Human-in-the-loop: Provably Efficient Preference-based Reinforcement Learning with General Function Approximation

Xiaoyu Chen [* 1] **Han Zhong** [* 2 3] **Zhuoran Yang** [4] **Zhaoran Wang** [5] **Liwei Wang** [1 2]

Abstract

We study human-in-the-loop reinforcement learning (RL) with trajectory preferences, where instead of receiving a numeric reward at each step, the RL agent only receives preferences over trajectory pairs from a human overseer. The goal of the RL agent is to learn the optimal policy which is most preferred by the human overseer. Despite the empirical successes, the theoretical understanding of preference-based RL (PbRL) is only limited to the tabular case. In this paper, we propose the first optimistic model-based algorithm for PbRL with general function approximation, which estimates the model using value-targeted regression and calculates the exploratory policies by solving an optimistic planning problem. We prove that our algorithm achieves the regret bound of $\tilde{O}(\mathrm{poly}(dH)\sqrt{K})$, where d is the complexity measure of the transition and preference model depending on the Eluder dimension and log-covering numbers, H is the planning horizon, K is the number of episodes, and $\tilde{O}(\cdot)$ omits logarithmic terms. Our lower bound indicates that our algorithm is near-optimal when specialized to the linear setting. Furthermore, we extend the PbRL problem by formulating a novel problem called RL with n-wise comparisons, and provide the first sample-efficient algorithm for this new setting. To the best of our knowledge, this is the first theoretical result for PbRL with (general) function approximation.

[*]Equal contribution [1]Key Laboratory of Machine Perception, MOE, School of Artificial Intelligence, Peking University [2]Center for Data Science, Peking University [3]Peng Cheng Laboratory [4]Department of Statistics and Data Science, Yale University [5]Department of Industrial Engineering and Management Sciences, Northwestern University. Correspondence to: Xiaoyu Chen <cxy30@pku.edu.cn>, Han Zhong <hanzhong@stu.pku.edu.cn>, Zhuoran Yang <zhuoran.yang@yale.edu>, Zhaoran Wang <zhaoranwang@gmail.com>, Liwei Wang <wanglw@cis.pku.edu.cn>.

Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

1. Introduction

Reinforcement learning (RL) is concerned with sequential decision-making problems in which the agent interacts with the environment to maximize its cumulative rewards. This framework has achieved tremendous successes in various fields such as Atari games (Mnih et al., 2013), Go (Silver et al., 2017), and StarCraft (Vinyals et al., 2019). While these empirical successes are encouraging, one limitation of the standard RL paradigm is that the learning algorithms and the policies crucially depend on the prior knowledge encoded in the definition of the reward function. In many real-world applications such as autonomous driving, healthcare, and robotics, reward functions might not be readily available or difficult to design, which leads to well-known challenges such as reward shaping (Ng et al., 1999) and reward hacking (Amodei et al., 2016; Berkenkamp et al., 2021).

If we have enough demonstrations of the desired task, one possible solution to address the above problems is to extract a reward function using inverse reinforcement learning (Ng et al., 2000; Abbeel & Ng, 2004). This reward function can be further used to train an agent with reinforcement learning algorithms. More directly, we can use imitation learning (Ho & Ermon, 2016; Hussein et al., 2017; Osa et al., 2018) to clone the demonstrated behavior. However, these approaches are not applicable to situations where the demonstration data from experts are expensive to obtain, or behaviors are difficult for humans to demonstrate.

Another popular alternative to handle the lack of reward functions is called Preference-based Reinforcement Learning (PbRL) (Busa-Fekete et al., 2014; Wirth et al., 2017). In PbRL, instead of observing the reward information on the encountered state-action pairs, the agent only receives 1 bit preference feedback over a trajectory pair from an expert or a human overseer. Such preference feedback is often more natural and straightforward to specify in many RL applications, especially those involving human evaluations. This learning paradigm has been widely applied to multiple areas, including robot training (Jain et al., 2013; 2015; Christiano et al., 2017), game playing (Wirth & Fürnkranz, 2012; 2014) and clinical trials (Zhao et al., 2011).

Despite its promising application in various areas, the theoretical understanding of PbRL is limited to only the tabular RL setting. Novoseller et al. (2020) proposes the Double Posterior Sampling method using Bayesian linear regression with an asymptotic regret sublinear in T (the number of time steps). Xu et al. (2020b) presents the first finite-time analysis for PbRL problems with near-optimal sample complexity bounds. Pacchiano et al. (2021) studies the regret minimization problem for PbRL with linearly-parameterized preference function. However, all the previous algorithms are restricted to the tabular setting, and their complexity bounds scale polynomial dependence on the cardinality of the state-action space. Therefore, their algorithms can fail in the more practical scenario where the state space is extremely large.

Recently, there have been tremendous results studying standard RL problem with general function approximation (e.g. Du et al. (2019); Yang & Wang (2019); Ayoub et al. (2020); Wang et al. (2020b); Jin et al. (2020); Zanette et al. (2020); Wang et al. (2020a); Zhou et al. (2021b); Chen et al. (2021b); Zhou et al. (2021a)). However, their algorithms cannot be directly applied to PbRL setting due to the following two reasons. Firstly, most of their algorithms estimate the value function by utilizing the Bellman update with general function approximation. However, we cannot estimate the value of certain policies since the reward values are hidden and unidentifiable up to shifts in rewards. Secondly, since the preference feedback in PbRL depends on the utility of the whole trajectories and can be even non-Markovian, the optimal policy for PbRL problems can be possibly history-dependent. This violates the fundamental requirement of the Markovian policy class in the standard RL setting.

In this work, we tackle the regret minimization problem for preference-based reinforcement learning with general function approximation. Specifically, we study the PbRL problem where both the unknown transition model and the unknown preference function are known to belong to given function spaces. The function spaces are general sets of functions, which may be either finitely parameterized or nonparametric. This setting is more general than the previous theoretical results for PbRL (Novoseller et al., 2020; Xu et al., 2020b; Pacchiano et al., 2021). Our contributions are summarized as follows:

- We propose a statistically efficient algorithm called Preference-based Optimistic Planning (PbOP) for PbRL with general function approximation. We prove that the regret of our algorithm is $\tilde{O}\left(\mathrm{poly}(dH)\sqrt{K}\right)$, where d is the complexity measure of the transition and preference model depending on the Eluder dimension (Russo & Van Roy, 2013) and log-covering numbers, H is the planning horizon, and K is the number of episodes. Additionally, we find that our algorithm can be almost directly applied to a setting called RL with once-per-episode feedback (Efroni et al., 2020; Chatterji et al., 2021), which is another reinforcement learning problem dealing with the lack of reward functions.

- We present a reduction from the setting of RL with once-per-episode feedback to PbRL. When specialized to the linear case, we prove a nearly-matching lower bound for PbRL based on this reduction. The lower bound indicates that our algorithm is near-optimal in the case of linear function approximation.

- We formulate a novel setting called RL with n-wise comparisons to cover situations where multiple trajectories are sampled and compared with each other in each episode. This setting is more general than the standard PbRL setting and covers many real situations including robotics and clinical trails. Based on our PbOP algorithm, we also propose an algorithm with near-optimal regret.

2. Related Work

Preference-based RL We refer readers to Wirth et al. (2017) for an overview of Preference-based RL. Overall, there are three different types of preference feedback in the PbRL literature. Firstly, the preferences can be defined on the action space where the labeler tells which action is better for a given state. Secondly, state preference determines the preferred state between state pairs, which indicates that there is an action in the preferred state that is better than all actions available in the other state. Lastly, trajectory preference compares between trajectory pairs and specifies that a trajectory should be preferred over the other ones. Trajectory preference is the most general form of preference-based feedback, which is also the main focus of this work. As discussed in the introduction, previous theoretical results studying PbRL with trajectory feedback mainly focus on the tabular RL setting with finite state and action space (Novoseller et al., 2020; Xu et al., 2020b; Pacchiano et al., 2021). Besides PbRL, preference-based learning has also been well-explored in bandit setting under the notion of "dueling bandits" (Yue et al., 2012; Falahatgar et al., 2017a;b; Busa-Fekete et al., 2018; Xu et al., 2020a; Busa-Fekete et al., 2018), which can be regarded as a special case of PbRL with single state and horizon $H = 1$.

RL with Function Approximation Recently, there are a large number of theoretical results about provably efficient exploration in the standard RL problem with function approximation, The most basic and frequently explored setting is RL with linear function approximation. See e.g., Du et al. (2019); Yang & Wang (2019); Jin et al. (2020); Cai et al. (2020); Zanette et al. (2020); Wang et al. (2020a); Zhou et al. (2021a;b) and references therein. Beyond linear setting, there are also results studying RL with general function approximation (Jiang et al., 2017; Wang et al., 2020b; Kong et al., 2021; Foster et al., 2020; Jin et al., 2021; Du et al., 2021). Our work is mostly relevant to previous works studying provably efficient model-based RL with general function approximation. For example, Osband & Van Roy (2014) makes explicit model-based assumption that the transition operator and the reward function lie in a given function class, and analyse the regret of Thompson sampling when applied to RL with general function approximation. Ayoub et al. (2020) proposes an algorithm for episodic model-based RL based on value-targeted regression. Recently, Chen et al. (2021a) extends the model-based algorithm for episodic RL to the infinite-horizon setting.

RL with Once-per-episode Feedback RL with once-per-episode feedback is another reinforcement learning paradigm to deal with the lack of a reward function in various real-world scenarios, in which the agent only receives non-Markovian feedback based on the whole trajectory at the end of an episode. Efroni et al. (2020) firstly studies the setting with the assumption of inherent Markovian rewards. They propose a hybrid optimistic-Thompson Sampling approach with a $\sqrt{K}$ regret. Chatterji et al. (2021) removes the Markovian rewards assumption and provide optimistic algorithms based on the well-known UCBVI algorithm (Azar et al., 2017). Though aimed to tackle the similar problem, the setting of RL with once-per-episode feedback is relatively independently studied compared with the PbRL, and this is the first work that points out the connections between two different settings.

3. Preliminaries

Throughout this paper, we use the following notations. For any positive integer n, we use $[n]$ to denote $\{1, 2, \cdots, n\}$. We denote by $\Delta(\mathcal{A})$ the set of probability distributions on a set $\mathcal{A}$.

3.1. PbRL with Trajectory Preferences

We study the episodic finite horizon Markov decision process (MDP), which is defined by a tuple $(\mathcal{S}, \mathcal{A}, H, \mathbb{P})$, where $\mathcal{S}$ and $\mathcal{A}$ are state and action spaces, respectively, H is the horizon of the MDP, and $\mathbb{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition kernel. In the episodic setting, the agent interacts with the environment for K episodes. Each episode consists of H steps. For ease of presentation, we assume the initial state of each episode is a fixed state $s_1 \in \mathcal{S}$. We remark that this setting can be generalized to the setting that the initial state is sampled from a fixed distribution. At the beginning of the k-th episode, the agent determines two policies $(\pi_{k,1}, \pi_{k,2})$. After executing these two policies, the agent obtains two trajectories $\{\tau_{k,i} = (s_{k,1,i}, a_{k,1,i}, s_{k,2,i}, a_{k,2,i}, \cdots, s_{k,H,i}, a_{k,H,i})\}_{i=1,2}$. In PbRL, unlike the standard RL where the agent can receive reward signals, the agent can only obtain the preference o_k between two trajectories $(\tau_{k,1}, \tau_{k,2})$. Here o_k is a Bernoulli random variable with $\Pr(o_k = 1) = \Pr(\tau_{k,1} > \tau_{k,2})$. For ease of presentation, we denote $\mathbb{T}(\tau_1, \tau_2) = \Pr(\tau_1 > \tau_2)$ for any two trajectories τ_1 and τ_2.

We define Π to be the set containing all history-dependent policies, and we use Traj to denote the set of H-step trajectories. For any transition $\mathbb{P}$ and policy π, we also use $\tau \sim (\mathbb{P}, \pi)$ to denote that the trajectory τ is sampled using policy π from the MDP with transition $\mathbb{P}$. With a slight abuse of notation, we define $\mathbb{T}(\pi_1, \pi_2) = \mathbb{E}_{\tau_1 \sim (\mathbb{P}, \pi_1), \tau_2 \sim (\mathbb{P}, \pi_2)} \mathbb{T}(\tau_1, \tau_2)$. Throughout this paper, we make the following assumption.

Assumption 3.1. There exists a policy π^*, such that $\mathbb{T}(\pi^*, \pi_0) \geq \frac{1}{2}, \forall \pi_0 \in \Pi$.

This is an extension of the optimal-arm assumption in dueling bandits (Busa-Fekete et al., 2018). It is also more general than Assumption 1 in (Xu et al., 2020b).

We study the regret minimization problem, where the regret is defined as: $Reg(K) = \sum_{k=1}^{K} \sum_{i=1}^{2} \left(\mathbb{T}(\pi^*, \pi_{k,i}) - \frac{1}{2}\right)$. By Assumption 3.1, we have the regret is non-negative, and our goal is to design algorithms with sub-linear regret guarantees.

3.2. General Function Approximation

In this subsection, we introduce notions that can characterize the complexity of the function class.

Covering Number When the function class has infinite elements, we usually use the covering number to capture the complexity.

Definition 3.2 (Covering Number)**.** The ϵ-covering number of a set $\mathcal{F}$ under metric d, denoted as $\mathcal{N}(\mathcal{F}, \epsilon, d)$, is the minimum integer m such that there exists a subset $\mathcal{F}' \subset \mathcal{F}$ with $|\mathcal{F}| = m$, and for any $f \in \mathcal{F}$, there exists some $f' \in \mathcal{F}'$ satisfying $d(f, f') \leq \epsilon$.

Eluder Dimension We use the concept of Eluder dimension introduced by Russo & Van Roy (2013) to characterize the complexity of different function classes in RL.

Definition 3.3 (α-independent)**.** Let $\mathcal{F}$ be a function class defined in $\mathcal{X}$, and $\{x_1, x_2, \cdots, x_n\} \in \mathcal{X}$. We say $x \in \mathcal{X}$ is α-independent of $\{x_1, x_2, \cdots, x_n\}$ with respect to $\mathcal{F}$ if there exists $f_1, f_2 \in \mathcal{F}$ such that $\sqrt{\sum_{i=1}^{n} (f_1(x_i) - f_2(x_i))^2} \leq \alpha$, but $f_1(x) - f_2(x) \geq \alpha$.

Definition 3.4 (Eluder Dimension)**.** Suppose $\mathcal{F}$ is a function class defined in $\mathcal{X}$, the α-Eluder dimension is the longest sequence $\{x_1, x_2, \cdots, x_n\} \in \mathcal{X}$ such that there exists $\alpha' \geq \alpha$ where x_i is α'-independent of $\{x_1, \cdots, x_{i-1}\}$ for all $i \in [n]$.

Preference Function We assume the function $\mathbb{T}(\tau_1, \tau_2)$ belongs to the function space $\mathcal{F}_{\mathbb{T}}$, which is defined as

$$\mathcal{F}_{\mathbb{T}} = \{f(\tau_1, \tau_2) \in [0,1], f(\tau_1, \tau_2) + f(\tau_2, \tau_1) = 1\} . \tag{1}$$

We assume the α-Eluder dimension of the function class $\mathcal{F}_{\mathbb{T}}$ is bounded by $d_{\mathbb{T}}$. Here α is a parameter which will be specified later.

Following the analysis by Russo & Van Roy (2013), we can show that the function apace $\mathcal{F}_{\mathbb{T}}$ has bounded Eluder dimension in linear and generalized linear cases.

Remark 3.5 (Linear Preference Models). Consider the case of d-dimensional linear preference models $f(\tau_1, \tau_2) = \psi(\tau_1, \tau_2)^\top \theta$ where $\psi : \mathrm{Traj} \times \mathrm{Traj} \to \mathbb{R}^d$ is a known feature map satisfying $\|\psi(\tau_1, \tau_2)\|_2 \leq L$ and $\theta \in \mathbb{R}^d$ is an unknown parameter with $\|\theta\|_2 \leq S$. Then the α-Eluder dimension of $\mathcal{F}_{\mathbb{T}}$ is at most $O(d \log(LS/\alpha))$.

Remark 3.6 (Generalized Linear Preference Models). For the case of d-dimensional generalized linear models $f(\tau_1, \tau_2) = g(\langle \phi(\tau_1, \tau_2), \theta \rangle)$ where g is an increasing Lipschitz continuous function, $\psi : \mathrm{Traj} \times \mathrm{Traj} \to \mathbb{R}^d$ is a known feature map satisfying $\|\psi(\tau_1, \tau_2)\|_2 \leq L$ and $\theta \in \mathbb{R}^d$ is an unknown parameter with $\|\theta\|_2 \leq S$. Set $\bar{h} = \sup_{\tilde{\theta}, \tau_1, \tau_2} g'(\langle \phi(\tau_1, \tau_2), \tilde{\theta} \rangle)$, $\underline{h} = \inf_{\tilde{\theta}, \tau_1, \tau_2} g'(\langle \phi(\tau_1, \tau_2), \tilde{\theta} \rangle)$ and $r = \bar{h}/\underline{h}$. Then the α-Eluder dimension of $\mathcal{F}_{\mathbb{T}}$ is at most $O(dr^2 \log(rLS\bar{h}/\alpha))$. Therefore, our results subsume the setting of logistic preference functions (Pacchiano et al., 2021) as a spacial case.

Model Complexity Similar to Ayoub et al. (2020), we use Eluder dimension to characterize the complexity of the model class. We denote $\mathcal{V} = \{f : \mathcal{S} \to [0,1]\}$. We assume the real transition $\mathbb{P}$ belongs to a transition set $\mathcal{P}$, and we define the function space $\mathcal{F}$ as the collections of functions $f : \mathcal{S} \times \mathcal{A} \times \mathcal{V} \to \mathbb{R}$:

$$\begin{aligned} \mathcal{F}_{\mathbb{P}} = \Big\{ f \mid \exists \mathbb{P} \in \mathcal{P}, s.t. \forall (s,a,v) \in \mathcal{S} \times \mathcal{A} \times \mathcal{V}, \\ f(s,a,v) = \int \mathbb{P}\left(ds' \mid s,a\right) v\left(s'\right) \Big\}. \end{aligned} \tag{2}$$

We define $d_{\mathbb{P}} = dim_{\mathcal{E}}(\mathcal{F}_{\mathbb{P}}, \alpha)$ to be the α-Eluder dimension of $\mathcal{F}_{\mathbb{P}}$.

Remark 3.7 (Linear Mixture Models). Such a model subsumes linear mixture models as a special case (Ayoub et al., 2020). Specifically, we say an MDP is a linear mixture MDP if

$$\mathcal{P} = \{\psi(s,a,s')^\top \theta : \theta \in \Theta\},$$

where $\psi : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$ is a known feature map satisfying $\| \sum_{s'} \psi(s,a,s') V(s') \|_2 \leq 1, \forall s \in \mathcal{S}, a \in \mathcal{A}, V \in \mathcal{V}$, and $\theta \in \Theta$ satisfies $\|\theta\|_2 \leq B$ for a constant B. For linear mixture models, the α-Eluder dimension of $\mathcal{F}_{\mathbb{P}}$ is of order $O(d \log(B/\alpha))$.

3.3. RL with Once-per-episode Feedback

In this subsection, we introduce another related RL setting called RL with once-per-episode feedback.

In the k-th episode, the agent executes a policy π_k and obtains a trajectory τ_k induced by π_k. At the end of the episode, the agent receives a feedback $y_k \in \{0,1\}$, where $y_k = g^*(\tau_k)$ for some unknown function g^*. We assume g^* belongs to the function space $\mathcal{F}_{\mathbb{G}}$, which is defined as $\mathcal{F}_{\mathbb{G}} = \{g : \mathrm{Traj} \to [0,1]\}$. We assume the α-Eluder dimension of the function class $\mathcal{F}_{\mathbb{G}}$ is bounded by $d_{\mathbb{G}}$. Here α is a parameter which will be specified later. With slight abuse of notations, we denote $g^*(\pi) = \mathbb{E}_{\tau \sim (\mathbb{P}, \pi)}[g^*(\tau)]$. For any policy π, we can define the preference value function as $V_1^\pi(s_1) = \mathbb{E}_{\tau \sim (\mathbb{P}, \pi)}[g^*(\tau)]$. Our goal is to minimize the regret $Reg(K) = \sum_{k=1}^K V_1^*(s_1) - V_1^{\pi_k}(s_1)$, where $V_1^*(s_1) = \max_\pi V_1^\pi(s_1)$ is the value of the optimal policy.

Remark 3.8. Similar to Remarks 3.5 and 3.6, our setting incorporates (generalized) linear case. Hence, our following results can be naturally applied to the scenario considered by Chatterji et al. (2021).

4. Main Results for Preference-based RL

In this section, we present the main results for preference-based RL. We first propose a novel algorithm called Preference-based Optimistic Planning (PbOP) and establish the regret upper bound for it. To show the sharpness of our result, we also prove an information-theoretic lower bound in the linear case.

4.1. Algorithm

The algorithm is formally defined in Algorithm 1. Overall, we employ the standard least-squares regression to learn the transition dynamics and the preference function. In each episode, we first update the model estimation based on the history samples till episode $k-1$. We define the confidence sets and calculate the confidence bonuses for the transition and preference estimations, respectively. Based on the confidence sets and the bonus terms, we maintain a policy set in which all policies are near-optimal with minor sub-optimality gap with high probability. Finally, we execute the most exploratory policy pair in the policy set and observe the preference between the trajectories sampled using these two policies.

Confidence Sets and Bonuses We first explain our construction of the confidence sets and bonus terms in Algorithm 1. For the preference function, the estimation $\hat{\mathbb{T}}_k$ is defined as the minimizer of the following least-squares loss:

$$\hat{\mathbb{T}}_k = \underset{\mathbb{T}' \in \mathcal{F}_{\mathbb{T}}}{\arg\min} \sum_{t=1}^{k-1} \left(\mathbb{T}'(\tau_{t,1}, \tau_{t,2}) - o_t\right)^2 . \tag{3}$$

Algorithm 1 PbOP: Preference-based Optimistic Planning

1: Set $\beta_{\mathbb{T}} = \beta_{\mathbb{P}} = 8\log(2K\mathcal{N}(\mathcal{F}_{\mathbb{T}}, 1/K, \|\cdot\|_\infty)/\delta)$
2: **for** episode $k = 1, \cdots, K$ **do**
3: Calculate the estimation $\hat{\mathbb{T}}_k$ and $\hat{\mathbb{P}}_k$ using least-squares regression (Eqn. (3) and (5))
4: Construct the high-confidence set $\mathcal{B}_{\mathbb{T},k}$ for the preference $\mathbb{T}$ (Eqn. (4))
5: Construct the high-confidence set $\mathcal{B}_{\mathbb{P},k}$ for transition $\mathbb{P}$ (Eqn. (7))
6: Define the preference bonus $b_{\mathbb{T},k}(\tau_1,\tau_2) = \max_{f_1,f_2\in\mathcal{B}_{\mathbb{T},k}} |f_1(\tau_1,\tau_2) - f_2(\tau_1,\tau_2)|$
7: Define the transition bonus $b_{\mathbb{P},k}(s,a,V) = \max_{\mathbb{P}_1,\mathbb{P}_2\in\mathcal{B}_{\mathbb{P},k}} (\mathbb{P}_1 - \mathbb{P}_2)V(s,a)$
8: Set $b_{\mathbb{P},k}(s,a) = \max_{V\in\mathcal{V}} b_{\mathbb{P},k}(s,a,V)$
9: Construct the policy set $\mathcal{S}_k$ as Eqn. (9)
10: Compute policies $(\pi_{k,1}, \pi_{k,2})$ as Eqn. (10)
11: Execute the policy $\pi_{k,1}$ and $\pi_{k,2}$ for one episode, respectively, and then observe the trajectory $\tau_{k,1}$ and $\tau_{k,2}$
12: Receive the preference o_k between $\tau_{k,1}$ and $\tau_{k,2}$
13: **end for**

We can guarantee that the real preference function $\mathbb{T}$ is contained in the following high-probability set for $\mathbb{T}$:

$$\mathcal{B}_{\mathbb{T},k} = \left\{ \mathbb{T}' \mid \sum_{t=1}^{k-1} \left(\hat{\mathbb{T}}_k - \mathbb{T}'\right)^2 (\tau_{t,1}, \tau_{t,2}) \le \beta_{\mathbb{T}} \right\}. \quad (4)$$

We define the exploration bonus $b_{\mathbb{T},k}(\tau_1,\tau_2)$ to be the "width" of the confidence set $\mathcal{B}_{\mathbb{T},k}$, i.e., $b_{\mathbb{T},k}(\tau_1,\tau_2) = \max_{f_1,f_2\in\mathcal{B}_{\mathbb{T},k}} |f_1(\tau_1,\tau_2) - f_2(\tau_1,\tau_2)|$. This bonus measures the uncertainty of a certain trajectory pair (τ_1,τ_2) w.r.t. the confidence set $\mathcal{B}_{\mathbb{T},k}$, which will be used to calculate the near-optimal policy set and the most exploratory policy pairs later.

When it comes to the confidence set and the bonus term for the transition estimation, the situation becomes slightly complicated. Recall that in the definition of the transition function $f_{\mathbb{P}} \in \mathcal{F}_{\mathbb{P}}$, the input variables includes both the state-action pair (s,a) and the next-step target function V. Given history samples $\{\tau_{t,i} = (s_{t,1,i}, a_{t,1,i}, s_{t,2,i}, a_{t,2,i}, \cdots, s_{t,H,i}, a_{t,H,i})\}_{i=1,2,t\in[k-1]}$, we can possibly estimate the transition model by minimizing the least-squares loss with respect to certain value target $\{V_{t,h,i}\}_{i=1,2,h\in[H],t\in[k-1]}$:

$$\hat{\mathbb{P}}_k = \mathrm{argmin}_{\mathbb{P}'\in\mathcal{P}} \sum_{i=1}^{2}\sum_{t=1}^{k-1}\sum_{h=1}^{H} \big(\langle \mathbb{P}'(\cdot \mid s_{t,h,i}, a_{t,h,i}), V_{k,h,i}\rangle - V_{k,h,i}(s_{k,h+1,i})\big)^2. \quad (5)$$

Now the remaining problem is how to define the target function $V_{t,h,i}$. In the problem of standard RL with general function approximation, Ayoub et al. (2020) uses the optimistic value estimation as the target function in each episode. However, in the PbRL setting, since the reward information is hidden, we cannot calculate the value estimation for each given state-action pairs. To tackle this problem, we define the bonus $b_{\mathbb{P},k}(s,a,V)$ for any $s\in\mathcal{S}, a\in\mathcal{A}$ and the target function $V\in\mathcal{V}$, and use $V = \arg\max_{V\in\mathcal{V}} b_{\mathbb{P},k}(s,a,V)$ as the regression target for the state-action pair (s,a). Similar ideas have also been applied to the problem of reward-free exploration for linear mixture MDPs (Zhang et al., 2021; Chen et al., 2021b).

To be more specific, we define $L_k(\mathbb{P}_1,\mathbb{P}_2)$ as

$$L_k(\mathbb{P}_1,\mathbb{P}_2) = \sum_{i=1}^{2}\sum_{t=1}^{k-1}\sum_{h=1}^{H} \big(\langle \mathbb{P}_1(\cdot \mid s_{t,h,i}, a_{t,h,i}) - \mathbb{P}_2(\cdot \mid s_{t,h,i}, a_{t,h,i}), V_{t,h,i}\rangle\big)^2. \quad (6)$$

We construct the high confidence set for transition $\mathbb{P}$:

$$\mathcal{B}_{\mathbb{P},k} = \left\{ \mathbb{P}' \mid L_k(\mathbb{P}', \hat{\mathbb{P}}_k) \le \beta_{\mathbb{P}} \right\}, \quad (7)$$

The exploration bonus $b_{\mathbb{P},k}(s,a,V)$ for the transition estimation measures the uncertainty of the confidence set $\mathcal{B}_{\mathbb{P},k}$:

$$b_{\mathbb{P},k}(s,a,V) = \max_{\mathbb{P}_1,\mathbb{P}_2\in\mathcal{B}_{\mathbb{P},k}} (\mathbb{P}_1 - \mathbb{P}_2)V(s,a). \quad (8)$$

Suppose $V_{max,k,s,a} = \arg\max_{V\in\mathcal{V}} b_{\mathbb{P},k}(s,a,V)$, then we use $V_{max,t,s_{t,h,i},a_{t,h,i}}$ as the online target for the history sample $(s_{t,h,i}, a_{t,h,i}, s_{t,h+1,i})$. With a slight abuse of notation, we use $b_{\mathbb{P},k}(s,a) = \max_{V\in\mathcal{V}} b_{\mathbb{P},k}(s,a,V)$ to denote the maximum uncertainty for a given state-action pair (s,a).

Near-optimal Policy Set With the estimated model, algorithms for standard RL calculate the optimistic value function using Bellman backup to balance the exploration and exploitation. However, in PbRL, we cannot update the value function through Bellman update since the environment can be non-Markovian. Even worse, since we only get feedback about the preference between trajectory pairs, we cannot directly evaluate a single policy. Inspired from a recent work for PbRL in the tabular setting (Pacchiano et al., 2021), we construct a near-optimal policy set using the preference information.

Define $b_{\mathbb{P},k}(\tau) = \sum_{(s,a)\in\tau} b_{\mathbb{P},k}(s,a)$. With the constructed confidence set and the bonus terms, then we construct the following set $\mathcal{S}_k$:

$$\mathcal{S}_k = \Big\{ \pi \mid \mathbb{E}_{\tau\sim(\hat{\mathbb{P}}_k,\pi),\tau_0\sim(\hat{\mathbb{P}}_k,\pi_0)} \big(\hat{\mathbb{T}}_k(\tau,\tau_0) + b_{\mathbb{T},k}(\tau,\tau_0) + b_{\mathbb{P},k}(\tau) + b_{\mathbb{P},k}(\tau_0)\big) \ge \frac{1}{2}, \forall \pi_0 \in \Pi \Big\}. \quad (9)$$

Intuitively speaking, $\mathcal{S}_k$ consists of policies that there is no other policy significantly outperforms it. By executing policies in $\mathcal{S}_k$ in episode k, We can guarantee that the regret suffered in episode k to be less than the summation of the bonuses of the trajectories $(\tau_{k,1}, \tau_{k,2})$, thus solve the exploitation problem in PbRL. With our concentration analysis, we can guarantee that the optimal policy $\pi^* \in \mathcal{S}_k$ for any $k \in [K]$ with high probability.

Exploratory Policies Now we explain how to deal with the exploration problem in PbRL with general function approximation. Since we have already defined the uncertainty $b_{\mathbb{T},k}$ and $b_{\mathbb{P},k}$ for trajectory pairs, we can choose two policies in $\mathcal{S}_k$ that maximize the uncertainty, and thus encourage exploration:

$$
\begin{aligned}
(\pi_{k,1}, \pi_{k,2}) = &\arg\max_{\pi_1,\pi_2 \in \mathcal{S}_k} \mathbb{E}_{\tau_1 \sim (\hat{\mathbb{P}}_k, \pi_1), \tau_2 \sim (\hat{\mathbb{P}}_k, \pi_2)} \\
&\big(b_{\mathbb{T},k}(\tau_1, \tau_2) + b_{\mathbb{P},k}(\tau_1) + b_{\mathbb{P},k}(\tau_2)\big).
\end{aligned} \tag{10}
$$

4.2. Regret Upper Bound

Theorem 4.1. *With probability at least $1 - \delta$, the regret of Algorithm 1 is upper bounded by*

$$
\begin{aligned}
Reg(K) \le \tilde{O}(&\sqrt{d_{\mathbb{P}} H K \log(\mathcal{N}(\mathcal{F}_{\mathbb{P}}, 1/K, \|\cdot\|_\infty)/\delta)} \\
&+ \sqrt{d_{\mathbb{T}} K \log(\mathcal{N}(\mathcal{F}_{\mathbb{T}}, 1/K, \|\cdot\|_\infty)/\delta)}),
\end{aligned}
$$

where $d_{\mathbb{P}}$ and $d_{\mathbb{T}}$ is the $1/K$-Eluder dimension of $\mathcal{F}_{\mathbb{P}}$ and $\mathcal{F}_{\mathbb{T}}$, respectively.

The regret bound in Theorem 4.1 has polynomial dependence on the Eluder dimension of the function class $\mathcal{F}_{\mathbb{P}}$ and $\mathcal{F}_{\mathbb{T}}$, and has no dependence on the cardinality of the state-action space. We defer the proof of Theorem 4.1 to Appendix A. When specialized to linear setting, the regret of Algorithm 1 can be bounded by the following corollary.

Corollary 4.2 (Linear Mixture Model and Linear Preference Function)**.** *For the setting of linear mixture models and linear preference functions defined in Remark 3.5 and 3.7, the regret of Algorithm 1 is upper bounded by*

$$
\begin{aligned}
Reg(K) \le O(&\tilde{d}_{\mathbb{P}} \sqrt{HK \log(BK) \log(BK/\delta)} \\
&+ \tilde{d}_{\mathbb{T}} \sqrt{K \log(LSK) \log(LSK/\delta)}),
\end{aligned}
$$

where $\tilde{d}_{\mathbb{P}}$ and $\tilde{d}_{\mathbb{T}}$ are the feature dimension of linear mixture models and linear preference functions, respectively.

RL with once-per-episode feedback (cf. Section 3.3) is another paradigm introduced to tackle the problem of the lack of reward functions. Though the setting is different, we find that our Algorithm 1 can be almost directly applied to this setting with near-optimal regret. Since this is not the main focus of this work, we refer the interested readers to Appendix B for detailed algorithm and proof.

Theorem 4.3. *With probability at least $1 - \delta$, the regret of Algorithm 4 for RL with once-per-episode feedback is upper bounded by*

$$
\begin{aligned}
Reg(K) \le \tilde{O}(&\sqrt{d_{\mathbb{P}} H K \log(\mathcal{N}(\mathcal{F}_{\mathbb{P}}, 1/K, \|\cdot\|_\infty)/\delta)} \\
&+ \sqrt{d_{\mathbb{G}} K \log(\mathcal{N}(\mathcal{F}_{\mathbb{G}}, 1/K, \|\cdot\|_\infty)/\delta)}),
\end{aligned}
$$

where $d_{\mathbb{P}}$ and $d_{\mathbb{G}}$ is the $1/K$-Eluder dimension of $\mathcal{F}_{\mathbb{P}}$ and $\mathcal{F}_{\mathbb{G}}$, respectively.

To the best of our knowledge, this is the first provably efficient algorithm for the problem of RL with once-per-episode feedback with general function approximation, which covers the result for RL with once-per-episode feedback in the tabular case (Chatterji et al., 2021).

4.3. Information-Theoretic Lower Bound

In this subsection, we establish the lower bound for PbRL in the linear setting, which is derived using the reduction from the problem of RL with once-per-episode feedback.

Firstly, we show the reduction from the problem of RL with once-per-episode feedback setting to the PbRL setting. Specifically, suppose we have an algorithm $\mathcal{ALG}$ for PbRL problems, we design a reduction protocol to solve the RL with once-per-episode feedback problem using Algorithm $\mathcal{ALG}$.

Algorithm 2 Reduction Protocol

1: **for** episode $k = 1, \cdots, K/2$ **do**
2: Invoke Algorithm $\mathcal{ALG}$ and obtain the policy $\pi_{k,1}$ and $\pi_{k,2}$.
3: Execute the policy $\pi_{k,1}$, and then observe the trajectory $\tau_{k,1}$ and the corresponding feedback $y_{k,1}$.
4: Execute the policy $\pi_{k,2}$, and then observe the trajectory $\tau_{k,2}$ and the corresponding feedback $y_{k,2}$.
5: Define the preference $o_k = \mathbb{1}(y_{k,1} > y_{k,2})$ if $y_{k,1} \neq y_{k,2}$, and $o_k = \begin{cases} 1 & w.p.\ 1/2 \\ 0 & w.p.\ 1/2 \end{cases}$ otherwise.
6: Send the information $(\tau_{k,1}, \tau_{k,2})$ and o_k to Algorithm $\mathcal{ALG}$.
7: **end for**

The reduction protocol is described in Algorithm 2. In each episode, the agent invokes Algorithm $\mathcal{ALG}$ to obtain the policy $\pi_{k,1}$ and $\pi_{k,2}$ based on the history data. The agent then executes these two policies, observes the trajectory $\tau_{k,1}, \tau_{k,2}$, and receives the corresponding feedback $y_{k,1}, y_{k,2}$. We define the preference $o_k = \mathbb{1}(y_{k,1} > y_{k,2})$ if $y_{k,1} \neq y_{k,2}$, and $o_k = \begin{cases} 1 & w.p.\ 1/2 \\ 0 & w.p.\ 1/2 \end{cases}$ otherwise. Note that in our design, we have $\Pr\{o_k = 1\} = \frac{\Pr(y_{k,1}=1) - \Pr(y_{k,2}=1) + 1}{2}$. Finally, we send the information $(\tau_{k,1}, \tau_{k,2})$ and o_k obtained in this episode to Algorithm $\mathcal{ALG}$. We have the following proposition based on the reduction protocol.

Proposition 4.4. *Suppose there exists a PbRL algorithm $\mathcal{ALG}$ with regret $poly(K, H, d_{\mathbb{P}}, d_{\mathbb{T}}, \mathcal{N}_{\mathbb{P}}, \mathcal{N}_{\mathbb{T}}, 1/\delta)$. For RL with once-per-episode feedback, there exists an algorithm with regret $poly(K, H, d_{\mathbb{P}}, d_{\mathcal{F}}, \mathcal{N}_{\mathbb{P}}, \mathcal{N}_{\mathcal{F}}, 1/\delta)$, where $d_{\mathcal{F}}$ and $\mathcal{N}_{\mathcal{F}}$ is the Eluder dimension and covering number of the function space*

$$\mathcal{F} = \{f(\tau_1, \tau_2) : \mathrm{Traj} \times \mathrm{Traj} \to [0,1],$$
$$s.t.\ f(\tau_1, \tau_2) = \frac{g(\tau_1) - g(\tau_2) + 1}{2}, g(\tau) \in \mathcal{F}_{\mathbb{G}}\}.$$

We state the lower bound for RL with once-per-episode feedback in Theorem 4.5, and defer the proof to Appendix C.

Theorem 4.5 (Lower Bound for RL with Once-per-episode Feedback)**.** *For any algorithm for RL with once-per-episode feedback problem, there exists a $\tilde{d}_{\mathbb{P}}$-dimensional linear mixture MDP with a $\tilde{d}_{\mathbb{G}}$-dimensional linear preference function such that the regret incurred by this algorithms is at least $\Omega(\tilde{d}_{\mathbb{P}}\sqrt{K} + \tilde{d}_{\mathbb{G}}\sqrt{K})$.*

The following lower bound for PbRL is implied by Proposition 4.4 and Theorem 4.5. This lower bound matches the upper bound in Corollary 4.2 w.r.t. the feature dimensions $\tilde{d}_{\mathbb{P}}, \tilde{d}_{\mathbb{T}}$ and the number of episode K except for logarithmic factors, which indicates that our algorithm is near-optimal when specialized to the linear setting.

Corollary 4.6 (Lower Bound for PbRL)**.** *For any algorithm for PbRL, there exists a $\tilde{d}_{\mathbb{P}}$-dimensional linear mixture MDP with a $\tilde{d}_{\mathbb{T}}$-dimensional linear preference function such that the regret incurred by this algorithm is at least $\Omega(\tilde{d}_{\mathbb{P}}\sqrt{K} + \tilde{d}_{\mathbb{T}}\sqrt{K})$.*

5. RL with n-wise Comparisons

In the previous section, we propose a sample-efficient algorithm with near-optimal regret for the problem of PbRL with trajectory feedback. However, this setting cannot cover some other RL situations with preference feedback. For example, in robotics, sampling new trajectories can be expensive and time-consuming compared with labeling preferences among trajectories. The human overseer may sample multiple trajectories in a distributed manner and compare all these trajectories with each other. In clinical trails, different medical treatments can be evaluated simultaneously, and the feedback is a pairwise comparison or a ranking among them. In this section, we propose a new setting called RL with n-wise comparisons, which is an extension of PbRL with trajectory feedback. We describe the setup and learning objective, followed by the algorithm and theoretical guarantees.

5.1. Setup and Learning Objective

Compared with the problem of PbRL, the main difference is that the agent needs to execute n policies in each episode. Specifically, in the k-th episode, the agent executes n policies $\{\pi_{k,i}\}_{i\in[n]}$, and obtains n trajectories $\{\tau_{k,i}\}_{i\in[n]}$. The agent receives the feedback of $n(n-1)/2$ pairwise comparisons $\{o_{k,i,j}\}_{1\le i<j\le n}$, where $o_{k,i,j}$ is a Bernoulli random variable such that $\Pr(o_{k,i,j} = 1) = \Pr(\tau_{k,i} > \tau_{k,j})$. Recall that we use the notations $\mathbb{T}(\tau_1, \tau_2) = \Pr(\tau_1 > \tau_2)$ and $\mathbb{T}(\pi_1, \pi_2) = \mathbb{E}_{\tau_1\sim(\mathbb{P},\pi_1),\tau_2\sim(\mathbb{P},\pi_2)}\mathbb{T}(\tau_1, \tau_2)$. For such a problem, our goal is to minimize the regret, which is defined as $Reg(K) = \sum_{k=1}^{K}\sum_{i=1}^{n}\left(\mathbb{T}(\pi^*, \pi_{k,i}) - \frac{1}{2}\right)$, where π^* is defined in Assumption 3.1. When $n = 2$, this regret reduces to the regret in the standard PbRL setting.

5.2. Algorithm

The algorithm, which is formally described in Algorithm 3, shares the similar framework with Algorithm 1. The main difference is on the construction of confidence sets and bonus terms. Similar to the PbRL setting, we use least-squares regression to estimate the preference function $\mathbb{T}$:

$$\hat{\mathbb{T}}_k = \arg\min_{\mathbb{T}'\in\mathcal{F}_T} \sum_{t=1}^{k-1}\sum_{i=1}^{n}\sum_{j=i+1}^{n}\left(\mathbb{T}'(\tau_{t,i}, \tau_{t,j}) - o_{t,i,j}\right)^2. \tag{11}$$

Notably, we have $n(n-1)/2$ samples instead of one sample in each episode.We construct the confidence set centered at $\hat{\mathbb{T}}_k$ by

$$\mathcal{B}_{\mathbb{T},k} = \left\{\tilde{\mathbb{T}} \mid \sum_{t=1}^{k-1}\sum_{i=1}^{n}\sum_{j=i+1}^{n}\left(\hat{\mathbb{T}}_k - \tilde{\mathbb{T}}\right)^2(\tau_{t,i}, \tau_{t,j}) \le \beta_{\mathbb{T}}\right\}. \tag{12}$$

Given the confidence set $\mathcal{B}_{\mathbb{T},k}$, we also use the function $b_{\mathbb{T},k}(\tau_1, \tau_2) = \max_{f_1,f_2\in\mathcal{B}_{\mathbb{T},k}} |f_1(\tau_1, \tau_2) - f_2(\tau_1, \tau_2)|$ to measure its uncertainty.

For estimating the transition dynamics, we utilize historical trajectories $\{\tau_{t,i}\}_{(t,i)\in[k-1]\times[n]}$ to perform the least-squares regression:

$$\hat{\mathbb{P}}_k = \mathrm{argmin}_{\mathbb{P}'\in\mathcal{P}} \sum_{i=1}^{n}\sum_{t=1}^{k-1}\sum_{h=1}^{H}\big(\langle\mathbb{P}'(\cdot \mid s_{t,h,i}, a_{t,h,i}), V_{k,h,i}\rangle - V_{k,h,i}(s_{k,h+1,i})\big)^2, \tag{13}$$

where $V_{k,h,i}$ is the target function defined as follows. Specifically, we construct the high confidence set for transition $\mathbb{P}$, which is defined as

$$\mathcal{B}_{\mathbb{P},k} = \left\{\mathbb{P}' \mid L_k(\mathbb{P}', \hat{\mathbb{P}}_k) \le \beta_{\mathbb{P}}\right\}, \tag{14}$$

Algorithm 3 PbOP+: Pairwise Preference-based Optimistic Planning

1: Set $\beta_{\mathbb{T}} = 8\log(2K\mathcal{N}\left(\mathcal{F}_{\mathbb{T}}, 1/(Kn^2), \|\cdot\|_\infty\right)/\delta)$ and $\beta_{\mathbb{P}} = 8\log(2K\mathcal{N}\left(\mathcal{F}_{\mathbb{P}}, 1/(Kn), \|\cdot\|_\infty\right)/\delta)$
2: **for** episode $k = 1, \cdots, K$ **do**
3: Calculate the estimation $\hat{\mathbb{T}}_k$ and $\hat{\mathbb{P}}_k$ using least-squares regression (Eqn. (11) and (13))
4: Construct the high-confidence set $\mathcal{B}_{\mathbb{T},k}$ for the preference $\mathbb{T}$ (Eqn. (12))
5: Construct the high-confidence set $\mathcal{B}_{\mathbb{P},k}$ for transition $\mathbb{P}$ (Eqn. (14))
6: Define the bonus term $b_{\mathbb{T},k}(\tau_1, \tau_2) = \max_{f_1, f_2 \in \mathcal{B}_{\mathbb{T},k}} |f_1(\tau_1, \tau_2) - f_2(\tau_1, \tau_2)|$
7: Define the bonus term $b_{\mathbb{P},k}(s,a) = \max_{\mathbb{P}_1, \mathbb{P}_2 \in \mathcal{B}_{\mathbb{P},k}} \max_{V \in \mathcal{V}} (\mathbb{P}_1 - \mathbb{P}_2) V(s,a)$
8: Set $b_{\mathbb{P},k}(\tau) = \sum_{(s,a)\in\tau} b_{\mathbb{P},k}(s,a)$
9: Compute the policy set $\mathcal{S}_k$ as Eqn. (16)
10: Compute policy $(\pi_{k,1}, \pi_{k,2}, \cdots, \pi_{k,n})$ as Eqn. (17)
11: Execute the policies $(\pi_{k,1}, \pi_{k,2}, \cdots, \pi_{k,n})$ for one episode, respectively, and then observe the trajectories $(\tau_{k,1}, \tau_{k,2}, \cdots, \pi_{k,n})$
12: Receive the preference $o_{k,i,j}$ between $\tau_{k,i}$ and $\tau_{k,j}$ for all $(i,j) \in \{(i,j) | 1 \le i < j \le n\}$
13: **end for**

where $L_k(\cdot,\cdot)$ is defined by

$$L_k(\mathbb{P}_1, \mathbb{P}_2) = \sum_{i=1}^{n} \sum_{t=1}^{k-1} \sum_{h=1}^{H} \left(\langle \mathbb{P}_1(\cdot \mid s_{t,h,i}, a_{t,h,i}) - \mathbb{P}_2(\cdot \mid s_{t,h,i}, a_{t,h,i}), V_{t,h,i} \rangle \right)^2. \quad (15)$$

Given any $V \in \mathcal{V}$, we choose the associated bonus $b_{\mathbb{P},k}(s,a,V)$ as $b_{\mathbb{P},k}(s,a,V) = \max_{\mathbb{P}_1,\mathbb{P}_2} (\mathbb{P}_1 - \mathbb{P}_2)V(s,a)$. Such a bonus function measures the uncertainty of the confidence set $\mathcal{B}_{\mathbb{P},k}$. We also choose the target value function $V_{t,h,i}$ as the function that can maximize the uncertainty. Formally, let $V_{\max,k,s,a} = \arg\max_{V\in\mathcal{V}} b_{\mathbb{P},k}(s,a,V)$, then we use $V_{\max,t,s_{t,h,i},a_{t,h,i}}$ as the online target for the historical sample $(s_{t,h,i}, a_{t,h,i}, s_{t,h+1,i})$. Also we denote $b_{\mathbb{P},k}(s,a) = \max_{V\in\mathcal{V}} b_{\mathbb{P},k}(s,a,V)$ and $b_{\mathbb{P},k}(\tau) = \sum_{(s,a)\in\tau} b_{\mathbb{P},k}(s,a)$.

Given the estimated transition $\hat{\mathbb{P}}_k$, bonus for transition $b_{\mathbb{P},k}$ and bonus for preference function $b_{\mathbb{T},k}$, we can construct the near-optimal set $\mathcal{S}_k$ like Eqn. (9):

$$\mathcal{S}_k = \Big\{ \pi \mid \mathbb{E}_{\tau\sim(\hat{\mathbb{P}}_k,\pi),\tau_0\sim(\hat{\mathbb{P}}_k,\pi_0)} \big(\hat{\mathbb{T}}_k(\tau,\tau_0) + b_{\mathbb{T},k}(\tau,\tau_0) + b_{\mathbb{P},k}(\tau) + b_{\mathbb{P},k}(\tau_0) \big) \ge \frac{1}{2}, \forall \pi_0 \in \Pi \Big\}. \quad (16)$$

Finally, we choose the exploratory polices $(\pi_{k,1}, \pi_{k,2}, \cdots, \pi_{k,n})$ that can maximize the pairwise uncertainty. We luckily find that the summation of bonuses exactly characterize the uncertainty of the policy tuple:

$$(\pi_{k,1}, \pi_{k,2}, \cdots, \pi_{k,n}) = \operatorname*{arg\,max}_{\pi_1,\pi_2,\cdots,\pi_n \in \mathcal{S}_k} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \mathbb{E}_{\tau_i\sim(\hat{\mathbb{P}}_k,\pi_i),\tau_j\sim(\hat{\mathbb{P}}_k,\pi_j)} \left(b_{\mathbb{T},k}(\tau_i,\tau_j) + b_{\mathbb{P},k}(\tau_i) + b_{\mathbb{P},k}(\tau_j) \right). \quad (17)$$

5.3. Theoretical Guarantees

In the following theorem, we establish the regret upper bound for Algorithm 3. The proof of this theorem is deferred to Appendix D.

Theorem 5.1. *With probability at least $1-\delta$, the regret of Algorithm 3 is upper bounded by*

$$Reg(K) \le \tilde{O}\Big(\sqrt{d_{\mathbb{P}} HnK \log(\mathcal{N}(\mathcal{F}_{\mathbb{P}}, 1/(Kn), \|\cdot\|_\infty)/\delta)} + \sqrt{d_{\mathbb{T}} K \log(\mathcal{N}(\mathcal{F}_{\mathbb{T}}, 1/(Kn^2), \|\cdot\|_\infty)/\delta)}\Big),$$

where $d_{\mathbb{P}}$ and $d_{\mathbb{T}}$ is the $1/K$-Eluder dimension of $\mathcal{F}_{\mathbb{P}}$ and $\mathcal{F}_{\mathbb{T}}$, respectively.

By replacing K with K/n in Theorem 5.1, we obtain a bound of $\tilde{O}(\sqrt{d_{\mathbb{P}} HK \log(\mathcal{N}(\mathcal{F}_{\mathbb{P}}, 1/(Kn), \|\cdot\|_\infty)/\delta)} + \sqrt{d_{\mathbb{T}} K/n \log(\mathcal{N}(\mathcal{F}_{\mathbb{T}}, 1/(Kn^2), \|\cdot\|_\infty)/\delta)})$. This improves the regret bound in Theorem 4.1 by a factor of $\sqrt{n}$ on the second term, which is the benefit of additional information from n-wise comparisons.

6. Conclusion

This paper studies the regret minimization problem of PbRL with trajectory feedback and general function approximation. Based on the value-targeted regression and optimistic planning methods, we propose a novel RL algorithm called PbOP with regret $\tilde{O}(\text{poly}(dH)\sqrt{K})$. Our lower bound indicates that our regret upper bound is tight w.r.t. the feature dimension and the number of episodes when specialized to the linear setting. Furthermore, we formulate a novel setting called RL with n-wise comparisons and provide the first sample efficient algorithm in this setting.

A few problems still remain open. Firstly, there is still a gap of $\sqrt{H}$ between Corollaries 4.2 and 4.6. We conjecture that our upper bound is not tight, which can possibly be improved with more refined concentration analysis based on Bernstein bounds. Secondly, our algorithm is computationally inefficient due to non-Markovian feedback. It is tempting to design both statistically and computationally efficient algorithms in a relaxed PbRL setting. Finally, our setting of RL with n-wise comparisons does not cover the case where the feedback among n trajectories is a n-wise ranking (Negahban et al., 2018), which is also an interesting problem to be explored in future research.

Acknowledgement

Liwei Wang was supported by National Key R&D Program of China (2018YFB1402600), Exploratory Research Project of Zhejiang Lab (No. 2022RC0AN02), BJNSF (L172037), Project 2020BD006 supported by PKUBaidu Fund, the major key project of PCL (PCL2021A12).

References

Abbeel, P. and Ng, A. Y. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 1, 2004.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

Ayoub, A., Jia, Z., Szepesvari, C., Wang, M., and Yang, L. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pp. 463–474. PMLR, 2020.

Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pp. 263–272. PMLR, 2017.

Berkenkamp, F., Krause, A., and Schoellig, A. P. Bayesian optimization with safety constraints: safe and automatic parameter tuning in robotics. *Machine Learning*, pp. 1–35, 2021.

Busa-Fekete, R., Szörényi, B., Weng, P., Cheng, W., and Hüllermeier, E. Preference-based reinforcement learning: evolutionary direct policy search using a preference-based racing algorithm. *Machine Learning*, 97(3):327–351, 2014.

Busa-Fekete, R., Hüllermeier, E., and Mesaoudi-Paul, A. E. Preference-based online learning with dueling bandits: A survey. *arXiv preprint arXiv:1807.11398*, 2018.

Cai, Q., Yang, Z., Jin, C., and Wang, Z. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pp. 1283–1294. PMLR, 2020.

Chatterji, N. S., Pacchiano, A., Bartlett, P. L., and Jordan, M. I. On the theory of reinforcement learning with once-per-episode feedback. *arXiv preprint arXiv:2105.14363*, 2021.

Chen, X., Hu, J., Jin, C., Li, L., and Wang, L. Understanding domain randomization for sim-to-real transfer. *arXiv preprint arXiv:2110.03239*, 2021a.

Chen, X., Hu, J., Yang, L. F., and Wang, L. Near-optimal reward-free exploration for linear mixture mdps with plug-in solver. *arXiv preprint arXiv:2110.03244*, 2021b.

Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. *arXiv preprint arXiv:1706.03741*, 2017.

Du, S. S., Luo, Y., Wang, R., and Zhang, H. Provably efficient q-learning with function approximation via distribution shift error checking oracle. *arXiv preprint arXiv:1906.06321*, 2019.

Du, S. S., Kakade, S. M., Lee, J. D., Lovett, S., Mahajan, G., Sun, W., and Wang, R. Bilinear classes: A structural framework for provable generalization in rl. *arXiv preprint arXiv:2103.10897*, 2021.

Efroni, Y., Merlis, N., and Mannor, S. Reinforcement learning with trajectory feedback. *arXiv preprint arXiv:2008.06036*, 2020.

Falahatgar, M., Hao, Y., Orlitsky, A., Pichapati, V., and Ravindrakumar, V. Maxing and ranking with few assumptions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 7063–7073, 2017a.

Falahatgar, M., Orlitsky, A., Pichapati, V., and Suresh, A. T. Maximum selection and ranking under noisy comparisons. In *International Conference on Machine Learning*, pp. 1088–1096. PMLR, 2017b.

Foster, D. J., Rakhlin, A., Simchi-Levi, D., and Xu, Y. Instance-dependent complexity of contextual bandits and reinforcement learning: A disagreement-based perspective. *arXiv preprint arXiv:2010.03104*, 2020.

Ho, J. and Ermon, S. Generative adversarial imitation learning. *Advances in neural information processing systems*, 29:4565–4573, 2016.

Hussein, A., Gaber, M. M., Elyan, E., and Jayne, C. Imitation learning: A survey of learning methods. *ACM Computing Surveys (CSUR)*, 50(2):1–35, 2017.

Jain, A., Wojcik, B., Joachims, T., and Saxena, A. Learning trajectory preferences for manipulators via iterative improvement. *arXiv preprint arXiv:1306.6294*, 2013.

Jain, A., Sharma, S., Joachims, T., and Saxena, A. Learning preferences for manipulation tasks from online coactive feedback. *The International Journal of Robotics Research*, 34(10):1296–1313, 2015.

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pp. 1704–1713. PMLR, 2017.

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

Jin, C., Liu, Q., and Miryoosefi, S. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *arXiv preprint arXiv:2102.00815*, 2021.

Kong, D., Salakhutdinov, R., Wang, R., and Yang, L. F. Online sub-sampling for reinforcement learning with general function approximation. *arXiv preprint arXiv:2106.07203*, 2021.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Negahban, S., Oh, S., Thekumparampil, K. K., and Xu, J. Learning from comparisons and choices. *The Journal of Machine Learning Research*, 19(1):1478–1572, 2018.

Ng, A. Y., Harada, D., and Russell, S. Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml*, volume 99, pp. 278–287, 1999.

Ng, A. Y., Russell, S. J., et al. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, pp. 2, 2000.

Novoseller, E., Wei, Y., Sui, Y., Yue, Y., and Burdick, J. Dueling posterior sampling for preference-based reinforcement learning. In *Conference on Uncertainty in Artificial Intelligence*, pp. 1029–1038. PMLR, 2020.

Osa, T., Pajarinen, J., Neumann, G., Bagnell, J. A., Abbeel, P., and Peters, J. An algorithmic perspective on imitation learning. *arXiv preprint arXiv:1811.06711*, 2018.

Osband, I. and Van Roy, B. Model-based reinforcement learning and the eluder dimension. *arXiv preprint arXiv:1406.1853*, 2014.

Pacchiano, A., Saha, A., and Lee, J. Dueling rl: Reinforcement learning with trajectory preferences. *arXiv preprint arXiv:2111.04850*, 2021.

Russo, D. and Van Roy, B. Eluder dimension and the sample complexity of optimistic exploration. In *NIPS*, pp. 2256–2264. Citeseer, 2013.

Russo, D. and Van Roy, B. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39 (4):1221–1243, 2014.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575 (7782):350–354, 2019.

Wang, R., Du, S. S., Yang, L. F., and Salakhutdinov, R. On reward-free reinforcement learning with linear function approximation. *arXiv preprint arXiv:2006.11274*, 2020a.

Wang, R., Salakhutdinov, R., and Yang, L. F. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. *arXiv preprint arXiv:2005.10804*, 2020b.

Wirth, C. and Fürnkranz, J. First steps towards learning from game annotations. 2012.

Wirth, C. and Fürnkranz, J. On learning from game annotations. *IEEE Transactions on Computational Intelligence and AI in Games*, 7(3):304–316, 2014.

Wirth, C., Akrour, R., Neumann, G., Fürnkranz, J., et al. A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research*, 18(136): 1–46, 2017.

Xu, Y., Joshi, A., Singh, A., and Dubrawski, A. Zeroth order non-convex optimization with dueling-choice bandits. In *Conference on Uncertainty in Artificial Intelligence*, pp. 899–908. PMLR, 2020a.

Xu, Y., Wang, R., Yang, L. F., Singh, A., and Dubrawski, A. Preference-based reinforcement learning with finite-time guarantees. *arXiv preprint arXiv:2006.08910*, 2020b.

Yang, L. and Wang, M. Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*, pp. 6995–7004. PMLR, 2019.

Yue, Y., Broder, J., Kleinberg, R., and Joachims, T. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012.

Zanette, A., Lazaric, A., Kochenderfer, M., and Brunskill, E. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, pp. 10978–10989. PMLR, 2020.

Zhang, W., Zhou, D., and Gu, Q. Reward-free model-based reinforcement learning with linear function approximation. *Advances in Neural Information Processing Systems*, 34, 2021.

Zhao, Y., Zeng, D., Socinski, M. A., and Kosorok, M. R. Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. *Biometrics*, 67(4):1422–1433, 2011.

Zhou, D., Gu, Q., and Szepesvari, C. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pp. 4532–4576. PMLR, 2021a.

Zhou, D., He, J., and Gu, Q. Provably efficient reinforcement learning for discounted mdps with feature mapping. In *International Conference on Machine Learning*, pp. 12793–12802. PMLR, 2021b.

A. Proof of Theorem 4.1

Lemma A.1. *Fix $\delta \in (0,1)$, with probability at least $1-\delta$, for all $k \in [K]$,*

$$\sum_{t=1}^{k-1} \left(\hat{\mathbb{T}}_k - \mathbb{T}\right)^2 (\tau_{t,1}, \tau_{t,2}) \le \beta_{\mathbb{T}}. \tag{18}$$

Proof. This lemma can be proved by the direct application of Lemma E.1. □

By Lemma A.1, we know that the true preference $\mathbb{T}(\tau_1, \tau_2) \in \mathcal{B}_{\mathbb{T},k}$ with high probability.

Lemma A.2. *Fix $\delta \in (0,1)$, with probability at least $1-\delta$, for all $k \in [K]$,*

$$L_k\left(\mathbb{P}, \hat{\mathbb{P}}_k\right) = \sum_{i=1}^{2}\sum_{t=1}^{k-1}\sum_{h=1}^{H}\left(\left\langle \mathbb{P}\left(\cdot \mid s_{t,h,i}, a_{t,h,i}\right) - \hat{\mathbb{P}}_k\left(\cdot \mid s_{t,h,i}, a_{t,h,i}\right), V_{t,h,i}\right\rangle\right)^2 \le \beta_{\mathbb{P}}. \tag{19}$$

Proof. This lemma can be proved by the direct application of Lemma E.1. □

By Lemma A.2, we know that the true transition kernel $\mathbb{P}(s'|s,a) \in \mathcal{B}_{\mathbb{P},k}$ with high probability.

We denote the high-probability event in Lemmas A.1 and A.2 as $\mathcal{E}$.

Lemma A.3. *Under event $\mathcal{E}$, for any two policies π_1, π_2 and scalar function $f : \mathrm{Traj} \times \mathrm{Traj} \to [0,1]$,*

$$\mathbb{E}_{\tau_1 \sim (\mathbb{P},\pi_1), \tau_2 \sim (\mathbb{P},\pi_2)}[f(\tau_1,\tau_2)] - \mathbb{E}_{\tau_1 \sim (\hat{\mathbb{P}}_k,\pi_1), \tau_2 \sim (\hat{\mathbb{P}}_k,\pi_2)}[f(\tau_1,\tau_2)] \le \mathbb{E}_{\tau_1 \sim (\hat{\mathbb{P}}_k,\pi_1), \tau_2 \sim (\hat{\mathbb{P}}_k,\pi_2)}[b_{\mathbb{P},k}(\tau_1) + b_{\mathbb{P},k}(\tau_2)] \tag{20}$$

Proof. We use $\tau \sim (\hat{\mathbb{P}}_k, \pi)_h$ to denote that the first h steps in the trajectory τ is sampled using policy π from the MDP with transition $\hat{\mathbb{P}}_k$, and the state-action pairs from step $h+1$ up until the last step is sampled using policy π from the MDP with the true transition $\mathbb{P}$. Therefore,

$$\begin{aligned}
&\mathbb{E}_{\tau_1 \sim (\mathbb{P},\pi_1), \tau_2 \sim (\mathbb{P},\pi_2)}[f(\tau_1,\tau_2)] - \mathbb{E}_{\tau_1 \sim (\hat{\mathbb{P}}_k,\pi_1), \tau_2 \sim (\hat{\mathbb{P}}_k,\pi_2)}[f(\tau_1,\tau_2)] && (21)\\
=&\mathbb{E}_{\tau_1 \sim (\hat{\mathbb{P}}_k,\pi_1)_0, \tau_2 \sim (\hat{\mathbb{P}}_k,\pi_2)_0}[f(\tau_1,\tau_2)] - \mathbb{E}_{\tau_1 \sim (\hat{\mathbb{P}}_k,\pi_1)_H, \tau_2 \sim (\hat{\mathbb{P}}_k,\pi_2)_H}[f(\tau_1,\tau_2)] && (22)\\
=&\sum_{h=1}^{H} \mathbb{E}_{\tau_1 \sim (\hat{\mathbb{P}}_k,\pi_1)_{h-1}, \tau_2 \sim (\hat{\mathbb{P}}_k,\pi_2)_{h-1}}[f(\tau_1,\tau_2)] - \mathbb{E}_{\tau_1 \sim (\hat{\mathbb{P}}_k,\pi_1)_h, \tau_2 \sim (\hat{\mathbb{P}}_k,\pi_2)_h}[f(\tau_1,\tau_2)] && (23)
\end{aligned}$$

For a given h, suppose $d_h(\mathbb{P},\pi)$ denotes the state-action distribution in step h when the agent interacts with the MDP with transition $\mathbb{P}$ using policy π, we have

$$\begin{aligned}
&\mathbb{E}_{\tau_1 \sim (\hat{\mathbb{P}}_k,\pi_1)_{h-1}, \tau_2 \sim (\hat{\mathbb{P}}_k,\pi_2)_{h-1}}[f(\tau_1,\tau_2)] - \mathbb{E}_{\tau_1 \sim (\hat{\mathbb{P}}_k,\pi_1)_h, \tau_2 \sim (\hat{\mathbb{P}}_k,\pi_2)_h}[f(\tau_1,\tau_2)] && (24)\\
\le&\mathbb{E}_{(s_{h,1},a_{h,1}) \sim d_h(\hat{\mathbb{P}}_k,\pi_1), (s_{h,2},a_{h,2}) \sim d_h(\hat{\mathbb{P}}_k,\pi_2)}\left[\max_V\left(\left(\mathbb{P}-\hat{\mathbb{P}}_k\right)V(s_{h,1},a_{h,1})\right) + \max_V\left(\left(\mathbb{P}-\hat{\mathbb{P}}_k\right)V(s_{h,2},a_{h,2})\right)\right] && (25)\\
\le&\mathbb{E}_{(s_{h,1},a_{h,1}) \sim d_h(\hat{\mathbb{P}}_k,\pi_1), (s_{h,2},a_{h,2}) \sim d_h(\hat{\mathbb{P}}_k,\pi_2)}\left[b_{\mathbb{P},k}(s_{h,1},a_{h,1}) + b_{\mathbb{P},k}(s_{h,2},a_{h,2})\right], && (26)
\end{aligned}$$

where the last inequality is due to $\mathbb{P} \in \mathcal{B}_{\mathbb{P},k}$ under event $\mathcal{E}$ by Lemma A.2. Summing over all $h \in [H]$, we can prove that

$$\mathbb{E}_{\tau_1 \sim (\mathbb{P},\pi_1), \tau_2 \sim (\mathbb{P},\pi_2)}[f(\tau_1,\tau_2)] - \mathbb{E}_{\tau_1 \sim (\hat{\mathbb{P}}_k,\pi_1), \tau_2 \sim (\hat{\mathbb{P}}_k,\pi_2)}[f(\tau_1,\tau_2)] \le \mathbb{E}_{\tau_1 \sim (\hat{\mathbb{P}}_k,\pi_1), \tau_2 \sim (\hat{\mathbb{P}}_k,\pi_2)}[b_{\mathbb{P},k}(\tau_1) + b_{\mathbb{P},k}(\tau_2)]. \tag{27}$$

□

Lemma A.4. *Under event $\mathcal{E}$, we have $\pi^* \in \mathcal{S}_k$.*

Proof. By Assumption 3.1, we know that

$$\mathbb{E}_{\tau_0 \sim (\mathbb{P},\pi_0), \tau^* \sim (\mathbb{P},\pi^*)} \mathbb{T}\left(\tau^*, \tau_0\right) \ge \frac{1}{2}, \forall \pi_0. \tag{28}$$

We decompose the LHS of the above inequality into the following three terms:

$$\begin{aligned}\mathbb{E}_{\tau_0\sim(\mathbb{P},\pi_0),\tau^*\sim(\mathbb{P},\pi^*)}\mathbb{T}\left(\tau^*,\tau_0\right) =&\mathbb{E}_{\tau_0\sim(\mathbb{P},\pi_0),\tau^*\sim(\mathbb{P},\pi^*)}\mathbb{T}\left(\tau^*,\tau_0\right) - \mathbb{E}_{\tau_0\sim(\hat{\mathbb{P}}_k,\pi_0),\tau^*\sim(\hat{\mathbb{P}}_k,\pi^*)}\mathbb{T}\left(\tau^*,\tau_0\right) &(29)\\ &+ \mathbb{E}_{\tau_0\sim(\hat{\mathbb{P}}_k,\pi_0),\tau^*\sim(\hat{\mathbb{P}}_k,\pi^*)}\mathbb{T}\left(\tau^*,\tau_0\right) - \mathbb{E}_{\tau_0\sim(\hat{\mathbb{P}}_k,\pi_0),\tau^*\sim(\hat{\mathbb{P}}_k,\pi^*)}\hat{\mathbb{T}}_k\left(\tau^*,\tau_0\right) &(30)\\ &+ \mathbb{E}_{\tau_0\sim(\hat{\mathbb{P}}_k,\pi_0),\tau^*\sim(\hat{\mathbb{P}}_k,\pi^*)}\hat{\mathbb{T}}_k\left(\tau^*,\tau_0\right). &(31)\end{aligned}$$

By Lemma A.3, we can upper bound the first term in the following way:

$$\mathbb{E}_{\tau_0\sim(\mathbb{P},\pi_0),\tau^*\sim(\mathbb{P},\pi^*)}\mathbb{T}\left(\tau^*,\tau_0\right) - \mathbb{E}_{\tau_0\sim(\hat{\mathbb{P}}_k,\pi_0),\tau^*\sim(\hat{\mathbb{P}}_k,\pi^*)}\mathbb{T}\left(\tau^*,\tau_0\right) \leq \mathbb{E}_{\tau_1\sim(\hat{\mathbb{P}}_k,\pi_1),\tau_2\sim(\hat{\mathbb{P}}_k,\pi_2)}[b_{\mathbb{P},k}(\tau_1)+b_{\mathbb{P},k}(\tau_2)]. \tag{32}$$

By Lemma A.1, we know that $\mathbb{T}(\tau_1,\tau_2)\in\mathcal{B}_{\mathbb{T},k}$ under event $\mathcal{E}$. Therefore,

$$\begin{aligned}&\mathbb{E}_{\tau_0\sim(\hat{\mathbb{P}}_k,\pi_0),\tau^*\sim(\hat{\mathbb{P}}_k,\pi^*)}\mathbb{T}\left(\tau^*,\tau_0\right) - \mathbb{E}_{\tau_0\sim(\hat{\mathbb{P}}_k,\pi_0),\tau^*\sim(\hat{\mathbb{P}}_k,\pi^*)}\hat{\mathbb{T}}_k\left(\tau^*,\tau_0\right) &(33)\\ \leq&\mathbb{E}_{\tau_0\sim(\hat{\mathbb{P}}_k,\pi_0),\tau^*\sim(\hat{\mathbb{P}}_k,\pi^*)}\max_{f_1,f_2\in\mathcal{B}_{T,k}}\left|f_1\left(\tau_1,\tau_2\right)-f_2\left(\tau^*,\tau_0\right)\right| &(34)\\ =&\mathbb{E}_{\tau_0\sim(\hat{\mathbb{P}}_k,\pi_0),\tau^*\sim(\hat{\mathbb{P}}_k,\pi^*)}b_{\mathbb{T},k}(\tau^*,\tau_0). &(35)\end{aligned}$$

Together with Inq. (28), we have

$$\mathbb{E}_{\tau_0\sim(\hat{\mathbb{P}}_k,\pi_0),\tau^*\sim(\hat{\mathbb{P}}_k,\pi^*)}\left(\hat{\mathbb{T}}_k\left(\tau^*,\tau_0\right)+b_{\mathbb{T},k}(\tau^*,\tau_0)+b_{\mathbb{P},k}(\tau^*)+b_{\mathbb{P},k}(\tau_0)\right)\geq\frac{1}{2},\forall\pi_0, \tag{36}$$

which indicates that $\pi^*\in\mathcal{S}_k$. ☐

Lemma A.5. *Under event $\mathcal{E}$, it holds that*

$$\sum_{k=1}^{K}b_{\mathbb{T},k}(\tau_{k,1},\tau_{k,2})\leq O(\sqrt{dK\log(K\mathcal{N}\left(\mathcal{F}_{\mathbb{T}},1/K,\|\cdot\|_\infty\right)/\delta)}). \tag{37}$$

Proof. This lemma follows from the direct application of Lemma E.2. Note that under event $\mathcal{E}$, we have $\max_{1\leq k\leq K}\operatorname{diam}(\mathcal{B}_{\mathbb{T},k}|_{(\tau_1,\tau_2)_{1:k}})\leq 2\sqrt{\beta_{\mathbb{T}}}$ by Lemma A.1, and $f(\tau_1,\tau_2)\in[0,1],\forall f\in\mathcal{F}_{\mathbb{T}}$. Therefore,

$$\sum_{k=1}^{K}b_{\mathbb{T},k}(\tau_{k,1},\tau_{k,2})=\sum_{k=1}^{K}\operatorname{diam}\left(\mathcal{B}_{\mathbb{T},k}|_{(\tau_{k,1},\tau_{k,2})}\right)\leq O(\sqrt{dK\log(K\mathcal{N}\left(\mathcal{F}_{\mathbb{T}},1/K,\|\cdot\|_\infty\right)/\delta)}). \tag{38}$$

☐

Lemma A.6. *Under event $\mathcal{E}$,*

$$\sum_{k=1}^{K}(b_{\mathbb{P},k}(\tau_{k,1})+b_{\mathbb{P},k}(\tau_{k,2}))\leq O(\sqrt{d_{\mathbb{P}}HK\log(K\mathcal{N}\left(\mathcal{F}_{\mathbb{T}},1/K,\|\cdot\|_\infty\right)/\delta)}). \tag{39}$$

Proof. This lemma follows from the direct application of Lemma E.2. Note that under event $\mathcal{E}$, we have $\sum_{i=1}^{2}\sum_{t=1}^{k}\sum_{h=1}^{H}\left(\left\langle\mathbb{P}\left(\cdot\mid s_{t,h,i},a_{t,h,i}\right)-\hat{\mathbb{P}}_k\left(\cdot\mid s_{t,h,i},a_{t,h,i}\right),V_{k,h,i}\right\rangle\right)^2\leq\beta_{\mathbb{P}}$ by Lemma A.2, and $f(s,a,V)\in[0,1],\forall f\in\mathcal{F}_{\mathbb{P}}$. Therefore,

$$\sum_{k=1}^{K}b_{\mathbb{P},k}(\tau_{k,1})+b_{\mathbb{P},k}(\tau_{k,2})=\sum_{i=1}^{2}\sum_{k=1}^{K}\sum_{h=1}^{H}b_{\mathbb{P},k}(s_{k,h,i},a_{k,h,i})\leq O(\sqrt{d_{\mathbb{P}}HK\log(K\mathcal{N}\left(\mathcal{F}_{\mathbb{T}},1/K,\|\cdot\|_\infty\right)/\delta)}). \tag{40}$$

☐

Proof of Theorem 4.1. By the definition of regret, we have

$$
\begin{aligned}
Reg(K) = &\sum_{k=1}^{K} \left(\mathbb{T}(\pi^*, \pi_{k,1}) + \mathbb{T}(\pi^*, \pi_{k,2}) - 1\right) && (41)\\
= &\sum_{k=1}^{K} \Big(\mathbb{E}_{\tau^*\sim(\hat{\mathbb{P}}_k,\pi^*),\tau_1\sim(\hat{\mathbb{P}}_k,\pi_{k,1})} \hat{\mathbb{T}}_k(\tau^*, \tau_1) + \mathbb{E}_{\tau^*\sim(\hat{\mathbb{P}}_k,\pi^*),\tau_2\sim(\hat{\mathbb{P}}_k,\pi_{k,2})} \hat{\mathbb{T}}_k(\tau^*, \tau_2) - 1\Big) && (42)\\
&+ \sum_{k=1}^{K} \Big(\mathbb{E}_{\tau^*\sim(\mathbb{P},\pi^*),\tau_1\sim(\mathbb{P},\pi_{k,1})} \mathbb{T}(\tau^*, \tau_1) - \mathbb{E}_{\tau^*\sim(\hat{\mathbb{P}}_k,\pi^*),\tau_1\sim(\hat{\mathbb{P}}_k,\pi_{k,1})} \mathbb{T}(\tau^*, \tau_1)\Big) && (43)\\
&+ \sum_{k=1}^{K} \Big(\mathbb{E}_{\tau^*\sim(\mathbb{P},\pi^*),\tau_2\sim(\mathbb{P},\pi_{k,2})} \mathbb{T}(\tau^*, \tau_2) - \mathbb{E}_{\tau^*\sim(\hat{\mathbb{P}}_k,\pi^*),\tau_2\sim(\hat{\mathbb{P}}_k,\pi_{k,2})} \mathbb{T}(\tau^*, \tau_2)\Big) && (44)\\
&+ \sum_{k=1}^{K} \mathbb{E}_{\tau^*\sim(\hat{\mathbb{P}}_k,\pi^*),\tau_1\sim(\hat{\mathbb{P}}_k,\pi_{k,1})} \Big(\mathbb{T}(\tau^*, \tau_1) - \hat{\mathbb{T}}_k(\tau^*, \tau_1)\Big) && (45)\\
&+ \sum_{k=1}^{K} \mathbb{E}_{\tau^*\sim(\hat{\mathbb{P}}_k,\pi^*),\tau_2\sim(\hat{\mathbb{P}}_k,\pi_{k,2})} \Big(\mathbb{T}(\tau^*, \tau_2) - \hat{\mathbb{T}}_k(\tau^*, \tau_2)\Big). && (46)
\end{aligned}
$$

Since $\pi_{k,1} \in \mathcal{S}_k$, we have

$$
\mathbb{E}_{\tau^*\sim(\hat{\mathbb{P}}_k,\pi^*),\tau_1\sim(\hat{\mathbb{P}}_k,\pi_{k,1})} \hat{\mathbb{T}}_k(\tau^*, \tau_1) - \frac{1}{2} \le \mathbb{E}_{\tau_1\sim(\hat{\mathbb{P}}_k,\pi_{k,1}),\tau^*\sim(\hat{\mathbb{P}}_k,\pi^*)} \left(b_{\mathbb{T},k}(\tau_1, \tau^*) + b_{\mathbb{P},k}(\tau_1) + b_{\mathbb{P},k}(\tau^*)\right), \qquad (47)
$$

where the inequality is due to the definition of $\mathcal{S}_k$.

By Lemma A.3, we know that

$$
\mathbb{E}_{\tau^*\sim(\mathbb{P},\pi^*),\tau_1\sim(\mathbb{P},\pi_{k,1})} \mathbb{T}(\tau^*, \tau_1) - \mathbb{E}_{\tau^*\sim(\hat{\mathbb{P}}_k,\pi^*),\tau_1\sim(\hat{\mathbb{P}}_k,\pi_{k,1})} \mathbb{T}(\tau^*, \tau_1) \le \mathbb{E}_{\tau^*\sim(\hat{\mathbb{P}}_k,\pi^*),\tau_1\sim(\hat{\mathbb{P}}_k,\pi_{k,1})} [b_{\mathbb{P},k}(\tau^*) + b_{\mathbb{P},k}(\tau_1)]. \qquad (48)
$$

From the definition of $b_{\mathbb{T},k}(\tau_1, \tau_2)$, we also know that

$$
\mathbb{E}_{\tau^*\sim(\hat{\mathbb{P}}_k,\pi^*),\tau_1\sim(\hat{\mathbb{P}}_k,\pi_{k,1})} \Big(\mathbb{T}(\tau^*, \tau_1) - \hat{\mathbb{T}}_k(\tau^*, \tau_1)\Big) \le \mathbb{E}_{\tau^*\sim(\hat{\mathbb{P}}_k,\pi^*),\tau_2\sim(\hat{\mathbb{P}}_k,\pi_{k,2})} b_{\mathbb{T},k}(\tau^*, \tau_1) \qquad (49)
$$

Similarly, for the policy $\pi_{k,2}$, we also have

$$
\begin{aligned}
&\mathbb{E}_{\tau^*\sim(\hat{\mathbb{P}}_k,\pi^*),\tau_2\sim(\hat{\mathbb{P}}_k,\pi_{k,2})} \hat{\mathbb{T}}_k(\tau^*, \tau_1) - \frac{1}{2} \le \mathbb{E}_{\tau_2\sim(\hat{\mathbb{P}}_k,\pi_{k,2}),\tau^*\sim(\hat{\mathbb{P}}_k,\pi^*)} \left(b_{\mathbb{T},k}(\tau_2, \tau^*) + b_{\mathbb{P},k}(\tau_2) + b_{\mathbb{P},k}(\tau^*)\right) && (50)\\
&\mathbb{E}_{\tau^*\sim(\mathbb{P},\pi^*),\tau_2\sim(\mathbb{P},\pi_{k,2})} \mathbb{T}(\tau^*, \tau_2) - \mathbb{E}_{\tau^*\sim(\hat{\mathbb{P}}_k,\pi^*),\tau_2\sim(\hat{\mathbb{P}}_k,\pi_{k,2})} \mathbb{T}(\tau^*, \tau_2) \le \mathbb{E}_{\tau^*\sim(\hat{\mathbb{P}}_k,\pi^*),\tau_2\sim(\hat{\mathbb{P}}_k,\pi_{k,2})} [b_{\mathbb{P},k}(\tau^*) + b_{\mathbb{P},k}(\tau_2)] && (51)\\
&\mathbb{E}_{\tau^*\sim(\hat{\mathbb{P}}_k,\pi^*),\tau_2\sim(\hat{\mathbb{P}}_k,\pi_{k,2})} \Big(\mathbb{T}(\tau^*, \tau_2) - \hat{\mathbb{T}}_k(\tau^*, \tau_2)\Big) \le \mathbb{E}_{\tau^*\sim(\hat{\mathbb{P}}_k,\pi^*),\tau_2\sim(\hat{\mathbb{P}}_k,\pi_{k,2})} b_{\mathbb{T},k}(\tau^*, \tau_2). && (52)
\end{aligned}
$$

Plugging the above inequalities back to Inq. (41), we have

$$
\begin{aligned}
Reg(K) \leq & \sum_{k=1}^{K} \mathbb{E}_{\tau_1 \sim (\hat{\mathbb{P}}_k, \pi_{k,1}), \tau^* \sim (\hat{\mathbb{P}}_k, \pi^*)} \left(b_{\mathbb{T},k}(\tau_1, \tau^*) + b_{\mathbb{P},k}(\tau_1) + b_{\mathbb{P},k}(\tau^*) \right) && (53)\\
& + \sum_{k=1}^{K} \mathbb{E}_{\tau_1 \sim (\hat{\mathbb{P}}_k, \pi_{k,1}), \tau^* \sim (\hat{\mathbb{P}}_k, \pi^*)} \left(b_{\mathbb{T},k}(\tau^*, \tau_1) + b_{\mathbb{P},k}(\tau_1) + b_{\mathbb{P},k}(\tau^*) \right) && (54)\\
& + \sum_{k=1}^{K} \mathbb{E}_{\tau_2 \sim (\hat{\mathbb{P}}_k, \pi_{k,2}), \tau^* \sim (\hat{\mathbb{P}}_k, \pi^*)} \left(b_{\mathbb{T},k}(\tau_2, \tau^*) + b_{\mathbb{P},k}(\tau_2) + b_{\mathbb{P},k}(\tau^*) \right) && (55)\\
& + \sum_{k=1}^{K} \mathbb{E}_{\tau_2 \sim (\hat{\mathbb{P}}_k, \pi_{k,2}), \tau^* \sim (\hat{\mathbb{P}}_k, \pi^*)} \left(b_{\mathbb{T},k}(\tau^*, \tau_2) + b_{\mathbb{P},k}(\tau_2) + b_{\mathbb{P},k}(\tau^*) \right) && (56)\\
\leq & \sum_{k=1}^{K} 2\mathbb{E}_{\tau_1 \sim (\hat{\mathbb{P}}_k, \pi_{k,1}), \tau_2 \sim (\hat{\mathbb{P}}_k, \pi_{k,2})} \left(b_{\mathbb{T},k}(\tau_1, \tau_2) + b_{\mathbb{P},k}(\tau_1) + b_{\mathbb{P},k}(\tau_2) \right), && (57)
\end{aligned}
$$

where the second inequality follows the fact that $\pi_{k,1}$ and $\pi_{k,2}$ are the maximizer of

$$
\mathbb{E}_{\tau_1 \sim (\hat{\mathbb{P}}_k, \pi_1), \tau_2 \sim (\hat{\mathbb{P}}_k, \pi_2)} \left(b_{\mathbb{T},k}(\tau_1, \tau_2) + b_{\mathbb{P},k}(\tau_1) + b_{\mathbb{P},k}(\tau_2) \right).
$$

By definition, we have $0 \leq b_{\mathbb{T},k}(\tau_1, \tau_2) \leq 1$ and $0 \leq b_{\mathbb{P},k}(\tau) \leq 1$. By Azuma's inequality, the following inequality holds with probability at least $1 - \delta/2$,

$$
\begin{aligned}
& \sum_{k=1}^{K} 2\mathbb{E}_{\tau_1 \sim (\hat{\mathbb{P}}_k, \pi_{k,1}), \tau_2 \sim (\hat{\mathbb{P}}_k, \pi_{k,2})} \left(b_{\mathbb{T},k}(\tau_1, \tau_2) + b_{\mathbb{P},k}(\tau_1) + b_{\mathbb{P},k}(\tau_2) \right) && (58)\\
\leq & \sum_{k=1}^{K} 2 \left(b_{\mathbb{T},k}(\tau_{k,1}, \tau_{k,2}) + b_{\mathbb{P},k}(\tau_{k,1}) + b_{\mathbb{P},k}(\tau_{k,2}) \right) + 4\sqrt{K \log(4/\delta)}. && (59)
\end{aligned}
$$

By Lemma A.5 and Lemma A.6, we can finally upper bound the total regret:

$$
Reg(K) \leq O\left(\sqrt{d_{\mathbb{P}} H K \log(\mathcal{N}\left(\mathcal{F}_{\mathbb{P}}, 1/K, \|\cdot\|_\infty\right)/\delta)} + \sqrt{d_{\mathbb{T}} K \log(\mathcal{N}\left(\mathcal{F}_{\mathbb{T}}, 1/K, \|\cdot\|_\infty\right)/\delta)} \right). \tag{60}
$$

□

B. RL with Once-per-episode Feedback

B.1. Algorithm

We estimate g^* by solving the following least-squares regression problem:

$$
\hat{g}_k = \arg\min_{g \in \mathcal{F}_{\mathbb{G}}} \sum_{t=1}^{k-1} [g(\tau_t) - y_t]^2. \tag{61}
$$

Then we construct the high-probability set for g^*:

$$
\mathcal{B}_{\mathbb{G},k} = \left\{ g \mid \sum_{t=1}^{k-1} \left(\hat{g}_k(\tau_t) - y_t \right)^2 \leq \beta_{\mathbb{G}} \right\}. \tag{62}
$$

The transition estimation $\hat{\mathbb{P}}_k$ is the minimizer of the least-square loss:

$$
\hat{\mathbb{P}}_k = \operatorname{argmin}_{\mathbb{P}' \in \mathcal{P}} \sum_{t=1}^{k-1} \sum_{h=1}^{H} \left(\left\langle \mathbb{P}'\left(\cdot \mid s_{t,h}, a_{t,h} \right), V_{k,h} \right\rangle - V_{k,h}(s_{k,h+1}) \right)^2, \tag{63}
$$

where the value target $V_{k,h}$ defined in Line 8 of Algorithm 4 is the value function that maximizes the uncertainty in state-action pair $(s_{k,h}, a_{k,h})$.

We define $L_k(\mathbb{P}_1, \mathbb{P}_2)$ as

$$L_k\left(\mathbb{P}_1, \mathbb{P}_2\right)=\sum_{t=1}^{k-1} \sum_{h=1}^{H}\left(\left\langle\mathbb{P}_1\left(\cdot \mid s_{t,h}, a_{t,h}\right)-\mathbb{P}_2\left(\cdot \mid s_{t,h}, a_{t,h}\right), V_{t,h}\right\rangle\right)^2. \tag{64}$$

We construct the high confidence set for transition $\mathbb{P}$, which is defined as

$$\mathcal{B}_{\mathbb{P},k}=\left\{\tilde{\mathbb{P}} \mid L_k(\tilde{\mathbb{P}}, \hat{\mathbb{P}}_k) \leq \beta_{\mathbb{P}}\right\}. \tag{65}$$

Similar with Algorithm 1, we calculate the policy set $\mathcal{S}_k$, which contains near-optimal policies with minor sub-optimality gap. Finally, we execute the most exploratory policy in $\mathcal{S}_k$.

Algorithm 4 RL with Trajectory Feedback

1: Set $\beta_{\mathbb{G}}=\beta_{\mathbb{P}}=8\log(2K\mathcal{N}\left(\mathcal{F}_{\mathbb{T}}, 1/K, \|\cdot\|_\infty\right)/\delta)$
2: **for** episode $k=1,\cdots,K$ **do**
3: Calculate the estimation g_k using least-squares regression (Eqn. (61))
4: Construct the high-confidence set $\mathcal{B}_{\mathbb{G},k}$ for the feedback function g^* (Eqn. (62))
5: Calculate the estimation $\hat{\mathbb{P}}_k$ using least-square regression (Eqn. 63).
6: Construct the high-confidence set $\mathcal{B}_{\mathbb{P},k}$ for transition $\mathbb{P}$ (Eqn. 65)
7: Define the bonus term $b_{\mathbb{P},k}(s,a)=\max_{\mathbb{P}_1,\mathbb{P}_2\in\mathcal{B}_{\mathbb{P},k}}\max_{V\in\mathcal{V}}(\mathbb{P}_1-\mathbb{P}_2)V(s,a)$, and $b_{\mathbb{P},k}(\tau)=\sum_{(s,a)\in\tau} b_{\mathbb{P},k}(s,a)$.
8: Define $V_{max,k,s,a}=\arg\max_{V\in\mathcal{V}}\max_{\mathbb{P}_1,\mathbb{P}_2\in\mathcal{B}_{\mathbb{P},k}}(\mathbb{P}_1-\mathbb{P}_2)V(s,a)$
9: Define the bonus term $b_{\mathbb{G},k}(\tau)=\max_{g_1,g_2\in\mathcal{B}_{\mathbb{G},k}}|g_1(\tau)-g_2(\tau)|$
10: Set $\mathcal{S}_k=\left\{\pi \mid \mathbb{E}_{\tau\sim(\hat{\mathbb{P}}_k,\pi),\tau_0\sim(\hat{\mathbb{P}}_k,\pi_0)}\left(\hat{g}_k(\tau)-\hat{g}_k(\tau_0)+b_{\mathbb{G},k}(\tau)+b_{\mathbb{G},k}(\tau_0)+b_{\mathbb{P},k}(\tau)+b_{\mathbb{P},k}(\tau_0)\right)\geq 0, \forall\pi_0\in\Pi\right\}$
11: Compute policy $\pi_k=\arg\max_{\pi\in\mathcal{S}_k}\mathbb{E}_{\tau\sim(\hat{\mathbb{P}}_k,\pi)}\left(b_{\mathbb{G},k}(\tau)+b_{\mathbb{P},k}(\tau)\right)$
12: Execute the policy π_k for one episode, then observe the trajectory τ_k and the feedback y_k
13: **end for**

B.2. Theoretical Results

Theorem B.1 (Restatement of Theorem 4.3). *With probability at least $1-\delta$, the regret of Algorithm 4 is upper bounded by*

$$Reg(K)\leq\tilde{O}(\sqrt{d_{\mathbb{P}}HK\log(\mathcal{N}\left(\mathcal{F}_{\mathbb{P}}, 1/K, \|\cdot\|_\infty\right)/\delta)}+\sqrt{d_{\mathbb{G}}K\log(\mathcal{N}\left(\mathcal{F}_{\mathbb{G}}, 1/K, \|\cdot\|_\infty\right)/\delta)}).$$

Proof. The proof shares almost the same idea with the analysis for PbRL. Therefore, we only explain the differences. Similar to Lemmas A.1 and A.2, we can show the following events happen with probability at least $1-\delta$,

$$\sum_{t=1}^{k-1}\left(\hat{g}_k-g^*\right)^2\left(\tau_t\right)\leq\beta_{\mathbb{T}}, \quad \sum_{t=1}^{k-1}\sum_{h=1}^{H}\left(\left\langle\mathbb{P}\left(\cdot \mid s_{t,h}, a_{t,h}\right)-\hat{\mathbb{P}}_k\left(\cdot \mid s_{t,h}, a_{t,h}\right), V_{t,h}\right\rangle\right)^2\leq\beta_{\mathbb{P}}. \tag{66}$$

Denote the above event as $\mathcal{E}$. Under event $\mathcal{E}$, we also know that $\pi^*\in\mathcal{S}_k$. Therefore, we upper bound the regret in the following way:

$$Reg(K) = \sum_{k=1}^{K} V^*(s_1) - V^{\pi_k}(s_1) \tag{67}$$

$$= \sum_{k=1}^{K} \mathbb{E}_{\tau^* \sim (\hat{\mathbb{P}}_k, \pi^*), \tau \sim (\hat{\mathbb{P}}_k, \pi_k)} \left(\hat{g}_k(\tau^*) - \hat{g}_k(\tau)\right) \tag{68}$$

$$+ \sum_{k=1}^{K} \mathbb{E}_{\tau^* \sim (\mathbb{P}, \pi^*), \tau \sim (\mathbb{P}, \pi_k)} \left(g^*(\tau^*) - g^*(\tau)\right) - \mathbb{E}_{\tau^* \sim (\hat{\mathbb{P}}_k, \pi^*), \tau \sim (\hat{\mathbb{P}}_k, \pi_k)} \left(g^*(\tau^*) - g^*(\tau)\right) \tag{69}$$

$$+ \sum_{k=1}^{K} \mathbb{E}_{\tau^* \sim (\hat{\mathbb{P}}_k, \pi^*), \tau \sim (\hat{\mathbb{P}}_k, \pi_k)} \left(\left(g^*(\tau^*) - g^*(\tau)\right) - \left(\hat{g}_k(\tau^*) - \hat{g}_k(\tau)\right)\right). \tag{70}$$

Since $\pi_k \in \mathcal{S}_k$, we have

$$\mathbb{E}_{\tau \sim (\hat{\mathbb{P}}_k, \pi_k), \tau^* \sim (\hat{\mathbb{P}}_k, \pi^*)} \left(\hat{g}_k(\tau) - \hat{g}_k(\tau^*)\right) \leq \mathbb{E}_{\tau \sim (\hat{\mathbb{P}}_k, \pi), \tau^* \sim (\hat{\mathbb{P}}_k, \pi^*)} \left(b_{\mathbb{G},k}(\tau) + b_{\mathbb{G},k}(\tau^*) + b_{\mathbb{P},k}(\tau) + b_{\mathbb{P},k}(\tau^*)\right) \tag{71}$$

Similarly, by Lemma A.3, we know that

$$\sum_{k=1}^{K} \mathbb{E}_{\tau^* \sim (\mathbb{P}, \pi^*), \tau \sim (\mathbb{P}, \pi_k)} \left(g^*(\tau^*) - g^*(\tau)\right) - \mathbb{E}_{\tau^* \sim (\hat{\mathbb{P}}_k, \pi^*), \tau \sim (\hat{\mathbb{P}}_k, \pi_k)} \left(g^*(\tau^*) - g^*(\tau)\right) \tag{72}$$

$$\leq \mathbb{E}_{\tau^* \sim (\hat{\mathbb{P}}_k, \pi^*), \tau \sim (\hat{\mathbb{P}}_k, \pi_k)} [b_{\mathbb{P},k}(\tau^*) + b_{\mathbb{P},k}(\tau)]. \tag{73}$$

From the definition of $b_{\mathbb{G},k}(\tau)$, we also know that

$$\mathbb{E}_{\tau \sim (\hat{\mathbb{P}}_k, \pi_k)} \left(g^*(\tau) - \hat{g}(\tau)\right) \leq \mathbb{E}_{\tau \sim (\hat{\mathbb{P}}_k, \pi_k)} b_{\mathbb{G},k}(\tau), \tag{74}$$

$$\mathbb{E}_{\tau^* \sim (\hat{\mathbb{P}}_k, \pi^*)} \left(g^*(\tau^*) - \hat{g}(\tau^*)\right) \leq \mathbb{E}_{\tau \sim (\hat{\mathbb{P}}_k, \pi_k)} b_{\mathbb{G},k}(\tau^*). \tag{75}$$

Plugging the above inequalities back to Inq. (67), we have

$$Reg(K) \leq \sum_{k=1}^{K} \mathbb{E}_{\tau \sim (\hat{\mathbb{P}}_k, \pi_k), \tau^* \sim (\hat{\mathbb{P}}_k, \pi^*)} \left(b_{\mathbb{G},k}(\tau) + b_{\mathbb{G},k}(\tau^*) + b_{\mathbb{P},k}(\tau) + b_{\mathbb{P},k}(\tau^*)\right) \tag{76}$$

$$+ \sum_{k=1}^{K} \mathbb{E}_{\tau \sim (\hat{\mathbb{P}}_k, \pi_k), \tau^* \sim (\hat{\mathbb{P}}_k, \pi^*)} \left(b_{\mathbb{G},k}(\tau) + b_{\mathbb{G},k}(\tau^*) + b_{\mathbb{P},k}(\tau) + b_{\mathbb{P},k}(\tau^*)\right) \tag{77}$$

$$\leq \sum_{k=1}^{K} 4\mathbb{E}_{\tau \sim (\hat{\mathbb{P}}_k, \pi_k)} \left(b_{\mathbb{G},k}(\tau) + b_{\mathbb{P},k}(\tau)\right), \tag{78}$$

where the second inequality follows the fact that π_k is the maximizer of $\mathbb{E}_{\tau \sim (\hat{\mathbb{P}}_k, \pi)} \left(b_{\mathbb{G},k}(\tau) + b_{\mathbb{P},k}(\tau)\right)$. By definition, we have $0 \leq b_{\mathbb{G},k}(\tau) \leq 1$ and $0 \leq b_{\mathbb{P},k}(\tau) \leq 1$. By Azuma's inequality, the following inequality holds with probability at least $1 - \delta/2$,

$$\sum_{k=1}^{K} 4\mathbb{E}_{\tau \sim (\hat{\mathbb{P}}_k, \pi_k)} \left(b_{\mathbb{G},k}(\tau) + b_{\mathbb{P},k}(\tau)\right) \tag{79}$$

$$\leq \sum_{k=1}^{K} 4\left(b_{\mathbb{G},k}(\tau_k) + b_{\mathbb{P},k}(\tau_k)\right) + 8\sqrt{K \log(4/\delta)}. \tag{80}$$

We upper bound the summation of bonus with the help of Lemma E.2. Finally, we have

$$Reg(K) \leq O\left(\sqrt{d_{\mathbb{P}} HK \log(\mathcal{N}\left(\mathcal{F}_{\mathbb{P}}, 1/K, \|\cdot\|_{\infty}\right)/\delta)} + \sqrt{d_{\mathbb{G}} K \log(\mathcal{N}\left(\mathcal{F}_{\mathbb{T}}, 1/K, \|\cdot\|_{\infty}\right)/\delta)}\right). \tag{81}$$

□

C. Proof of Theorem 4.5

Proof. For any $\tau = (s_1, a_1, \cdots, s_H, a_H)$, let $g^*(\tau) = \sum_{h=1}^{H} r(s_h, a_h)$. Then the RL with once-per-episode feedback problem reduces to the traditional RL (RL with reward signals). Thus, the lower bound $\Omega(\tilde{d}_{\mathbb{P}}\sqrt{K})$[1] established in Zhou et al. (2021a) immediately implies the same regret lower bound for the RL with once-per-episode feedback problem. Meanwhile, by regarding the trajectory as an "arm", the lower bound $\Omega(\tilde{d}_{\mathbb{T}}\sqrt{K})$ for linear bandits implies the lower bound $\Omega(\tilde{d}_{\mathbb{T}}\sqrt{K})$ for our setting. Putting these two lower bound together, we obtain $Reg(K) \geq \Omega(\max\{\tilde{d}_{\mathbb{P}}\sqrt{K}, \tilde{d}_{\mathbb{T}}\sqrt{K}\})$, which equivalents to $Reg(K) \geq \Omega(\tilde{d}_{\mathbb{P}}\sqrt{K} + \tilde{d}_{\mathbb{T}}\sqrt{K})$. Therefore, we finish the proof. $\square$

D. Proof of Theorem 5.1

We also need the following concentration lemmas to guarantee that the true preference $\mathbb{T}(\cdot,\cdot) \in \mathcal{B}_{\mathbb{T},k}$ with high probability and the true transition kernel $\mathbb{P}(s'|s,a) \in \mathcal{B}_{\mathbb{P},k}$ with high probability, respectively.

Lemma D.1. *Fix $\delta \in (0,1)$, with probability at least $1-\delta$, for all $k \in [K]$,*

$$\sum_{t=1}^{k-1}\sum_{i=1}^{n}\sum_{j=i+1}^{n}\left(\hat{\mathbb{T}}_k - \mathbb{T}\right)^2(\tau_{t,i}, \tau_{t,j}) \leq \beta_{\mathbb{T}}. \tag{82}$$

Proof. This lemma can be proved by the direct application of Lemma E.1. $\square$

Lemma D.2. *Fix $\delta \in (0,1)$, with probability at least $1-\delta$, for all $k \in [K]$,*

$$L_k\left(\mathbb{P}, \hat{\mathbb{P}}_k\right) = \sum_{i=1}^{n}\sum_{t=1}^{k-1}\sum_{h=1}^{H}\left(\left\langle \mathbb{P}\left(\cdot \mid s_{t,h,i}, a_{t,h,i}\right) - \hat{\mathbb{P}}_k\left(\cdot \mid s_{t,h,i}, a_{t,h,i}\right), V_{t,h,i}\right\rangle\right)^2 \leq \beta_{\mathbb{P}}. \tag{83}$$

Proof. This lemma can be proved by the direct application of Lemma E.1. $\square$

With slight abuse of notation, we denote the high-probability event in Lemmas D.1 and D.2 as $\mathcal{E}$.

Lemma D.3. *Under event $\mathcal{E}$, for any two policies π_1, π_2 and scalar function $f : \mathrm{Traj} \times \mathrm{Traj} \to [0,1]$, we have*

$$\mathbb{E}_{\tau_1\sim(\mathbb{P},\pi_1),\tau_2\sim(\mathbb{P},\pi_2)}[f(\tau_1,\tau_2)] - \mathbb{E}_{\tau_1\sim(\hat{\mathbb{P}}_k,\pi_1),\tau_2\sim(\hat{\mathbb{P}}_k,\pi_2)}[f(\tau_1,\tau_2)] \leq \mathbb{E}_{\tau_1\sim(\hat{\mathbb{P}}_k,\pi_1),\tau_2\sim(\hat{\mathbb{P}}_k,\pi_2)}[b_{\mathbb{P},k}(\tau_1) + b_{\mathbb{P},k}(\tau_2)]. \tag{84}$$

Proof. The proof is the same as that of Lemma A.3 and we omit it here to avoid repetition. $\square$

Lemma D.4. *Under event $\mathcal{E}$, we have $\pi^* \in \mathcal{S}_k$.*

Proof. The proof is the same as that of Lemma A.4 and we omit it here to avoid repetition. $\square$

With these lemmas, we are ready to provide the proof of Theorem 5.1.

[1] Their lower bound is $\Omega(\tilde{d}_{\mathbb{P}}\sqrt{H^3K})$ because they consider the setting that the total reward is bounded by H and the transition kernel is time-inhomogeneous.

Proof of Theorem 5.1. By the definition of regret, we have

$$
\begin{aligned}
Reg(K) &= \sum_{k=1}^{K}\sum_{i=1}^{n}\left(\mathbb{T}(\pi^*, \pi_{k,i}) - \frac{1}{2}\right) && (85)\\
&= \sum_{k=1}^{K}\sum_{i=1}^{n}\left(\mathbb{E}_{\tau^*\sim(\hat{\mathbb{P}}_k,\pi^*),\tau_i\sim(\hat{\mathbb{P}}_k,\pi_{k,i})}\hat{\mathbb{T}}(\tau^*,\tau_i) - \frac{1}{2}\right) && (86)\\
&\quad + \sum_{k=1}^{K}\sum_{i=1}^{n}\left(\mathbb{E}_{\tau^*\sim(\mathbb{P},\pi^*),\tau_i\sim(\mathbb{P},\pi_{k,i})}\mathbb{T}(\tau^*,\tau_i) - \mathbb{E}_{\tau^*\sim(\hat{\mathbb{P}}_k,\pi^*),\tau_i\sim(\hat{\mathbb{P}}_k,\pi_{k,i})}\mathbb{T}(\tau^*,\tau_i)\right) && (87)\\
&\quad + \sum_{k=1}^{K}\sum_{i=1}^{n}\mathbb{E}_{\tau^*\sim(\hat{\mathbb{P}}_k,\pi^*),\tau_i\sim(\hat{\mathbb{P}}_k,\pi_{k,i})}\left(\mathbb{T}(\tau^*,\tau_i) - \hat{\mathbb{T}}(\tau^*,\tau_i)\right). && (88)
\end{aligned}
$$

Note that $\pi_{k,i}\in\mathcal{S}_k$ for all $i\in[n]$, we have

$$
\mathbb{E}_{\tau^*\sim(\hat{\mathbb{P}}_k,\pi^*),\tau_i\sim(\hat{\mathbb{P}}_k,\pi_{k,i})}\hat{\mathbb{T}}(\tau^*,\tau_i) - \frac{1}{2} \le \mathbb{E}_{\tau_i\sim(\hat{\mathbb{P}}_k,\pi_{k,1}),\tau^*\sim(\hat{\mathbb{P}}_k,\pi^*)}\left(b_{\mathbb{T},k}(\tau_i,\tau^*) + b_{\mathbb{P},k}(\tau_i) + b_{\mathbb{P},k}(\tau^*)\right), \tag{89}
$$

By Lemma D.3, we have that

$$
\mathbb{E}_{\tau^*\sim(\mathbb{P},\pi^*),\tau_i\sim(\mathbb{P},\pi_{k,i})}\mathbb{T}(\tau^*,\tau_i) - \mathbb{E}_{\tau^*\sim(\hat{\mathbb{P}}_k,\pi^*),\tau_i\sim(\hat{\mathbb{P}}_k,\pi_{k,i})}\mathbb{T}(\tau^*,\tau_i) \le \mathbb{E}_{\tau^*\sim(\hat{\mathbb{P}}_k,\pi^*),\tau_i\sim(\hat{\mathbb{P}}_k,\pi_{k,i})}[b_{\mathbb{P},k}(\tau^*) + b_{\mathbb{P},k}(\tau_i)]. \tag{90}
$$

By the definition of $b_{\mathbb{T},k}$, we also know that

$$
\mathbb{E}_{\tau^*\sim(\hat{\mathbb{P}}_k,\pi^*),\tau_i\sim(\hat{\mathbb{P}}_k,\pi_{k,i})}\left(\mathbb{T}(\tau^*,\tau_i) - \hat{\mathbb{T}}(\tau^*,\tau_i)\right) \le b_{\mathbb{T},k}(\tau^*,\tau_i). \tag{91}
$$

Plugging the above inequalities back to Inq. (85), we have

$$
\begin{aligned}
Reg(K) &\le \sum_{k=1}^{K}\sum_{i=1}^{n}\mathbb{E}_{\tau_i\sim(\hat{\mathbb{P}}_k,\pi_{k,i}),\tau^*\sim(\hat{\mathbb{P}}_k,\pi^*)}\left(b_{\mathbb{T},k}(\tau_i,\tau^*) + b_{\mathbb{P},k}(\tau_i) + b_{\mathbb{P},k}(\tau^*)\right) && (92)\\
&\quad + \sum_{k=1}^{K}\sum_{i=1}^{n}\mathbb{E}_{\tau_i\sim(\hat{\mathbb{P}}_k,\pi_{k,i}),\tau^*\sim(\hat{\mathbb{P}}_k,\pi^*)}\left(b_{\mathbb{T},k}(\tau^*,\tau_i) + b_{\mathbb{P},k}(\tau_i) + b_{\mathbb{P},k}(\tau^*)\right) && (93)\\
&= 2\sum_{k=1}^{K}\sum_{i=1}^{n}\mathbb{E}_{\tau_i\sim(\hat{\mathbb{P}}_k,\pi_{k,i}),\tau^*\sim(\hat{\mathbb{P}}_k,\pi^*)}\left(b_{\mathbb{T},k}(\tau^*,\tau_i) + b_{\mathbb{P},k}(\tau_i) + b_{\mathbb{P},k}(\tau^*)\right), && (94)
\end{aligned}
$$

For any $(k,i)\in[K]\times[n]$, we have

$$
\begin{aligned}
&\sum_{i=1}^{n}\mathbb{E}_{\tau_i\sim(\hat{\mathbb{P}}_k,\pi_{k,i}),\tau^*\sim(\hat{\mathbb{P}}_k,\pi^*)}\left(b_{\mathbb{T},k}(\tau^*,\tau_i) + b_{\mathbb{P},k}(\tau_i) + b_{\mathbb{P},k}(\tau^*)\right) && (95)\\
=&\frac{1}{n-1}\cdot\sum_{i=1}^{n}\sum_{j\ne i}\mathbb{E}_{\tau_j\sim(\hat{\mathbb{P}}_k,\pi_{k,j}),\tau^*\sim(\hat{\mathbb{P}}_k,\pi^*)}\left(b_{\mathbb{T},k}(\tau^*,\tau_j) + b_{\mathbb{P},k}(\tau_j) + b_{\mathbb{P},k}(\tau^*)\right). && (96)
\end{aligned}
$$

Note that

$$
\begin{aligned}
(\pi_{k,1},\pi_{k,2},\cdots,\pi_{k,n}) &= \mathop{\arg\max}_{\pi_1,\pi_2,\cdots,\pi_n\in\mathcal{S}_k}\sum_{i=1}^{n}\sum_{j=i+1}^{n}\mathbb{E}_{\tau_i\sim(\hat{\mathbb{P}}_k,\pi_i),\tau_j\sim(\hat{\mathbb{P}}_k,\pi_j)}\left(b_{\mathbb{T},k}(\tau_i,\tau_j) + b_{\mathbb{P},k}(\tau_i) + b_{\mathbb{P},k}(\tau_j)\right) && (97)\\
&= \mathop{\arg\max}_{\pi_1,\pi_2,\cdots,\pi_n\in\mathcal{S}_k}\frac{1}{2}\sum_{i=1}^{n}\sum_{j\ne i}\mathbb{E}_{\tau_i\sim(\hat{\mathbb{P}}_k,\pi_i),\tau_j\sim(\hat{\mathbb{P}}_k,\pi_j)}\left(b_{\mathbb{T},k}(\tau_i,\tau_j) + b_{\mathbb{P},k}(\tau_i) + b_{\mathbb{P},k}(\tau_j)\right) && (98)\\
&= \mathop{\arg\max}_{\pi_1,\pi_2,\cdots,\pi_n\in\mathcal{S}_k}\sum_{i=1}^{n}\sum_{j\ne i}\mathbb{E}_{\tau_i\sim(\hat{\mathbb{P}}_k,\pi_i),\tau_j\sim(\hat{\mathbb{P}}_k,\pi_j)}\left(b_{\mathbb{T},k}(\tau_i,\tau_j) + b_{\mathbb{P},k}(\tau_i) + b_{\mathbb{P},k}(\tau_j)\right), && (99)
\end{aligned}
$$

together with the fact that $\pi^* \in \mathcal{S}_k$ (Lemma D.4), we have for any $i \in [n]$,

$$\sum_{j\neq i} \mathbb{E}_{\tau_j\sim(\hat{\mathbb{P}}_k,\pi_{k,j}),\tau^*\sim(\hat{\mathbb{P}}_k,\pi^*)} \left(b_{\mathbb{T},k}(\tau^*,\tau_j) + b_{\mathbb{P},k}(\tau_j) + b_{\mathbb{P},k}(\tau^*)\right) \tag{100}$$

$$+ \sum_{j,l\neq i} \mathbb{E}_{\tau_j\sim(\hat{\mathbb{P}}_k,\pi_{k,j}),\tau_l\sim(\hat{\mathbb{P}}_k,\pi_{k,l})} \left(b_{\mathbb{T},k}(\tau_j,\tau_l) + b_{\mathbb{P},k}(\tau_j) + b_{\mathbb{P},k}(\tau_l)\right) \tag{101}$$

$$\leq \sum_{i=1}^{n}\sum_{j\neq i} \mathbb{E}_{\tau_i\sim(\hat{\mathbb{P}}_k,\pi_{k,i}),\tau_j\sim(\hat{\mathbb{P}}_k,\pi_j)} \left(b_{\mathbb{T},k}(\tau_i,\tau_j) + b_{\mathbb{P},k}(\tau_i) + b_{\mathbb{P},k}(\tau_j)\right), \tag{102}$$

which equivalents to

$$\sum_{j\neq i} \mathbb{E}_{\tau_j\sim(\hat{\mathbb{P}}_k,\pi_{k,j}),\tau^*\sim(\hat{\mathbb{P}}_k,\pi^*)} \left(b_{\mathbb{T},k}(\tau^*,\tau_j) + b_{\mathbb{P},k}(\tau_j) + b_{\mathbb{P},k}(\tau^*)\right) \tag{103}$$

$$\leq \sum_{j\neq i} \mathbb{E}_{\tau_j\sim(\hat{\mathbb{P}}_k,\pi_{k,j}),\tau_i\sim(\hat{\mathbb{P}}_k,\pi_i)} \left(b_{\mathbb{T},k}(\tau_i,\tau_j) + b_{\mathbb{P},k}(\tau_i) + b_{\mathbb{P},k}(\tau_i)\right). \tag{104}$$

Taking summation over $i \in [n]$ gives that

$$\sum_{i=1}^{n}\sum_{j\neq i} \mathbb{E}_{\tau_i\sim(\hat{\mathbb{P}}_k,\pi_{k,j}),\tau^*\sim(\hat{\mathbb{P}}_k,\pi^*)} \left(b_{\mathbb{T},k}(\tau^*,\tau_j) + b_{\mathbb{P},k}(\tau_j) + b_{\mathbb{P},k}(\tau^*)\right) \tag{105}$$

$$\leq \sum_{i=1}^{n}\sum_{j\neq i} \mathbb{E}_{\tau_i\sim(\hat{\mathbb{P}}_k,\pi_{k,j}),\tau_j\sim(\hat{\mathbb{P}}_k,\pi_j)} \left(b_{\mathbb{T},k}(\tau_i,\tau_j) + b_{\mathbb{P},k}(\tau_j) + b_{\mathbb{P},k}(\tau_i)\right). \tag{106}$$

By definition, we have $0 \leq b_{\mathbb{T},k}(\tau_1,\tau_2) \leq H$ and $0 \leq b_{\mathbb{P},k}(\tau) \leq 1$. By Azuma's inequality, the following inequality holds with probability at least $1-\delta/2$,

$$\sum_{k=1}^{K}\sum_{i=1}^{n}\sum_{j\neq i} \mathbb{E}_{\tau_i\sim(\hat{\mathbb{P}}_k,\pi_{k,j}),\tau_j\sim(\hat{\mathbb{P}}_k,\pi_j)} \left(b_{\mathbb{T},k}(\tau_i,\tau_j) + b_{\mathbb{P},k}(\tau_j) + b_{\mathbb{P},k}(\tau_i)\right) \tag{107}$$

$$\leq \sum_{k=1}^{K}\sum_{i=1}^{n}\sum_{j\neq i} \left(b_{\mathbb{T},k}(\tau_{k,i},\tau_{k,j}) + b_{\mathbb{P},k}(\tau_{k,j}) + b_{\mathbb{P},k}(\tau_{k,i})\right) + 2Hn\sqrt{K\log(4/\delta)} \tag{108}$$

$$=2\sum_{k=1}^{K}\sum_{i=1}^{n}\sum_{j=i+1}^{n} \left(b_{\mathbb{T},k}(\tau_{k,i},\tau_{k,j}) + b_{\mathbb{P},k}(\tau_{k,j}) + b_{\mathbb{P},k}(\tau_{k,i})\right) + 2Hn\sqrt{K\log(4/\delta)} \tag{109}$$

Lemma D.5. *Under event $\mathcal{E}$, it holds that*

$$\sum_{k=1}^{K}\sum_{i=1}^{n}\sum_{j=i+1}^{n} b_{\mathbb{T},k}(\tau_{k,i},\tau_{k,j}) \leq O(\sqrt{dKn^2\log(K\mathcal{N}\left(\mathcal{F}_{\mathbb{T}},1/(Kn^2),\|\cdot\|_\infty\right)/\delta)}). \tag{110}$$

Proof. This lemma follows from the direct application of Lemma E.2. Note that under event $\mathcal{E}$, we have $\max_{1\leq k\leq K}\operatorname{diam}(\mathcal{B}_{\mathbb{T},k}|_{(\tau_1,\tau_2)_{1:k}}) \leq 2\sqrt{\beta_{\mathbb{T}}}$ by Lemma D.1, and $f(\tau_1,\tau_2) \in [0,1], \forall f \in \mathcal{F}_{\mathbb{T}}$. Therefore,

$$\sum_{k=1}^{K}\sum_{i=1}^{n}\sum_{j=i+1}^{n} b_{\mathbb{T},k}(\tau_{k,i},\tau_{k,j}) = \sum_{k=1}^{K}\sum_{i=1}^{n}\sum_{j\neq i} \operatorname{diam}\left(\mathcal{B}_{\mathbb{T},k}|_{(\tau_{k,1},\tau_{k,2})}\right) \leq O(\sqrt{dKn^2\log(K\mathcal{N}\left(\mathcal{F}_{\mathbb{T}},1/(Kn^2),\|\cdot\|_\infty\right)/\delta)}). \tag{111}$$

□

Lemma D.6. *Under event $\mathcal{E}$, it holds that*

$$\sum_{k=1}^{K}\sum_{i=1}^{n} b_{\mathbb{P},k}(\tau_{k,i}) \leq O(\sqrt{d_{\mathbb{P}}HKn\log(K\mathcal{N}\left(\mathcal{F}_{\mathbb{P}},1/(Kn),\|\cdot\|_\infty\right)/\delta)}). \tag{112}$$

Proof. This lemma follows from the direct application of Lemma E.2. Note that under event $\mathcal{E}$, we have $\sum_{i=1}^{n}\sum_{t=1}^{k}\sum_{h=1}^{H}\left(\left\langle \mathbb{P}\left(\cdot \mid s_{t,h,i},a_{t,h,i}\right)-\hat{\mathbb{P}}_k\left(\cdot \mid s_{t,h,i},a_{t,h,i}\right),V_{k,h,i}\right\rangle\right)^2 \leq \beta_{\mathbb{P}}$ by Lemma D.2, and $f(s,a,V) \in [0,1], \forall f \in \mathcal{F}_{\mathbb{P}}$. Therefore,

$$\sum_{k=1}^{K}\sum_{i=1}^{n} b_{\mathbb{P},k}(\tau_{k,i}) = \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{h=1}^{H} b_{\mathbb{P},k}(s_{k,h,i},a_{k,h,i}) \leq O(\sqrt{d_{\mathbb{P}}HKn\log(K\mathcal{N}\left(\mathcal{F}_{\mathbb{P}},1/(Kn),\|\cdot\|_{\infty}\right)/\delta)}). \tag{113}$$

□

By Lemmas D.5 and D.6, we can finally upper bound the total regret:

$$Reg(K) \leq O\left(\sqrt{d_{\mathbb{P}}HnK\cdot\log(\mathcal{N}\left(\mathcal{F}_{\mathbb{P}},1/(Kn),\|\cdot\|_{\infty}\right)/\delta)} + \sqrt{d_{\mathbb{T}}K\cdot\log(\mathcal{N}\left(\mathcal{F}_{\mathbb{T}},1/(Kn^2),\|\cdot\|_{\infty}\right)/\delta)}\right). \tag{114}$$

□

E. Auxiliary Lemmas

Let $(X_p,Y_p)_{p=1,2,\dots}$ be a sequence of random elements, $X_p \in \mathcal{X}$ for some measurable set $\mathcal{X}$ and $Y_p \in \mathbb{R}$. Let $\mathcal{F}$ be a subset of the set of real-valued measurable functions with domain $\mathcal{X}$. Let $\mathbb{F} = (\mathbb{F}_p)_{p=0,1,\cdots}$ be a filtration such that for all $p \geq 1$, $(X_1,Y_1,\cdots,X_{p-1},Y_{p-1},X_p)$ is $\mathbb{F}_{p-1}$ measurable and such that there exists some function $f_\star \in \mathcal{F}$ such that $\mathbb{E}\left[Y_p \mid \mathbb{F}_{p-1}\right] = f_*\left(X_p\right)$ holds for all $p \geq 1$. The (nonlinear) least square predictor given $(X_1,Y_1,\cdots,X_t,Y_t)$ is $\hat{f}_t = \operatorname{argmin}_{f\in\mathcal{F}}\sum_{p=1}^{t}\left(f\left(X_p\right)-Y_p\right)^2$. We say that Z is conditionally ρ-subgaussion given the σ-algebra $\mathbb{F}$ is for all $\lambda \in \mathbb{R}$, $\log\mathbb{E}[\exp(\lambda Z) \mid \mathbb{F}] \leq \frac{1}{2}\lambda^2\rho^2$. For $\alpha > 0$, let N_α be the $\|\cdot\|_\infty$-covering number of $\mathcal{F}$ at scale α. For $\beta > 0$, define

$$\mathcal{F}_t(\beta) = \left\{f \in \mathcal{F} : \sum_{p=1}^{t}\left(f\left(X_p\right)-\hat{f}_t\left(X_p\right)\right)^2 \leq \beta\right\}. \tag{115}$$

Lemma E.1. *(Theorem 5 of (Ayoub et al., 2020)). Let $\mathbb{F}$ be the filtration defined above and assume that the functions in $\mathcal{F}$ are bounded by the positive constant $C > 0$. Assume that for each $s \geq 1$, $\left(Y_p - f_*\left(X_p\right)\right)$ is conditionally σ-subgaussian given $\mathbb{F}_{p-1}$. Then, for any $\alpha > 0$, with probability $1-\delta$, for all $t \geq 1$, $f_* \in \mathcal{F}_t\left(\beta_t(\delta,\alpha)\right)$, where*

$$\beta_t(\delta,\alpha) = 8\sigma^2\log\left(2N_\alpha/\delta\right) + 4t\alpha\left(C + \sqrt{\sigma^2\log(4t(t+1)/\delta)}\right). \tag{116}$$

Lemma E.2. *(Lemma 5 of (Russo & Van Roy, 2014)). Let $\mathcal{F} \in B_\infty(\mathcal{X},C)$ be a set of functions bounded by $C > 0$, $(\mathcal{F}_t)_{t\geq 1}$ and $(x_t)_{t\geq 1}$ be sequences such that $\mathcal{F}_t \subset \mathcal{F}$ and $x_t \in \mathcal{X}$ hold for $t \geq 1$. Let $\left.\mathcal{F}\right|_{x_{1:t}} = \left\{\left(f\left(x_1\right),\ldots,f\left(x_t\right)\right) : f \in \mathcal{F}\right\}\left(\subset \mathbb{R}^t\right)$ and for $S \subset \mathbb{R}^t$, let $\operatorname{diam}(S) = \sup_{u,v\in S}\|u-v\|_2$ be the diameter of S. Then, for any $T \geq 1$ and $\alpha > 0$ it holds that*

$$\sum_{t=1}^{T}\operatorname{diam}\left(\left.\mathcal{F}_t\right|_{x_t}\right) \leq \alpha + C(d\wedge T) + 2\delta_T\sqrt{dT}, \tag{117}$$

where $\delta_T = \max_{1\leq t\leq T}\operatorname{diam}\left(\left.\mathcal{F}_t\right|_{x_{1:t}}\right)$ and $d = \dim_{\mathcal{E}}(\mathcal{F},\alpha)$.