$.

6: **end for**

7: Output $\hat{\pi} = \text{Unif}(\pi^{(1:T)})$.

221 enforcing optimism by balancing the estimated value $f_1(s_1)$ and the Bellman residual loss
222 $\mathcal{L}_{\mathcal{D}}^{\text{BR}}(f)$.

223 2. (Data collection) Based on the optimistic estimate $f^{(t)}$, selects $\pi^{(t)} := \pi_{f^{(t)}}$ and collect a trajectory
224 $\tau^{(t)} = (s_1^{(t)}, a_1^{(t)}, \dots, s_H^{(t)}, a_H^{(t)})$ with outcome reward $r^{(t)}$.

225 Compared to [Algorithm 1](#), [Algorithm 2](#) has the several advantages. First, it does not rely on the
226 reward function class \mathcal{R} and the comparator function class \mathcal{G} , and the Bellman residual loss $\mathcal{L}_{\mathcal{D}}^{\text{BR}}$
227 is much simpler than the Bellman loss $\mathcal{L}_{\mathcal{D}}^{\text{BE}}$, thanks to the deterministic nature of the underlying
228 MDP. Therefore, [Algorithm 2](#) is more amenable to computationally efficient implementation, as it
229 replaces the max-min optimization problem (6) in [Algorithm 1](#) with a much simpler maximization
230 problem (9). Further, for every round t , the algorithm only needs to collect one episode from the
231 greedy policy $\pi^{(t)}$.

232 **Theoretical analysis.** We present the upper bound of [Algorithm 2](#) in terms of the following notion
233 of coverability,

$$C'_{\text{cov}}(\Pi) := \mathbb{E}_{s_1 \sim \rho} C_{\text{cov}}(\Pi; M_{s_1}^*),$$

234 where M^* is the underlying MDP, $M_{s_1}^*$ is the MDP with deterministic initial state s_1 and the same
235 transition dynamics as M^* , and $\Pi = \Pi_{\mathcal{F}}$ is the policy class induced by \mathcal{F} . In general, $C'_{\text{cov}}(\Pi)$ is
236 always an upper bound on the coverability $C_{\text{cov}}(\Pi)$, and the guarantee of [Algorithm 2](#) scales with
237 $C'_{\text{cov}}(\Pi)$ as it avoids the layer-wise exploration strategy of [Algorithm 1](#). We also denote

$$\log N_{\mathcal{F}, T} := \inf_{\alpha \geq 0} \left(\max_{h \in [H]} N(\mathcal{F}_h, \alpha) + T\alpha \right).$$

238 **Theorem 2.** Let $\delta \in (0, 1)$. Suppose that [Assumption 1](#) holds, and the parameters are chosen as

$$\lambda = c_0 \max \left\{ \frac{H^3 \log(N_{\mathcal{F}, T}/\delta)}{\varepsilon}, T\varepsilon_{\text{app}} \right\}, \quad T \geq c_1 \frac{C'_{\text{cov}}(\Pi) H^4 \log(T)}{\varepsilon^2} \cdot \log(N_{\mathcal{F}, T}/\delta), \quad (10)$$

239 where $c_0, c_1 > 0$ are absolute constants. Then with probability at least $1 - \delta$, [Algorithm 2](#) achieves

$$\frac{1}{T} \sum_{t=1}^T \left(V^*(s_1^{(t)}) - V^{\pi^{(t)}}(s_1^{(t)}) \right) \leq \varepsilon + O(C'_{\text{cov}}(\Pi) H \log(T) \cdot \varepsilon_{\text{app}}).$$

240 The above upper bound provides the PAC guarantee through the standard online-to-batch conversion,
241 and its proof is deferred to [Appendix D](#). It is worth noting that [Theorem 2](#) only relies on realiz-
242 ability assumption on the Q-function class \mathcal{F} , significantly relaxing the assumptions of realizability
243 ([Assumption 1](#)) and completeness ([Assumption 2](#)) in [Theorem 1](#).

244 4 Preference-based Reinforcement Learning

245 The goal of preference-based learning is to find a near-optimal policy only through interacting with
246 the environment that provides *preference feedback*. As an extension of our results presented in
247 [Section 3](#), in this section we present a similar algorithm for preference-based RL with the same
248 sample complexity guarantee.

249 **Preference-based learning in MDP.** In preference-based RL, the interaction protocol of the learner
 250 with the environment is specified as follows. For each round $t = 1, 2, \dots$,

- 251 • The learner selects policy $\pi^{(t,+)}$ and $\pi^{(t,-)}$.
- 252 • The learner receives trajectories $\tau^{(t,+)} \sim \pi^{(t,+)}$, $\tau^{(t,-)} \sim \pi^{(t,-)}$, and *preference feedback* $y^{(t)} \sim$
 253 $\text{Bern}(\mathbb{C}(\tau^{(t,+)}, \tau^{(t,-)}))$, where \mathbb{C} is a comparison function.

254 Intuitively, for any trajectory pair (τ^+, τ^-) , the comparison function $\mathbb{C}(\tau^+, \tau^-) = \mathbb{P}(\tau^+ \succ \tau^-)$
 255 measures the probability that τ^+ is more preferred. In this paper, we mainly focus on the Bradley-
 256 Terry-Luce (BTR) model (Bradley and Terry, 1952), which is widely used on RLHF literature. We
 257 expect that our algorithm and analysis techniques apply to a broader class of preference models.

258 **Definition 3** (BTR model). *The comparison function \mathbb{C} is specified as*

$$\mathbb{C}(\tau^+, \tau^-) = \frac{\exp(\beta R^*(\tau^+))}{\exp(\beta R^*(\tau^+)) + \exp(\beta R^*(\tau^-))},$$

259 where R^* is the ground-truth reward function, $\beta > 0$ is a parameter.

260 Under BTR model, the preference feedback in fact contains information of the outcome rewards.
 261 Hence, in this sense, preference-based RL can be regarded as an extension of outcome-based RL with
 262 weaker feedback.

263 **Algorithm for preference-based RL.** To extend Algorithm 1, we need to modify the reward model
 264 loss $\mathcal{L}_{\mathcal{D}}^{\text{RM}}$ (defined in (3)) to incorporate preference feedback. For any dataset $\mathcal{D} = \{(\tau^+, \tau^-, y)\}$
 265 consisting of (trajectories, preference) pair, we introduce the following preference-based reward
 266 model loss $\mathcal{L}_{\mathcal{D}}^{\text{PbRM}}$:

$$\mathcal{L}_{\mathcal{D}}^{\text{PbRM}}(R) := \sum_{(\tau^+, \tau^-, y) \in \mathcal{D}} L(R(\tau^+) - R(\tau^-), y), \quad (11)$$

267 where $L(w, y) := -\beta w y + \log(1 + e^{\beta w})$ is the logistic loss. It is well-known that under BTR
 268 model (Definition 3), the ground-truth reward R^* is the population minimizer of $\mathcal{L}_{\mathcal{D}}^{\text{PbRM}}$, and any
 269 approximate minimizer of $\mathcal{L}_{\mathcal{D}}^{\text{PbRM}}$ can serve as a proxy for R^* . Therefore, with the loss function
 270 $\mathcal{L}_{\mathcal{D}}^{\text{PbRM}}$, we propose the following algorithm (detailed description in Appendix E), which generalizes
 271 Algorithm 1 to handle preference feedback: For each iteration $t = 1, 2, \dots, T$, the algorithm
 272 performs the following two steps.

- 273 1. (Optimism) Compute optimistic estimates of (Q^*, R^*) through solving the following joint maxi-
 274 mization problem with the dataset \mathcal{D} consisting of all previously observed (trajectories, feedback)
 275 pairs:

$$(f^{(t)}, R^{(t)}) = \max_{f \in \mathcal{F}, R \in \mathcal{R}} \lambda \left[f_1(s_1) - \widehat{V}_{\mathcal{D}, R}^{\text{ref}} \right] - \mathcal{L}_{\mathcal{D}}^{\text{BE}}(f; R) - \mathcal{L}_{\mathcal{D}}^{\text{PbRM}}(R), \quad (12)$$

276 where the Bellman loss $\mathcal{L}_{\mathcal{D}}^{\text{BE}}$ is defined in (5), $\widehat{V}_{\mathcal{D}, R}^{\text{ref}}$ is the estimated value function of π_{ref} defined
 277 as

$$\widehat{V}_{\mathcal{D}, R}^{\text{ref}} := \frac{1}{|\mathcal{D}|} \sum_{(\tau^+, \tau^-, y) \in \mathcal{D}} R(\tau^-). \quad (13)$$

278 The term $f_1(s_1) - \widehat{V}_{\mathcal{D}, R}^{\text{ref}}$ can be regarded as an estimate of the advantage of π_f over π_{ref} under
 279 (f, R) . It is introduced to avoid over-estimating the optimal value, as the preference feedback
 280 only provide information between the *difference* between two trajectories.

- 281 2. (Data collection) The algorithm selects the greedy policy $\pi^{(t)} := \pi_{f^{(t)}}$. For each $h \in [H]$, the
 282 algorithm sets $\pi^{(t, h, +)} := \pi \circ_h \pi_{\text{ref}}$ and $\pi^{(t, h, -)} := \pi_{\text{ref}}$, executes $(\pi^{(t, h, +)}, \pi^{(t, h, -)})$ to collect
 283 trajectories $(\tau^{(t, h, +)}, \tau^{(t, h, -)})$ and the preference feedback $y^{(t, h)}$.

284 We provide the following sample complexity guarantee of the algorithm above.

285 **Theorem 3.** Let $\delta \in (0, 1)$. Suppose that [Assumption 1](#) and [Assumption 2](#) hold, and the parameters
286 are chosen as

$$\lambda = c_0 \max \left\{ \frac{H \log(N_{TH^2}/\delta)}{\varepsilon}, TH\varepsilon_{\text{app}} \right\}, \quad T \geq \tilde{O} \left(\frac{C_{\text{cov}} H^2}{\varepsilon^2} \cdot \log(N_{TH^2}/\delta) \right), \quad (14)$$

287 where $c_0 > 0$ is an absolute constant, and $\tilde{O}(\cdot)$ omits poly-logarithmic factors and constant depending
288 on β . Then, with probability at least $1 - \delta$, the output policy $\hat{\pi}$ of [Algorithm 1](#) satisfies

$$V^*(s_1) - V^{\hat{\pi}}(s_1) \leq \varepsilon + \tilde{O}(C_{\text{cov}} H^2 \varepsilon_{\text{app}}).$$

289 The proof of [Theorem 3](#) is deferred to [Section E.1](#).

290 5 Lower Bounds

291 [Theorem 1](#) from the last section indicates that as long as the coverability of the MDP is bounded,
292 finding a nearly optimal policy within a polynomial number of outcome-based samples is possible, if
293 both the function class and the reward class satisfy the realizability and completeness assumption.
294 Compared to per-step-based samples, the sample complexity of our results matches (up to polynomial
295 factors) the sample complexity of Algorithm GOLF in [Xie et al. \(2022\)](#). This indicates a statistical
296 equivalence between learning with outcome-based samples and learning with per-step-based samples.
297 Additionally, the results in [Jia et al. \(2025\)](#) indicate that outcome-based samples are statistically
298 equivalent to per-step-based samples in the setting of offline reinforcement learning. Hence, we
299 would like to ask the following similar question in the online reinforcement learning setting:

300 *Is outcome-based samples are statistically equivalent to per-step-based samples in the online RL*
301 *setting?*

302 In the following, we answer the above question negatively. We provide a special case in which finding
303 a nearly optimal policy is impossible with a polynomial number of outcome-based samples, but
304 possible with a polynomial number of per-step-based samples.

305 **Theorem 4.** For any positive integer d , there exists a class \mathcal{M} of two-layer MDPs that realizes the
306 ground truth model, such that

(a) *There exists an algorithm for per-step-based samples, whose regret has the following upper bound*

$$R(T) = \tilde{O}(d\sqrt{T}).$$

(b) *For any algorithm that takes outcome-based samples, the regret have the following lower bound*

$$R(T) = \Omega(\min\{T, e^{\Omega(d)}\}).$$

307 The above algorithm states in terms of the regret. According to the online-to-batch conversion,
308 when fixing d , there exists a class \mathcal{M} of MDPs such that the PAC sample complexity of learning
309 through per-step-based samples is $\tilde{O}(d^2/\varepsilon^2)$, while the sample complexity of learning through
310 outcome-based samples is $\Omega(e^{\Omega(d)}/\varepsilon)$. Hence there is an exponential separation between learning
311 with per-step-based samples and learning with outcome-based samples.

312 6 Conclusion

313 In this work, we develop a model-free, sample-efficient algorithm for outcome-based reinforcement
314 learning that relies solely on trajectory-level rewards and achieves theoretical guarantees bounded
315 by coverability under function approximation. From the lower bound side, we show that joint
316 optimization of reward and value functions is essential, and establish a fundamental exponential
317 gap between outcome-based and per-step feedback. For deterministic MDPs, we propose a simpler,
318 more efficient variant, and extend our approach to preference-based feedback, demonstrating that it
319 preserves the same statistical efficiency.

References

- Philip Amortila, Dylan J Foster, Nan Jiang, Ayush Sekhari, and Tengyang Xie. Harnessing density ratios for online reinforcement learning. *arXiv preprint arXiv:2401.09681*, 2024a.
- Philip Amortila, Dylan J Foster, and Akshay Krishnamurthy. Scalable online exploration via coverability. *arXiv preprint arXiv:2403.06571*, 2024b.
- András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71:89–129, 2008.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.
- Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 19–26. JMLR Workshop and Conference Proceedings, 2011.
- Mohak Bhardwaj, Tengyang Xie, Byron Boots, Nan Jiang, and Ching-An Cheng. Adversarial model for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2023.
- Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.
- Asaf Cassel, Haipeng Luo, Aviv Rosenberg, and Dmitry Sotnikov. Near-optimal regret in linear mdps with aggregate bandit feedback. *arXiv preprint arXiv:2405.07637*, 2024.
- Shicong Cen, Jincheng Mei, Katayoon Goshvadi, Hanjun Dai, Tong Yang, Sherry Yang, Dale Schuurmans, Yuejie Chi, and Bo Dai. Value-incentivized preference optimization: A unified approach to online and offline rlhf. *arXiv preprint arXiv:2405.19320*, 2024.
- Niladri Chatterji, Aldo Pacchiano, Peter Bartlett, and Michael Jordan. On the theory of reinforcement learning with once-per-episode feedback. *Advances in Neural Information Processing Systems*, 34: 3401–3412, 2021.
- Fan Chen, Song Mei, and Yu Bai. Unified algorithms for rl with decision-estimation coefficients: pac, reward-free, preference-based learning, and beyond. *arXiv preprint arXiv:2209.11745*, 2022a.
- Fan Chen, Constantinos Daskalakis, Noah Golowich, and Alexander Rakhlin. Near-optimal learning and planning in separated latent mdps. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 995–1067. PMLR, 2024.
- Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pages 1042–1051. PMLR, 2019.
- Xiaoyu Chen, Han Zhong, Zhuoran Yang, Zhaoran Wang, and Liwei Wang. Human-in-the-loop: Provably efficient preference-based reinforcement learning with general function approximation. In *International Conference on Machine Learning*, pages 3773–3793. PMLR, 2022b.
- Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 208–214. JMLR Workshop and Conference Proceedings, 2011.
- Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. In *21st Annual Conference on Learning Theory*, number 101, pages 355–366, 2008.
- Nirjhar Das, Souradip Chakraborty, Aldo Pacchiano, and Sayak Ray Chowdhury. Provably sample efficient rlhf via active preference optimization. *arXiv preprint arXiv:2402.10500*, 2024.
- Kefan Dong, Jiaqi Yang, and Tengyu Ma. Provable model-based nonlinear bandit and reinforcement learning: Shelve optimism, embrace virtual curvature. *Advances in Neural Information Processing Systems*, 34:26168–26182, 2021.

367 Simon Du, Sham Kakade, Jason Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong
368 Wang. Bilinear classes: A structural framework for provable generalization in rl. In *International*
369 *Conference on Machine Learning*, pages 2826–2836. PMLR, 2021.

370 Yonathan Efroni, Nadav Merlis, and Shie Mannor. Reinforcement learning with trajectory feedback.
371 In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 7288–7295,
372 2021.

373 Amir-massoud Farahmand, Csaba Szepesvári, and Rémi Munos. Error propagation for approximate
374 policy and value iteration. *Advances in neural information processing systems*, 23, 2010.

375 Dylan J Foster, Sham M Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of
376 interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

377 Dylan J Foster, Alexander Rakhlin, Ayush Sekhari, and Karthik Sridharan. On the complexity of
378 adversarial decision making. *Advances in Neural Information Processing Systems*, 35:35404–
379 35417, 2022.

380 Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec
381 Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint*
382 *arXiv:2412.16720*, 2024.

383 Zeyu Jia, Alexander Rakhlin, and Tengyang Xie. Do we need to verify step by step? rethinking
384 process supervision from a theoretical perspective. *arXiv preprint arXiv:2502.10581*, 2025.

385 Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Context-
386 tual decision processes with low bellman rank are pac-learnable. In *International Conference on*
387 *Machine Learning*, pages 1704–1713. PMLR, 2017.

388 Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of rl
389 problems, and sample-efficient algorithms. *Advances in Neural Information Processing Systems*,
390 34:13406–13418, 2021a.

391 Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In
392 *International Conference on Machine Learning*, pages 5084–5096. PMLR, 2021b.

393 Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In
394 *ICML*, volume 2, pages 267–274, 2002.

395 Tal Lancewicki and Yishay Mansour. Near-optimal regret using policy optimization in online mdps
396 with aggregate bandit feedback. *arXiv preprint arXiv:2502.04004*, 2025.

397 Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

398 Gene Li, Pritish Kamath, Dylan J Foster, and Nati Srebro. Understanding the eluder dimension.
399 *Advances in Neural Information Processing Systems*, 35:23737–23750, 2022.

400 Fanghui Liu, Luca Viano, and Volkan Cevher. What can online reinforcement learning with function
401 approximation benefit from general coverage conditions? In *International Conference on Machine*
402 *Learning*, pages 22063–22091. PMLR, 2023.

403 Rémi Munos. Error bounds for approximate policy iteration. In *ICML*, volume 3, pages 560–567.
404 Citeseer, 2003.

405 Gergely Neu and Gábor Bartók. An efficient algorithm for learning with semi-bandit feedback. In
406 *International Conference on Algorithmic Learning Theory*, pages 234–248. Springer, 2013.

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

410 Ian Osband and Benjamin Van Roy. Model-based reinforcement learning and the eluder dimension.
411 In *Advances in Neural Information Processing Systems*, volume 27, pages 1466–1474, 2014.

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.

Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. In *Advances in Neural Information Processing Systems*, volume 26, pages 2256–2264, 2013.

Wen Sun, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Model-based RL in contextual decision processes: PAC bounds and exponential improvements over model-free approaches. In *Conference on learning theory*, pages 2898–2933. PMLR, 2019.

Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Yuanhao Wang, Qinghua Liu, and Chi Jin. Is rlhf more difficult than standard rl? *arXiv preprint arXiv:2306.14111*, 2023.

Runzhe Wu and Wen Sun. Making rl with preference-based feedback efficient via randomization. *arXiv preprint arXiv:2310.14554*, 2023.

Tengyang Xie and Nan Jiang. Batch value-function approximation with only realizability. In *International Conference on Machine Learning*, pages 11404–11413. PMLR, 2021.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34:6683–6694, 2021.

Tengyang Xie, Dylan J Foster, Yu Bai, Nan Jiang, and Sham M Kakade. The role of coverage in online reinforcement learning. *arXiv preprint arXiv:2210.04157*, 2022.

Tengyang Xie, Dylan J Foster, Akshay Krishnamurthy, Corby Rosset, Ahmed Awadallah, and Alexander Rakhlin. Exploratory preference optimization: Harnessing implicit q*-approximation for sample-efficient rlhf. *arXiv preprint arXiv:2405.21046*, 2024.

Yichong Xu, Ruosong Wang, Lin Yang, Aarti Singh, and Artur Dubrawski. Preference-based reinforcement learning with finite-time guarantees. *Advances in Neural Information Processing Systems*, 33:18784–18794, 2020.

Chenlu Ye, Wei Xiong, Yuheng Zhang, Nan Jiang, and Tong Zhang. A theoretical analysis of nash learning from human feedback under general kl-regularized preference. *arXiv preprint arXiv:2402.07314*, 2024.

Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, pages 10978–10989. PMLR, 2020.

Wenhao Zhan, Masatoshi Uehara, Nathan Kallus, Jason D Lee, and Wen Sun. Provable offline preference-based reinforcement learning. *arXiv preprint arXiv:2305.14816*, 2023.

Shenao Zhang, Donghan Yu, Hiteshi Sharma, Han Zhong, Zhihan Liu, Ziyi Yang, Shuohang Wang, Hany Hassan, and Zhaoran Wang. Self-exploring language models: Active preference elicitation for online alignment. *arXiv preprint arXiv:2405.19332*, 2024.

Tong Zhang. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2(Mar):527–550, 2002.

Banghua Zhu, Michael Jordan, and Jiantao Jiao. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *International Conference on Machine Learning*, pages 43037–43067. PMLR, 2023.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [NA].
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes]" is generally preferable to "[No]", it is perfectly acceptable to answer "[No]" provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No]" or "[NA]" is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading “NeurIPS Paper Checklist”,**
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers.**

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

Answer: [Yes]

Justification: Claims made in the abstract and introduction accurately reflect this paper’s contributions and scope. More supporting details are included in the main text.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: This paper discussed the limitations in the main body below each theorem.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: This paper provides the full set of assumptions in the main body and a complete (and correct) proof in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [NA]

Justification: This paper does not include experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: This paper does not include experiments requiring code.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [NA]

Justification: This paper does not include experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: This paper does not include experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [NA]

Justification: The paper does not include experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines?>

Answer: [Yes]

Justification: The authors have reviewed the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: This is a pure theoretical paper. There is no societal impact of the work performed.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: This paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: This paper does not use existing assets.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: This paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

767 Justification: This paper does not involve crowdsourcing and research with human subjects.

768 Guidelines:

- 769 • The answer NA means that the paper does not involve crowdsourcing nor research with
- 770 human subjects.
- 771 • Including this information in the supplemental material is fine, but if the main contribu-
- 772 tion of the paper involves human subjects, then as much detail as possible should be
- 773 included in the main paper.
- 774 • According to the NeurIPS Code of Ethics, workers involved in data collection, curation,
- 775 or other labor should be paid at least the minimum wage in the country of the data
- 776 collector.

777 **15. Institutional review board (IRB) approvals or equivalent for research with human**

778 **subjects**

779 Question: Does the paper describe potential risks incurred by study participants, whether

780 such risks were disclosed to the subjects, and whether Institutional Review Board (IRB)

781 approvals (or an equivalent approval/review based on the requirements of your country or

782 institution) were obtained?

783 Answer: [NA]

784 Justification: This paper does not involve research with human subjects.

785 Guidelines:

- 786 • The answer NA means that the paper does not involve crowdsourcing nor research with
- 787 human subjects.
- 788 • Depending on the country in which research is conducted, IRB approval (or equivalent)
- 789 may be required for any human subjects research. If you obtained IRB approval, you
- 790 should clearly state this in the paper.
- 791 • We recognize that the procedures for this may vary significantly between institutions
- 792 and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the
- 793 guidelines for their institution.
- 794 • For initial submissions, do not include any information that would break anonymity (if
- 795 applicable), such as the institution conducting the review.

796 **16. Declaration of LLM usage**

797 Question: Does the paper describe the usage of LLMs if it is an important, original, or

798 non-standard component of the core methods in this research? Note that if the LLM is used

799 only for writing, editing, or formatting purposes and does not impact the core methodology,

800 scientific rigor, or originality of the research, declaration is not required.

801 Answer: [NA]

802 Justification: The core method development in this research does not involve LLMs as any

803 important, original, or non-standard components.

804 Guidelines:

- 805 • The answer NA means that the core method development in this research does not
- 806 involve LLMs as any important, original, or non-standard components.
- 807 • Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for
- 808 what should or should not be described.

809 A More Related Works

810 We review more related works in this section.

811 The *coverability coefficient* has recently gained attention in the theory of online reinforcement
812 learning (Xie et al., 2022; Liu et al., 2023; Amortila et al., 2024a,b). This condition is in the same
813 spirit as the widely used concentrability coefficient (Munos, 2003; Antos et al., 2008; Farahmand
814 et al., 2010; Chen and Jiang, 2019; Jin et al., 2021b; Xie and Jiang, 2021; Xie et al., 2021; Bhardwaj
815 et al., 2023), a concept frequently employed in the theory of offline (or batch) reinforcement learning.
816 A well-known duality suggests that the coverability coefficient can be interpreted as the optimal
817 concentrability coefficient attainable by any offline data distribution. For further discussion, see Xie
818 et al. (2022).

819 A related body of theoretical work explores *reinforcement learning with trajectory feedback* (Neu
820 and Bartók, 2013; Efroni et al., 2021; Chatterji et al., 2021; Cassel et al., 2024; Lancewicki and
821 Mansour, 2025), where the learner receives only episode-level feedback at the end of each trajectory.
822 This category also encompasses preference-based reinforcement learning (Pacchiano et al., 2021;
823 Chen et al., 2022b; Zhu et al., 2023; Wu and Sun, 2023; Zhan et al., 2023), which relies on pairwise
824 comparisons between trajectories. While most prior work focuses on tabular or linear MDP settings,
825 we take a step further by studying learning with function approximation, and bound the complexity
826 by the coverability coefficient.

827 In the context of *Online Reinforcement Learning*, numerous prior works have investigated the
828 complexity of exploration and policy optimization, introducing various complexity measures such as
829 Bellman rank (Jiang et al., 2017), Eluder dimension (Russo and Van Roy, 2013; Osband and Van Roy,
830 2014), witness rank (Sun et al., 2019), Bellman-Eluder dimension (Jin et al., 2021a), the bilinear class
831 (Du et al., 2021), and decision-estimation coefficients (Foster et al., 2021). These complexity notions
832 characterize properties of the function or model class but are generally not instance-dependent. In
833 contrast, Xie et al. (2022) introduces the instance-dependent notion of *coverability coefficients* to
834 provide complexity bounds in online reinforcement learning. For further discussion on instance-
835 dependent complexity measures, we refer the reader to the discussions therein.

836 We further review some literatures on *online preference-based learning* or *online RLHF*. Xu et al.
837 (2020); Novoseller et al. (2020); Pacchiano et al. (2021); Wu and Sun (2023); Zhan et al. (2023); Das
838 et al. (2024) provides theoretical guarantees for tabular MDPs and linear MDPs. Ye et al. (2024)
839 studies RLHF with general function approximation for contextual bandits, which is equivalent to
840 the case where $H = 1$. Chen et al. (2022b); Wang et al. (2023) use the Eluder dimension type
841 complexity measures to characterize the sample complexity of online RLHF, which sometimes can
842 be too pessimistic. Xie et al. (2024); Cen et al. (2024); Zhang et al. (2024) proposed algorithms for
843 online RLHF with function approximation, but their complexity depends on the trajectory coverability
844 instead of the state coverability

845 B Technical tools

846 B.1 Uniform convergence with square loss

847 To prove the uniform convergence results with square loss, we frequently use the following version
848 Freedman’s inequality (see e.g., Beygelzimer et al., 2011).

849 **Lemma 5** (Freedman’s inequality). *Suppose that $Z^{(1)}, \dots, Z^{(T)}$ is a martingale difference sequence*
850 *that is adapted to the filtration $(\mathfrak{F}^{(t)})_{t=1}^T$, and $Z^{(t)} \leq C$ almost surely for all $t \in [T]$. Then for any*
851 *$\lambda \in [0, \frac{1}{C}]$, with probability at least $1 - \delta$, for all $n \leq T$,*

$$\sum_{t=1}^n Z^{(t)} \leq \lambda \sum_{t=1}^n \mathbb{E}[(Z^{(t)})^2 | \mathfrak{F}^{(t-1)}] + \frac{\log(1/\delta)}{\lambda}.$$

852 **Lemma 6.** *Suppose that $(x^{(1)}, y^{(1)}), \dots, (x^{(T)}, y^{(T)})$ is a sequence of random variable in $\mathcal{X} \times [0, C]$*
853 *that is adapted to the filtration $(\mathfrak{F}^{(t)})_{t=1}^T$, such that there exists a function $F^* : \mathcal{X} \rightarrow [0, 1]$ with*
854 *$F^*(x^{(t)}) = \mathbb{E}[y^{(t)} | \mathfrak{F}^{(t-1)}, x^{(t)}]$ almost surely. Then for any function $F : \mathcal{X} \rightarrow [0, C]$, it holds that*

855 with probability at least $1 - \delta$, for all $n \in [T]$,

$$\sum_{t=1}^n (F(x^{(t)}) - y^{(t)})^2 - \sum_{t=1}^n (F^*(x^{(t)}) - y^{(t)})^2 \geq \frac{1}{2} \sum_{t=1}^n \mathbb{E} \left[(F(x^{(t)}) - F^*(x^{(t)}))^2 \middle| \mathfrak{F}^{(t-1)} \right] - 10C^2 \log(1/\delta).$$

856 Conversely, it holds that with probability at least $1 - \delta$, for all $n \in [T]$,

$$\sum_{t=1}^n (F(x^{(t)}) - y^{(t)})^2 - \sum_{t=1}^n (F^*(x^{(t)}) - y^{(t)})^2 \leq 2 \sum_{t=1}^n \mathbb{E} \left[(F(x^{(t)}) - F^*(x^{(t)}))^2 \middle| \mathfrak{F}^{(t-1)} \right] + 5C^2 \log(1/\delta).$$

857 **Proof of Lemma 6.** Denote

$$\begin{aligned} W^{(t)} &:= (F(x^{(t)}) - y^{(t)})^2 - (F^*(x^{(t)}) - y^{(t)})^2 \\ &= (F(x^{(t)}) - F^*(x^{(t)}))^2 + 2(F(x^{(t)}) - F^*(x^{(t)}))(F^*(x^{(t)}) - y^{(t)}). \end{aligned}$$

858 Note that

$$\mathbb{E} [W^{(t)} | \mathfrak{F}^{(t-1)}] = \mathbb{E} \left[(F(x^{(t)}) - F^*(x^{(t)}))^2 \middle| \mathfrak{F}^{(t-1)} \right],$$

859 and

$$Z^{(t)} := W^{(t)} - \mathbb{E} [W^{(t)} | \mathfrak{F}^{(t-1)}] \leq W^{(t)} \leq C^2.$$

860 Therefore, using Freedman's inequality (Lemma 5), for any fixed $\lambda \in [0, \frac{1}{C^2}]$, we have with probability at least $1 - \delta$,

$$\sum_{t=1}^n Z^{(t)} \leq \lambda \sum_{t=1}^n \mathbb{E} [(Z^{(t)})^2 | \mathfrak{F}^{(t-1)}] + \frac{\log(1/\delta)}{\lambda}, \quad \forall n \in [T].$$

862 Note that

$$\begin{aligned} \mathbb{E} [(Z^{(t)})^2 | \mathfrak{F}^{(t-1)}] &\leq \mathbb{E} [(W^{(t)})^2 | \mathfrak{F}^{(t-1)}] \\ &= \mathbb{E} \left[(F(x^{(t)}) - F^*(x^{(t)}))^4 + 4(F(x^{(t)}) - F^*(x^{(t)}))^2 (F^*(x^{(t)}) - y^{(t)})^2 \middle| \mathfrak{F}^{(t-1)} \right] \\ &\leq 5C^2 \mathbb{E} \left[(F(x^{(t)}) - F^*(x^{(t)}))^2 \middle| \mathfrak{F}^{(t-1)} \right]. \end{aligned}$$

863 Therefore, by setting $\lambda = \frac{1}{5C^2}$, we get the desired upper bound.

864 Similarly, for the lower bound, we can apply Freedman's inequality with $(-Z^{(t)})$ to show that for
865 $\lambda = \frac{1}{10C^2}$, with probability at least $1 - \delta$,

$$\begin{aligned} -\sum_{t=1}^n Z^{(t)} &\leq \lambda \sum_{t=1}^n \mathbb{E} [(Z^{(t)})^2 | \mathfrak{F}^{(t-1)}] + \frac{\log(1/\delta)}{\lambda} \\ &\leq \frac{1}{2} \sum_{t=1}^n \mathbb{E} \left[(F(x^{(t)}) - F^*(x^{(t)}))^2 \middle| \mathfrak{F}^{(t-1)} \right] + 10C^2 \log(1/\delta), \quad \forall n \in [T]. \end{aligned}$$

866 □

867 **Proposition 7.** Fix a parameter $\alpha \geq 0$. Under the assumption of Lemma 6, suppose that $\mathcal{H} \subseteq (\mathcal{X} \rightarrow [0, C])$ is a fixed function class, and $F^\sharp \in \mathcal{H}$ satisfies $\|F^* - F^\sharp\|_\infty \leq \varepsilon_{\text{app}}$. Define

$$\mathcal{L}_n(F) := \sum_{t=1}^n (F(x^{(t)}) - y^{(t)})^2, \quad \mathcal{E}_n(F) := \sum_{t=1}^n \mathbb{E} \left[(F(x^{(t)}) - F^*(x^{(t)}))^2 \middle| \mathfrak{F}^{(t-1)} \right].$$

869 Let $\kappa := 15C^2 \log(2N(\mathcal{H}, \alpha)/\delta) + 3Cn\alpha + 4n\varepsilon_{\text{app}}^2$. Then the following holds simultaneously with
870 probability at least $1 - \delta$:

871 (1) For each $n \in [T]$,

$$\mathcal{L}_n(F^\sharp) - \inf_{F' \in \mathcal{H}} \mathcal{L}_n(F') \leq \kappa.$$

872 (2) For each $n \in [T]$, for all $F \in \mathcal{H}$,

$$\frac{1}{2} \mathcal{E}_n(F) \leq \mathcal{L}_n(F) - \inf_{F' \in \mathcal{H}} \mathcal{L}_n(F') + \kappa.$$

Proof of Proposition 7. Denote $N := N(\mathcal{H}, \alpha)$. Let \mathcal{H}_α be a minimal α -covering of \mathcal{H} . Then applying Lemma 6 and the union bound, we have with probability at least $1 - \delta$, the following holds simultaneously for $n \in [T]$:

(1) For all $F' \in \mathcal{H}_\alpha$, it holds that

$$\frac{1}{2}\mathcal{E}_n(F') \leq \mathcal{L}_n(F') - \mathcal{L}_n(F^*) + 10C^2 \log(2N/\delta).$$

(2) It holds that

$$\mathcal{L}_n(F^\sharp) - \mathcal{L}_n(F^*) \leq 2\mathcal{E}_n(F^\sharp) + 5C^2 \log(2/\delta).$$

In the following, we condition on the above success event.

By definition, $\mathcal{E}_n(F^\sharp) \leq n\varepsilon_{\text{app}}^2$, and hence

$$\mathcal{L}_n(F^*) \geq \mathcal{L}_n(F^\sharp) - 4n\varepsilon_{\text{app}}^2 - 5C^2 \log(2/\delta). \quad (15)$$

Furthermore, for any $F \in \mathcal{H}$, there exists $F' \in \mathcal{H}_\alpha$ with $\|F - F'\|_\infty \leq \alpha$, which implies

$$|\mathcal{L}_n(F) - \mathcal{L}_n(F')| \leq 2Cn\alpha, \quad |\mathcal{E}_n(F) - \mathcal{E}_n(F')| \leq 2Cn\alpha.$$

Therefore, under the success event, we have

$$\frac{1}{2}\mathcal{E}_n(F) \leq \mathcal{L}_n(F) - \mathcal{L}_n(F^*) + 10C^2 \log(2N/\delta) + 3Cn\alpha.$$

holds for arbitrary $F \in \mathcal{F}$. Hence, by (15), we have

$$\frac{1}{2}\mathcal{E}_n(F) \leq \mathcal{L}_n(F) - \mathcal{L}_n(F^\sharp) + 15C^2 \log(2N/\delta) + 3Cn\alpha + 4n\varepsilon_{\text{app}}^2.$$

Noting that $\mathcal{L}_n(F^\sharp) \geq \inf_{F' \in \mathcal{H}} \mathcal{L}_n(F')$ completes the proof of (2). Furthermore, using $\mathcal{E}_n(F) \geq 0$, we also have

$$\mathcal{L}_n(F^\sharp) \leq \inf_{F' \in \mathcal{H}} \mathcal{L}_n(F') + 15C^2 \log(2N/\delta) + 3Cn\alpha + 4n\varepsilon_{\text{app}}^2.$$

This completes the proof of (1). \square

B.2 Uniform convergence with log-loss

We prove the following result, which is a direct extension of the standard MLE guarantee (Zhang, 2002).

Proposition 8. Suppose that $\{P_\theta(y|x)\}_{\theta \in \Theta} \subseteq (\mathcal{X} \rightarrow \Delta(\mathcal{Y}))$ is a class of condition densities parametrized by an abstract parameter class Θ . Without loss of generality, we assume \mathcal{Y} is discrete.

A α -covering of Θ is a subset $\Theta' \subseteq \Theta$ such that for any $\theta \in \Theta$, there exists $\theta' \in \Theta'$ such that $|\log P_\theta(y|x) - \log P_{\theta'}(y|x)| \leq \alpha$ for all $x \in \mathcal{X}, y \in \mathcal{Y}$. We define the covering number of Θ under log-loss as

$$N_{\log}(\Theta, \alpha) := \min\{|\Theta'| : \Theta' \text{ is a } \alpha\text{-covering of } \Theta\}.$$

Suppose that $(x^{(1)}, y^{(1)}), \dots, (x^{(T)}, y^{(T)})$ is a sequence of random variables adapted to the filtration $(\mathfrak{F}^{(t)})_{t=1}^T$, such that there exists $\theta^* \in \Theta$ so that $\mathbb{P}(y^{(t)} = \cdot | x^{(t)}, \mathfrak{F}^{(t-1)}) = P_{\theta^*}(y^{(t)} = \cdot | x^{(t)})$ almost surely for $t \in [T]$. Then it holds that for all $n \in [T]$, for all $\theta \in \Theta$,

$$\begin{aligned} \sum_{t=1}^n \mathbb{E} \left[D_{\text{H}}^2(P_\theta(\cdot | x^{(t)}), P_{\theta^*}(\cdot | x^{(t)})) | \mathfrak{F}^{(t-1)} \right] &\leq -\frac{1}{2} \sum_{t=1}^n [\log P_\theta(y^{(t)} | x^{(t)}) - \log P_{\theta^*}(y^{(t)} | x^{(t)})] \\ &\quad + \log N_{\log}(\Theta, \alpha) + 2n\alpha. \end{aligned}$$

897 **Proof of Proposition 8.** Let $\Theta' \subseteq \Theta$ be a minimal α -covering, and let $N := |\Theta'| = N_{\log}(\Theta, \alpha)$.
 898 For each $\theta \in \Theta$, we consider

$$L^{(t)}(\theta) := \log P_{\theta}(y^{(t)}|x^{(t)}) - \log P_{\theta^*}(y^{(t)}|x^{(t)}).$$

899 Then it holds that

$$\begin{aligned} \mathbb{E} \left[\exp \left(\frac{1}{2} L^{(t)}(\theta) \right) \middle| x^{(t)}, \mathfrak{F}^{(t-1)} \right] &= \mathbb{E}_{y \sim P_{\theta^*}(\cdot|x^{(t)})} \sqrt{\frac{P_{\theta}(y|x^{(t)})}{P_{\theta^*}(y|x^{(t)})}} \\ &= \sum_{y \in \mathcal{Y}} \sqrt{P_{\theta}(y|x^{(t)})P_{\theta^*}(y|x^{(t)})} \\ &= 1 - D_{\text{H}}^2(P_{\theta}(\cdot|x^{(t)}), P_{\theta^*}(\cdot|x^{(t)})). \end{aligned}$$

900 Therefore, applying Lemma 9 and using union bound over $\theta \in \Theta'$, we have the following bound:
 901 with probability at least $1 - \delta$, for any $\theta' \in \Theta'$, $n \in [T]$,

$$\sum_{t=1}^n -\log \left[\exp \left(\frac{1}{2} L^{(t)}(\theta') \right) \middle| \mathcal{F}^{(t-1)} \right] \leq -\frac{1}{2} \sum_{t=1}^n L^{(t)}(\theta') + \log(N/\delta).$$

902 In the following, we condition on the above event. Fix any $\theta \in \Theta$. Then, there exists $\theta' \in \Theta'$ such
 903 that $|\log P_{\theta}(y|x) - \log P_{\theta'}(y|x)| \leq \alpha$ for all $x \in \mathcal{X}, y \in \mathcal{Y}$, and hence $|L^{(t)}(\theta) - L^{(t)}(\theta')| \leq \alpha$
 904 almost surely. Therefore, combining the results above and using $\log w \leq w - 1$ for $w > 0$, we have

$$\begin{aligned} \sum_{t=1}^n \mathbb{E} [D_{\text{H}}^2(P_{\theta}(\cdot|x^{(t)}), P_{\theta^*}(\cdot|x^{(t)})) | \mathfrak{F}^{(t-1)}] &\leq \sum_{t=1}^n -\log \left[\exp \left(\frac{1}{2} L^{(t)}(\theta) \right) \middle| \mathcal{F}^{(t-1)} \right] \\ &\leq n\alpha + \sum_{t=1}^n -\log \left[\exp \left(\frac{1}{2} L^{(t)}(\theta') \right) \middle| \mathcal{F}^{(t-1)} \right] \\ &\leq n\alpha - \frac{1}{2} \sum_{t=1}^n L^{(t)}(\theta') + \log(N/\delta) \\ &\leq 2n\alpha - \frac{1}{2} \sum_{t=1}^n L^{(t)}(\theta) + \log(N/\delta). \end{aligned}$$

905 By the arbitrariness of $\theta \in \Theta$, the proof is completed. \square

Lemma 9 (Foster et al. (2021), Lemma A.4). *For any sequence of real-valued random variables $X^{(1)}, \dots, X^{(T)}$ adapted to a filtration $(\mathfrak{F}^{(t)})_{t=1}^T$, it holds that with probability at least $1 - \delta$, for all $n \in [T]$,*

$$\sum_{t=1}^n -\log [\exp(-X^{(t)}) | \mathcal{F}^{(t-1)}] \leq \sum_{t=1}^n X^{(t)} + \log(1/\delta).$$

906 C Missing Proofs in Section 3.1

907 C.1 Proof of Theorem 1

908 We first present a more detailed statement of the upper bound of Theorem 1, as follows.

909 **Theorem 10.** *Let $\delta \in (0, 1)$, $\rho \in [0, 1]$, and we denote $C_{\text{cov}} = C_{\text{cov}}(\Pi_{\mathcal{F}})$, where $\Pi_{\mathcal{F}} = \{\pi_f : f \in$
 910 $\mathcal{F}\}$ is the policy class induced by \mathcal{F} . Suppose that Assumption 1 and Assumption 2 hold. Then with
 911 probability at least $1 - \delta$, the output policy $\hat{\pi}$ of Algorithm 1 satisfies*

$$\begin{aligned} V^*(s_1) - V^{\hat{\pi}}(s_1) &= \frac{1}{T} \sum_{t=1}^T \left(V^*(s_1) - V^{\pi^{(t)}}(s_1) \right) \\ &\leq O(H) \cdot \left[\frac{\log(N(\rho)/\delta) + TH^2(\rho + \varepsilon_{\text{app}}^2)}{\lambda} + \frac{\lambda C_{\text{cov}} \log(T)}{T} \right]. \end{aligned}$$

Therefore, for any $\varepsilon \in (0, 1)$, with the optimally-tuned parameter λ , it holds that $V^*(s_1) - V^{\hat{\pi}}(s_1) \leq \varepsilon + \tilde{O}(\sqrt{C_{\text{cov}}} H^2 \varepsilon_{\text{app}})$, as long as

$$T \geq \tilde{O}\left(\frac{C_{\text{cov}} H^2}{\varepsilon^2} \cdot \log N(\varepsilon^2 / (C_{\text{cov}} H^4))\right).$$

Recall that we let $Q^\sharp \in \mathcal{Q}$, $R^\sharp \in \mathcal{R}$ be such that

$$\max_{h \in [H]} \|Q_h^\sharp - Q_h^*\|_\infty \leq \varepsilon_{\text{app}}, \quad \max_{h \in [H]} \|R_h^\sharp - R_h^*\|_\infty \leq \varepsilon_{\text{app}}.$$

For each $t \in [T]$, we write $\mathcal{D}^{(t)}$ to be the dataset maintained by [Algorithm 1](#) at the end of the t th iteration, i.e.,

$$\mathcal{D}^{(t)} = \{(\tau^{(k,h)}, r^{(k,h)})\}_{k \leq t, h \in [H]}.$$

We summarize the uniform concentration results for the loss $\mathcal{L}_{\mathcal{D}^{(t)}}^{\text{BE}}$ and $\mathcal{L}_{\mathcal{D}^{(t)}}^{\text{RM}}$ as follows. We note that these concentration bounds are fairly standard (see e.g. [Jin et al. \(2021a\)](#)), and for completeness, we present the proof in [Appendix C.3](#).

Proposition 11. *Let $\delta \in (0, 1)$, $\rho \geq 0$. Suppose that [Assumption 1](#) and [Assumption 2](#) holds. Then with probability at least $1 - \delta$, for all $t \in [T]$, $f \in \mathcal{F}$, $R \in \mathcal{R}$, it holds that*

$$\begin{aligned} \frac{1}{2} \sum_{k \leq t} \sum_{h=1}^H \mathbb{E}^{\pi^{(k)} \circ_h \pi_{\text{ref}}} (R(\tau) - R^*(\tau))^2 &\leq \mathcal{L}_{\mathcal{D}^{(t)}}^{\text{RM}}(R) - \mathcal{L}_{\mathcal{D}^{(t)}}^{\text{RM}}(R^\sharp) + H\kappa, \\ \frac{1}{2} \sum_{k \leq t} \sum_{h=1}^H \mathbb{E}^{\pi^{(k)}} (f_h(s_h, a_h) - [\mathcal{T}_{R,h} f_{h+1}](s_h, a_h))^2 &\leq \mathcal{L}_{\mathcal{D}^{(t)}}^{\text{BE}}(f; R) - \mathcal{L}_{\mathcal{D}^{(t)}}^{\text{BE}}(Q^\sharp; R^\sharp) + H\kappa, \end{aligned}$$

where

$$\kappa = C(\log N(\rho) + \log(H/\delta) + TH^2(\varepsilon_{\text{app}}^2 + \rho)),$$

and $C > 0$ is an absolute constant.

Performance difference decomposition. Denote $V^\sharp(s_1) := \max_{a \in \mathcal{A}} Q^\sharp(s_1, a)$. Then it is clear that $|V^\sharp(s_1) - V^*(s_1)| \leq \varepsilon_{\text{app}}$. Therefore, for any $t \in [T]$, by optimism (the definition of $(f^{(t)}, R^{(t)})$), it holds that

$$\begin{aligned} V^*(s_1) - \varepsilon_{\text{app}} &\leq V^\sharp(s_1) \\ &= V^\sharp(s_1) - \frac{\mathcal{L}_{\mathcal{D}^{(t-1)}}^{\text{BE}}(Q^\sharp, R^\sharp) + \mathcal{L}_{\mathcal{D}^{(t-1)}}^{\text{RM}}(R^\sharp)}{\lambda} + \frac{\mathcal{L}_{\mathcal{D}^{(t-1)}}^{\text{BE}}(Q^\sharp, R^\sharp) + \mathcal{L}_{\mathcal{D}^{(t-1)}}^{\text{RM}}(R^\sharp)}{\lambda} \\ &\leq f_1^{(t)}(s_1, \pi^{(t)}) - \frac{\mathcal{L}_{\mathcal{D}^{(t-1)}}^{\text{BE}}(f^{(t)}, R^{(t)}) + \mathcal{L}_{\mathcal{D}^{(t-1)}}^{\text{RM}}(R^{(t)})}{\lambda} + \frac{\mathcal{L}_{\mathcal{D}^{(t-1)}}^{\text{BE}}(Q^\sharp, R^\sharp) + \mathcal{L}_{\mathcal{D}^{(t-1)}}^{\text{RM}}(R^\sharp)}{\lambda}. \end{aligned}$$

Furthermore, by the standard performance difference lemma ([Kakade and Langford, 2002](#)), it holds that

$$f_1^{(t)}(s_1, \pi^{(t)}) - V^{\pi^{(t)}}(s_1) = \sum_{h=1}^H \mathbb{E}^{\pi^{(t)}} [f_h^{(t)}(s_h, a_h) - [\mathcal{T}_h^* f_{h+1}^{(t)}](s_h, a_h)]. \quad (16)$$

Based on (16), the existing approaches (with per-step reward feedback) bound the expectation of the Bellman error $e_h^{(t)}(s_h, a_h) := f_h^{(t)}(s_h, a_h) - [\mathcal{T}_h^* f_{h+1}^{(t)}](s_h, a_h)$ through various arguments (e.g., eluder argument ([Jin et al., 2021a](#)) and coverability argument ([Xie et al., 2022](#))). However, in outcome reward model, it is possible that $\mathbb{E}^{\pi^{(t)}} [f_h^{(t)}(s_h, a_h) - [\mathcal{T}_h^* f_{h+1}^{(t)}](s_h, a_h)]$ is large even when the sub-optimality of $\pi^{(t)}$ is small, because the outcome reward is invariant under shifting of the ground-truth reward function R^* .

Therefore, we consider the following refined decomposition:

$$\begin{aligned} f_1^{(t)}(s_1) - V^{\pi^{(t)}}(s_1) &= \sum_{h=1}^H \mathbb{E}^{\pi^{(t)}} [f_h^{(t)}(s_h, a_h) - [\mathcal{T}_{R^{(t)}} f_{h+1}^{(t)}](s_h, a_h)] \\ &\quad + \mathbb{E}^{\pi^{(t)}} \left[\sum_{h=1}^H R_h^{(t)}(s_h, a_h) - \sum_{h=1}^H R_h^*(s_h, a_h) \right]. \end{aligned} \quad (17)$$

936 **Coverability argument.** Following the coverability argument of [Xie et al. \(2022\)](#), we have the
 937 following upper bound on the expected Bellman errors.

938 **Proposition 12.** Denote $e_h^{(t)} := f_h^{(t)} - \mathcal{T}_h^* f_{h+1}$. Then, for each $h \in [H]$, it holds that

$$\sum_{t=1}^T \mathbb{E}^{\pi^{(t)}} |e_h^{(t)}(s_h, a_h)| \leq \sqrt{2C_{\text{cov}} \log \left(1 + \frac{C_{\text{cov}} T}{\kappa} \right) \cdot \left[2T\kappa + \sum_{1 \leq k < t \leq T} \mathbb{E}^{\pi^{(k)}} e_h^{(t)}(s_h, a_h)^2 \right]}.$$

939 Following the ideas of [Jia et al. \(2025\)](#), we prove the following upper bound on the reward errors.

940 **Proposition 13.** It holds that

$$\begin{aligned} & \sum_{t=1}^T \mathbb{E}^{\pi^{(t)}} |R^{(t)}(\tau) - R^*(\tau)| \\ & \leq \sqrt{8HC_{\text{cov}} \log \left(1 + \frac{C_{\text{cov}} T}{\kappa} \right) \cdot \left[HT\kappa + \sum_{1 \leq k < t \leq T} \sum_{h=1}^H \mathbb{E}^{\pi^{(k)} \circ_h \pi_{\text{ref}}} (R^{(t)}(\tau) - R^*(\tau))^2 \right]}. \end{aligned}$$

941 Based on the results above, we finalize the proof of [Theorem 1](#).

942 **Proof of Theorem 1.** By optimism and the decomposition (17), it holds that for $t \in [T]$,

$$\begin{aligned} V^*(s_1) - V^{\pi^{(t)}}(s_1) - \varepsilon_{\text{app}} & \leq \sum_{h=1}^H \mathbb{E}^{\pi^{(t)}} [e_h^{(t)}(s_h, a_h)] - \frac{\mathcal{L}_{\mathcal{D}^{(t-1)}}^{\text{BE}}(f^{(t)}, R^{(t)}) - \mathcal{L}_{\mathcal{D}^{(t-1)}}^{\text{BE}}(Q^\#, R^\#)}{\lambda} \\ & \quad + \mathbb{E}^{\pi^{(t)}} [R^{(t)}(\tau) - R^*(\tau)] - \frac{\mathcal{L}_{\mathcal{D}^{(t-1)}}^{\text{RM}}(R^{(t)}) - \mathcal{L}_{\mathcal{D}^{(t-1)}}^{\text{RM}}(R^\#)}{\lambda}. \end{aligned}$$

943 Then, under the success event of [Proposition 11](#), we have

$$\begin{aligned} & V^*(s_1) - V^{\pi^{(t)}}(s_1) - \varepsilon_{\text{app}} - \frac{2H\kappa}{\lambda} \\ & \leq \sum_{h=1}^H \left(\mathbb{E}^{\pi^{(t)}} [e_h^{(t)}(s_h, a_h)] - \frac{1}{2\lambda} \sum_{k < t} \mathbb{E}^{\pi^{(k)}} e_h^{(t)}(s_h, a_h)^2 \right) \\ & \quad + \mathbb{E}^{\pi^{(t)}} [R^{(t)}(\tau) - R^*(\tau)] - \frac{1}{2\lambda} \sum_{k < t} \sum_{h=1}^H \mathbb{E}^{\pi^{(k)} \circ_h \pi_{\text{ref}}} (R^{(t)}(\tau) - R^*(\tau))^2. \end{aligned} \tag{18}$$

944 Applying [Proposition 12](#) and Cauchy inequality gives for all $h \in [H]$,

$$\sum_{t=1}^T \mathbb{E}^{\pi^{(t)}} |e_h^{(t)}(s_h, a_h)| \leq \lambda C_{\text{cov}} \log \left(1 + \frac{C_{\text{cov}} T}{\kappa} \right) + \frac{1}{2\lambda} \left[2T\kappa + \sum_{1 \leq k < t \leq T} \mathbb{E}^{\pi^{(k)}} e_h^{(t)}(s_h, a_h)^2 \right],$$

945 and similarly, applying [Proposition 13](#) and Cauchy inequality gives

$$\begin{aligned} & \sum_{t=1}^T \mathbb{E}^{\pi^{(t)}} |R^{(t)}(\tau) - R^*(\tau)| \\ & \leq 4\lambda H C_{\text{cov}} \log \left(1 + \frac{C_{\text{cov}} T}{\kappa} \right) + \frac{1}{2\lambda} \left[HT\kappa + \sum_{1 \leq k < t \leq T} \sum_{h=1}^H \mathbb{E}^{\pi^{(k)} \circ_h \pi_{\text{ref}}} (R^{(t)}(\tau) - R^*(\tau))^2 \right]. \end{aligned}$$

946 Therefore, we take summation of (18) over $t \in [T]$, and combining the inequalities above gives

$$\sum_{t=1}^T \left(V^*(s_1) - V^{\pi^{(t)}}(s_1) \right) \leq T\varepsilon_{\text{app}} + \frac{4TH\kappa}{\lambda} + 5\lambda H C_{\text{cov}} \log \left(1 + \frac{C_{\text{cov}} T}{\kappa} \right).$$

947 This is the desired upper bound. \square

948 C.2 Proof of Proposition 12 and Proposition 13

949 The following proposition is an generalized version of the results in Xie et al. (2022, Appendix D).
 950 For proof, see e.g. Chen et al. (2024).

951 **Proposition 14** (Xie et al. (2022)). *Let $C \geq 1$ be a parameter. Suppose that $p^{(1)}, \dots, p^{(T)}$ is a*
 952 *sequence of distributions over \mathcal{X} , and there exists $\mu \in \Delta(\mathcal{X})$ such that $p^{(t)}(x)/\mu(x) \leq C$ for all*
 953 *$x \in \mathcal{X}, t \in [T]$. Then for any sequence $\psi^{(1)}, \dots, \psi^{(T)}$ of functions $\mathcal{X} \rightarrow [0, 1]$ and constant $B \geq 1$,*
 954 *it holds that*

$$\sum_{t=1}^T \mathbb{E}_{x \sim p^{(t)}} \psi^{(t)}(x) \leq \sqrt{2C \log \left(1 + \frac{CT}{B} \right) \left[2TB + \sum_{t=1}^T \sum_{k < t} \mathbb{E}_{x \sim p^{(k)}} \psi^{(t)}(x)^2 \right]}.$$

955 As a warm-up, we prove Proposition 12 by directly invoking Proposition 14.

956 **Proof of Proposition 12.** Fix a $h \in [H]$. To apply Proposition 14, we consider $\mathcal{X} = \mathcal{S} \times \mathcal{A}$, and
 957 define

$$p^{(t)} := \mathbb{P}^{\pi^{(t)}}((s_h, a_h) = \cdot) \in \Delta(\mathcal{S} \times \mathcal{A}), \quad \psi^{(t)} := |e_h^{(t)}| \in (\mathcal{S} \times \mathcal{A} \rightarrow [0, 1]).$$

958 By the definition of coverability (Definition 1), for $C = C_{\text{cov}}(\Pi)$, there exists $\mu \in \Delta(\mathcal{S} \times \mathcal{A})$
 959 such that $p^{(t)}(s, a)/\mu(s, a) \leq C$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$. Therefore, applying Proposition 14 with
 960 $B = \kappa \geq 1$ gives the desired upper bound. \square

961 Next, we proceed to prove Proposition 13. Our key proof technique is summarized in the following
 962 proposition, which is inspired by the (rather sophisticated) analysis of Jia et al. (2025).

963 **Proposition 15.** *Recall that for any $D = (D_h : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R})$, we denote*

$$D(\tau) = \sum_{h=1}^H D_h(s_h, a_h), \quad \forall \tau = (s_1, a_1, \dots, s_H, a_H) \in (\mathcal{S} \times \mathcal{A})^H.$$

964 *Fix a Markov policy π_{ref} , we denote $\bar{D}_1(s) = \bar{D}_{H+1}(s) = 0$, and*

$$\bar{D}_h(s) := \mathbb{E}^{\pi_{\text{ref}}} \left[\sum_{\ell=h}^H D_\ell(s_\ell, a_\ell) \middle| s_h = s \right], \quad \forall 1 < h \leq H, s \in \mathcal{S}.$$

965 *Then for any policy π , it holds that*

$$\sum_{h=1}^H \mathbb{E}^{\pi} (D_h(s_h, a_h) + \bar{D}_{h+1}(s_{h+1}) - \bar{D}_h(s_h))^2 \leq 4 \sum_{h=1}^H \mathbb{E}^{\pi \circ_h \pi_{\text{ref}}} D(\tau)^2.$$

966 The proof of Proposition 15 is deferred to the end of this subsection. With Proposition 15, we prove
 967 Proposition 13 as follows.

968 **Proof of Proposition 13.** To apply Proposition 15, for each $t \in [T]$, we consider

$$\Delta_h^{(t)}(s) := \mathbb{E}^{\pi_{\text{ref}}} \left[\sum_{\ell=h}^H R_\ell^{(t)}(s_\ell, a_\ell) - \sum_{\ell=h}^H R_\ell^*(s_\ell, a_\ell) \middle| s_h = s \right], \quad \forall h = 2, \dots, H, s \in \mathcal{S},$$

969 Then, by Proposition 15, it holds that for any policy π ,

$$\begin{aligned} & \sum_{h=1}^H \mathbb{E}^{\pi} (R_h^{(t)}(s_h, a_h) - R_h^*(s_h, a_h) + \Delta_{h+1}^{(t)}(s_{h+1}) - \Delta_h^{(t)}(s_h))^2 \\ & \leq 4 \sum_{h=1}^H \mathbb{E}^{\pi \circ_h \pi_{\text{ref}}} (R^{(t)}(\tau) - R^*(\tau))^2. \end{aligned} \tag{19}$$

970 Furthermore,

$$\begin{aligned} \mathbb{E}^\pi |R^{(t)}(\tau) - R^*(\tau)| &= \mathbb{E}^\pi \left| \sum_{h=1}^H [R_h^{(t)}(s_h, a_h) - R_h^*(s_h, a_h) + \Delta_{h+1}^{(t)}(s_{h+1}) - \Delta_h^{(t)}(s_h)] \right| \\ &\leq \sum_{h=1}^H \mathbb{E}^\pi |R_h^{(t)}(s_h, a_h) - R_h^*(s_h, a_h) + \Delta_{h+1}^{(t)}(s_{h+1}) - \Delta_h^{(t)}(s_h)|. \end{aligned} \quad (20)$$

971 Therefore, to apply [Proposition 14](#), we consider the space $\mathcal{X} = \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, and for each $h \in [H]$,
972 we define

$$\begin{aligned} p_h^{(t)} &:= \mathbb{P}^{\pi^{(t)}}((s_h, a_h, s_{h+1}) = \cdot) \in \Delta(\mathcal{S} \times \mathcal{A} \times \mathcal{S}), \\ \psi_h^{(t)}(s, a, s') &:= |R_h^{(t)}(s, a) - R_h^*(s, a) + \Delta_{h+1}^{(t)}(s') - \Delta_h^{(t)}(s)|. \end{aligned}$$

973 Note that for any h , we have $\psi_h^{(t)} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$. Further, let $\mu_h \in \Delta(\mathcal{S} \times \mathcal{A})$ be the
974 distribution such that $\|d_h^\pi / \mu_h\|_\infty \leq C_{\text{cov}}$ for any policy π . Then we can consider the distribution
975 $\bar{\mu}_h \in \Delta(\mathcal{S} \times \mathcal{A} \times \mathcal{S})$ given by $\bar{\mu}_h(s, a, s') = \mu_h(s, a) \mathbb{T}_h(s'|s, a)$. Then it holds that

$$\left\| \frac{p_h^{(t)}}{\bar{\mu}_h} \right\|_\infty = \sup_{s, a, s'} \frac{p_h^{(t)}(s, a, s')}{\bar{\mu}_h(s, a, s')} = \sup_{s, a, s'} \frac{d_h^\pi(s, a) \mathbb{T}_h(s'|s, a)}{\mu_h(s, a) \mathbb{T}_h(s'|s, a)} \leq C_{\text{cov}}.$$

976 Therefore, for $h \in [H]$, applying [Proposition 14](#) on the sequence $(p_h^{(1)}, \dots, p_h^{(T)})$ and $(\psi_h^{(1)}, \dots, \psi_h^{(T)})$
977 gives

$$\sum_{t=1}^T \mathbb{E}_{x \sim p_h^{(t)}} \psi_h^{(t)}(x) \leq \sqrt{2C_{\text{cov}} \log \left(1 + \frac{C_{\text{cov}} T}{\kappa} \right) \left[2T\kappa + \sum_{t=1}^T \sum_{k < t} \mathbb{E}_{x \sim p_h^{(k)}} \psi_h^{(t)}(x)^2 \right]}.$$

978 To conclude, we combine the inequalities above and bound

$$\begin{aligned} &\sum_{t=1}^T \mathbb{E}^{\pi^{(t)}} |R^{(t)}(\tau) - R^*(\tau)| \\ &\leq \sum_{t=1}^T \sum_{h=1}^H \mathbb{E}^{\pi^{(t)}} |R_h^{(t)}(s_h, a_h) - R_h^*(s_h, a_h) + \Delta_{h+1}^{(t)}(s_{h+1}) - \Delta_h^{(t)}(s_h)| \\ &= \sum_{t=1}^T \sum_{h=1}^H \mathbb{E}_{x \sim p_h^{(t)}} \psi_h^{(t)}(x) = \sum_{h=1}^H \sum_{t=1}^T \mathbb{E}_{x \sim p_h^{(t)}} \psi_h^{(t)}(x) \\ &\leq \sum_{h=1}^H \sqrt{2C_{\text{cov}} \log \left(1 + \frac{C_{\text{cov}} T}{\kappa} \right) \left[2T\kappa + \sum_{t=1}^T \sum_{k < t} \mathbb{E}_{x \sim p_h^{(k)}} \psi_h^{(t)}(x)^2 \right]} \\ &\leq \sqrt{2HC_{\text{cov}} \log \left(1 + \frac{C_{\text{cov}} T}{\kappa} \right) \left[2TH\kappa + \sum_{h=1}^H \sum_{t=1}^T \sum_{k < t} \mathbb{E}_{x \sim p_h^{(k)}} \psi_h^{(t)}(x)^2 \right]} \\ &\leq \sqrt{8HC_{\text{cov}} \log \left(1 + \frac{C_{\text{cov}} T}{\kappa} \right) \cdot \left[HT\kappa + \sum_{1 \leq k < t \leq T} \sum_{h=1}^H \mathbb{E}^{\pi^{(k)} \circ_h \pi_{\text{ref}}} (R^{(t)}(\tau) - R^*(\tau))^2 \right]}, \end{aligned}$$

979 where the first inequality follows from (20), the second last line follows from Cauchy inequality, and
980 last inequality follows from the definition of $(p_h^{(t)}, \psi_h^{(t)})$ and (19). \square

981 **Proof of Proposition 15.** For $\tau = (s_1, a_1, \dots, s_H, a_H) \in (\mathcal{S} \times \mathcal{A})^H$, we denote $\tau_h =$
982 $(s_1, a_1, \dots, s_h, a_h, s_{h+1})$ to be the prefix sequence of τ for each $h \in [H]$. Then, we note that

$$\mathbb{E}^{\pi \circ_h \pi_{\text{ref}}} [D(\tau) | \tau_h] = \sum_{\ell=1}^h D_\ell(s_\ell, a_\ell) + \mathbb{E}^{\pi \circ_h \pi_{\text{ref}}} \left[\sum_{\ell=h+1}^H D_\ell(s_\ell, a_\ell) \middle| \tau_h \right]$$

$$= \sum_{\ell=1}^h D_\ell(s_\ell, a_\ell) + \bar{D}_{h+1}(s_{h+1}),$$

983 because the policy $\pi \circ_h \pi_{\text{ref}}$ executes the Markov policy π_{ref} starting at the $(h+1)$ -th step. Therefore,
 984 it holds that

$$\begin{aligned} \mathbb{E}_{\tau_h \sim \pi} \left(\sum_{\ell=1}^h D_\ell(s_\ell, a_\ell) + \bar{D}_{h+1}(s_{h+1}) \right)^2 &= \mathbb{E}_{\tau_h \sim \pi \circ_h \pi_{\text{ref}}} (\mathbb{E}^{\pi \circ_h \pi_{\text{ref}}} [D(\tau) | \tau_h])^2 \\ &\leq \mathbb{E}_{\tau \sim \pi \circ_h \pi_{\text{ref}}} D(\tau)^2 = \mathbb{E}^{\pi \circ_h \pi_{\text{ref}}} D(\tau)^2, \end{aligned}$$

985 where the first equality follows from the fact that the policy $\pi \circ_h \pi_{\text{ref}}$ executes π for the first h steps.
 986 Therefore, for $h > 1$, it holds that

$$\begin{aligned} &\mathbb{E}^\pi (D_h(s_h, a_h) + \bar{D}_{h+1}(s_{h+1}) - \bar{D}_h(s_h))^2 \\ &= \mathbb{E}_{\tau_h \sim \pi} \left(\sum_{\ell=1}^h D_\ell(s_\ell, a_\ell) + \bar{D}_{h+1}(s_{h+1}) - \sum_{\ell=1}^{h-1} D_\ell(s_\ell, a_\ell) - \bar{D}_h(s_h) \right)^2 \\ &\leq 2\mathbb{E}_{\tau_h \sim \pi} \left(\sum_{\ell=1}^h D_\ell(s_\ell, a_\ell) + \bar{D}_{h+1}(s_{h+1}) \right)^2 + \mathbb{E}_{\tau_{h-1} \sim \pi} \left(\sum_{\ell=1}^{h-1} D_\ell(s_\ell, a_\ell) + \bar{D}_h(s_h) \right)^2 \\ &\leq 2\mathbb{E}^{\pi \circ_h \pi_{\text{ref}}} D(\tau)^2 + \mathbb{E}^{\pi \circ_{h-1} \pi_{\text{ref}}} D(\tau)^2. \end{aligned}$$

987 For $h = 1$, because $\bar{D}_1(s) = 0$, we already have

$$\mathbb{E}^\pi (D_1(s_1, a_1) + \bar{D}_2(s_2) - \bar{D}_1(s_1))^2 = \mathbb{E}^\pi (D_1(s_1, a_1) + \bar{D}_2(s_2))^2 \leq \mathbb{E}^{\pi \circ_1 \pi_{\text{ref}}} D(\tau)^2.$$

988 Taking summation over $h \in [H]$ completes the proof. \square

989 C.3 Proof of Proposition 11

990 We prove Proposition 11 in Lemma 16 and Lemma 17 separately. Recall again that under Assumption
 991 1, the function $Q^\sharp \in \mathcal{F}$, $R^\sharp \in \mathcal{R}$ satisfy

$$\max_{h \in [H]} \|Q_h^\sharp - Q_h^*\|_\infty \leq \varepsilon_{\text{app}}, \quad \max_{h \in [H]} \|R_h^\sharp - R_h^*\|_\infty \leq \varepsilon_{\text{app}}.$$

992 **Lemma 16.** Under Assumption 1, with probability at least $1 - \delta$, for any $t \in [T]$, for all $R \in \mathcal{R}$, it
 993 holds that

$$\begin{aligned} \frac{1}{2} \sum_{k \leq t} \sum_{h=1}^H \mathbb{E}^{\pi^{(k)} \circ_h \pi_{\text{ref}}} (R(\tau) - R^*(\tau))^2 &\leq \mathcal{L}_{\mathcal{D}^{(t)}}^{\text{RM}}(R) - \mathcal{L}_{\mathcal{D}^{(t)}}^{\text{RM}}(R^\sharp) \\ &\quad + 15H \log N(\rho) + 15(2/\delta) + 2TH^3 \varepsilon_{\text{app}}^2 + 4TH^2 \rho. \end{aligned}$$

994 **Proof of Lemma 16.** To apply Proposition 7, we consider the whole history

$$\{(\tau^{(t,h)}, r^{(t,h)})\}_{t \in [T], h \in [H]}.$$

995 generated by executing Algorithm 1, and recall that $\mathcal{D}^{(t-1)} = \{(\tau^{(k,h)}, r^{(k,h)})\}_{k < t, h \in [H]}$ is the history
 996 up to the t -th iteration. Note that $(\tau^{(t,1)}, r^{(t,1)}), \dots, (\tau^{(t,H)}, r^{(t,H)})$ are pairwise independent given
 997 $\mathcal{D}^{(t-1)}$, with

$$\tau^{(t,h)} \sim \pi^{(t)} \circ_h \pi_{\text{ref}}, \quad \mathbb{E}[r^{(t,h)} | \mathcal{D}^{(t-1)}, \tau^{(t,h)}] = R^*(\tau^{(t,h)}).$$

998 Also note that $r \in [0, 1]$ almost surely, and we regard $\mathcal{R} \subseteq ((\mathcal{S} \times \mathcal{A})^H \rightarrow [0, 1])$, and $R^\sharp \in \mathcal{R}$
 999 satisfies $|R^\sharp(\tau) - R^*(\tau)| \leq H\varepsilon_{\text{app}}$ for all $\tau \in (\mathcal{S} \times \mathcal{A})^H$.

1000 Therefore, applying Proposition 7 on the function class \mathcal{R} and the sequence
 1001 $\{(\tau^{(t,h)}, r^{(t,h)})\}_{t \in [T], h \in [0, H]}$ gives that with probability at least $1 - \delta$, for all $R \in \mathcal{R}$, $t \in [T]$, it holds
 1002 that

$$\frac{1}{2} \sum_{k=1}^t \sum_{h=1}^H \mathbb{E}^{\pi^{(k)} \circ_h \pi_{\text{ref}}} (R(\tau) - R^*(\tau))^2 = \frac{1}{2} \sum_{k=1}^t \sum_{h=1}^H \mathbb{E} \left[(R(\tau^{(k,h)}) - R^*(\tau^{(k,h)}))^2 \middle| \mathcal{D}^{(k-1)} \right]$$

$$\leq \mathcal{L}_{\mathcal{D}^{(t)}}^{\text{RM}}(R) - \mathcal{L}_{\mathcal{D}^{(t)}}^{\text{RM}}(R^\sharp) + 15 \log(2N(\mathcal{R}, H\rho)/\delta) + 2TH^3\varepsilon_{\text{app}}^2 + 4TH^2\rho.$$

1003 Finally, we note that $\log N(\mathcal{R}, H\rho) \leq \sum_{h=1}^H \log N(\mathcal{R}_h, \rho) \leq H \log N(\rho)$. This gives the desired
1004 upper bound. \square

1005 Similarly, we prove [Proposition 11](#) (2) as follows, following [Jin et al. \(2021a\)](#).

1006 **Lemma 17.** Fix $h \in [H]$ and $\delta \in (0, 1)$, $\rho \geq 0$. Suppose that [Assumption 1](#) and [Assumption 2](#) holds.
1007 Then with probability at least $1 - \delta$, the following holds:

1008 (1) For each $t \in [T]$,

$$\mathcal{E}_{\mathcal{D}^{(t)},h}(Q_h^\sharp, Q_{h+1}^\sharp; R_h^\sharp) - \inf_{g_h \in \mathcal{G}_h} \mathcal{E}_{\mathcal{D}^{(t)},h}(g_h, Q_{h+1}^\sharp; R_h^\sharp) \leq O(TH\varepsilon_{\text{app}}^2 + TH\rho + \log(N(\rho)/\delta)).$$

1009 (2) For each $t \in [T]$, for all $f_h \in \mathcal{F}_h$, $f_{h+1} \in \mathcal{F}_{h+1}$, and $R_h \in \mathcal{R}_h$,

$$\begin{aligned} & \frac{1}{2} \sum_{k=1}^t \mathbb{E}^{\pi^{(k)}} (f_h(s_h, a_h) - [\mathcal{T}_{R,h} f_{h+1}](s_h, a_h))^2 \\ & \leq \mathcal{E}_{\mathcal{D}^{(t)},h}(f_h, f_{h+1}; R_h) - \inf_{g_h \in \mathcal{G}_h} \mathcal{E}_{\mathcal{D}^{(t)},h}(g_h, f_{h+1}; R_h) + O(TH\varepsilon_{\text{app}}^2 + TH\rho + \log(N(\rho)/\delta)), \end{aligned}$$

1010 where we use $O(\cdot)$ to hide absolute constant for simplicity.

1011 **Proof of Lemma 17.** Fix $h \in [H]$ and denote $N := N(\rho)$. We let \mathcal{F}'_{h+1} be a minimal ρ -covering
1012 of \mathcal{F}_{h+1} , and let \mathcal{R}'_h be a minimal ρ -covering of \mathcal{R}_h . By definition, $|\mathcal{F}'_{h+1}| \leq N$, $|\mathcal{R}'_h| \leq N$.

1013 In the following, we adopt the notation of the proof of [Lemma 16](#). Recall that conditional on $\mathcal{D}^{(t-1)}$,

$$\tau^{(t,\ell)} = (s_1^{(t,\ell)}, a_1^{(t,\ell)}, \dots, s_H^{(t,\ell)}, a_H^{(t,\ell)}) \sim \pi^{(t)} \circ_{\ell} \pi_{\text{ref}},$$

1014 and $\tau^{(t,1)}, \dots, \tau^{(t,H)}$ are independent conditional on $\mathcal{D}^{(t-1)}$. For simplicity, we denote $x^{(t,\ell)} :=$
1015 $(s_h^{(t,\ell)}, a_h^{(t,\ell)})$.

1016 Fix $f_{h+1} \in \mathcal{F}'_{h+1} \cup \{Q_{h+1}^\sharp\}$ and $R_h \in \mathcal{R}'_h \cup \{R_h^\sharp\}$, we consider

$$y^{(t,\ell)} := f_{h+1}(s_{h+1}^{(t,\ell)}) + R_h(s_h^{(t,\ell)}, a_h^{(t,\ell)}).$$

1017 and it holds that

$$\mathbb{E}[y^{(t,\ell)} | \mathcal{D}^{(t-1)}, x^{(t,\ell)}] = [\mathcal{T}_{R,h} f_{h+1}](x^{(t,\ell)}).$$

1018 Then, for any $g_h \in \mathcal{G}_h$, it holds that

$$\mathcal{E}_{\mathcal{D}^{(t)},h}(g_h, f_{h+1}; R_h) = \sum_{k=1}^t \sum_{\ell=1}^H (g_h(x^{(k,\ell)}) - y^{(k,\ell)})^2,$$

1019 and we also have

$$\sum_{k=1}^t \sum_{\ell=1}^H \mathbb{E}^{\pi^{(k)} \circ_{\ell} \pi_{\text{ref}}} (g_h(s_h, a_h) - [\mathcal{T}_{R,h} f_{h+1}](s_h, a_h))^2 = \sum_{k=1}^t \sum_{\ell=1}^H \mathbb{E} \left[(g_h(x^{(k,\ell)}) - [\mathcal{T}_{R,h} f_{h+1}](x^{(k,\ell)}))^2 \middle| \mathcal{D}^{(k-1)} \right]$$

1020 Then, applying [Proposition 7](#) with the function class $\mathcal{H} = \mathcal{G}_h$ yields that with probability at least
1021 $1 - \frac{\delta}{2N}$, the following holds:

1022 (a) For each $t \in [T]$, for any $g_h \in \mathcal{G}_h$,

$$\begin{aligned} & \frac{1}{2} \sum_{k=1}^t \mathbb{E}^{\pi^{(k)}} (g_h(s_h, a_h) - [\mathcal{T}_{R,h} f_{h+1}](s_h, a_h))^2 \\ & \leq \mathcal{E}_{\mathcal{D}^{(t)},h}(g_h, f_{h+1}; R_h) - \inf_{g'_h \in \mathcal{G}_h} \mathcal{E}_{\mathcal{D}^{(t)},h}(g'_h, f_{h+1}; R_h) + O(\log(N/\delta) + TH\varepsilon_{\text{app}}^2 + TH\rho). \end{aligned}$$

1023 (b) When $f_{h+1} = Q_{h+1}^\sharp$ and $R_h = R_h^\sharp$, the function $Q_h^\sharp \in \mathcal{F}_h \subseteq \mathcal{G}_h$ satisfies the inequality

1024 $\|Q_h^\sharp - \mathcal{T}_{R^\sharp,h} Q_{h+1}^\sharp\|_\infty \leq 3\varepsilon_{\text{app}}$, and thus

$$\mathcal{E}_{\mathcal{D}^{(t)},h}(Q_h^\sharp, Q_{h+1}^\sharp; R_h^\sharp) \leq \inf_{g_h \in \mathcal{G}_h} \mathcal{E}_{\mathcal{D}^{(t)},h}(g_h, Q_{h+1}^\sharp; R_h^\sharp) + O(\log(N/\delta) + TH\varepsilon_{\text{app}}^2 + TH\rho).$$

Therefore, taking the union bound, we know that the inequalities (a) and (b) above hold simultaneously with probability at least $1 - \delta$ for all $f_{h+1} \in \mathcal{F}'_{h+1} \cup \{Q^\sharp_{h+1}\}$ and $R_h \in \mathcal{R}'_h \cup \{R^\sharp_h\}$. In particular, we have completed the proof of (1).

To prove (2), we only need to note that $\mathcal{G}_h \subseteq \mathcal{F}_h$, and for any $f_{h+1} \in \mathcal{F}_{h+1}$, $R_h \in \mathcal{R}_h$, there exists $f'_{h+1} \in \mathcal{F}'_{h+1}$, $R'_h \in \mathcal{R}'_h$ such that $\|f_{h+1} - f'_{h+1}\|_\infty \leq \rho$, $\|R_h - R'_h\|_\infty \leq \rho$. Therefore, by the standard covering argument and the fact that $\pi^{(k)} \circ_H \pi_{\text{ref}} = \pi^{(k)}$, we have also shown (2). \square

D Proof of Theorem 2

In this section, we provide the proof of Theorem 2, which is a direct adaption of the proof of Theorem 1 in Appendix C. We first present a more detailed statement of the upper bound (with any parameter $\lambda > 0$).

Theorem 18. Suppose that Assumption 1 holds. Then with probability at least $1 - \delta$, Algorithm 2 achieves

$$\frac{1}{T} \sum_{t=1}^T \left(V^\star(s_1^{(t)}) - V^{\pi^{(t)}}(s_1^{(t)}) \right) \leq \varepsilon_{\text{app}} + O(1) \cdot \left[\frac{H^3 \log(N_{\mathcal{F},T}/\delta) + T\varepsilon_{\text{app}}^2}{\lambda} + \frac{\lambda H C'_{\text{cov}}(\Pi)}{T} \right],$$

We also work with a slightly relaxed version of Assumption 3.

Assumption 4. Under any policy π , for each $h \in [H]$, it holds that almost surely

$$Q_h^\star(s_h, a_h) = R_h^\star(s_h, a_h) + V_{h+1}^\star(s_{h+1}).$$

Further, to simplify the notation, for each $f \in \mathcal{F}$, we recall that the induced reward model R^f is defined as $R_1^f(s, a) := f_1(s, a)$, $R_h^f(s, a) = f_h(s, a) - f_h(s)$, which implies

$$R^f(\tau) = \sum_{h=1}^H f_h(s_h, a_h) - f_{h+1}(s_{h+1}).$$

Uniform convergence. For each $t \in [T]$, we define $\mathcal{D}^{(t-1)} := \{(\tau^{(k)}, r^{(k)})\}_{k < t}$ be the data collected before t th iteration. We also recall that by definition (8), we have

$$\mathcal{L}_{\mathcal{D}^{(t)}}^{\text{BR}}(f) := \sum_{k=1}^t \left(\sum_{h=1}^H [f_h(s_h^{(k)}, a_h^{(k)}) - f_{h+1}(s_{h+1}^{(k)})] - r^{(k)} \right)^2 = \sum_{k=1}^t (R^f(\tau^{(k)}) - r^{(k)})^2.$$

Therefore, a direct instantiation of Proposition 7 on the class $\mathcal{R} := \{R^f : f \in \mathcal{F}\}$ yields the following proposition.

Proposition 19. Let $\delta \in (0, 1)$, $\rho \geq 0$. Suppose that Assumption 1 and Assumption 4 holds. Then with probability at least $1 - \delta$, for all $t \in [T]$, $f \in \mathcal{F}$, it holds that

$$\frac{1}{2} \sum_{k=1}^t \mathbb{E}^{\pi^{(k)}} (R^f(\tau) - R^\star(\tau))^2 \leq \mathcal{L}_{\mathcal{D}^{(t)}}^{\text{BR}}(f) - \mathcal{L}_{\mathcal{D}^{(t)}}^{\text{BE}}(Q^\sharp) + \kappa,$$

where

$$\kappa = C(H^3 \log(N_{\mathcal{F}}(\alpha)/\delta) + TH\alpha + T\varepsilon_{\text{app}}^2),$$

$C > 0$ is an absolute constant, and we denote $N_{\mathcal{F}}(\alpha) := \max_{h \in [H]} N(\mathcal{F}_h, \alpha)$ for any $\alpha \geq 0$.

Performance difference decomposition. In this setting, we can rewrite the decomposition (16) as

$$\begin{aligned} f_1^{(t)}(s_1) - V^{\pi^{(t)}}(s_1) &= \sum_{h=1}^H \mathbb{E}^{\pi^{(t)}} [f_h^{(t)}(s_h, a_h) - R_h^\star(s_h, a_h) - f_{h+1}^{(t)}(s_{h+1})] \\ &= \mathbb{E}^{\pi^{(t)}} \left[\sum_{h=1}^H [f_h^{(t)}(s_h, a_h) - f_{h+1}^{(t)}(s_{h+1})] - \sum_{h=1}^H R_h^\star(s_h, a_h) \right] \\ &= \mathbb{E}^{\pi^{(t)}} [R^{(t)}(\tau) - R^\star(\tau)], \end{aligned} \tag{21}$$

where we denote $R^{(t)} := R^{f^{(t)}}$, which is a reward model given by

$$R_1^{(t)}(s, a) := f_1^{(t)}(s, a), \quad R_h^{(t)}(s, a) = f_h^{(t)}(s, a) - f_h^{(t)}(s).$$

1051 **Optimism.** Similar to [Appendix C.1](#), we use the fact that from (9),

$$f^{(t)} = \max_{f \in \mathcal{F}} \lambda f_1(s_1^{(t)}) - \mathcal{L}_{\mathcal{D}^{(t-1)}}^{\text{BR}}(f),$$

1052 and hence

$$\lambda f_1^{(t)}(s_1^{(t)}) - \mathcal{L}_{\mathcal{D}^{(t-1)}}^{\text{BR}}(f^{(t)}) \geq \lambda V_1^\#(s_1^{(t)}) - \mathcal{L}_{\mathcal{D}^{(t-1)}}^{\text{BR}}(f^{(t)}).$$

1053 Using $|V_1^\#(s_1^{(t)}) - V_1^*(s_1^{(t)})| \leq \varepsilon_{\text{app}}$, (21) and [Proposition 19](#), we now deduce that

$$\begin{aligned} V^*(s_1^{(t)}) - V^{\pi^{(t)}}(s_1^{(t)}) &\leq \varepsilon_{\text{app}} + \mathbb{E}^{\pi^{(t)}}[R^{(t)}(\tau) - R^*(\tau)] - \frac{\mathcal{L}_{\mathcal{D}^{(t-1)}}^{\text{BR}}(f^{(t)}) - \mathcal{L}_{\mathcal{D}^{(t-1)}}^{\text{BE}}(Q^\#)}{\lambda} \\ &\leq \varepsilon_{\text{app}} + \frac{\kappa}{\lambda} + \mathbb{E}^{\pi^{(t)}}[R^{(t)}(\tau) - R^*(\tau)] - \frac{1}{2\lambda} \sum_{k=1}^{t-1} \mathbb{E}^{\pi^{(k)}}(R^{(t)}(\tau) - R^*(\tau))^2. \end{aligned} \quad (22)$$

1054 Therefore, it remains to prove an analogue to [Proposition 13](#).

1055 **Coverability argument.** We strength [Proposition 13](#) using the deterministic nature of the underlying
1056 MDP. For each $s \in \mathcal{S}$ and $h \in [H]$, we define

$$\mathcal{S}_h(s; \Pi) := \{(s', a) : \exists \pi \in \Pi, \text{ under } \pi \text{ and } s_1 = s, \text{ it holds that } s_h = s', a_h = a\},$$

1057 and $N_h(s; \Pi) := |\mathcal{S}_h(s; \Pi)|$.

1058 **Proposition 20.** Let $B \geq 1$. For any initial state $s_1 \in \mathcal{S}$, any sequence of reward functions
1059 $R^{(1)}, \dots, R^{(T)}$ and any sequence of policies $\pi^{(1)}, \dots, \pi^{(T)}$, it holds that

$$\begin{aligned} &\sum_{t=1}^T \mathbb{E}^{\pi^{(t)}}[R^{(t)}(\tau) - R^*(\tau) | s_1] \\ &\leq \sqrt{2N(s_1) \log\left(1 + \frac{4TH}{B}\right) \cdot \left[2TB + \sum_{1 \leq k < t \leq T} \mathbb{E}^{\pi^{(k)}}[(R^{(t)}(\tau) - R^*(\tau))^2 | s_1]\right]}, \end{aligned}$$

1060 where $N(s_1) := \sum_{h=1}^H N_h(s_1; \Pi)$, and the conditional distribution $\mathbb{E}^{\pi^{(t)}}[\cdot | s_1]$ is taken over the
1061 expectation of τ generated by executing policy π starting with the initial state s_1 .

1062 The proof of [Proposition 20](#) is deferred to the end of this section.

1063 **Finalizing the proof.** With the above preparation, we now finalize the proof of [Theorem 2](#). Taking
1064 summation of (22) over $t = 1, 2, \dots, T$, we have

$$\begin{aligned} &\sum_{t=1}^T V^*(s_1^{(t)}) - V^{\pi^{(t)}}(s_1^{(t)}) \\ &\leq T\varepsilon_{\text{app}} + \frac{T\kappa}{\lambda} + \sum_{t=1}^T \mathbb{E}^{\pi^{(t)}}[R^{(t)}(\tau) - R^*(\tau)] - \frac{1}{2\lambda} \sum_{1 \leq k < t \leq T} \mathbb{E}^{\pi^{(k)}}(R^{(t)}(\tau) - R^*(\tau))^2 \\ &= T\varepsilon_{\text{app}} + \frac{T\kappa}{\lambda} + \mathbb{E}_{s_1 \sim \rho} \left[\sum_{t=1}^T \mathbb{E}^{\pi^{(t)}}[R^{(t)}(\tau) - R^*(\tau) | s_1] - \frac{1}{2\lambda} \sum_{1 \leq k < t \leq T} \mathbb{E}^{\pi^{(k)}}[(R^{(t)}(\tau) - R^*(\tau))^2 | s_1] \right] \\ &\leq T\varepsilon_{\text{app}} + \frac{2T\kappa}{\lambda} + \mathbb{E}_{s_1 \sim \rho} \left[N(s_1) \lambda \log\left(1 + \frac{TH}{\kappa}\right) \right], \end{aligned}$$

1065 where the last inequality follows from [Proposition 20](#) and Cauchy inequality. This is the desired
1066 upper bound. \square

1067 **Proof of Proposition 20.** In the following proof, we assume $s_1 \in \mathcal{S}$ is fixed. Consider

$$\mathcal{I} := \{(h, s, a) : h \in [H], (s, a) \in \mathcal{S}_h(s_1; \Pi)\} \subseteq [H] \times \mathcal{S} \times \mathcal{A}.$$

1068 Note that $|\mathcal{I}| = \sum_{h=1}^H N_h(s_1; \Pi) = N(s_1)$. By definition, for any policy π , there is a unique
 1069 pair $(s_h^\pi, a_h^\pi) \in \mathcal{S}_h(s_1; \Pi)$, such that under π and starting from s_1 , we have $s_h = s_h^\pi, a_h = a_h^\pi$
 1070 deterministically.

1071 For each $t \in [T]$, we consider the following vectors indexed by \mathcal{I} :

$$\begin{aligned} \psi^{(t)} &:= [R_h^{(t)}(s, a) - R_h^*(s, a)]_{(h,s,a) \in \mathcal{I}} \in \mathbb{R}^{\mathcal{I}}, \\ \phi^{(t)} &:= [\mathbb{P}^{\pi^{(t)}}(s_h = s, a_h = a | s_1)]_{(h,s,a) \in \mathcal{I}} = \sum_{h=1}^H e_{(h, s_h^{\pi^{(t)}}, a_h^{\pi^{(t)}})} \in \mathbb{R}^{\mathcal{I}}. \end{aligned}$$

1072 With this definition, it holds that for any $k, t \in [T]$,

$$\mathbb{E}^{\pi^{(k)}} [R^{(t)}(\tau) - R^*(\tau) | s_1] = \sum_{h=1}^H [R^{(t)}(s_h^{\pi^{(k)}}, a_h^{\pi^{(k)}}) - R^*(s_h^{\pi^{(k)}}, a_h^{\pi^{(k)}})] = \langle \phi^{(k)}, \psi^{(t)} \rangle.$$

1073 Therefore, we apply the elliptical potential argument (Lattimore and Szepesvári, 2020). Let $V_t :=$
 1074 $\sum_{k < t} \phi^{(k)} (\phi^{(k)})^\top + BI$. Then it holds that

$$\begin{aligned} \sum_{t=1}^T |\langle \phi^{(t)}, \psi^{(t)} \rangle| &\leq \sum_{t=1}^T \min\{\|\phi^{(t)}\|_{V_t^{-1}}, 1\} \cdot \max\{\|\psi^{(t)}\|_{V_t}, 1\} \\ &\leq \sqrt{\sum_{t=1}^T \min\{\|\phi^{(t)}\|_{V_t^{-1}}^2, 1\}} \cdot \sqrt{\sum_{t=1}^T \max\{\|\psi^{(t)}\|_{V_t}^2, 1\}}. \end{aligned}$$

1075 Note that

$$\begin{aligned} \sum_{t=1}^T \max\{\|\psi^{(t)}\|_{V_t}^2, 1\} &\leq \sum_{t=1}^T \left[1 + B \|\psi^{(t)}\|^2 + \sum_{k=1}^{t-1} \langle \phi^{(k)}, \psi^{(t)} \rangle^2 \right] \\ &\leq T(1 + 4B|\mathcal{I}|) + \sum_{1 \leq k < t \leq T} \mathbb{E}^{\pi^{(k)}} \left[(R^{(t)}(\tau) - R^*(\tau))^2 | s_1 \right], \end{aligned}$$

1076 and by Lattimore and Szepesvári (2020), we have

$$\sum_{t=1}^T \min\{\|\phi^{(t)}\|_{V_t^{-1}}^2, 1\} \leq 2|\mathcal{I}| \log \left(1 + \frac{TH}{|\mathcal{I}|B} \right).$$

1077 Combining the inequalities above and rescale $B \leftarrow \frac{B}{4|\mathcal{I}|}$ completes the proof. \square

1078 E Proofs from Section 4

1079 We present the full description of our algorithm or preference-based RL as follows.

1080 E.1 Proof of Theorem 3

1081 For each $t \in [T]$, we write $\mathcal{D}^{(t)}$ to be the dataset maintained by Algorithm 1 at the end of the t th
 1082 iteration, i.e.,

$$\mathcal{D}^{(t)} = \{(\tau^{(k,h,+)}, \tau^{(k,h,-)}, y^{(k,h)})\}_{k \leq t, h \in [H]}.$$

1083 Note that for each $t \in [T], h \in [H]$, we have $\pi^{(t,h,-)} = \pi_{\text{ref}}$. Therefore, for each $R \in \mathcal{R}$, we define
 1084 $V_R^{\text{ref}} := \mathbb{E}^{\pi_{\text{ref}}} [R(\tau)]$ and recall that

$$\widehat{V}_{\mathcal{D}, R}^{\text{ref}} := \frac{1}{|\mathcal{D}|} \sum_{(\tau^+, \tau^-, y) \in \mathcal{D}} R(\tau^-).$$

1085 The following lemma follows from the standard uniform convergence rate with Hoeffding's inequality
 1086 and the union bound.

Algorithm 3 Outcome-Based Exploration for Preference-based RL

input: Function class \mathcal{F} , parameter $\lambda > 0$, reference policy π_{ref} .

initialize: $\mathcal{D} \leftarrow \emptyset$.

1: **for** $t = 1, 2, \dots, T$ **do**

2: Compute the optimistic estimates through (12):

$$(f^{(t)}, R^{(t)}) = \max_{f \in \mathcal{F}, R \in \mathcal{R}} \lambda \left[f_1(s_1) - \widehat{V}_{\mathcal{D}, R}^{\text{ref}} \right] - \mathcal{L}_{\mathcal{D}}^{\text{BE}}(f; R) - \mathcal{L}_{\mathcal{D}}^{\text{PbRM}}(R),$$

3: Select policy $\pi^{(t)} \leftarrow \pi_{f^{(t)}}$.

4: **for** $h = 1, 2, \dots, H$ **do**

5: Execute $\pi^{(t)} \circ_h \pi_{\text{ref}}$ for two episode and obtain two trajectories $(\tau^{(t, h, +)}, \tau^{(t, h, -)})$ and preference feedback $y^{(t, h)}$.

6:

7: Update dataset: $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\tau^{(t, h, +)}, \tau^{(t, h, -)}, y^{(t, h)})\}$.

8: **end for**

9: **end for**

10: Output $\widehat{\pi} = \text{Unif}(\pi^{(1:T)})$.

1087 **Lemma 21.** Let $\delta \in (0, 1)$, $\rho \geq 0$. Suppose that [Assumption 1](#) and [Assumption 2](#) holds. Then with
1088 probability at least $1 - \delta$, for all $t \in [T]$, $R \in \mathcal{R}$, it holds that

$$\left| \widehat{V}_{\mathcal{D}^{(t)}, R}^{\text{ref}} - V_R^{\text{ref}} \right| \leq \sqrt{\frac{\log(2TN(\rho)/\delta)}{t}} + H\rho.$$

1089 We summarize the uniform concentration results for the loss $\mathcal{L}_{\mathcal{D}^{(t)}}^{\text{BE}}$ and $\mathcal{L}_{\mathcal{D}^{(t)}}^{\text{PbRM}}$ as follows. The proof
1090 is analogous to [Proposition 11](#) and is provided in [Appendix E.2](#).

1091 **Proposition 22.** Let $\delta \in (0, 1)$, $\rho \geq 0$. Suppose that [Assumption 1](#) and [Assumption 2](#) holds. Then
1092 with probability at least $1 - \delta$, for all $t \in [T]$, $f \in \mathcal{F}$, $R \in \mathcal{R}$, it holds that

$$\left| \widehat{V}_{\mathcal{D}^{(t)}, R}^{\text{ref}} - V_R^{\text{ref}} \right| \leq \sqrt{\frac{\kappa}{tH}},$$

$$\sum_{k \leq t} \sum_{h=1}^H \mathbb{E}^{\pi^{(k, h, +)}, \pi^{(k, h, -)}} \left([R(\tau^+) - R(\tau^-)] - [R^*(\tau^+) - R^*(\tau^-)] \right)^2 \leq C_\beta [\mathcal{L}_{\mathcal{D}^{(t)}}^{\text{PbRM}}(R) - \mathcal{L}_{\mathcal{D}^{(t)}}^{\text{PbRM}}(R^\#)] + C_\beta H\kappa,$$

$$\sum_{k \leq t} \sum_{h=1}^H \mathbb{E}^{\pi^{(k)}} (f_h(s_h, a_h) - [\mathcal{T}_{R, h} f_{h+1}](s_h, a_h))^2 \leq 2[\mathcal{L}_{\mathcal{D}^{(t)}}^{\text{BE}}(f; R) - \mathcal{L}_{\mathcal{D}^{(t)}}^{\text{BE}}(Q^\#; R^\#)] + H\kappa,$$

1093 where $C_\beta = \frac{4e^{2\beta}}{\beta^2}$,

$$\kappa = C(\log N(\rho) + \log(TH/\delta) + TH^2(\beta + 1)(\varepsilon_{\text{app}}^2 + \rho)),$$

1094 and $C > 0$ is an absolute constant.

1095 In the following, we condition on the success event of [Proposition 22](#). Note that $\pi^{(t, h, -)} \equiv \pi_{\text{ref}}$, and
1096 hence [Proposition 22](#) implies that for all $R \in \mathcal{R}$, $t \in [T]$,

$$\sum_{k \leq t} \sum_{h=1}^H \mathbb{E}^{\pi^{(k)} \circ_h \pi_{\text{ref}}} ([R(\tau) - R^*(\tau)] - [V_R^{\text{ref}} - V_{R^*}^{\text{ref}}])^2 \leq C_\beta [\mathcal{L}_{\mathcal{D}^{(t)}}^{\text{PbRM}}(R) - \mathcal{L}_{\mathcal{D}^{(t)}}^{\text{PbRM}}(R^\#)] + C_\beta H\kappa.$$

1097 Therefore, for any reward function R , we define \tilde{R} as $\tilde{R}_1(s, a) = R_1(s, a) - V_R^{\text{ref}}$ and $\tilde{R}_h(s, a) =$
1098 $R_h(s, a)$ for $h > 1$. Then it is clear that $\tilde{R}(\tau) = R(\tau) - V_R^{\text{ref}}$, and for all $R \in \mathcal{R}$, $t \in [T]$, we have

$$\sum_{k \leq t} \sum_{h=1}^H \mathbb{E}^{\pi^{(k)} \circ_h \pi_{\text{ref}}} \left(\tilde{R}(\tau) - \tilde{R}^*(\tau) \right)^2 \leq C_\beta [\mathcal{L}_{\mathcal{D}^{(t)}}^{\text{PbRM}}(R) - \mathcal{L}_{\mathcal{D}^{(t)}}^{\text{PbRM}}(R^\#)] + C_\beta H\kappa. \quad (23)$$

1099 **Performance difference decomposition.** In this setting, we re-write (17) as follows:

$$\begin{aligned}
f_1^{(t)}(s_1) - V^{\pi^{(t)}}(s_1) &= \sum_{h=1}^H \mathbb{E}^{\pi^{(t)}} [f_h^{(t)}(s_h, a_h) - [\mathcal{T}_{R^{(t)}} f_{h+1}^{(t)}](s_h, a_h)] \\
&\quad + \mathbb{E}^{\pi^{(t)}} \left[\sum_{h=1}^H R_h^{(t)}(s_h, a_h) - \sum_{h=1}^H R_h^*(s_h, a_h) \right] \\
&= \sum_{h=1}^H \mathbb{E}^{\pi^{(t)}} e_h^{(t)}(s_h, a_h) + \mathbb{E}^{\pi^{(t)}} [\tilde{R}^{(t)}(\tau) - \tilde{R}^*(\tau)] + V_{R^{(t)}}^{\text{ref}} - V_{R^*}^{\text{ref}},
\end{aligned}$$

1100 where we recall that we denote $e_h^{(t)} := f_h^{(t)} - \mathcal{T}_{R^{(t)}} f_{h+1}^{(t)}$. Therefore, we re-organize the equality as

$$[f_1^{(t)}(s_1) - V_{R^{(t)}}^{\text{ref}}] - [V^{\pi^{(t)}}(s_1) - V_{R^*}^{\text{ref}}] = \mathbb{E}^{\pi^{(t)}} [\tilde{R}^{(t)}(\tau) - \tilde{R}^*(\tau)] + \sum_{h=1}^H \mathbb{E}^{\pi^{(t)}} e_h^{(t)}(s_h, a_h). \quad (24)$$

1101 With the above preparation, we present the proof of [Theorem 3](#), which closely follows the proof of
1102 [Theorem 1](#) in [Appendix C.1](#).

1103 **Proof of Theorem 3.** By definition, for each $t \in [T]$,

$$(f^{(t)}, R^{(t)}) = \max_{f \in \mathcal{F}, R \in \mathcal{R}} \lambda [f_1(s_1) - \hat{V}_{\mathcal{D}^{(t-1)}, R}^{\text{ref}}] - \mathcal{L}_{\mathcal{D}^{(t-1)}}^{\text{BE}}(f; R) - \mathcal{L}_{\mathcal{D}^{(t-1)}}^{\text{PbRM}}(R).$$

1104 Therefore, using $Q^\sharp \in \mathcal{F}$, $R^\sharp \in \mathcal{R}$, we have

$$\begin{aligned}
&\left[f_1^{(t)}(s_1) - \hat{V}_{\mathcal{D}^{(t-1)}, R^{(t)}}^{\text{ref}} \right] - \left[V_1^\sharp(s_1) - \hat{V}_{\mathcal{D}^{(t-1)}, R^\sharp}^{\text{ref}} \right] \\
&\leq - \frac{\mathcal{L}_{\mathcal{D}^{(t-1)}}^{\text{BE}}(f^{(t)}; R^{(t)}) - \mathcal{L}_{\mathcal{D}^{(t-1)}}^{\text{BE}}(Q^\sharp; R^\sharp)}{\lambda} - \frac{\mathcal{L}_{\mathcal{D}^{(t-1)}}^{\text{PbRM}}(R^{(t)}) - \mathcal{L}_{\mathcal{D}^{(t-1)}}^{\text{PbRM}}(R^\sharp)}{\lambda}.
\end{aligned}$$

1105 Using the decomposition (24), [Proposition 22](#), and the fact that $|V_1^\sharp(s_1) - V_1^*(s_1)| \leq \varepsilon_{\text{app}}$,

1106 $|R^\sharp(\tau) - R^*(\tau)| \leq H\varepsilon_{\text{app}}$, we have

$$\begin{aligned}
V_1^*(s_1) - V^{\pi^{(t)}}(s_1) &\leq (H+1)\varepsilon_{\text{app}} + \left| V_{R^{(t)}}^{\text{ref}} - \hat{V}_{\mathcal{D}^{(t-1)}, R^{(t)}}^{\text{ref}} \right| + \left| V_{R^\sharp}^{\text{ref}} - \hat{V}_{\mathcal{D}^{(t-1)}, R^\sharp}^{\text{ref}} \right| + \frac{2H\kappa}{\lambda} \\
&\quad + \sum_{h=1}^H \left(\mathbb{E}^{\pi^{(t)}} [e_h^{(t)}(s_h, a_h)] - \frac{1}{C_\beta \lambda} \sum_{k < t} \mathbb{E}^{\pi^{(k)}} e_h^{(t)}(s_h, a_h)^2 \right) \\
&\quad + \mathbb{E}^{\pi^{(t)}} [\tilde{R}^{(t)}(\tau) - \tilde{R}^*(\tau)] - \frac{1}{2\lambda} \sum_{k < t} \sum_{h=1}^H \mathbb{E}^{\pi^{(k)} \circ_h \pi_{\text{ref}}} (\tilde{R}^{(t)}(\tau) - \tilde{R}^*(\tau))^2.
\end{aligned} \quad (25)$$

1107 Taking summation over $t = 1, 2, \dots, T$ and apply [Proposition 12](#), [Proposition 13](#), and [Lemma 21](#)
1108 yields

$$\sum_{t=1}^T V_1^*(s_1) - V^{\pi^{(t)}}(s_1) \leq O(1) \cdot \left[H(\varepsilon_{\text{app}} + \rho) + \sqrt{T\kappa} + \frac{TH\kappa}{\lambda} + C_\beta \lambda H C_{\text{cov}} \log \left(1 + \frac{C_{\text{cov}} T}{\kappa} \right) \right].$$

1109 This is the desired upper bound. \square

1110 E.2 Proof of [Proposition 22](#)

1111 The inequality involving $\mathcal{L}_{\mathcal{D}^{(t)}}^{\text{BE}}$ is implied by [Proposition 11](#) and proven in [Appendix C.3](#). In the
1112 following, we only need to prove the inequality involving $\mathcal{L}_{\mathcal{D}^{(t)}}^{\text{PbRM}}$ by invoking [Proposition 8](#).

1113 Consider the class $\Theta = \mathcal{R} \cup \{R^*\}$, $\mathcal{X} = (\mathcal{S} \times \mathcal{A})^H \times (\mathcal{S} \times \mathcal{A})^H$, and $\mathcal{Y} = \{0, 1\}$. For any $R \in \Theta$,
 1114 we define

$$P_R(1|\tau^+, \tau^-) = \frac{\exp(\beta R(\tau^+))}{\exp(\beta R(\tau^+)) + \exp(\beta R(\tau^-))}, \quad P_R(0|\tau^+, \tau^-) = \frac{\exp(\beta R(\tau^-))}{\exp(\beta R(\tau^+)) + \exp(\beta R(\tau^-))},$$

1115 following Definition 3.

1116 Recall that $\mathcal{D}^{(t-1)} = \{(\tau^{(k,h,+)}, \tau^{(k,h,-)}, y^{(k,h)})\}_{k \leq t, h \in [H]}$ is the history up to the t -th iteration. For
 1117 simplicity, we denote $x^{(t,h)} := (\tau^{(t,h,+)}, \tau^{(t,h,-)})$. Note that $(x^{(t,1)}, y^{(t,1)}), \dots, (x^{(t,H)}, y^{(t,H)})$. Then
 1118 it is clear that for all $t \in [T]$, $h \in [H]$,

$$\mathbb{P}(y^{(t,h)} | x^{(t,h)}, \mathcal{D}^{(t-1)}) = P_{R^*}(y^{(t,h)} | x^{(t,h)}),$$

1119 and it also holds that

$$L(R(\tau^+) - R(\tau^-), y) = -\log P_R(y|\tau^+, \tau^-), \quad \forall y \in \{0, 1\}.$$

1120 Further, noting that $N_{\log}(\Theta, 2H\beta\rho) \leq N(\mathcal{R}, \rho) + 1$. Therefore, applying Proposition 8 gives the
 1121 following result: with probability at least $1 - \frac{\delta}{2}$, for any $R \in \mathcal{R}$, $t \in [T]$,

$$\begin{aligned} & \sum_{k=1}^t \sum_{h=1}^H \mathbb{E}^{\pi^{(k,h,+)}, \pi^{(k,h,-)}} D_H^2(P_R(\cdot|\tau^+, \tau^-), P_{R^*}(\cdot|\tau^+, \tau^-)) \\ & \leq \frac{1}{2} \sum_{(\tau^+, \tau^-, y)} [L(R(\tau^+) - R(\tau^-), y) - L(R^*(\tau^+) - R^*(\tau^-), y)] \\ & \quad + \log(N(\mathcal{R}, H\rho) + 1) + \log(2/\delta) + TH^3\beta\rho \\ & \leq \frac{1}{2} [\mathcal{L}_{\mathcal{D}^{(t)}}^{\text{PbRM}}(R) - \mathcal{L}_{\mathcal{D}^{(t)}}^{\text{PbRM}}(R^*)] + \frac{1}{2} H\kappa, \end{aligned}$$

1122 where the second inequality uses the fact that $|R^\#(\tau) - R^*(\tau)| \leq H\varepsilon_{\text{app}}$. Finally, note that
 1123 $D_H^2(\text{Bern}(p), \text{Bern}(q)) \geq \frac{1}{2}(p - q)^2$ and

$$\left| \frac{1}{e^{\beta w} + 1} - \frac{1}{e^{\beta w'} + 1} \right| \geq \frac{\beta}{2e^\beta} |w - w'|, \quad \forall w, w' \in [-1, 1].$$

1124 Therefore, using the definition of P_R completes the proof. \square

1125 F Proofs of Lower Bounds

1126 F.1 Hard Case of Learning Rewards and Value Functions Separately

1127 As mentioned in Section 3.1, in Algorithm 1 the learner has to optimize over the reward class and
 1128 value function class jointly. In the following, we argue that if the learner first learns a fitted reward
 1129 model in the reward class, then optimizes the value function with the fitted rewards, the output
 1130 policies at each iteration never converge to the optimal policy.

In detail, we consider algorithms in the form of Algorithm 4, where the learner fits the reward model
 $R^{(t)}$ at iteration t first, then the learner calls algorithm `alg`, which takes per-step rewards data as
 input and outputs a policy π_t at each iteration. To align with the structure of Algorithm 1, we take
`alg` to be a single iteration of the GOLF algorithm in Jin et al. (2021a), i.e. $\pi^{(t)} = \pi_{f^{(t)}}$ where
 $f^{(t)} = \arg\max_{f \in \mathcal{F}^{(t)}} f(x_1, \pi_f(x_1))$. Here the confidence set $\mathcal{F}^{(t)}$ is defined as

$$\mathcal{F}^{(t)} = \{f \in \mathcal{F} : \mathcal{L}_{\mathcal{D}}^{\text{BE}}(f; R) \leq \beta\}$$

1131 with $\mathcal{L}_{\mathcal{D}}^{\text{BE}}$ defined in Eq. (5).

1132 Then we have the following proposition, which shows that this approach outputs suboptimal policies
 1133 at every iteration in some special hard cases.

1134 **Proposition 23.** Consider Algorithm 4 with `alg` to a single iteration of the GOLF algorithm. After
 1135 running T iterations, the learner averages over all policies to output a policy. There exists an MDP
 1136 class that realizes the ground truth MDP, such that the above algorithm outputs a policy which is at
 1137 least 0.01-suboptimal.

Algorithm 4 RL with imputation of reward

input: Algorithm alg, reward regression oracle O .

- 1: **Initialize** $\mathcal{D}_h^{(0)} = \emptyset$ for every $h \in [H]$
- 2: **for** $t = 1, 2, \dots, T$ **do**
- 3: Receive $\pi^{(t)}$ from alg
- 4: Execute $\pi^{(t)}$ and receive $(\tau^{(t)}, r^{(t)})$, where $\tau^{(t)} = (s_1^{(t)}, a_1^{(t)}, \dots, s_H^{(t)}, a_H^{(t)})$
- 5: Receive the fitted reward function from O :

$$R^{(t)} = \min_{R \in \mathcal{R}} \sum_{k=1}^t (R(\tau^{(k)}) - r^{(k)})^2.$$

- 6: Let $\mathcal{D}_h^{(t)} = \mathcal{D}_h^{(t-1)} \cup \{(s_1^{(t)}, a_1^{(t)}, R_1^{(t)})\}$ for every $h \in [H]$
 - 7: Feed $\mathcal{D}_1^{(t)}, \dots, \mathcal{D}_H^{(t)}$ to alg and output policy $\pi^{(t)}$
 - 8: **end for**
-

Proof of Proposition 23. We consider the following class of two-layer MDP, where $\mathcal{S}_1 = \{s_1\}$, $\mathcal{S}_2 = \{s_2\}$, and the action space to be $\mathcal{A} = \{a_1, a_2\}$. The transition models \mathbb{T} are identical across the class, and have the following form:

$$\mathbb{T}(s_2 \mid s_1, a_i) = 1, \quad \forall i \in \{1, 2\}.$$

1138 The reward class is defined as $\mathcal{R} = \{R^1, R^2\}$, where

$$\begin{aligned} R^1(s_1, a_1) &= R^1(s_1, a_2) = 0.20, & R^1(s_2, a_1) &= 0.20, & R^1(s_2, a_2) &= 0.19, \\ \text{and } R^2(s_1, a_1) &= R^2(s_1, a_2) = 0.00, & R^2(s_2, a_1) &= 0.38, & R^2(s_2, a_2) &= 0.39. \end{aligned}$$

1139 The Q -function class \mathcal{Q} is defined as $\mathcal{Q} = \{Q^1, Q^2, Q^3, Q^4\}$, which takes value in Table 1 respectively. Notice that in all possible reward models and Q -functions, the values at (s_1, a_1) and at (s_1, a_2)

Table 1: Value of Q^1, \dots, Q^4

	(s_1, a_1)	(s_1, a_2)	(s_2, a_1)	(s_2, a_2)
Q^1	0.40	0.40	0.20	0.19
Q^2	0.20	0.20	0.20	0.19
Q^3	0.59	0.59	0.38	0.39
Q^4	0.39	0.39	0.38	0.39

1140

1141 are the same. In the following, when without ambiguity we simply use $R(s_1)$ to denote $R(s_1, a_1)$
 1142 and $R(s_1, a_2)$, and use $Q(s_1)$ to denote $Q(s_1, a_1)$ and $Q(s_2, a_2)$.

1143 We further suppose the ground truth model reward satisfies $R = R^1$, then we can verify that the
 1144 optimal Q -function is Q^1 . It is easy to verify that sets \mathcal{Q} and \mathcal{R} satisfy the completeness assumption.
 1145 Hence, sets \mathcal{Q} and \mathcal{R} satisfy the realizability assumption Assumption 1 and the completeness
 1146 assumption Assumption 2 with $\mathcal{G} = \mathcal{Q}$.

To see why this is a hard-case for GOLF type algorithms, we first notice that for any trajectory $\tau = (s_1, \tilde{a}_1, s_2, \tilde{a}_2)$ with outcome reward $r = R^1(s_1, \tilde{a}_1) + R^1(s_2, \tilde{a}_2)$ collected by the algorithm, we always have

$$r = R^2(s_1) + R^2(s_2, \tilde{a}_2).$$

Hence as long as \mathcal{D} does not contain state-action pair (s_2, a_1) , when fitting the reward function using the following ERM oracle:

$$R = \operatorname{argmin}_{R \in \mathcal{R}} \sum_{(\tau, r) \in \mathcal{D}} (r(\tau) - r)^2,$$

1147 the reward model R^2 always achieves the minimum. In the worst case, we assume the fitted reward
 1148 models encountered by the learner at such rounds are always R^2 .

In the following, we verify that by running the GOLF algorithm, the learner will not encounter the state-action pair (s_2, a_1) at any round. We notice that the optimal policies of Q^3 and Q^4 all take a_2

at state s_2 , and also that

$$Q^3(s_1) \geq Q^1(s_1) \quad \text{and} \quad Q^3(s_1) \geq Q^2(s_1).$$

1149 Hence, to verify that the algorithm never chooses a_2 at state s_2 , we only need to verify that if either
1150 Q^1 or Q^2 belongs to the confidence set, then Q^3 also belongs to the confidence set.

When the learner collects a new trajectory, two new pieces of data will be added to the dataset \mathcal{D} . If the trajectory does not pass through the state-action pair (s_2, a_1) , these two pieces of data will be in the following form:

$$(s_1, a_1, R^2(s_1)), \quad (s_2, a_2, R^2(s_2, a_2)) \quad \text{or} \quad (s_1, a_2, R^2(s_1)), \quad (s_2, a_2, R^2(s_2, a_2)).$$

1151 No matter which one of these two, we have the following inequality for the sum of squared Bellman
1152 error across these two pieces of data

$$\begin{aligned} \mathcal{E}_1(Q^1)^2 + \mathcal{E}_2(Q^1)^2 &= 0.20^2 + 0.20^2 \geq 0.20^2 = \mathcal{E}_1(Q^3)^2 + \mathcal{E}_2(Q^3)^2, \\ \mathcal{E}_1(Q^2)^2 + \mathcal{E}_2(Q^2)^2 &= 0.01^2 + 0.28^2 \geq 0.20^2 = \mathcal{E}_1(Q^3)^2 + \mathcal{E}_2(Q^3)^2. \end{aligned}$$

1153 According to the construction of the confidence set, if either Q^1 or Q^2 belongs to the confidence set,
1154 then Q^3 belongs to the confidence set as well.

Therefore, no matter how many rounds the algorithm runs, the optimistic policy always takes action a_2 at state s_2 . Hence the average policy $\hat{\pi}$ also takes a_2 at s_2 , which implies that

$$J(\pi^*) - J(\hat{\pi}) \geq 0.01.$$

1155

□

1156 F.2 Proof of Theorem 4

We define the class $\mathcal{M} = \{M^{v,b} : v \in B_2(1), b \in [0, 1]\}$ of two-layer MDPs as follows: for each MDP instance in the class, the state space is given by $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2$, where $\mathcal{S}_1 = \{s_1\}$ and $\mathcal{S}_2 = \{s_\theta : \theta \in B_2(1)\}$ are disjoint subsets of states. The action space is denoted by $\mathcal{A} = \{a_\theta : \theta \in B_2(1)\}$, and the transition dynamics are the same across all MDPs in this class:

$$\mathbb{T}(s_{\theta'} | s_1, a_\theta) = \mathbb{I}[\theta = \theta'], \quad \forall \theta, \theta' \in B_2(1).$$

1157 Every MDP $M^{v,b}$ in \mathcal{M} is parameterized by a vector $v \in B_2(1)$ and bias constant $b \in [0, 1]$, with
1158 the reward model $R^{v,b}$ defined as follows:

$$\begin{aligned} R^{v,b}(s_1, a_\theta) &= \text{ReLU}(\langle \theta, v \rangle - b) + \langle \theta, v \rangle & \forall \theta \in B_2(1), \\ R^{v,b}(s_\theta, a_{\theta'}) &= -\langle \theta, v \rangle & \forall \theta, \theta' \in B_2(1). \end{aligned}$$

1159 The per-step reward of a sample (s, a) is according to the normal distribution with mean to be
1160 $R^{v,b}(s, a)$ and variance to be 1. We next show that this class of MDP \mathcal{M} can be learned with
1161 per-step-based samples, but cannot be learned with outcome-based samples.

Exponential Lower Bound for Outcome-Based Samples: . When executing a policy π in MDP M^v , suppose the agent selects a_θ in the first layer and action $a_{\theta'}$ in the second layer, the data observed are in the following form of trajectory together outcome-based rewards:

$$((s_1, a_\theta, s_\theta, a_{\theta'}), R),$$

where the trajectory reward R is sampled according to the Gaussian distribution $\mathcal{N}(\text{ReLU}(\langle \theta, v \rangle - b), 2)$. For policy π which takes action a_θ at state s_1 , its value function equals to $\text{ReLU}(\langle \theta, v \rangle)$, hence,

$$J(\pi^*) - J(\pi) = \max_{\theta^* \in B_2(1)} \text{ReLU}(\langle \theta^*, v \rangle - b) - \text{ReLU}(\langle \theta, v \rangle - b).$$

Notice that θ' does not have influence on either the outcome $\text{ReLU}(\langle \theta, v \rangle - b)$ or the performance of the policy. The above setup is equivalent to the ReLU bandit setting in Dong et al. (2021); Li et al. (2022); Foster et al. (2021), in the sense that as long as there exists an algorithm that achieves regret $R(T)$ in the ReLU bandit setting, then there also exists an algorithm that achieves regret $R(T)$ in the above MDP setting, and vice versa. According to (Foster et al., 2021, Proposition 6.6), we have the following lower bound on the regret:

$$R(T) = \Omega(\min\{T, e^{\Omega(d)}\}).$$

Polynomial Upper Bound for Per-Step-Based Samples: . We first notice that for fixed v, b , the optimal policy of $M^{v,b}$ always takes action a_v at state s_1 . Hence for any policy π which takes a_θ at s_1 for some $\theta \in B_2(1)$, we have

$$J(\pi^*) - J(\pi) = \text{ReLU}(\langle v, v \rangle - b) - \text{ReLU}(\langle \theta, v \rangle - b) \leq \langle v, v \rangle - \langle v, \theta \rangle = \max_{\theta^* \in B_2(1)} \langle v, \theta^* \rangle - \langle v, \theta \rangle.$$

The right-hand side is the objective of the linear bandit (Dani et al., 2008; Chu et al., 2011). Further notice that in the per-step-based model by taking action a_θ at s_1 and an arbitrary action $a_{\theta'}$ in the second layer, the reward we observe in the second layer has mean $-\langle \theta, v \rangle$. Hence, the piece of feedback observation also aligns with the feedback in the linear bandit setting. Therefore, according to the regret upper bound for the linear bandit in Dani et al. (2008), there exists an algorithm for the per-step-based learning which has regret upper bound

$$R(T) = \tilde{O}(d\sqrt{T}).$$

1162

□