

# Sample-Efficient Reinforcement Learning from Human Feedback via Information-Directed Sampling

Han Qi<sup>\*†‡</sup>

Haochen Yang<sup>\*§</sup>

Qiaosheng Zhang<sup>†</sup>

Zhuoran Yang<sup>¶</sup>

## Abstract

We study the problem of reinforcement learning from human feedback (RLHF), a critical problem in training large language models, from a theoretical perspective. Our main contribution is the design of novel sample-efficient RLHF algorithms based on information-directed sampling (IDS), an online decision-making principle inspired by information theory. Our algorithms maximize the sum of the value function and a mutual information term that encourages exploration of the unknown environment (which quantifies the information gained about the environment through observed human feedback data). To tackle the challenge of large state spaces and improve sample efficiency, we construct a simplified *surrogate environment* and introduce a novel distance measure (named the  $\ell_g$ -distance), enabling our IDS-based algorithm to achieve a Bayesian regret upper bound of order  $O(H^{3/2} \sqrt{\log(K(\epsilon))T})$ , where  $H$  is the episode length,  $T$  is the number of episode and  $K(\epsilon)$  is related to the covering number of the environment. Specializing to the tabular settings, this regret bound is of order  $\tilde{O}(H^2 \sqrt{SAT})$ , where  $S$  and  $A$  are the numbers of states and actions. Finally, we propose an Approximate-IDS algorithm that is computationally more efficient while maintaining nearly the same sample efficiency. The design principle of this approximate algorithm is not only effective in RLHF settings but also applicable to the standard RL framework. Moreover, our work showcases the value of information theory in reinforcement learning and in the training of large language models.

## 1 Introduction

Reinforcement learning from human feedback (RLHF) is a key technique for aligning large language models (LLMs) to human values [Ouyang et al. \(2022\)](#), and has also shown immense potential in many other fields, such as stock prediction, robot training, medical treatments [Zhu et al. \(2023\)](#). It can be viewed as an extension of standard reinforcement learning (RL) in the sense that feedback is not given as a numerical reward, but as a one-bit preference over a trajectory pair. Compared to standard RL, this preference-based setting is often more aligned with real-world scenarios, especially for tasks involving human evaluations [Chen et al. \(2022\)](#). However, a key challenge for applying RLHF algorithms is their reliance on extensive human feedback data, which is usually expensive and time-intensive to collect. To address this challenge, recent works on RLHF mainly focus on developing online learning methods that encourage exploration to improve sample efficiency, thereby reducing the amount of human feedback needed [Xie et al. \(2024\)](#). This brings the RLHF problem back to a fundamental question in RL: *how to effectively balance the trade-off between exploration and exploitation to improve sample efficiency?*

---

<sup>\*</sup>Equal contribution.

<sup>†</sup>Shanghai AI Laboratory. Email: [zhangqiaosheng@pjlab.org.cn](mailto:zhangqiaosheng@pjlab.org.cn)

<sup>‡</sup>Xi'an Jiaotong University. Email: [qihan19@stu.xjtu.edu.cn](mailto:qihan19@stu.xjtu.edu.cn)

<sup>§</sup>Peking University. Email: [hcyang@stu.pku.edu.cn](mailto:hcyang@stu.pku.edu.cn)

<sup>¶</sup>Yale University. Email: [zhuoran.yang@yale.edu](mailto:zhuoran.yang@yale.edu)

To tackle this trade-off, two major design principles have been introduced. The first approach, *Optimism in the Face of Uncertainty* (OFU), typically relies on constructing confidence sets that include the true environment with high probability to construct corresponding policies, one example of which is the Upper Confidence Bound (UCB) approach [Tossou et al. \(2019\)](#); [Ye et al. \(2024\)](#). In this paper, however, we focus on the less explored second approach, *Posterior Sampling*, which adopts the Bayesian framework and treats the environment as a random variable. The Bayesian reinforcement learning is a natural framework for studying in-context RL with LLMs (Transformer-based policies). One classical posterior sampling algorithm is Thompson Sampling (TS), which has been proved to be sample-efficient and enjoy sublinear Bayesian regret upper bounds in both RL [Moradipari et al. \(2023\)](#) and RLHF settings [Wu and Sun \(2023\)](#).

Apart from TS, *information-directed sampling* (IDS) emerges as a novel and principled online decision-making approach. By incorporating a mutual information term into the policy selection procedure, IDS manages to further encourage exploration about the unknown environment, thus tackling the exploration-exploitation tradeoff to a certain extent [Hao and Lattimore \(2022\)](#); [Russo and Van Roy \(2014\)](#). Compared with UCB and TS, IDS is more adept at learning complex information-regret structures, and is more flexible and robust to observation noise [Zhang et al. \(2024\)](#). In addition, empirical evidence has demonstrated that IDS performs exceptionally well across a range of scenarios, such as sparse linear bandits [Hao et al. \(2021\)](#), bandits with graph feedback [Hao et al. \(2022\)](#), Markov Decision Processes (MDPs) [Hao and Lattimore \(2022\)](#).

Despite their theoretical and empirical advantages, existing IDS-based algorithms are restricted to RL problems with explicitly observable rewards, and are not applicable to RLHF settings. In the LLM era, there is a pressing need for sample-efficient RLHF algorithms, particularly for scenarios with large state spaces. To tackle these challenges, we first introduce the concept of *surrogate environment*, a compressed (simplified) representation of the potentially complex environment, which helps address the issue of large state spaces. Building on this, and inspired by rate-distortion theory, we design IDS-based RLHF algorithms that are not only theoretically sample-efficient but also computationally easy to implement.

**Main contributions:** The contribution of this paper can be summarized as follows.

1. We first introduce a basic IDS-based algorithm for the RLHF setting where the reward is unobservable and only preference feedback is available (see Sec. 4.1). In each episode, it follows the Bayesian posterior sampling paradigm, and solves an optimization problem that maximizes the sum of an expected value term (exploitation) and a mutual information term (exploration). Here, the mutual information quantifies the amount of information about a learning target (e.g., the environment) that can be gained through the trajectories and preference.
2. To tackle the challenge posed by large state spaces, we construct a simplified surrogate environment as the learning target in our algorithm. Using tools from information theory and posterior consistency theory, we prove that our IDS-based algorithm with surrogate environment (Algorithm 1) achieves a Bayesian regret bound of  $O(H^{3/2}\sqrt{\log(K(\epsilon))T})$ , where  $H$  is the episode length,  $T$  is the number of episode, and  $K(\epsilon)$  is related to the *covering number* of the environment. We also specialize our algorithm and results to the tabular RLHF, linear RLHF, and contextual dueling bandit settings, and demonstrate the advantages of our algorithm over existing ones.
3. In the literature on information-directed sampling (IDS), to the best of our knowledge, there is no efficient implementation of IDS that learns a surrogate environment, which hinders its practical application. We propose an Approximate-IDS algorithm (Algorithm 2) that is computationally more efficient than Algorithm 1 while maintaining nearly the same sample efficiency. The

advantage of this algorithm is that it does not need to construct the surrogate environment. This algorithm selects policies using an alternative optimization objective that can be optimized with standard RL techniques, such as PPO [Schulman et al. \(2017\)](#). Furthermore, we note that the design principle of Algorithm 2 is not only effective for preference-based learning but is also applicable to general RL tasks.

**Highlights on technical novelty:** In the process of constructing the surrogate environment, we introduce a novel distance measure, the  $\ell_g$ -distance, to quantify the discrepancy between two probability measures (see Eqn. (4.2) in Sec. 4.2). The newly proposed distance measure facilitates the design of our computationally efficient algorithm (Algorithm 2), while the KL divergence and  $\ell_1$ -distance are unable to achieve this. However, it simultaneously introduces new challenges for theoretical analysis. We overcome this difficulty by investigating the unique properties of the new distance measure, as elaborated in Appendices A and B.

**Comparisons with related works:** First, we note that most existing works on RLHF assume deterministic rewards, whereas our work considers a more general framework where both transitions and rewards are stochastic. Among existing RLHF algorithms, the most relevant to ours is the TS-based algorithm by [Wu and Sun \(2023\)](#). Although in the general setting their algorithm’s regret bound is not directly comparable to ours (as theirs depends on the *eluder dimension*, while ours depends on the covering number), we note that in the tabular setting, our bound is superior if we coarsely substitute the dimension  $d$  with  $SA$  in their linear setting. When comparing with prior works on standard RL, we note that our regret bound<sup>1</sup>  $\tilde{O}(H^2\sqrt{SAT})$  is superior to the regret bound  $\tilde{O}(H^2\sqrt{S^2A^2T})$  of the surrogate-IDS algorithm by [Hao and Lattimore \(2022\)](#), even though we consider a more challenging RLHF setting where we rely only on human feedback to learn the reward model. Moreover, compared to a prior work on TS for standard RL [Moradipari et al. \(2023\)](#), our analysis method removes a technical assumption that almost all optimal policies visit almost all state action pairs.

## 2 Related Works

**Reinforcement Learning from Human Feedback (RLHF):** RLHF has emerged as a critical approach in aligning AI systems with human values, especially in complex tasks where human feedback plays a crucial role [Achiam et al. \(2023\)](#); [Touvron et al. \(2023\)](#). The RLHF framework typically involves a three-stage process: supervised fine-tuning (SFT), reward modeling (RM), and reinforcement learning (RL) using algorithms like Proximal Policy Optimization (PPO)[Ouyang et al. \(2022\)](#); [Ziegler et al. \(2019\)](#). Direct Preference Optimization (DPO) [Rafailov et al. \(2024\)](#) is another approach that directly uses generative models as reward models and trains them using preference data.

The practical success of RLHF has also sparked a variety of theoretical studies. According to the type of preference feedback, these works can be roughly divided into two categories: *action preference* [Fürnkranz et al. \(2012\)](#); [Saha \(2021\)](#); [Ji et al. \(2024\)](#); [Sekhari et al. \(2024\)](#); [Li et al. \(2024\)](#); [Bai et al. \(2025\)](#) and *trajectory preference* [Busa-Fekete et al. \(2014\)](#); [Xu et al. \(2020\)](#); [Pacchiano et al. \(2021\)](#); [Chen et al. \(2022\)](#); [Taranovic et al. \(2022\)](#); [Wu and Sun \(2023\)](#). The literature on action preferences is generally referred to as the *contextual dueling bandits*. In this paper, we focus on the trajectory preference. Most of the existing work in this area follow the OFU principle with the exception of [Wu and Sun \(2023\)](#); [Li et al. \(2024\)](#) and [Li et al. \(2024\)](#), who investigate a well-known Bayesian method—TS. Note that [Wu and Sun \(2023\)](#) uses trajectory preferences and can be applied to the general function approximation framework while [Li et al. \(2024\)](#) focuses on contextual dueling

---

<sup>1</sup>We say  $f(n) = \tilde{O}(g(n))$  if  $f(n) = O(g(n) \cdot \text{polylog}(n))$ .

bandits. We also use the posterior sampling method, but unlike TS, our method follows the principle of information-directed sampling.

**Information-Directed Sampling (IDS):** IDS is a design principle for sequential decision-making problems, which balances exploration and exploitation by evaluating the information gain from each action or trajectory. Ref. [Russo and Van Roy \(2014\)](#) first introduces the IDS principle in the bandit setting. They decompose the Bayesian regret into a information ratio term and a cumulative information gain term, and bound the regret by tools from information theory. Based on their work, many studies use this method to analyze the regret of the TS algorithm in bandit settings [Russo and Van Roy \(2016\)](#); [Dong and Van Roy \(2018\)](#); [Bubeck and Sellke \(2020\)](#); [Liu et al. \(2018\)](#); [Kirschner et al. \(2021\)](#); [Hao et al. \(2021, 2022\)](#).

Recently, [Hao and Lattimore \(2022\)](#); [Moradipari et al. \(2023\)](#) study the Bayesian regret of IDS and TS without any prior assumptions for MDP settings. Ref. [Moradipari et al. \(2023\)](#) focuses on analyzing TS in general settings while [Hao and Lattimore \(2022\)](#) proposes a regularized-IDS algorithm for tabular and linear settings. Ref. [Zhang et al. \(2024\)](#) uses the principle of IDS to design a set of algorithms for multi-agent reinforcement learning. They both use the surrogate environment as the learning target to get a sharper bound. However, implementing the surrogate version of the algorithm is a challenge. In this paper, we introduce IDS into RLHF for general MDP settings. We propose an easy-to-implement surrogate algorithm and prove that the regret upper bound has the same order as the original version.

### 3 Preliminaries

#### 3.1 Notations

For any positive integer  $n$ , we use  $[n]$  to denote the set  $\{1, 2, \dots, n\}$ . For a measurable space  $\mathcal{X}$  and a probability measure  $\mu$  on it, we let  $\Delta(\mathcal{X}, \mu)$  denote the set of all possible probability distributions over  $\mathcal{X}$  that are absolutely continuous with respect to  $\mu$ . When  $\mu$  is clear from the context, we use  $\Delta(\mathcal{X})$  for brevity. For two probability densities  $p, q$  on  $\mathcal{X}$ , we denote their Kullback-Leibler (KL) divergence  $D_{\text{KL}}$  as

$$D_{\text{KL}}(p\|q) \triangleq \int_{\mathcal{X}} p(x) \cdot \log\left(\frac{p(x)}{q(x)}\right) dx.$$

For two random variables  $X$  and  $Y$ , their mutual information  $\mathbb{I}(X; Y)$  is defined as

$$\mathbb{I}(X; Y) \triangleq D_{\text{KL}}(\mathbb{P}((X, Y) \in \cdot) \| \mathbb{P}(X \in \cdot) \times \mathbb{P}(Y \in \cdot)).$$

The conditional mutual information of  $X$  and  $Y$ , given another random variable  $Z$ , is defined as

$$\mathbb{I}(X; Y|Z) \triangleq \mathbb{E}_Z[D_{\text{KL}}(\mathbb{P}((X, Y) \in \cdot | Z) \| \mathbb{P}(X \in \cdot | Z) \times \mathbb{P}(Y \in \cdot | Z))].$$

#### 3.2 Finite-horizon MDPs

The environment is denoted as  $\mathcal{E} = (\mathcal{S}, \mathcal{A}, H, \{P_h\}_{h=1}^H, \{R_h\}_{h=1}^H)$ , where  $\mathcal{S}$  and  $\mathcal{A}$  are the measurable state and action spaces respectively, and  $H$  is the episode length. For each step  $h \in [H]$ ,  $P_h : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S}, \mu_{\mathcal{S}})$  is the transition probability kernel, where  $\mu_{\mathcal{S}}$  is the base probability measure on  $\mathcal{S}$ ;  $R_h : \mathcal{S} \times \mathcal{A} \rightarrow \Delta([0, 1], \text{Lebesgue})$  is the reward function. Since we mostly deal with the mean value of the reward, we define  $r_h(s, a) \triangleq \mathbb{E}_x[R_h(x|s, a)] = \int_0^1 x R_h(x|s, a) dx$ . We assume that  $\mathcal{S}, \mathcal{A}$  are known while the transition kernels  $\{P_h\}_{h=1}^H$  and rewards  $\{R_h\}_{h=1}^H$  are unknown and random.

We consider a Bayesian framework, where we treat the environment  $\mathcal{E}$  as a random variable and have a prior belief on  $\mathcal{E}$ . For each step  $h \in [H]$ , let  $\Theta_h^P$  and  $\Theta_h^R$  be the function spaces of  $P_h$

and  $R_h$  respectively, and let  $\Theta_h \triangleq \Theta_h^P \times \Theta_h^R$ . The spaces  $\Theta_h^P$  and  $\Theta_h^R$  are assumed to be equipped with prior probability measures, denoted as  $\rho_h^P$  and  $\rho_h^R$  respectively. Define the full function spaces  $\Theta^P \triangleq \prod_{h=1}^H \Theta_h^P$ ,  $\Theta^R \triangleq \prod_{h=1}^H \Theta_h^R$ ,  $\Theta \triangleq \prod_{h=1}^H \Theta_h$ , which parameterize the set of all environments and also induce the product prior probability measure  $\rho^P \triangleq \prod_{h=1}^H \rho_h^P$  for  $\Theta^P$ ,  $\rho^R \triangleq \prod_{h=1}^H \rho_h^R$  for  $\Theta^R$ , and  $\rho \triangleq \rho^P \otimes \rho^R$  being the prior of environments. Notice that this setting ensures the independence of the priors over different layers. Since the notion of the convex combination of environments will be used in our analysis, without loss of generality, we assume  $\Theta$  is convex.

### 3.3 Interaction protocol

The process of an agent interacting with a finite-horizon MDP is as follows. The agent starts at an initial state  $s_1^t$ , which is assumed to be fixed for all episodes  $t \in [T]$ . In each episode  $t \in [T]$ , the agent selects two policies  $(\pi_0^t, \pi_1^t)$  from the set of all possible policies  $\Pi$ , where a policy  $\pi$  is denoted by stochastic maps  $(\pi_1, \dots, \pi_H)$  with each  $\pi_h : \mathcal{S} \rightarrow \Delta(\mathcal{A})$ . Note that by this definition we assume the policy to be stationary, i.e., depends only on the current state and layer. At layer  $h$  in episode  $t$ , for  $i = 0, 1$ , the agent observes state pair  $(s_h^{t,0}, s_h^{t,1})$ , separately executes  $\pi_i^t$  on  $s_h^{t,i}$  to obtain action pair  $(a_h^{t,0}, a_h^{t,1})$  with probability  $\pi_i^t(a_h^{t,i} | s_h^{t,i})$ , takes the actions and changes to the next random state  $s_{h+1}^{t,i}$  with probability  $P_h(s_{h+1}^{t,i} | s_h^{t,i}, a_h^{t,i})$ . At state  $s_{H+1}$ , the agent stops acting and obtains two trajectories  $\tau_0^t$  and  $\tau_1^t$ , where

$$\tau_i^t \triangleq (s_1^{t,i}, a_1^{t,i}, \dots, s_H^{t,i}, a_H^{t,i}).$$

In the RLHF setting, the agent cannot directly receive a numerical reward, but only receives a *preference signal*  $o_t$  over trajectory pair  $(\tau_0^t, \tau_1^t)$ , where  $o_t$  is a Bernoulli random variable with  $\mathbb{P}(o_t = 1 | \tau_0^t, \tau_1^t) \triangleq \mathbb{P}(\tau_1^t \text{ is preferred to } \tau_0^t)$ . We assume the preference follows the *Bradley-Terry (BT) model* [Bradley and Terry \(1952\)](#), which has been widely used in existing works on RLHF. The BT model assumes the probability of humans preferring one choice to the other is proportional to the exponential of the value of cumulative reward:

$$\mathbb{P}(o_t = 1 | \tau_0^t, \tau_1^t) = \sigma(r(\tau_1^t) - r(\tau_0^t)),$$

where  $r(\tau^t) \triangleq \sum_{h=1}^H r_h(s_h^t, a_h^t)$  for  $\tau^t = (s_1^t, a_1^t, \dots, s_H^t, a_H^t)$ , and  $\sigma(x) \triangleq 1/(1 + e^{-x})$  is the sigmoid function.

Let  $\mathcal{H}_t \triangleq (\tau_0^t, \tau_1^t, o_t)$  be the history of episode  $t$  that includes both trajectories and preference feedback, and let  $\mathcal{D}_t \triangleq (\mathcal{H}_1, \dots, \mathcal{H}_{t-1})$  be the entire history up to episode  $t$ . The history of episode  $t$  up to layer  $h$  is denoted as

$$\mathcal{H}_{t,h} \triangleq (s_1^{t,i}, a_1^{t,i}, \dots, s_h^{t,i}, a_h^{t,i})_{i \in \{0,1\}}.$$

In the Bayesian setting, we often need to take conditional expectations with regard to  $\mathcal{D}_t$ . For brevity, we follow the standard notation in [Hao and Lattimore \(2022\)](#), letting  $\mathbb{P}_t(\cdot) \triangleq \mathbb{P}(\cdot | \mathcal{D}_t)$ , and  $\mathbb{E}_t[\cdot] \triangleq \mathbb{E}[\cdot | \mathcal{D}_t]$ . The mean environment  $\bar{\mathcal{E}}_t$  is defined to satisfy  $P_h^{\bar{\mathcal{E}}_t}(\cdot | s, a) = \mathbb{E}_t[P_h^{\mathcal{E}}(\cdot | s, a)]$  and  $R_h^{\bar{\mathcal{E}}_t}(\cdot | s, a) = \mathbb{E}_t[R_h^{\mathcal{E}}(\cdot | s, a)]$  for all  $s \in \mathcal{S}$  and  $a \in \mathcal{A}$ . Finally, let  $\mathcal{R}_{t,h} \triangleq (r_1^{t,i}, \dots, r_h^{t,i})_{i \in \{0,1\}}$  denote the corresponding potential unobserved rewards, where each  $r_h^{t,i}$  is a random variable satisfying  $r_h^{t,i} \sim R_h(\cdot | s_h^{t,i}, a_h^{t,i})$ .

### 3.4 Value function and Bayesian regret

Define the value function  $V_{h,\pi}^{\mathcal{E}} : \mathcal{S} \rightarrow [0, H]$  as the expected cumulative rewards received under policy  $\pi$  interacting with  $\mathcal{E}$  at layer  $h$ :

$$V_{h,\pi}^{\mathcal{E}}(s) \triangleq \mathbb{E}_{\pi}^{\mathcal{E}} \left[ \sum_{h'=h}^H r_{h'}(s_{h'}, a_{h'}) | s_h = s \right],$$

where  $\mathbb{E}_{\pi}^{\mathcal{E}}$  denotes the expectation over the trajectory generated under policy  $\pi$  and environment  $\mathcal{E}$ . We set  $V_{H+1,\pi}^{\mathcal{E}}(\cdot) \triangleq 0$ . For environment  $\mathcal{E}$ , let  $\pi_{\mathcal{E}}^*$  be the optimal policy that satisfies  $\pi_{\mathcal{E}}^* = \max_{\pi} V_{h,\pi}^{\mathcal{E}}(s)$  for all  $s \in \mathcal{S}$  and  $h \in [H]$ . Note that under Bayesian settings,  $\pi_{\mathcal{E}}^*$  is a function of  $\mathcal{E}$ , which is also a random variable.

Finally, for a sequence of policies  $\pi = (\pi_t)_{t \in [T]}$  over  $T$  episodes, we define the *regret* of  $\pi$  in environment  $\mathcal{E}$  as

$$R_T(\mathcal{E}, \pi) \triangleq \sum_{t=1}^T V_{1,\pi_{\mathcal{E}}^*}^{\mathcal{E}}(s_1^t) - V_{1,\pi^t}^{\mathcal{E}}(s_1^t). \quad (3.1)$$

Since this work focuses on the Bayesian setting, we also define the *Bayesian regret* as

$$BR_T(\pi) \triangleq \mathbb{E}_{\mathcal{E} \sim \rho}[R_T(\mathcal{E}, \pi)]. \quad (3.2)$$

The task of finding a policy  $\pi$  with minimal Bayesian regret, in the context of a finite-horizon MDP, is called a Bayesian RLHF problem.

## 4 The basic IDS Algorithm

This section introduces a basic IDS algorithm for RLHF settings. In Sec. 4.1, we present the generic form of our algorithm with an abstract learning target. Sec. 4.2 suggests constructing a discrete surrogate environment as the learning target and then describes an IDS algorithm with the surrogate environment (Algorithm 1). Sec. 4.3 provides the Bayesian regret bound for Algorithm 1, while Sec. 4.4 specializes this result to tabular RLHF, linear RLHF, and contextual dueling bandits.

### 4.1 Algorithm description: a generic form

At the beginning of episode  $t$ , based on the prior distribution  $\rho$  and history data  $\mathcal{D}_t$ , the agent first computes the posterior distribution of the environment  $\mathcal{E} \sim \mathbb{P}(\cdot | \mathcal{D}_t)$ , or equivalently, the transition  $P$  and reward  $R$ . Then, the agent chooses a stochastic policy  $\pi_{\text{IDS}}^t$  by maximizing a weighted sum of an expected value term and a mutual information term:

$$\pi_{\text{IDS}}^t = \arg \max_{\pi \in \Pi} \mathbb{E}_t[V_{1,\pi}^{\mathcal{E}}(s_1)] + \frac{\lambda}{2} \cdot \mathbb{I}_t^{\pi}(\chi; (\mathcal{H}_t, \mathcal{R}_{t,H})), \quad (4.1)$$

where  $\lambda > 0$  is a tunable parameter. Here,  $\chi$  is called the *learning target*, which is a random variable and is usually selected as the whole environment  $\mathcal{E}$  when the state space is not too large. However, in Sec. 4.2 where we consider large state space cases, we will construct a surrogate environment as the learning target to achieve tighter regret bounds.

The subscript  $t$  in  $\mathbb{I}_t^{\pi}(\chi; (\mathcal{H}_t, \mathcal{R}_{t,H}))$  in Eqn. (4.1) means that the distributions of  $\chi$  and  $(\mathcal{H}_t, \mathcal{R}_{t,H})$  are both conditioned on  $\mathcal{D}_t$ , and the superscript  $\pi$  means that  $(\mathcal{H}_t, \mathcal{R}_{t,H})$  are obtained by executing the policy  $\pi$ . Intuitively, a larger value of  $\mathbb{I}_t^{\pi}(\chi; (\mathcal{H}_t, \mathcal{R}_{t,H}))$  indicates that the data obtained at episode

$t$  contains more information about the learning target  $\chi$ . Accordingly, the introduction of mutual information in the policy selection procedure further encourages exploration about the unknown environment, while the expected value term  $\mathbb{E}_t[V_{1,\pi}^{\mathcal{E}}(s_1)]$  promotes exploitation. In this way, our algorithm manages to tackle the exploration-exploitation tradeoff to a certain extent.

## 4.2 Constructing surrogate environments as learning targets

In real-world scenarios, the environment is often too complex to be fully included as the agent's learning target  $\chi$ , thus it is better for the agent to focus only on the significant parts of the environment. In this subsection, we construct a discrete surrogate environment and propose an IDS algorithm with this surrogate environment as the learning target.

### 4.2.1 A new distance measure

The discrete environment is constructed using a covering argument with suitable distance measures. Unlike previous works on standard RL settings [Hao and Lattimore \(2022\)](#); [Moradipari et al. \(2023\)](#) that use either  $\ell_1$ -distance or KL-divergence, we propose a new distance measure between two probability measures, called the  $\ell_g$ -distance, which is better suited to our RLHF framework:

$$\ell_g(P, Q) \triangleq \sup_{o \in \mathcal{O}} \|\log P(\cdot|o) - \log Q(\cdot|o)\|_1 = \sup_{o \in \mathcal{O}} \int_{x \in \mathcal{X}} \left| \log \frac{P(x|o)}{Q(x|o)} \right| d\mu_{\mathcal{X}}, \quad (4.2)$$

where  $\mathcal{O} = \mathcal{S} \times \mathcal{A}$ .

**Remark 4.1.** For any two vector-valued maps  $P, Q$ , we define

$$\ell_g(P, Q) \triangleq \sup_{o \in \mathcal{O}} \int_{x \in \mathcal{X}} \sum_i \left| \log \frac{P_i(x|o)}{Q_i(x|o)} \right| d\mu_{\mathcal{X}} \quad (4.3)$$

where  $P_i$  and  $Q_i$  are the  $i$ -th component of  $P$  and  $Q$  respectively. This generalization of one-dimension case is useful for the analysis of linear RLHF problems (Theorem 4.16).

**Remark 4.2.** To guarantee  $\ell_g$  is well-defined, we let  $\log \frac{0}{0} \triangleq 0$ . Similar to the KL divergence, we allow for taking infinite values of  $\ell_g$ , e.g., if there exists a subset  $\mathcal{X}' \subset \mathcal{X}$  with positive measure such that  $Q(x|o) = 0$  but  $P(x|o)$  is nonzero on  $\mathcal{X}'$ , by definition we have  $\ell_g(P, Q) = \infty$ .

Although similar to the KL divergence, one of the fundamental properties of  $\ell_g$  is that  $\ell_g$  is a distance metric, which is more convenient for analysis.

**Lemma 4.3.**  $\ell_g$  is a distance metric.

*Proof.* By definition, it is easy to see that  $\ell_g(P, Q) = \ell_g(Q, P)$  and  $\ell_g(P, Q) = 0 \Leftrightarrow P = Q$ . It then suffices to show the triangle inequality. For any three probability distributions  $P, Q, R$ , we have

$$\begin{aligned} \ell_g(P, Q) &= \sup_o \int_{x \in \mathcal{X}} \left| \log \frac{P(x|o)}{Q(x|o)} \right| = \sup_o \int_{x \in \mathcal{X}} \left| \log \frac{P(x|o)}{R(x|o)} - \log \frac{Q(x|o)}{R(x|o)} \right| \\ &\leq \sup_o \int_{x \in \mathcal{X}} \left| \log \frac{P(x|o)}{R(x|o)} \right| + \int_{x \in \mathcal{X}} \left| \log \frac{Q(x|o)}{R(x|o)} \right| \\ &= \ell_g(P, R) + \ell_g(Q, R), \end{aligned}$$

which completes the proof of Lemma 4.3.  $\square$

To guarantee the existence of a finite coverage, we need the following assumptions:

**Assumption 4.4.**  $(\Theta, \tau_{\ell_g})$  is a compact topological space, where  $\tau_{\ell_g}$  is the topology generated by the metric  $\ell_g$ .

**Assumption 4.5.** For any  $P \in \Theta$ , there exist  $\beta, B > 0$  such that

$$\beta \leq \inf_{o,x} \{P(x|o) : P(x|o) \neq 0\} \leq \sup_{o,x} \{P(x|o)\} \leq B.$$

Note that, our assumption permits the probability density to be zero, but restricts the non-zero support to have a lower bound of  $\beta$ . This lower bound  $\beta$  is only required for Algorithm 2 in Section 5; Algorithm 1 does not rely on this assumption. In the regret upper bound of Algorithm 2, the term involving  $\beta$  is  $\log \frac{1}{\beta}$ . Consequently,  $\beta$  can be chosen to be extremely small. For instance, if  $\beta = e^{-100}$  (a value far beyond the floating-point precision of modern computers), then  $\log \frac{1}{\beta} = 100$ . Even in this case, the upper bound is only affected by a constant factor relative to the original bound.

Given the new distance  $\ell_g$ , we introduce the definition of  $\epsilon$ -covering number.

**Definition 4.6** ( $\epsilon$ -covering number). For a set  $\mathcal{G}$ , the  $\epsilon$ -covering number of  $\mathcal{G}$  with respect to  $\ell_g$  is the size  $K(\mathcal{G}, \epsilon)$  of the smallest set  $\{G_1, \dots, G_{K(\mathcal{G}, \epsilon)}\} \subset \mathcal{G}$  such that

$$\forall P \in \mathcal{G}, \exists P' \in \{G_1, \dots, G_{K(\mathcal{G}, \epsilon)}\} : \ell_g(P, P') \leq \epsilon. \quad (4.4)$$

#### 4.2.2 Partition of the environment

First, we introduce the concept of  $\epsilon$ -value partition, which must exist based on Assumption 4.4.

**Definition 4.7** ( $\epsilon$ -value partition). Given any  $\epsilon > 0$ , we say a partition  $\{\Theta_k^\epsilon\}_{k=1}^K$  over  $\Theta$  is an  $\epsilon$ -value partition for a RLHF problem if for any  $k \in [K]$  and  $\mathcal{E}, \mathcal{E}' \in \Theta_k^\epsilon$ ,

$$V_{1,\pi_{\mathcal{E}}^*}^{\mathcal{E}}(s_1) - V_{1,\pi_{\mathcal{E}'}^*}^{\mathcal{E}'}(s_1) \leq \epsilon. \quad (4.5)$$

We now provide a concrete construction of the  $\epsilon$ -value partition as follows. For any  $\mathcal{E}_0 \in \Theta$ , we define the  $\epsilon$ -ball centered at  $\mathcal{E}_0$  as

$$B(\mathcal{E}_0, \epsilon) \triangleq \{\mathcal{E} \in \Theta : \ell_g(\mathcal{E}, \mathcal{E}_0) \leq \epsilon\}. \quad (4.6)$$

Let  $\delta_P \triangleq \epsilon/6BH^2$  and  $\delta_R \triangleq \epsilon/6BH$ . Let  $K(\Theta_h^P, \delta_P)$  and  $K(\Theta_h^R, \delta_R)$  be the  $\delta_P$ -covering and  $\delta_R$ -covering numbers of  $\Theta_h^P$  and  $\Theta_h^R$  respectively. We denote  $\{B_h^P(i, \delta_P)\}_{i=1}^{K(\Theta_h^P, \delta_P)}$  and  $\{B_h^R(j, \delta_R)\}_{j=1}^{K(\Theta_h^R, \delta_R)}$  as the corresponding  $\epsilon$ -balls that cover  $\Theta_h^P$  and  $\Theta_h^R$ . For each  $i_h \in [K(\Theta_h^P, \delta_P)]$  and  $j_h \in [K(\Theta_h^R, \delta_R)]$ , we define

$$\Theta_{h,i_h,j_h}^\epsilon \triangleq \{\mathcal{E} \in \Theta \mid P_h^{\mathcal{E}} \in B_h^P(i_h, \delta_P), R_h^{\mathcal{E}} \in B_h^R(j_h, \delta_R)\}. \quad (4.7)$$

Setting  $K(\epsilon) \triangleq \prod_{h=1}^H K(\Theta_h^P, \delta_P) \times K(\Theta_h^R, \delta_R)$ , we can then find a bijective mapping from  $(h, i_h, j_h)$  to  $[K(\epsilon)]$ , and we obtain an  $\epsilon$ -value partition that satisfies  $\cup_{k=1}^{K(\epsilon)} \Theta_k^\epsilon = \Theta$ .<sup>2</sup> Now, we prove that for any  $\mathcal{E}, \mathcal{E}'$  belonging to the same partition,

$$V_{1,\pi_{\mathcal{E}}^*}^{\mathcal{E}}(s_1) - V_{1,\pi_{\mathcal{E}'}^*}^{\mathcal{E}'}(s_1) \leq \epsilon.$$

---

<sup>2</sup>If an environment  $\mathcal{E} \in \Theta$  belongs to more than one partition, we will ensure it only appears in a single partition by truncating the other partitions.

By Lemma C.1, we have

$$\begin{aligned}
& V_{1,\pi_{\mathcal{E}}^*}^{\mathcal{E}}(s_1) - V_{1,\pi_{\mathcal{E}}^*}^{\mathcal{E}'}(s_1) \\
&= \sum_{h=1}^H \mathbb{E}_{\pi_{\mathcal{E}}^*}^{\mathcal{E}'} \left[ \mathbb{E}_{s' \sim P_h^{\mathcal{E}}(\cdot|s_h, a_h)} \left[ V_{h+1,\pi_{\mathcal{E}}^*}^{\mathcal{E}}(s') \right] - \mathbb{E}_{s' \sim P_h^{\mathcal{E}'}(\cdot|s_h, a_h)} \left[ V_{h+1,\pi_{\mathcal{E}}^*}^{\mathcal{E}'}(s') \right] \right] + \sum_{h=1}^H \mathbb{E}_{\pi_{\mathcal{E}}^*}^{\mathcal{E}'} \left[ R_h^{\mathcal{E}}(s_h, a_h) - R_h^{\mathcal{E}'}(s_h, a_h) \right] \\
&\leq \sum_{h=1}^H \mathbb{E}_{\pi_{\mathcal{E}}^*}^{\mathcal{E}'} \left[ \int_{\mathcal{S}} \left| P_h^{\mathcal{E}}(s'|s_h, a_h) - P_h^{\mathcal{E}'}(s'|s_h, a_h) \right| \cdot V_{h+1,\pi_{\mathcal{E}}^*}^{\mathcal{E}'}(s') d\mu_{\mathcal{S}} + \int_{[0,1]} \left| x \left( R_h^{\mathcal{E}}(x|s_h, a_h) - R_h^{\mathcal{E}'}(x|s_h, a_h) \right) \right| dx \right] \\
&\leq \sum_{h=1}^H \mathbb{E}_{\pi_{\mathcal{E}}^*}^{\mathcal{E}'} \left[ HB \cdot \int_{\mathcal{S}} \left| \log \frac{P_h^{\mathcal{E}}(s'|s_h, a_h)}{P_h^{\mathcal{E}'}(s'|s_h, a_h)} \right| d\mu_{\mathcal{S}} + B \cdot \int_{[0,1]} \left| \log \frac{R_h^{\mathcal{E}}(x|s_h, a_h)}{R_h^{\mathcal{E}'}(x|s_h, a_h)} \right| dx \right] \\
&\leq \sum_{h=1}^H \mathbb{E}_{\pi_{\mathcal{E}}^*}^{\mathcal{E}'} [HB \cdot 2\delta_P + B \cdot 2\delta_R] = \frac{2\epsilon}{3} \leq \epsilon. \tag{4.8}
\end{aligned}$$

where the second inequality is due to the fact that  $V_{h+1,\pi_{\mathcal{E}}^*}^{\mathcal{E}'}(s') \leq H$ , and  $|a - b| \leq B \cdot |\log \frac{a}{b}|$  for any  $a, b \in (0, B)$ . The last inequality is due to the definition of  $\ell_g$ : since  $\mathcal{E}, \mathcal{E}'$  lie in the same  $\Theta_k^\epsilon$ , we have  $\ell_g(P_h^{\mathcal{E}}, P_h^{\mathcal{E}'}) \leq 2\delta_P$  and  $\ell_g(R_h^{\mathcal{E}}, R_h^{\mathcal{E}'}) \leq 2\delta_R$ . This shows that  $\{\Theta_k^\epsilon\}_{k=1}^K$  gives an  $\epsilon$ -value partition.

**Remark 4.8.** An important distinction between the  $\ell_g$ -distance and the  $\ell_1$ -distance is that the  $\epsilon$ -ball under the  $\ell_g$ -distance is not convex. In other words, there exist instances of probability measures for which the  $\epsilon$ -ball defined in Eq. (4.6) is non-convex. We provide a counterexample in the Appendix B to demonstrate this non-convexity. While the lack of convexity does not affect the partitioning of the environment, it does influence the construction of the surrogate environment in the subsequent analysis.

#### 4.2.3 Construct the surrogate environment

Based on the above  $\epsilon$ -value partition, we explicitly construct the *surrogate environment*  $\tilde{\mathcal{E}}_t^*$  for episode  $t$  as:

$$\tilde{\mathcal{E}}_t^* = \tilde{\mathcal{E}}_{k,t}^* \text{ iff } \mathcal{E} \in \Theta_k^\epsilon, \tag{4.9}$$

where  $\tilde{\mathcal{E}}_{k,t}^* \triangleq \mathbb{E}_t [\mathcal{E} | \mathcal{E} \in \Theta_k^\epsilon]$ . Since  $\cup_{k=1}^{K(\epsilon)} \Theta_k^\epsilon = \Theta$ , for any  $\mathcal{E}$ , there exists  $k \in [K(\epsilon)]$  such that  $\mathcal{E} \in \Theta_k^\epsilon$ . Hence, the surrogate environment is well defined. For this surrogate environment, we have the following result.

**Lemma 4.9.** Fix  $t \in [T]$  and environment  $\mathcal{E} \in \Theta$ . Given the  $\epsilon$ -value partition  $\{\Theta_k^\epsilon\}_{k=1}^{K(\epsilon)}$  (Eqn. (4.7)), the surrogate environment  $\tilde{\mathcal{E}}_t^*$  constructed by Eqn. (4.9) satisfies the following:

1. For any  $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$  and any instance  $(\tilde{\mathcal{E}}_t^*, \mathcal{E}) \sim \mathbb{P}_t(\tilde{\mathcal{E}}_t^*, \mathcal{E})$ , it holds that

$$\ell_g(P_h^{\tilde{\mathcal{E}}_t^*}, P_h^{\mathcal{E}}) \leq \frac{\epsilon}{2BH^2}, \quad \ell_g(R_h^{\tilde{\mathcal{E}}_t^*}, R_h^{\mathcal{E}}) \leq \frac{\epsilon}{2BH} \tag{4.10}$$

2. The following inequality holds:

$$\mathbb{E}_t [V_{1,\pi_{\mathcal{E}}^*}^{\mathcal{E}}(s_1^t) - V_{1,\pi_{\text{TS}}^t}^{\mathcal{E}}(s_1^t)] - \mathbb{E}_t [V_{1,\pi_{\mathcal{E}}^*}^{\tilde{\mathcal{E}}_t^*}(s_1^t) - V_{1,\pi_{\text{TS}}^t}^{\tilde{\mathcal{E}}_t^*}(s_1^t)] \leq \epsilon, \tag{4.11}$$

where  $\pi_{\text{TS}}^t \triangleq \arg \max_{\pi \in \Pi} V_{1,\pi}^{\mathcal{E}}(s_1^t)$  is the TS policy that depends on the random environment  $\mathcal{E}$ .

A necessary step in the regret analysis is to bound  $\ell_g(P_h^{\tilde{\mathcal{E}}_t^*}, P_h^{\mathcal{E}})$  by the radius of the  $\epsilon$ -balls, i.e., Eq. (4.10). Since the  $\epsilon$ -ball under  $\ell_1$  is convex, the posterior mean  $\tilde{\mathcal{E}}_{k,t}^*$  also lies in  $\Theta_k^\epsilon$ . We can directly derive that  $\ell_1(P_h^{\tilde{\mathcal{E}}_t^*}, P_h^{\mathcal{E}}) \leq 2\epsilon$ . However,  $\Theta_k^\epsilon$  is not convex under the new metric  $\ell_g$ , hence  $\tilde{\mathcal{E}}_t^*$  and  $\mathcal{E}$  may not lie in the same partition. Therefore, we are unable to immediately obtain Eq. (4.10). Although we cannot use the convexity of  $\epsilon$ -balls under  $\ell_g$ , a specific geometric property of  $\ell_g$  (Lemma B.1) significantly simplifies our proof.

*Proof.* (1) Based on our environment partitioning,  $\Theta_k^\epsilon$  is a ball. Let  $\mathcal{C}$  be the center of  $\Theta_k^\epsilon$ , by Lemma B.1, we have  $\ell_g(P_h^{\tilde{\mathcal{E}}_t^*}, P_h^{\mathcal{C}}) \leq 2\delta_P$ . Then, by triangle inequality of  $\ell_g$  (Lemma 4.3), we have

$$\ell_g(P_h^{\tilde{\mathcal{E}}_t^*}, P_h^{\mathcal{E}}) \leq \ell_g(P_h^{\tilde{\mathcal{E}}_t^*}, P_h^{\mathcal{C}}) + \ell_g(P_h^{\mathcal{E}}, P_h^{\mathcal{C}}) \leq 3\delta_P = \frac{\epsilon}{2BH^2}.$$

The analysis for the reward term  $\ell_g(R_h^{\tilde{\mathcal{E}}_t^*}, R_h^{\mathcal{E}})$  is exactly the same as above, which yields the proof of the first conclusion in Lemma 4.9.

(2) For the second property, we divide  $\mathbb{E}_t \left[ V_{1,\pi_{\mathcal{E}}^*}^{\mathcal{E}}(s_1^t) - V_{1,\pi_{\text{TS}}^t}^{\mathcal{E}}(s_1^t) \right] - \mathbb{E}_t \left[ V_{1,\pi_{\mathcal{E}}^*}^{\tilde{\mathcal{E}}_t^*}(s_1^t) - V_{1,\pi_{\text{TS}}^t}^{\tilde{\mathcal{E}}_t^*}(s_1^t) \right]$  into two parts.

- We first show that  $\mathbb{E}_t \left[ V_{1,\pi_{\text{TS}}^t}^{\mathcal{E}}(s_1^t) \right] = \mathbb{E}_t \left[ V_{1,\pi_{\text{TS}}^t}^{\tilde{\mathcal{E}}_t^*}(s_1^t) \right]$ . Let  $\mathcal{E}_t \sim \mathbb{P}(\cdot | \mathcal{D}_t)$  be an independent sample of  $\mathcal{E}$ . By the law of total expectation and the definition of  $\tilde{\mathcal{E}}_t^*$ , we have

$$\begin{aligned} \mathbb{E}_t \left[ V_{1,\pi_{\text{TS}}^t}^{\tilde{\mathcal{E}}_t^*}(s_1^t) \right] &= \sum_{k=1}^K \mathbb{P}(\mathcal{E} \in \Theta_k^\epsilon) \cdot \mathbb{E}_t \left[ V_{1,\pi_{\text{TS}}^t}^{\tilde{\mathcal{E}}_t^*}(s_1^t) \middle| \mathcal{E} \in \Theta_k^\epsilon \right] \\ &= \sum_{k=1}^K \mathbb{P}(\mathcal{E} \in \Theta_k^\epsilon) \cdot \mathbb{E}_t \left[ V_{1,\pi_{\text{TS}}^t}^{\tilde{\mathcal{E}}_{k,t}^*}(s_1^t) \right]. \end{aligned}$$

By the definition of  $\tilde{\mathcal{E}}_{k,t}^*$ ,

$$\mathbb{E}_t \left[ V_{1,\pi_{\text{TS}}^t}^{\tilde{\mathcal{E}}_{k,t}^*}(s_1^t) \right] = \int_{\mathcal{E}' \in \Theta_k^\epsilon} \mathbb{E}_t \left[ V_{1,\pi_{\text{TS}}^t}^{\mathcal{E}'}(s_1^t) \right] d\mathbb{P}(\mathcal{E}_t = \mathcal{E}' | \mathcal{E}_t \in \Theta_k^\epsilon).$$

Then, we have

$$\begin{aligned} \mathbb{E}_t \left[ V_{1,\pi_{\text{TS}}^t}^{\tilde{\mathcal{E}}_{k,t}^*}(s_1^t) \right] &= \sum_{k=1}^K \mathbb{P}(\mathcal{E} \in \Theta_k^\epsilon) \cdot \mathbb{E}_t \left[ V_{1,\pi_{\text{TS}}^t}^{\tilde{\mathcal{E}}_{k,t}^*}(s_1^t) \right] \\ &= \sum_{k=1}^K \mathbb{P}(\mathcal{E} \in \Theta_k^\epsilon) \cdot \int_{\mathcal{E}' \in \Theta_k^\epsilon} \mathbb{E}_t \left[ V_{1,\pi_{\text{TS}}^t}^{\mathcal{E}'}(s_1^t) \right] d\mathbb{P}(\mathcal{E}_t = \mathcal{E}' | \mathcal{E}_t \in \Theta_k^\epsilon) \\ &\stackrel{(a)}{=} \sum_{k=1}^K \mathbb{P}(\mathcal{E} \in \Theta_k^\epsilon) \cdot \int_{\mathcal{E}' \in \Theta_k^\epsilon} \mathbb{E}_t \left[ V_{1,\pi_{\text{TS}}^t}^{\mathcal{E}'}(s_1^t) \middle| \mathcal{E}_t \in \Theta_k^\epsilon \right] d\mathbb{P}(\mathcal{E}_t = \mathcal{E}' | \mathcal{E}_t \in \Theta_k^\epsilon) \\ &= \sum_{k=1}^K \mathbb{P}(\mathcal{E} \in \Theta_k^\epsilon) \cdot \mathbb{E}_t \left[ V_{1,\pi_{\text{TS}}^t}^{\mathcal{E}}(s_1^t) \middle| \mathcal{E}_t \in \Theta_k^\epsilon \right] \end{aligned}$$

---

**Algorithm 1** IDS for RLHF

---

1: **Input:** Priors  $\rho^P, \rho^R$ , baseline policy  $\pi_0$ ,  $\lambda > 0$ , surrogate environment partition tolerance  $\epsilon > 0$ .

2: **for**  $t = 1$  **to**  $T$  **do**

3:   Compute posteriors:

$$\rho_t^P(P) \propto \rho^P(P) \prod_{i=1}^{t-1} \prod_{h=1}^H P_h(s_{h+1}^{i,1} | s_h^{i,1}, a_h^{i,1}) \quad (4.13)$$

$$\rho_t^R(R) \propto \rho^R(R) \prod_{i=1}^{t-1} (o_i \sigma(r(\tau_1^i) - r(\tau_0^i)) + (1 - o_i) \sigma(r(\tau_0^i) - r(\tau_1^i))) \quad (4.14)$$

4:   Compute the surrogate environment  $\tilde{\mathcal{E}}_t^*$ , and update policy by

$$\pi_{\text{IDS}}^t = \arg \max_{\pi \in \Pi} \mathbb{E}_t[V_{1,\pi}^{\mathcal{E}}(s_1)] + \frac{\lambda}{2} \mathbb{I}_t^{\pi} (\tilde{\mathcal{E}}_t^*; (\mathcal{H}_t, \mathcal{R}_{t,H}))$$

5:   Sample  $\tau_0^t \sim \pi_0, \tau_1^t \sim \pi_{\text{IDS}}^t$ .

6:   Obtain preference feedback  $o_t$  on  $\{\tau_0^t, \tau_1^t\}$ .

7: **end for**

---

$$\stackrel{(b)}{=} \mathbb{E}_t \left[ V_{1,\pi_{\text{TS}}^t}^{\mathcal{E}}(s_1) \right].$$

where (a) uses the fact that  $\pi_{\text{TS}}^t = \pi_{\mathcal{E}_t^*}^*$ ,  $\mathcal{E}_t^*$  is independent of  $\mathcal{E}_t$ , (b) follows from  $\mathcal{E}_t$  is an independent sample of  $\mathcal{E}$ .

- Next, we show that  $\mathbb{E}_t \left[ V_{1,\pi_{\mathcal{E}}^*}^{\mathcal{E}}(s_1^t) \right] - \mathbb{E}_t \left[ V_{1,\pi_{\mathcal{E}}^*}^{\tilde{\mathcal{E}}_t^*}(s_1^t) \right] \leq \epsilon$ . Adopting the same decomposition trick as in Eqn. (4.8), we have

$$\begin{aligned} & V_{1,\pi_{\mathcal{E}}^*}^{\mathcal{E}}(s_1^t) - V_{1,\pi_{\mathcal{E}}^*}^{\tilde{\mathcal{E}}_t^*}(s_1^t) \\ & \leq \sum_{h=1}^H \mathbb{E}_{\pi_{\mathcal{E}}^*}^{\tilde{\mathcal{E}}_t^*} \left[ HB \cdot \int_{\mathcal{S}} \left| \log \frac{P_h^{\mathcal{E}}(s'|s_h, a_h)}{P_h^{\tilde{\mathcal{E}}_t^*}(s'|s_h, a_h)} \right| d\mu_{\mathcal{S}} + B \cdot \int_{[0,1]} \left| \log \frac{R_h^{\mathcal{E}}(x|s_h, a_h)}{R_h^{\tilde{\mathcal{E}}_t^*}(x|s_h, a_h)} \right| dx \right] \\ & \leq \sum_{h=1}^H \mathbb{E}_{\pi_{\mathcal{E}}^*}^{\tilde{\mathcal{E}}_t^*} \left[ HB \cdot \frac{\epsilon}{2BH^2} + B \cdot \frac{\epsilon}{2BH} \right] = \epsilon, \end{aligned} \quad (4.12)$$

where the second inequality is due to Eqn. (4.10). Adding up the two parts yields the proof of the second property in Lemma 4.9. By this, we have finished the proof of Lemma 4.9.

□

It is worth noting that Eqn. (4.10) represents a unique property of the surrogate environment, specifically attributed to our metric  $\ell_g$ , which distinguishes it from the KL divergence. It can be proven that the  $\ell_1$ -distance also possesses this property. However, as seen in Section 5, Proposition 5.1 cannot be guaranteed under the  $\ell_1$ -distance, making it difficult to design efficient approximation algorithms.

#### 4.2.4 IDS with surrogate environments

The pseudo-code of our IDS algorithm (with the learning target being the surrogate environment  $\tilde{\mathcal{E}}_t^*$ ) is shown in Algorithm 1. The algorithm requires priors  $\rho^P, \rho^R$  as input. Prior refers to the initial assumptions about model parameters before learning begins. Suitable priors can be derived from existing domain knowledge (e.g., robot dynamics parameters, user behavior patterns). For instance, a Gaussian prior might be adopted in robotic Haninger et al. (2022, 2023), while a Dirichlet prior could be used in multi-agent collaboration Wu et al. (2021). In our paper, an appropriate prior can be selected for the IDS algorithm according to practical applications. Roughly speaking, the agent, at each episode  $t$ , first computes the posterior distributions of the transition kernel  $P$  and reward function  $R$  (as shown in Eqns. (4.13)-(4.14)). Then, the agent computes the surrogate environment  $\tilde{\mathcal{E}}_t^*$  based on Eqn. (4.9), and chooses the policy

$$\pi_{\text{IDS}}^t = \arg \max_{\pi \in \Pi} \mathbb{E}_t[V_{1,\pi}^{\mathcal{E}}(s_1)] + \frac{\lambda}{2} \cdot \mathbb{I}_t^{\pi}(\tilde{\mathcal{E}}_t^*; (\mathcal{H}_t, \mathcal{R}_{t,H})).$$

We sample two trajectories from the baseline policy  $\pi_0$  and the IDS policy  $\pi_{\text{IDS}}^t$ , respectively, and then obtain a preference  $o_t$  regarding the two trajectories. Moreover, human feedback and state-action sequence data are added to the history data  $\mathcal{D}_t$  for updating the posterior distribution for the next episode.

### 4.3 Regret analysis of Algorithm 1

Before presenting our main results, we need to first introduce a notion of *value diameter*. For any  $\mathcal{E}$ , we define the corresponding value diameter  $\alpha_{\mathcal{E}}$  as

$$\alpha_{\mathcal{E}} \triangleq \max_{1 \leq h \leq H} \left\{ \sup_s V_{h,\pi_{\mathcal{E}}^*}^{\mathcal{E}}(s) - \inf_s V_{h,\pi_{\mathcal{E}}^*}^{\mathcal{E}}(s) \right\} + \max_{h,s,a} \left\{ r_h^{\sup}(s, a) - r_h^{\inf}(s, a) \right\}.$$

Since the reward is bounded by  $[0, 1]$ , we have  $\alpha_{\mathcal{E}} \leq H + 1$ . The *average value diameter* over  $\Theta$  is denoted by  $\alpha \triangleq \mathbb{E}_{\mathcal{E} \sim \rho}[\alpha_{\mathcal{E}}^2]^{1/2}$ . Similar to the prior work Moradipari et al. (2023), we need to make the following assumption about posterior consistency.

**Assumption 4.10** (Posterior Consistency). Under our preference model, the posterior distribution of environment is strongly consistent.

This means as the sample size approaches infinity, the posterior distribution of environment obtained through Eqns. (4.13)-(4.14) tends to concentrate around the true distribution. In other words, the posterior distribution will correctly identify the true environment that generates these trajectory data.

**Theorem 4.11.** Given a Bayesian RLHF problem, for any  $\epsilon > 0$  and sufficiently large  $T$ , by choosing  $\lambda = \sqrt{\alpha^2 TH / \log(K(\epsilon))}$ , we have

$$BR_T(\pi_{\text{IDS}}) \leq \alpha \sqrt{TH \log(K(\epsilon))} + T\epsilon + T_0, \quad (4.15)$$

where  $T_0$  is a fixed positive integer that is independent of  $T$ . Setting  $\epsilon = \frac{1}{T}$ , our regret upper bound is of order

$$O\left(H^{\frac{3}{2}} \sqrt{T \log(K(\frac{1}{T}))}\right).$$

The detailed proof of Theorem 4.11 is deferred to Appendix A.1. We point out that the existing TS-based RLHF algorithm Wu and Sun (2023) has an upper bound of order

$$\tilde{O}(H^2 \sqrt{T(\ell_P + \ell_R)} (\dim_1(P, 1/T)) + \dim_1(R, 1/T)),$$

where  $\dim_1(P, 1/T)$  and  $\dim_1(R, 1/T)$  are the  $\ell_1$ -norm eluder dimension of the transition and reward function class,  $\ell_P$  and  $\ell_R$  are the bracketing covering number of the transition and reward function class. Without considering the way to characterize the complexity of the the reward and the transition model (i.e., via covering number or eluder dimension), our bound is superior to theirs by a factor of  $\sqrt{H}$ .

**Remark 4.12.** The regret upper bounds in some related works Saha et al. (2023); Wu and Sun (2023) are related to the derivative bound of the link function. However, our upper bound is independent of the link function we use (which is the sigmoid function). This is because the posterior consistency assumption implicitly imposes requirements on the link function — the link function should be monotonically increasing to ensure that better trajectories correspond to higher preference probabilities. For example, if the link function is equal to a constant  $\frac{1}{2}$ , then the posterior distribution of rewards would not change (according to the posterior update rule in Eqn. (4.14)), and thus could not converge to the true distribution. Therefore, the assumption in previous work of a strictly positive lower bound on the link function's derivative is encompassed by our posterior consistency assumption.

#### 4.4 Applications

Finally, we show that our algorithm can be applied in multiple scenarios, such as tabular RLHF, linear RLHF, and contextual dueling bandits.

**Definition 4.13** (Tabular RLHF). We say a Bayesian RLHF problem is tabular if  $|\mathcal{S}| = S$  and  $|\mathcal{A}| = A$  are both finite.

**Definition 4.14** (Linear RLHF). Let  $\phi^P : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}^d$  and  $\phi^R : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}^d$  be known feature maps with bounded norms  $\|\phi^P(s, a)\|_2 \leq 1$  and  $\|\phi^R(s, a)\|_2 \leq 1$ . We say a Bayesian RLHF problem is linear if for any  $\mathcal{E} = \{(P_h^\mathcal{E}, R_h^\mathcal{E})\}_{h=1}^H \in \Theta$ , there exists vector-valued maps  $\psi_h^{P,\mathcal{E}}$  and  $\psi_h^{R,\mathcal{E}}$  with bounded  $\ell_2$ -norm such that for every  $(s, a) \in \mathcal{S} \times \mathcal{A}$ ,

$$P_h^\mathcal{E}(\cdot | s, a) = \langle \phi^P(s, a), \psi_h^{P,\mathcal{E}}(\cdot) \rangle,$$

$$R_h^\mathcal{E}(\cdot | s, a) = \langle \phi^R(s, a), \psi_h^{R,\mathcal{E}}(\cdot) \rangle.$$

We assume that each component of the vector-valued maps  $\psi_h^{P,\mathcal{E}}$  and  $\psi_h^{R,\mathcal{E}}$  belongs to some compact set  $\mathcal{F} \subset L^2$ , i.e.,  $\forall i \in [d]$ ,  $(\psi_h^{P,\mathcal{E}})_i \in \mathcal{F}$  and  $(\psi_h^{R,\mathcal{E}})_i \in \mathcal{F}$ .

Specializing Theorem 4.11 to tabular and linear Bayesian RLHF problems, we have the following Bayesian regret bounds. The proofs of Theorems 4.15 and 4.16 are deferred to Appendices A.2 and A.3 respectively.

**Theorem 4.15** (Tabular RLHF). Given a tabular Bayesian RLHF problem, for any  $\epsilon > 0$  and sufficiently large  $T$ , we have

$$BR_T(\pi_{\text{IDS}}) \leq \alpha H \sqrt{3SAT \log \left( \frac{6H^2 \sqrt{S}}{\epsilon} \right)} + T\epsilon + T_0,$$

where  $T_0$  is a fixed integer that is independent of  $T$ . Setting  $\epsilon = \frac{1}{T}$ , our regret bound is of order  $\tilde{O}(\sqrt{SAH^4T})$ .

Recall that the IDS algorithm proposed for the tabular RL setting [Hao and Lattimore \(2022\)](#) has a regret upper bound of order  $\tilde{O}(\sqrt{S^2A^2H^4T})$ . Compared to their result, our method relies on less informative data (preference feedback instead of directly observable rewards) but achieves a better regret bound by a factor of  $S$  and  $A$ . This improvement is primarily due to our refined analytical techniques, inspired by recent advancements in TS [Moradipari et al. \(2023\)](#).

**Theorem 4.16** (Linear RLHF). Let  $M \triangleq \sup_{i,s} \max\{(\psi_h^P(s))_i, (\psi_h^R(s))_i\}$  and  $K_{\mathcal{F}}(\epsilon)$  denote the  $\frac{\epsilon}{dMH^2}$ -covering number of  $\mathcal{F}$ . Given a linear RLHF problem, for any  $\epsilon > 0$  and sufficiently large  $T$ , we have

$$BR_T(\pi_{IDS}) \leq \alpha H \sqrt{dT \log(K_{\mathcal{F}}(\epsilon))} + T\epsilon + T_0, \quad (4.16)$$

where  $T_0$  is a fixed integer that is independent of  $T$ . Setting  $\epsilon = \frac{1}{T}$ , this upper bound is of order

$$O\left(H^2 \sqrt{dT \log(K_{\mathcal{F}}(\frac{1}{T}))}\right).$$

Compared to [Wu and Sun \(2023\)](#), which derives a regret upper bound of  $\tilde{O}(H^{11/2}d^{17/2}\sqrt{T})$  for their TS algorithm, our regret upper bound is better when the covering number of the linear MDP is not of exponential size. If we convert their result to the tabular setting by coarsely substituting  $d$  with  $SA$ , our regret bound is also better in terms of  $H, S, A$ . However, we also point out that the above comparison is not an apples-to-apples comparison, as we consider Bayesian regret, while they consider frequentist regret, and their algorithm also accounts for the number of queries.

**Corollary 4.17** (Contextual Dueling Bandits). Contextual dueling bandits are a simplified version of our MDP setting (with  $H = 1$ ) and have been extensively studied in previous RLHF research [Ye et al. \(2024\)](#); [Zhu et al. \(2023\)](#); [Li et al. \(2024\)](#). By setting  $H = 1$  in Theorem 4.16, the Bayesian regret for Algorithm 1 in the linear contextual dueling bandit problem satisfies

$$BR_T(\pi_{IDS}) \leq 2\sqrt{dT \log(K_{\mathcal{F}}(\epsilon))} + T\epsilon + T_0,$$

for any  $\epsilon > 0$  and sufficiently large  $T$ . Setting  $\epsilon = \frac{1}{T}$ , the regret upper bound is of order  $\tilde{O}(\sqrt{dT})$ .

Without considering the covering number of linear environment, our regret upper bound is better than  $\tilde{O}(d\sqrt{T})$  derived by [Li et al. \(2024\)](#). Another work [Saha \(2021\)](#) assumes a finite number of arms with a regret upper bound of  $\tilde{O}(\sqrt{dT})$ , while we assume that the parameter space of the linear MDP is compact.

## 5 The Approximate-IDS Algorithm

While the IDS algorithm (Algorithm 1) is principled and sample-efficient, it suffers from relatively high computational complexity. This is because the calculation of the surrogate environment  $\tilde{\mathcal{E}}_t^*$  (which depends on the construction of  $\epsilon$ -value partition) is challenging. As a remedy, we develop a computationally efficient algorithm, named *Approximate-IDS*, whose optimization objective is independent of the surrogate environment (thus avoids the partition of  $\Theta$  in computation) and has finer properties for analysis. This allows the algorithm to be computed efficiently by traditional RL algorithms from standard RL theory.

---

**Algorithm 2** Approximate-IDS for RLHF

---

- 1: **Input:** Priors  $\rho^P, \rho^r$ , baseline policy  $\pi_0, \lambda > 0$ .
  - 2: **for**  $t = 1$  **to**  $T$  **do**
  - 3:   Compute the posterior as Algorithm 1 (Line 3).
  - 4:    $\pi_{\text{app}}^t = \arg \max_{\pi \in \Pi} \mathbb{E}_{\pi}^{\bar{\mathcal{E}}_t} \left[ \sum_{h=1}^H \bar{r}_h(s_h, a_h) \right]$
  - 5:   Sample  $\tau_0^t \sim \pi_0, \tau_1^t \sim \pi_{\text{app}}^t$ .
  - 6:   Obtain preference feedback  $o_t$  on  $\{\tau_0^t, \tau_1^t\}$ .
  - 7: **end for**
- 

## 5.1 Algorithm description

The pseudo-code for Approximate-IDS is shown in Algorithm 2. For convenience of description, we define

$$\text{KL}_{s,a}^h(\mathcal{E}, \mathcal{E}') \triangleq D_{\text{KL}}((P_h^{\mathcal{E}} \otimes R_h^{\mathcal{E}})(\cdot|s, a) || (P_h^{\mathcal{E}'} \otimes R_h^{\mathcal{E}'})(\cdot|s, a)), \quad (5.1)$$

and

$$\bar{r}_h(s_h, a_h) \triangleq r_h(s_h, a_h) + \frac{\lambda}{2} \cdot \mathbb{E}_t [\text{KL}_{s_h, a_h}^h(\mathcal{E}, \bar{\mathcal{E}}_t)], \quad (5.2)$$

where  $\bar{\mathcal{E}}_t$  denotes the posterior mean of  $\mathcal{E}$  given  $\mathcal{D}_t$ , i.e.,  $P_h^{\bar{\mathcal{E}}_t}(\cdot|s, a) = \mathbb{E}_t[P_h^{\mathcal{E}}(\cdot|s, a)]$  and  $R_h^{\bar{\mathcal{E}}_t}(\cdot|s, a) = \mathbb{E}_t[R_h^{\mathcal{E}}(\cdot|s, a)]$ .

The overall procedure is similar to that of Algorithm 1, with the key difference being the selection of the IDS policy (Line 4). Intuitively,  $\mathcal{E}$  is sufficiently close to  $\bar{\mathcal{E}}_t^*$  under metric  $\ell_g$ , thus it is reasonable to use  $\mathcal{E}$  directly for mutual information computation. Given trajectories and rewards, the additional environmental information revealed by human feedback, i.e.,  $\mathbb{I}_t^\pi(\bar{\mathcal{E}}_t^*; o_t | (\mathcal{H}_{t,H}, \mathcal{R}_{t,H}))$ , can be disregarded. Thus, we use the entire environment  $\mathcal{E}$  instead of the surrogate environment  $\bar{\mathcal{E}}_t^*$  to compute the mutual information and discard the information of the trajectory generated by the baseline policy  $\pi_0$  and human feedback. Therefore, we replace the mutual information term  $\mathbb{I}_t^\pi(\bar{\mathcal{E}}_t^*; (\mathcal{H}_t, \mathcal{R}_{t,H}))$  by  $\sum_{h=1}^H \mathbb{E}_t[\mathbb{E}_{\pi}^{\bar{\mathcal{E}}_t}[\text{KL}_{s_h, a_h}^h(\mathcal{E}, \bar{\mathcal{E}}_t)]]$  (Eqn. (C.4) in Lemma C.2). We can compute the approximate IDS policy as follows:

$$\begin{aligned} \pi_{\text{app}}^t &= \arg \max_{\pi \in \Pi} \mathbb{E}_t[V_{1,\pi}^{\mathcal{E}}(s_1)] + \frac{\lambda}{2} \sum_{h=1}^H \mathbb{E}_t[\mathbb{E}_{\pi}^{\bar{\mathcal{E}}_t}[\text{KL}_{s_h, a_h}^h(\mathcal{E}, \bar{\mathcal{E}}_t)]] \\ &\stackrel{(a)}{=} \arg \max_{\pi \in \Pi} \mathbb{E}_{\pi}^{\bar{\mathcal{E}}_t} \left[ \sum_{h=1}^H \bar{r}_h(s_h, a_h) \right], \end{aligned} \quad (5.3)$$

where (a) uses the linearity of expectation and independence of priors over different layers.

Note that  $\bar{r}$  and  $\bar{\mathcal{E}}_t$  are both independent of the surrogate environment, and can be well approximated by Monte Carlo sampling. Therefore, by introducing  $\bar{r}$ , solving  $\pi_{\text{app}}^t$  at episode  $t$  is equivalent to finding an optimal policy based on MDP  $\{P_h^{\bar{\mathcal{E}}_t}, \bar{r}_h\}_{h=1}^H$ , which can be solved efficiently by the PPO algorithm Schulman et al. (2017).

## 5.2 Regret bounds for Approximate-IDS

We first introduce an auxiliary reward function  $r'_h$  for the convenience of regret analysis. It serves as a bridge connecting the approximated  $\bar{r}_h$  to the real mutual information term in Algorithm 1.

**Proposition 5.1.** For any  $\epsilon$ -value partition  $\{\Theta_k^\epsilon\}_{k=1}^{K(\epsilon)}$  (Eqn. (4.7)) and the surrogate environment  $\tilde{\mathcal{E}}_t^*$  constructed by Eqn. (4.9), Define

$$r'_h(s, a) \triangleq r_h(s, a) + \frac{\lambda}{2} \mathbb{E}_t [\text{KL}_{s,a}^h(\tilde{\mathcal{E}}_t^*, \bar{\mathcal{E}}_t)]. \quad (5.4)$$

Then, for any policy  $\pi$ , we have

$$\left| \mathbb{E}_{\pi}^{\bar{\mathcal{E}}_t} \left[ \sum_{h=1}^H r'_h(s_h, a_h) \right] - \mathbb{E}_{\pi}^{\bar{\mathcal{E}}_t} \left[ \sum_{h=1}^H \bar{r}_h(s_h, a_h) \right] \right| \leq \frac{\lambda}{2} \epsilon (1 + \log \frac{B}{\beta}). \quad (5.5)$$

The proof is deferred to Appendix A.4. The proof indicates that the proposition may fail to hold when the surrogate environment is constructed using the  $\ell_1$ -distance or KL divergence. To better understand this proposition, consider an extreme scenario: we divide the environment into the smallest units, with each  $\Theta_k^\epsilon$  containing only one environment. We have  $\mathcal{E} = \tilde{\mathcal{E}}_t^*$ . The left hand of Eqn. (5.5) equals 0, so Proposition 5.1 holds true. Since  $|a - b| \leq B |\log \frac{a}{b}|$  for any  $a, b \in (0, B)$ , we have  $\ell_1(P, Q) \leq B \ell_g(P, Q)$ . If we ignore the constant  $B$ , by fixing the  $\epsilon$ -value, our distance achieves a finer environmental partition. On this finer partition,  $\mathcal{E}$  and  $\tilde{\mathcal{E}}_t^*$  behave more similarly, allowing us to ensure that Proposition 5.1 holds. Then, using Proposition 5.1, we give the Bayesian regret bound for the Approximate-IDS algorithm.

**Theorem 5.2.** Given a Bayesian RLHF problem, for any  $\epsilon > 0$  and sufficiently large  $T$ , by choosing  $\lambda = \sqrt{\alpha^2 TH / 2 \log(K(\epsilon))}$ , we have the following regret upper bound for Algorithm 2:

$$BR_T(\pi_{\text{app}}) \leq \alpha \sqrt{2TH \log(K(\epsilon))} + \left( 1 + \frac{(1 + \log(B/\beta))}{2} \sqrt{\frac{\alpha^2 TH}{2 \log(K(\epsilon))}} \right) T\epsilon + T_0. \quad (5.6)$$

By choosing a small  $\epsilon$ , the regret upper bound is of order  $O(\sqrt{H^3 T \log(K(\epsilon))})$ , matching that of Algorithm 1 presented in Sec. 4.3.

*Proof.* First notice that

$$\mathbb{E}_{\pi_{\text{app}}^t}^{\bar{\mathcal{E}}_t} \left[ \sum_{h=1}^H \bar{r}_h(s_h, a_h) \right] \stackrel{(a)}{\geq} \mathbb{E}_{\pi_{\text{IDS}}^t}^{\bar{\mathcal{E}}_t} \left[ \sum_{h=1}^H \bar{r}_h(s_h, a_h) \right] \stackrel{(b)}{\geq} \mathbb{E}_{\pi_{\text{IDS}}^t}^{\bar{\mathcal{E}}_t} \left[ \sum_{h=1}^H r_h(s_h, a_h) \right] \stackrel{(c)}{=} \mathbb{E}_t \left[ V_{1, \pi_{\text{IDS}}^t}^{\mathcal{E}}(s_1^t) \right],$$

where (a) uses the optimality of  $\pi_{\text{app}}^t$ , (b) follows from Eqn. (5.2), (c) uses the definition of  $\bar{\mathcal{E}}_t$  and the linearity of expectation.

Therefore,

$$\begin{aligned} \mathbb{E}_t \left[ V_{1, \pi_{\text{IDS}}^t}^{\mathcal{E}}(s_1^t) \right] - \mathbb{E}_t \left[ V_{1, \pi_{\text{app}}^t}^{\mathcal{E}}(s_1^t) \right] &= \mathbb{E}_t \left[ V_{1, \pi_{\text{IDS}}^t}^{\mathcal{E}}(s_1^t) \right] - \mathbb{E}_{\pi_{\text{app}}^t}^{\bar{\mathcal{E}}_t} \left[ \sum_{h=1}^H r_h(s_h, a_h) \right] \\ &\leq \mathbb{E}_{\pi_{\text{app}}^t}^{\bar{\mathcal{E}}_t} \left[ \sum_{h=1}^H \bar{r}_h(s_h, a_h) - r_h(s_h, a_h) \right] \\ &\leq \frac{\lambda \epsilon (1 + \log(B/\beta))}{2} + \mathbb{E}_{\pi_{\text{app}}^t}^{\bar{\mathcal{E}}_t} \left[ \sum_{h=1}^H r'_h(s_h, a_h) - r_h(s_h, a_h) \right] \\ &\leq \frac{\lambda \epsilon (1 + \log(B/\beta))}{2} + \frac{\lambda}{2} \cdot \mathbb{I}_t^{\pi_{\text{app}}^t} \left( \tilde{\mathcal{E}}_t^*; (\mathcal{H}_t, \mathcal{R}_{t,H}) \right), \end{aligned} \quad (5.7)$$

where the first inequality is due to Eqn. (5.7), the second inequality is due to Proposition 5.1, and the last inequality is due to Lemma C.2. Taking expectation in Eqn. (5.7) with respect to  $\mathcal{D}_t$  and then summing over  $t \in [T]$ , we obtain

$$BR_T(\pi_{\text{app}}) - BR_T(\pi_{\text{IDS}}) \leq \frac{\lambda\epsilon(1 + \log(B/\beta))T}{2} + \frac{\lambda}{2}\log(K(\epsilon)), \quad (5.8)$$

where we use the same trick in Eqn. (A.5) to derive  $\log(K(\epsilon))$  as an upper bound for  $\sum_{t=1}^T \mathbb{I}_t^{\pi_{\text{app}}^t} (\tilde{\mathcal{E}}_t^*; (\mathcal{H}_t, \mathcal{R}_{t,H}))$ . Finally, plugging the upper bound for  $BR_T(\pi_{\text{IDS}})$  (Eqn. (A.24)) into Eqn. (5.8) and taking  $\lambda = \sqrt{\alpha^2 TH / 2 \log(K(\epsilon))}$  yields the proof of Theorem 5.2.  $\square$

### 5.3 Regret Analysis With $\ell_1$ -distance

Using the standard  $\ell_1$ -distance for both environment partitioning and surrogate environment construction, rather than the proposed  $\ell_g$ -distance, would necessitate significantly stronger and less realistic assumptions, and result in worse theoretical guarantees. To ensure the validity of Proposition 5.1, the following stronger assumptions should be additionally imposed.

**Assumption 5.3.** For any  $P \in \Theta$ , there exist  $\beta, B > 0$  such that

$$\beta \leq \inf_{o,x} \{P(x|o)\} \leq \sup_{o,x} \{P(x|o)\} \leq B. \quad (5.9)$$

This assumption excludes zero probability densities for transitions/rewards, limiting its practical applicability significantly. Under this assumption, Proposition 5.1 can be reformulated as follows.

**Proposition 5.4.** For any  $\epsilon$ -value partition  $\{\Theta_k^\epsilon\}_{k=1}^{K(\epsilon)}$  (partitioned by  $\ell_1$ -distance) and the surrogate environment  $\tilde{\mathcal{E}}_t^*$  (constructed by  $\ell_1$ -distance), we have

$$\left| \mathbb{E}_{\pi}^{\tilde{\mathcal{E}}_t} \left[ \sum_{h=1}^H r'_h(s_h, a_h) \right] - \mathbb{E}_{\pi}^{\tilde{\mathcal{E}}_t} \left[ \sum_{h=1}^H \bar{r}_h(s_h, a_h) \right] \right| \leq \frac{\lambda}{2} \epsilon \left( 1 + \frac{B}{\beta} \right). \quad (5.10)$$

The proof of Proposition 5.4 can be found in Appendix A.5. Using Proposition 5.4, the Bayesian regret bound for Algorithm 2 under  $\ell_1$ -distance is of order

$$\tilde{O}\left(\frac{1}{\beta} \sqrt{H^3 T}\right).$$

In comparison, the regret bound derived using  $\ell_g$  is of order

$$\tilde{O}\left((\log \frac{1}{\beta}) \sqrt{H^3 T}\right).$$

Although we have also derived a regret bound of sublinear dependence on  $H$  and  $T$  using the  $\ell_1$ -distance, this bound includes a factor of  $\frac{1}{\beta}$ . When  $\beta$  is small, the bound becomes excessively large. Moreover, the underlying assumption of this bound contradicts real-world scenarios, as common MDPs permit zero transition and reward probabilities in practical settings—such as the boundaries in Go or obstacles in robotic navigation. In light of these limitations, our proposed  $\ell_g$ -based approach demonstrates significant superiority.

## 6 Conclusion

In this paper, we introduced novel information-directed sampling (IDS) algorithms to address key challenges in the RLHF problem, a critical component of LLM training. Our method improves the sample efficiency by maximizing both the value function and the mutual information between the (surrogate) environment and trajectories. We also developed a computationally efficient Approximate-IDS algorithm suitable for real-world applications while maintaining the regret bound order of the original method. A potentially practical implication of our sample-efficient algorithms is their ability to align LLMs to human values with less human feedback while maintaining similar performance, thereby reducing the cost and time of LLM training. Additionally, our findings highlight the value of information theory in the rapidly evolving era of LLMs.

## References

- Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S. et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Bai, C., Zhang, Y., Qiu, S., Zhang, Q., Xu, K. and Li, X. (2025). Online preference alignment for language models via count-based exploration. In *International Conference on Learning Representations (ICLR)*.
- Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, **39** 324–345.
- Bubeck, S. and Sellke, M. (2020). First-order bayesian regret analysis of thompson sampling. In *Algorithmic Learning Theory*. PMLR.
- Busa-Fekete, R., Szörényi, B., Weng, P., Cheng, W. and Hüllermeier, E. (2014). Preference-based reinforcement learning: evolutionary direct policy search using a preference-based racing algorithm. *Machine learning*, **97** 327–351.
- Chen, X., Zhong, H., Yang, Z., Wang, Z. and Wang, L. (2022). Human-in-the-loop: Provably efficient preference-based reinforcement learning with general function approximation. In *International Conference on Machine Learning*. PMLR.
- Dong, S. and Van Roy, B. (2018). An information-theoretic analysis for thompson sampling with many actions. *Advances in Neural Information Processing Systems*, **31**.
- Foster, D. J., Kakade, S. M., Qian, J. and Rakhlin, A. (2021). The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*.
- Fürnkranz, J., Hüllermeier, E., Cheng, W. and Park, S.-H. (2012). Preference-based reinforcement learning: a formal framework and a policy iteration algorithm. *Machine learning*, **89** 123–156.
- Ghosal, S. and van der Vaart, A. W. (2017). *Fundamentals of nonparametric Bayesian inference*, vol. 44. Cambridge University Press.
- Haninger, K., Hegeler, C. and Peternele, L. (2022). Model predictive control with gaussian processes for flexible multi-modal physical human robot interaction. In *2022 international conference on robotics and automation (ICRA)*. IEEE.

- Haninger, K., Hegeler, C. and Peternel, L. (2023). Model predictive impedance control with gaussian processes for human and environment interaction. *Robotics and Autonomous Systems*, **165** 104431.
- Hao, B. and Lattimore, T. (2022). Regret bounds for information-directed reinforcement learning. *Advances in neural information processing systems*, **35** 28575–28587.
- Hao, B., Lattimore, T. and Deng, W. (2021). Information directed sampling for sparse linear bandits. *Advances in Neural Information Processing Systems*, **34** 16738–16750.
- Hao, B., Lattimore, T. and Qin, C. (2022). Contextual information-directed sampling. In *International Conference on Machine Learning*. PMLR.
- Ji, K., He, J. and Gu, Q. (2024). Reinforcement learning from human feedback with active queries. *arXiv preprint arXiv:2402.09401*.
- Kirschner, J., Lattimore, T., Vernade, C. and Szepesvári, C. (2021). Asymptotically optimal information-directed sampling. In *Conference on Learning Theory*. PMLR.
- Li, X., Zhao, H. and Gu, Q. (2024). Feel-good thompson sampling for contextual dueling bandits. *arXiv preprint arXiv:2404.06013*.
- Liu, F., Buccapatnam, S. and Shroff, N. (2018). Information directed sampling for stochastic bandits with graph feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32.
- Moradipari, A., Pedramfar, M., Shokrian Zini, M. and Aggarwal, V. (2023). Improved bayesian regret bounds for thompson sampling in reinforcement learning. *Advances in Neural Information Processing Systems*, **36** 23557–23569.
- Osband, I., Russo, D. and Van Roy, B. (2013). (more) efficient reinforcement learning via posterior sampling. *Advances in Neural Information Processing Systems*, **26**.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A. et al. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, **35** 27730–27744.
- Pacchiano, A., Saha, A. and Lee, J. (2021). Dueling rl: reinforcement learning with trajectory preferences. *arXiv preprint arXiv:2111.04850*.
- Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S. and Finn, C. (2024). Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, **36**.
- Russo, D. and Van Roy, B. (2014). Learning to optimize via information-directed sampling. *Advances in neural information processing systems*, **27**.
- Russo, D. and Van Roy, B. (2016). An information-theoretic analysis of thompson sampling. *Journal of Machine Learning Research*, **17** 1–30.
- Saha, A. (2021). Optimal algorithms for stochastic contextual preference bandits. *Advances in Neural Information Processing Systems*, **34** 30050–30062.
- Saha, A., Pacchiano, A. and Lee, J. (2023). Dueling rl: Reinforcement learning with trajectory preferences. In *International Conference on Artificial Intelligence and Statistics*. PMLR.

- Schulman, J., Wolski, F., Dhariwal, P., Radford, A. and Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- Sekhari, A., Sridharan, K., Sun, W. and Wu, R. (2024). Contextual bandits and imitation learning with preference-based active queries. *Advances in Neural Information Processing Systems*, **36**.
- Taranovic, A., Kupcsik, A. G., Freymuth, N. and Neumann, G. (2022). Adversarial imitation learning with preferences. In *The Eleventh International Conference on Learning Representations*.
- Tossou, A., Basu, D. and Dimitrakakis, C. (2019). Near-optimal optimistic reinforcement learning using empirical bernstein inequalities. *arXiv preprint arXiv:1905.12425*.
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S. et al. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Wu, R. and Sun, W. (2023). Making rl with preference-based feedback efficient via randomization. *arXiv preprint arXiv:2310.14554*.
- Wu, S. A., Wang, R. E., Evans, J. A., Tenenbaum, J. B., Parkes, D. C. and Kleiman-Weiner, M. (2021). Too many cooks: Bayesian inference for coordinating multi-agent collaboration. *Topics in Cognitive Science*, **13** 414–432.
- Xie, T., Foster, D. J., Krishnamurthy, A., Rosset, C., Awadallah, A. and Rakhlin, A. (2024). Exploratory preference optimization: Harnessing implicit  $q^*$ -approximation for sample-efficient rlhf. *arXiv preprint arXiv:2405.21046*.
- Xu, Y., Wang, R., Yang, L., Singh, A. and Dubrawski, A. (2020). Preference-based reinforcement learning with finite-time guarantees. *Advances in Neural Information Processing Systems*, **33** 18784–18794.
- Ye, C., Xiong, W., Zhang, Y., Jiang, N. and Zhang, T. (2024). A theoretical analysis of nash learning from human feedback under general kl-regularized preference. *arXiv preprint arXiv:2402.07314*.
- Zhang, Q., Bai, C., Hu, S., Wang, Z. and Li, X. (2024). Provably efficient information-directed sampling algorithms for multi-agent reinforcement learning. *arXiv preprint arXiv:2404.19292*.
- Zhu, B., Jordan, M. and Jiao, J. (2023). Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *International Conference on Machine Learning*. PMLR.
- Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P. and Irving, G. (2019). Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

## A Proofs of Theorem and Proposition

### A.1 Proof of Theorem 4.11

**Theorem.** Given a Bayesian RLHF problem, for any  $\epsilon > 0$  and sufficiently large  $T$ , by choosing  $\lambda = \sqrt{\alpha^2 TH / \log(K(\epsilon))}$ , we have

$$BR_T(\pi_{\text{IDS}}) \leq \alpha \sqrt{TH \log(K(\epsilon))} + T\epsilon + T_0, \quad (\text{A.1})$$

where  $T_0$  is a fixed positive integer that is independent of  $T$ .

*Proof.* We divide the proof into 5 steps. First, we point out that by the law of total expectation, we can rewrite the Bayesian regret as

$$BR_T(\pi_{\text{IDS}}) = \sum_{t=1}^T \mathbb{E}_{\mathcal{D}_t} \left[ \mathbb{E}_{\mathcal{E} \sim \mathbb{P}(\cdot | \mathcal{D}_t)} \left[ V_{1,\pi_{\mathcal{E}}^*}^{\mathcal{E}}(s_1^t) - V_{1,\pi_{\text{IDS}}^t}^{\mathcal{E}}(s_1^t) \right] \right], \quad (\text{A.2})$$

whose form is more convenient for analysis.

**Step 1.** Reduce  $BR_T(\pi_{\text{IDS}})$  to the surrogate environment, and convert  $BR_T(\pi_{\text{IDS}})$  into  $BR_T(\pi_{\text{TS}})$ . By Lemma 4.9 and the optimality of  $\pi_{\text{IDS}}$ , we have

$$\begin{aligned} BR_T(\pi_{\text{IDS}}) &= \sum_{t=1}^T \mathbb{E}_{\mathcal{D}_t} \left[ \mathbb{E}_{\mathcal{E} \sim \mathbb{P}(\cdot | \mathcal{D}_t)} \left[ V_{1,\pi_{\mathcal{E}}^*}^{\mathcal{E}}(s_1^t) - V_{1,\pi_{\text{IDS}}^t}^{\mathcal{E}}(s_1^t) \right] \right] \\ &= \sum_{t=1}^T \mathbb{E}_{\mathcal{D}_t} \left[ \mathbb{E}_t \left[ V_{1,\pi_{\mathcal{E}}^*}^{\mathcal{E}}(s_1^t) - V_{1,\pi_{\text{IDS}}^t}^{\mathcal{E}}(s_1^t) \right] - \epsilon - \frac{\lambda}{2} \mathbb{I}_t^{\pi_{\text{IDS}}^t} \left( \tilde{\mathcal{E}}_t^*; (\mathcal{H}_t, \mathcal{R}_{t,H}) \right) \right] \\ &\quad + \frac{\lambda}{2} \sum_{t=1}^T \mathbb{E}_{\mathcal{D}_t} \left[ \mathbb{I}_t^{\pi_{\text{IDS}}^t} \left( \tilde{\mathcal{E}}_t^*; (\mathcal{H}_t, \mathcal{R}_{t,H}) \right) \right] + T\epsilon \\ &\stackrel{(a)}{\leq} \sum_{t=1}^T \mathbb{E}_{\mathcal{D}_t} \left[ \mathbb{E}_t \left[ V_{1,\pi_{\mathcal{E}}^*}^{\mathcal{E}}(s_1^t) - V_{1,\pi_{\text{TS}}^t}^{\mathcal{E}}(s_1^t) \right] - \epsilon - \frac{\lambda}{2} \mathbb{I}_t^{\pi_{\text{TS}}^t} \left( \tilde{\mathcal{E}}_t^*; (\mathcal{H}_t, \mathcal{R}_{t,H}) \right) \right] \\ &\quad + \frac{\lambda}{2} \sum_{t=1}^T \mathbb{E}_{\mathcal{D}_t} \left[ \mathbb{I}_t^{\pi_{\text{IDS}}^t} \left( \tilde{\mathcal{E}}_t^*; (\mathcal{H}_t, \mathcal{R}_{t,H}) \right) \right] + T\epsilon \\ &\stackrel{(b)}{\leq} \sum_{t=1}^T \mathbb{E}_{\mathcal{D}_t} \left[ \mathbb{E}_t \left[ V_{1,\pi_{\mathcal{E}}^*}^{\tilde{\mathcal{E}}_t^*}(s_1^t) - V_{1,\pi_{\text{TS}}^t}^{\tilde{\mathcal{E}}_t^*}(s_1^t) \right] - \frac{\lambda}{2} \mathbb{I}_t^{\pi_{\text{TS}}^t} \left( \tilde{\mathcal{E}}_t^*; (\mathcal{H}_t, \mathcal{R}_{t,H}) \right) \right] \\ &\quad + \frac{\lambda}{2} \sum_{t=1}^T \mathbb{E}_{\mathcal{D}_t} \left[ \mathbb{I}_t^{\pi_{\text{IDS}}^t} \left( \tilde{\mathcal{E}}_t^*; (\mathcal{H}_t, \mathcal{R}_{t,H}) \right) \right] + T\epsilon, \end{aligned} \quad (\text{A.3})$$

where (a) uses the optimality of  $\pi_{\text{IDS}}^t$ , (b) uses Lemma 4.9.

For the first term in Eqn. (A.3), using the basic fact that  $A - \lambda B/2 \leq A^2/2\lambda B$  for  $B, \lambda \geq 0$ , we have

$$\mathbb{E}_t \left[ V_{1,\pi_{\mathcal{E}}^*}^{\tilde{\mathcal{E}}_t^*}(s_1^t) - V_{1,\pi_{\text{TS}}^t}^{\tilde{\mathcal{E}}_t^*}(s_1^t) \right] - \frac{\lambda}{2} \mathbb{I}_t^{\pi_{\text{TS}}^t} \left( \tilde{\mathcal{E}}_t^*; (\mathcal{H}_t, \mathcal{R}_{t,H}) \right) \leq \frac{1}{2\lambda} \frac{\left( \mathbb{E}_t \left[ V_{1,\pi_{\mathcal{E}}^*}^{\tilde{\mathcal{E}}_t^*}(s_1^t) - V_{1,\pi_{\text{TS}}^t}^{\tilde{\mathcal{E}}_t^*}(s_1^t) \right] \right)^2}{\mathbb{I}_t^{\pi_{\text{TS}}^t} \left( \tilde{\mathcal{E}}_t^*; (\mathcal{H}_t, \mathcal{R}_{t,H}) \right)} \triangleq \frac{1}{2\lambda} \Gamma_t^{\pi_{\text{TS}}^t}. \quad (\text{A.4})$$

where we introduce the tool of *information ratio*  $\Gamma_t^{\pi_{\text{TS}}^t}$  for ease of analysis.

Let  $\zeta$  be a discrete random variable taking values in  $\{1, \dots, K(\epsilon)\}$  such that  $\zeta = k$  if and only if  $\mathcal{E} \in \Theta_k^\epsilon$ . From the construction of the surrogate environment (Eqn. (4.9)), the distribution of  $\tilde{\mathcal{E}}_t^*$  depend on  $\mathcal{E}$  only through  $\zeta$ , i.e.,  $\tilde{\mathcal{E}}_t^*$  and  $\mathcal{E}$  are independent conditioning on  $\zeta$ .

For the second term in Eqn. (A.3), we have

$$\sum_{t=1}^T \mathbb{E}_{\mathcal{D}_t} \left[ \mathbb{I}_t^{\pi_{\text{IDS}}^t} \left( \tilde{\mathcal{E}}_t^*; (\mathcal{H}_t, \mathcal{R}_{t,H}) \right) \right] \leq \sum_{t=1}^T \mathbb{E}_{\mathcal{D}_t} \left[ \mathbb{I}_t^{\pi_{\text{IDS}}^t} (\zeta; (\mathcal{H}_t, \mathcal{R}_{t,H})) \right] = \mathbb{I}(\zeta; \mathcal{D}_{T+1}) \leq \mathbb{H}(\zeta) \leq \log(K(\epsilon)), \quad (\text{A.5})$$

where the first inequality is due to data processing inequality, the second equality is due to the chain rule of mutual information, and the last two inequalities follow from the basic definition of entropy. Therefore, we derive an upper bound for  $BR_T(\pi_{\text{IDS}})$  as follows

$$BR_T(\pi_{\text{IDS}}) \leq \frac{1}{2\lambda} \mathbb{E} \left[ \sum_{t=1}^T \Gamma_t^{\pi_{\text{TS}}^t} \right] + \frac{\lambda}{2} \log(K(\epsilon)) + T\epsilon. \quad (\text{A.6})$$

**Step 2** (Bound  $\Gamma_t^{\pi_{\text{TS}}^t}$ ). Before stepping into technical details, we need to introduce several concepts. First, the state-action occupancy function  $d_{h,\pi}^\mathcal{E} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  at step  $h$  under policy  $\pi$  and environment  $\mathcal{E}$ , is defined as the Radon-Nikodym derivative of the state-action occupancy measure  $\mathbb{P}_\pi^\mathcal{E}((s_h, a_h) = \cdot)$  with regard to the base probability measure  $\mu_{\mathcal{S} \times \mathcal{A}}$  on the product space  $\mathcal{S} \times \mathcal{A}$ , i.e.,

$$d_{h,\pi}^\mathcal{E}(s, a) \triangleq \frac{d\mathbb{P}_\pi^\mathcal{E}(s_h = s, a_h = a)}{d\mu_{\mathcal{S} \times \mathcal{A}}}.$$

For convenience of analysis, we assume that  $d_{h,\pi}^\mathcal{E}(s, a)$  is measurable and upper bounded for all  $\pi, \mathcal{E}, s, a, h$ . Recall that, the mean environment  $\bar{\mathcal{E}}_t$  is defined to satisfy  $P_h^{\bar{\mathcal{E}}_t}(\cdot|s, a) = \mathbb{E}_t[P_h^\mathcal{E}(\cdot|s, a)]$  and  $R_h^{\bar{\mathcal{E}}_t}(\cdot|s, a) = \mathbb{E}_t[R_h^\mathcal{E}(\cdot|s, a)]$  for all  $s \in \mathcal{S}$  and  $a \in \mathcal{A}$ . By the definition of  $d_{h,\pi}^\mathcal{E}$ , the following equality also holds:  $d_{h,\pi}^{\bar{\mathcal{E}}_t}(s, a) = \mathbb{E}_t[d_{h,\pi}^\mathcal{E}(s, a)]$ . One important property of the mean environment is that the posterior mean of the surrogate environment  $\mathbb{E}_t[\tilde{\mathcal{E}}_t^*]$  coincides with that of the whole environment  $\bar{\mathcal{E}}_t$ . To check this, using the property of conditional expectation:

$$\begin{aligned} \mathbb{E}_t[\tilde{\mathcal{E}}_t^*] &= \sum_{k=1}^K \mathbb{P}(\mathcal{E} \in \Theta_k^\epsilon) \cdot \mathbb{E}_t[\tilde{\mathcal{E}}_t^* | \mathcal{E} \in \Theta_k^\epsilon] \\ &\stackrel{(a)}{=} \sum_{k=1}^K \mathbb{P}(\mathcal{E} \in \Theta_k^\epsilon) \cdot \mathbb{E}_t[\tilde{\mathcal{E}}_{k,t}^*] \\ &\stackrel{(b)}{=} \sum_{k=1}^K \mathbb{P}(\mathcal{E} \in \Theta_k^\epsilon) \cdot \tilde{\mathcal{E}}_{k,t}^* \\ &\stackrel{(c)}{=} \sum_{k=1}^K \mathbb{P}(\mathcal{E} \in \Theta_k^\epsilon) \cdot \mathbb{E}_t[\mathcal{E} | \mathcal{E} \in \Theta_k^\epsilon] = \bar{\mathcal{E}}_t, \end{aligned} \quad (\text{A.7})$$

where (a) comes from the definition of  $\tilde{\mathcal{E}}_t^*$  and  $\tilde{\mathcal{E}}_{k,t}^*$ , (b) and (c) uses the following fact

$$\mathbb{E}_t[\tilde{\mathcal{E}}_{k,t}^*] = \mathbb{E}_t[\mathbb{E}_t[\mathcal{E} | \mathcal{E} \in \Theta_k^\epsilon]] = \mathbb{E}_t[\mathcal{E} | \mathcal{E} \in \Theta_k^\epsilon] = \tilde{\mathcal{E}}_t^*$$

Finally, we denote the value function difference as

$$\Delta_h^{\tilde{\mathcal{E}}_t^*}(s, a) \triangleq \mathbb{E}_{(s', r') \sim (P_h^{\tilde{\mathcal{E}}_t^*} \otimes R_h^{\tilde{\mathcal{E}}_t^*})(\cdot|s, a)} \left[ r' + V_{h+1, \pi_\mathcal{E}^*}^{\tilde{\mathcal{E}}_t^*}(s') \right] - \mathbb{E}_{(s', r') \sim (P_h^{\tilde{\mathcal{E}}_t} \otimes R_h^{\tilde{\mathcal{E}}_t})(\cdot|s, a)} \left[ r' + V_{h+1, \pi_\mathcal{E}^*}^{\tilde{\mathcal{E}}_t}(s') \right]. \quad (\text{A.8})$$

Now we are ready to give an upper bound for  $\Gamma_t^{\pi_{\text{TS}}^t}$ . We hope to use Lemma C.1 to rewrite the numerator

$$\left( \mathbb{E}_t \left[ V_{1, \pi_\mathcal{E}^*}^{\tilde{\mathcal{E}}_t^*}(s_1^t) - V_{1, \pi_{\text{TS}}^t}^{\tilde{\mathcal{E}}_t^*}(s_1^t) \right] \right)^2.$$

However, Lemma C.1 can only be applied to handle the difference between two value functions with the same policy and different environments, while in  $V_{1, \pi_\mathcal{E}^*}^{\tilde{\mathcal{E}}_t^*}(s_1^t)$  and  $V_{1, \pi_{\text{TS}}^t}^{\tilde{\mathcal{E}}_t^*}(s_1^t)$ , the environments are the same and the policies are different. For the purpose of “unifying” the policy, we use Eqn. (A.7) and note that  $\pi_{\text{TS}}$  is independent of  $\tilde{\mathcal{E}}_t^*$ , yielding

$$\mathbb{E}_t \left[ V_{1, \pi_{\text{TS}}^t}^{\tilde{\mathcal{E}}_t^*}(s_1^t) \right] = \mathbb{E}_t \left[ V_{1, \pi_\mathcal{E}^*}^{\tilde{\mathcal{E}}_t^*}(s_1^t) \right].$$

Furthermore, conditioned on  $\mathcal{D}_t$ ,  $\pi_{\text{TS}}^t$  and  $\pi_\mathcal{E}^*$  are identically distributed, and are both independent of  $\tilde{\mathcal{E}}_t$ . This implies

$$\mathbb{E}_t \left[ V_{1, \pi_{\text{TS}}^t}^{\tilde{\mathcal{E}}_t^*}(s_1^t) \right] = \mathbb{E}_t \left[ V_{1, \pi_\mathcal{E}^*}^{\tilde{\mathcal{E}}_t^*}(s_1^t) \right].$$

Therefore, by Lemma C.1, we have

$$\begin{aligned} \mathbb{E}_t \left[ V_{1, \pi_\mathcal{E}^*}^{\tilde{\mathcal{E}}_t^*}(s_1^t) - V_{1, \pi_{\text{TS}}^t}^{\tilde{\mathcal{E}}_t^*}(s_1^t) \right] &= \mathbb{E}_t \left[ V_{1, \pi_\mathcal{E}^*}^{\tilde{\mathcal{E}}_t^*}(s_1^t) - V_{1, \pi_\mathcal{E}^*}^{\tilde{\mathcal{E}}_t}(s_1^t) \right] \\ &= \sum_{h=1}^H \mathbb{E}_t \mathbb{E}_{\pi_\mathcal{E}^*}^{\tilde{\mathcal{E}}_t} \left[ \Delta_h^{\tilde{\mathcal{E}}_t^*}(s, a) \right] \\ &= \sum_{h=1}^H \mathbb{E}_t \left[ \int_{\mathcal{S} \times \mathcal{A}} d_{h, \pi_\mathcal{E}^*}^{\tilde{\mathcal{E}}_t}(s, a) \Delta_h^{\tilde{\mathcal{E}}_t^*}(s, a) d\mu_{\mathcal{S} \times \mathcal{A}} \right], \end{aligned} \quad (\text{A.9})$$

where the notation of  $d_{h, \pi_\mathcal{E}^*}^{\tilde{\mathcal{E}}_t}(s, a)$  and  $\Delta_h^{\tilde{\mathcal{E}}_t^*}(s, a)$  are introduced to simplify the formula. Following Moradipari et al. (2023), we define

$$\mathcal{I}^t \triangleq \sum_{h=1}^H \mathbb{E}_t \mathbb{E}_{\pi_{\text{TS}}^t}^{\tilde{\mathcal{E}}_t} \left[ \frac{\Delta_h^{\tilde{\mathcal{E}}_t^*}(s, a)^2}{\alpha_\mathcal{E}^2} \right], \quad \mathcal{T}^t \triangleq \sum_{h=1}^H \int_{\mathbb{E}_t[d_{h, \pi_\mathcal{E}^*}^{\tilde{\mathcal{E}}_t}(s, a)] \neq 0} \frac{\mathbb{E}_t \left[ \alpha_\mathcal{E}^2 \cdot d_{h, \pi_\mathcal{E}^*}^{\tilde{\mathcal{E}}_t}(s, a)^2 \right]}{\mathbb{E}_t \left[ d_{h, \pi_\mathcal{E}^*}^{\tilde{\mathcal{E}}_t}(s, a) \right]} d\mu_{\mathcal{S} \times \mathcal{A}}. \quad (\text{A.10})$$

By the Cauchy-Schwarz inequality, we have

$$\sum_{h=1}^H \mathbb{E}_t \left[ \int_{\mathcal{S} \times \mathcal{A}} d_{h, \pi_\mathcal{E}^*}^{\tilde{\mathcal{E}}_t}(s, a) \Delta_h^{\tilde{\mathcal{E}}_t^*}(s, a) d\mu_{\mathcal{S} \times \mathcal{A}} \right]$$

$$\begin{aligned}
&= \sum_{h=1}^H \mathbb{E}_t \left[ \int_{\mathbb{E}_t[d_{h,\pi_\mathcal{E}^*}^{\bar{\mathcal{E}}_t}(s,a)] \neq 0} d_{h,\pi_\mathcal{E}^*}^{\bar{\mathcal{E}}_t}(s,a) \Delta_h^{\bar{\mathcal{E}}_t^*}(s,a) d\mu_{\mathcal{S} \times \mathcal{A}} \right] \\
&\leq \left( \sum_{h=1}^H \mathbb{E}_t \int_{\mathbb{E}_t[d_{h,\pi_\mathcal{E}^*}^{\bar{\mathcal{E}}_t}(s,a)] \neq 0} \frac{\alpha_\mathcal{E}^2 \cdot d_{h,\pi_\mathcal{E}^*}^{\bar{\mathcal{E}}_t}(s,a)^2}{\mathbb{E}_t[d_{h,\pi_\mathcal{E}^*}^{\bar{\mathcal{E}}_t}(s,a)]} \right)^{1/2} \left( \sum_{h=1}^H \mathbb{E}_t \int_{\mathcal{S} \times \mathcal{A}} \mathbb{E}_t[d_{h,\pi_\mathcal{E}^*}^{\bar{\mathcal{E}}_t}(s,a)] \cdot \frac{\Delta_h^{\bar{\mathcal{E}}_t^*}(s,a)^2}{\alpha_\mathcal{E}^2} \right)^{1/2} \\
&= \sqrt{\mathcal{I}^t \cdot \mathcal{T}^t}, \tag{A.11}
\end{aligned}$$

where the first equality is due to the fact that  $\Delta_h^{\bar{\mathcal{E}}_t^*}(s,a)$  is bounded ( $\leq 2H$ ), and the second inequality is simply the Cauchy-Schwarz inequality with  $\sum_h \mathbb{E}_t \int_{\mathcal{S} \times \mathcal{A}}$  as an ‘‘integrated’’ integral over the space  $[H] \times \Theta \times \mathcal{S} \times \mathcal{A}$ . Let us briefly discuss why the third equality holds. For the term  $\mathcal{T}^t$ , the derivation is straightforward, since  $\mathbb{E}_t[X/\mathbb{E}_t[Y]] = \mathbb{E}_t[X]/\mathbb{E}_t[Y]$ . For the term  $\mathcal{I}^t$ , first recall that  $\mathbb{E}_t[d_{h,\pi_\mathcal{E}^*}^{\bar{\mathcal{E}}_t}(s,a)] = \mathbb{E}_t[d_{h,\pi_{\text{TS}}^t}^{\bar{\mathcal{E}}_t}(s,a)]$  due to the property of TS. Then, we can use the independence between  $d_{h,\pi_{\text{TS}}^t}^{\bar{\mathcal{E}}_t}(s,a)$  and  $\Delta_h^{\bar{\mathcal{E}}_t^*}(s,a)$  given  $\mathcal{D}_t$  to ‘‘extract’’ the expectation ( $\mathbb{E}[XY] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$  for  $X, Y$  mutually independent):

$$\sum_{h=1}^H \mathbb{E}_t \int_{\mathcal{S} \times \mathcal{A}} \mathbb{E}_t[d_{h,\pi_\mathcal{E}^*}^{\bar{\mathcal{E}}_t}(s,a)] \cdot \frac{\Delta_h^{\bar{\mathcal{E}}_t^*}(s,a)^2}{\alpha_\mathcal{E}^2} = \sum_{h=1}^H \mathbb{E}_t \int_{\mathcal{S} \times \mathcal{A}} d_{h,\pi_{\text{TS}}^t}^{\bar{\mathcal{E}}_t}(s,a) \cdot \frac{\Delta_h^{\bar{\mathcal{E}}_t^*}(s,a)^2}{\alpha_\mathcal{E}^2} = \mathcal{I}^t. \tag{A.12}$$

To summarize, by Eqn. (A.11) we have the following bound for  $\Gamma_t^{\pi_{\text{TS}}^t}$ :

$$\Gamma_t^{\pi_{\text{TS}}^t} \leq \frac{\mathcal{I}^t \cdot \mathcal{T}^t}{\mathbb{I}_t^{\pi_{\text{TS}}^t}(\bar{\mathcal{E}}_t^*; (\mathcal{H}_t, \mathcal{R}_{t,H}))}. \tag{A.13}$$

**Step 3** (Bound  $\mathcal{I}^t$ ). The key observation in this step is that  $\mathcal{I}^t$  is in the form of total variation, and thus can be upper bounded by mutual information (in the form of KL divergence) by Pinsker’s inequality. Specifically,

$$\begin{aligned}
\mathcal{I}^t &= \sum_{h=1}^H \mathbb{E}_t \mathbb{E}_{\pi_{\text{TS}}^t}^{\bar{\mathcal{E}}_t} \left( \mathbb{E}_{(s',r') \sim (P_h^{\bar{\mathcal{E}}_t^*} \otimes R_h^{\bar{\mathcal{E}}_t^*})(\cdot|s_h, a_h)} \left[ \frac{r' + V_{h+1,\pi_\mathcal{E}^*}^{\bar{\mathcal{E}}_t^*}(s')}{\alpha_\mathcal{E}} \right] - \mathbb{E}_{(s',r') \sim (P_h^{\bar{\mathcal{E}}_t} \otimes R_h^{\bar{\mathcal{E}}_t})(\cdot|s_h, a_h)} \left[ \frac{r' + V_{h+1,\pi_\mathcal{E}^*}^{\bar{\mathcal{E}}_t^*}(s')}{\alpha_\mathcal{E}} \right] \right)^2 \\
&\stackrel{(a)}{=} \sum_{h=1}^H \mathbb{E}_t \mathbb{E}_{\pi_{\text{TS}}^t}^{\bar{\mathcal{E}}_t} \left( \mathbb{E}_{(s',r') \sim (P_h^{\bar{\mathcal{E}}_t^*} \otimes R_h^{\bar{\mathcal{E}}_t^*})(\cdot|s_h, a_h)} \left[ \frac{r' - r_h^{\inf}(s_h, a_h) + V_{h+1,\pi_\mathcal{E}^*}^{\bar{\mathcal{E}}_t^*}(s') - \inf_s V_{h+1,\pi_\mathcal{E}^*}^{\bar{\mathcal{E}}_t^*}(s')}{\alpha_\mathcal{E}} \right] \right. \\
&\quad \left. - \mathbb{E}_{(s',r') \sim (P_h^{\bar{\mathcal{E}}_t} \otimes R_h^{\bar{\mathcal{E}}_t})(\cdot|s_h, a_h)} \left[ \frac{r' - r_h^{\inf}(s_h, a_h) + V_{h+1,\pi_\mathcal{E}^*}^{\bar{\mathcal{E}}_t}(s') - \inf_s V_{h+1,\pi_\mathcal{E}^*}^{\bar{\mathcal{E}}_t}(s')}{\alpha_\mathcal{E}} \right] \right)^2 \\
&\stackrel{(b)}{\leq} \frac{1}{2} \sum_{h=1}^H \mathbb{E}_t \mathbb{E}_{\pi_{\text{TS}}^t}^{\bar{\mathcal{E}}_t} \left[ D_{\text{KL}} \left( \left( P_h^{\bar{\mathcal{E}}_t^*} \otimes R_h^{\bar{\mathcal{E}}_t^*} \right)(\cdot|s_h, a_h) \middle\| \left( P_h^{\bar{\mathcal{E}}_t} \otimes R_h^{\bar{\mathcal{E}}_t} \right)(\cdot|s_h, a_h) \right) \right] \\
&\stackrel{(c)}{\leq} \frac{1}{2} \mathbb{I}_t^{\pi_{\text{TS}}^t}(\bar{\mathcal{E}}_t^*; (\mathcal{H}_t, \mathcal{R}_{t,H})), \tag{A.14}
\end{aligned}$$

where (a) adds and subtracts the constant term  $r_h^{\inf}(s_h, a_h)$  and  $\inf_s V_{h+1,\pi_\mathcal{E}^*}^{\bar{\mathcal{E}}_t^*}(s')$ , (b) uses the Pinsker’s

inequality, (c) uses Lemma C.2. Plugging into Eqn. (A.13), we derive that  $\Gamma_t^{\pi_{\text{TS}}^t} \leq \frac{1}{2}\mathcal{T}^t$ , and thus

$$BR_T(\pi_{\text{IDS}}) \leq \frac{1}{4\lambda} \mathbb{E} \left[ \sum_{t=1}^T \mathcal{T}^t \right] + \frac{\lambda}{2} \log(K(\epsilon)) + T\epsilon. \quad (\text{A.15})$$

**Step 4** (Bound  $\mathbb{E}[\mathcal{T}^t]$ ). The analysis tools used in this step is the Doob's consistency theorem, with more details discussed in Appendix C.2. Define the true environment as  $\mathcal{E}_0$ . For brevity of notations, we define

$$\mathcal{B}_{h,t} \triangleq \left\{ (s, a) \in \mathcal{S} \times \mathcal{A} \mid \mathbb{E} \left[ d_{h,\pi_{\mathcal{E}}}^{\bar{\mathcal{E}}_t}(s, a) \right] \neq 0 \right\},$$

so that we can write  $\mathcal{T}^t$  as

$$\mathcal{T}^t = \sum_{h=1}^H \int_{(s,a) \in \mathcal{B}_{h,t}} \frac{\mathbb{E}_t \left[ \alpha_{\mathcal{E}}^2 \cdot d_{h,\pi_{\mathcal{E}}}^{\bar{\mathcal{E}}_t}(s, a)^2 \right]}{\mathbb{E}_t \left[ d_{h,\pi_{\mathcal{E}}}^{\bar{\mathcal{E}}_t}(s, a) \right]} d\mu_{\mathcal{S} \times \mathcal{A}}. \quad (\text{A.16})$$

We now introduce another variable  $\mathcal{E}'$  to apply Lemma C.3. Let  $\mathcal{E}' \sim \mathbb{P}(\cdot | \mathcal{D}_t)$  be a random variable independent of  $\mathcal{E}$ . By definition of  $d_{h,\pi_{\mathcal{E}}}^{\mathcal{E}}$ , we have  $\mathbb{E}_t \left[ d_{h,\pi_{\mathcal{E}}}^{\bar{\mathcal{E}}_t}(s, a) \right] = \mathbb{E}_{\mathcal{E}, \mathcal{E}' \sim \mathbb{P}_t(\cdot)} \left[ d_{h,\pi_{\mathcal{E}}}^{\mathcal{E}'}(s, a) \right]$ , and

$$\begin{aligned} \mathbb{E}_t \left[ \alpha_{\mathcal{E}}^2 \cdot d_{h,\pi_{\mathcal{E}}}^{\bar{\mathcal{E}}_t}(s, a)^2 \right] &= \mathbb{E}_{\mathcal{E} \sim \mathbb{P}_t(\cdot)} \left[ \alpha_{\mathcal{E}}^2 \cdot \mathbb{E}_{\mathcal{E}' \sim \mathbb{P}_t(\cdot)} \left[ d_{h,\pi_{\mathcal{E}}}^{\mathcal{E}'}(s, a)^2 \right] \right] \\ &\leq \mathbb{E}_{\mathcal{E}, \mathcal{E}' \sim \mathbb{P}_t(\cdot)} \left[ \alpha_{\mathcal{E}}^2 \cdot d_{h,\pi_{\mathcal{E}}}^{\mathcal{E}'}(s, a)^2 \right], \end{aligned} \quad (\text{A.17})$$

where the inequality is due to the fact that  $\mathbb{E}[X^2] \geq (\mathbb{E}[X])^2$ . Therefore, we have

$$\mathcal{T}^t \leq \sum_{h=1}^H \int_{\mathcal{S} \times \mathcal{A}} \frac{\mathbb{E}_t \left[ \alpha_{\mathcal{E}}^2 \cdot d_{h,\pi_{\mathcal{E}}}^{\mathcal{E}'}(s, a)^2 \right]}{\mathbb{E}_t \left[ d_{h,\pi_{\mathcal{E}}}^{\mathcal{E}'}(s, a) \right]} \chi_{\mathcal{B}_{h,t}}(s, a) d\mu_{\mathcal{S} \times \mathcal{A}} = \sum_{h=1}^H \int_{\mathcal{S} \times \mathcal{A}} g_t(s, a, h, \mathcal{D}_t) d\mu_{\mathcal{S} \times \mathcal{A}}, \quad (\text{A.18})$$

where  $\chi_A(\cdot)$  is the indicator function, i.e.,  $\chi_A(x) = 1$  if  $x \in A$ ;  $\chi_A(x) = 0$  if  $x \notin A$ . Since  $d_{h,\pi}^{\mathcal{E}}(s, a)$  is assumed to be bounded, let

$$M_d \triangleq \sup_{s,a,h,\pi_{\mathcal{E}}} d_{h,\pi}^{\mathcal{E}}(s, a) < \infty.$$

This implies  $g_t(s, a, h, \mathcal{D}_t) \leq M_d H^2$  and  $\mathcal{T}^t \leq M_d H^3$ . By Lemma C.3, we have the following convergence results: for any  $(s, a) \in \mathcal{S} \times \mathcal{A}$ ,

$$\lim_{t \rightarrow \infty} \mathbb{E}_t \left[ \alpha_{\mathcal{E}}^2 d_{h,\pi_{\mathcal{E}}}^{\mathcal{E}'}(s, a)^2 \right] = \alpha_{\mathcal{E}_0}^2 d_{h,\pi_{\mathcal{E}_0}}^{\mathcal{E}_0}(s, a)^2, \quad (\text{A.19})$$

$$\lim_{t \rightarrow \infty} \mathbb{E}_t \left[ d_{h,\pi_{\mathcal{E}}}^{\mathcal{E}'}(s, a) \right] = d_{h,\pi_{\mathcal{E}_0}}^{\mathcal{E}_0}(s, a), \quad (\text{A.20})$$

and for any  $x$ ,

$$\lim_{t \rightarrow \infty} \chi_{\mathcal{B}_{h,t}}(x) = \chi_{\mathcal{B}_h}(x), \quad \text{where } \mathcal{B}_h \triangleq \left\{ (s, a) \in \mathcal{S} \times \mathcal{A} \mid \mathbb{E} \left[ d_{h,\pi_{\mathcal{E}_0}}^{\mathcal{E}_0}(s, a) \right] \neq 0 \right\}. \quad (\text{A.21})$$

From Eqn. (A.21), we can also derive that  $\lim_{t \rightarrow \infty} \chi_{\mathcal{B}_{h,t} \cap \mathcal{B}_h}(x) = 1$ , and  $\lim_{t \rightarrow \infty} \chi_{\mathcal{B}_{h,t} \setminus \mathcal{B}_h}(x) = 0$ . By

Lebesgue dominated convergence theorem, we have

$$\begin{aligned}
\lim_{t \rightarrow \infty} \mathbb{E} [\mathcal{T}^t | \mathcal{E}_0] &= \lim_{t \rightarrow \infty} \mathbb{E}_{\mathcal{D}_t \sim \mathbb{P}(\cdot | \mathcal{E}_0)} \sum_{h=1}^H \int_{\mathcal{S} \times \mathcal{A}} \frac{\mathbb{E}_t [\alpha_{\mathcal{E}}^2 \cdot d_{h, \pi_{\mathcal{E}}^*}^{\mathcal{E}'}(s, a)^2]}{\mathbb{E}_t [d_{h, \pi_{\mathcal{E}}^*}^{\mathcal{E}'}(s, a)]} \cdot \chi_{\mathcal{B}_{h,t}}(s, a) d\mu_{\mathcal{S} \times \mathcal{A}} \\
&\leq \lim_{t \rightarrow \infty} \mathbb{E}_{\mathcal{D}_t \sim \mathbb{P}(\cdot | \mathcal{E}_0)} \sum_{h=1}^H \int_{\mathcal{S} \times \mathcal{A}} \frac{\mathbb{E}_t [\alpha_{\mathcal{E}}^2 \cdot d_{h, \pi_{\mathcal{E}}^*}^{\mathcal{E}'}(s, a)^2]}{\mathbb{E}_t [d_{h, \pi_{\mathcal{E}}^*}^{\mathcal{E}'}(s, a)]} \cdot \chi_{\mathcal{B}_{h,t} \cap \mathcal{B}_h}(s, a) \\
&\quad + M_d H^2 \lim_{t \rightarrow \infty} \sum_{h=1}^H \int_{\mathcal{S} \times \mathcal{A}} \chi_{\mathcal{B}_{h,t} \setminus \mathcal{B}_h}(s, a) \\
&= \sum_{h=1}^H \int_{\mathcal{S} \times \mathcal{A}} \mathbb{E}_{\mathcal{D}_t \sim \mathbb{P}(\cdot | \mathcal{E}_0)} \left[ \lim_{t \rightarrow \infty} \frac{\mathbb{E}_t [\alpha_{\mathcal{E}}^2 \cdot d_{h, \pi_{\mathcal{E}}^*}^{\mathcal{E}'}(s, a)^2]}{\mathbb{E}_t [d_{h, \pi_{\mathcal{E}}^*}^{\mathcal{E}'}(s, a)]} \cdot \chi_{\mathcal{B}_{h,t} \cap \mathcal{B}_h}(s, a) \right] \\
&= \alpha_{\mathcal{E}_0}^2 \cdot \int_{\mathcal{S} \times \mathcal{A}} d_{h, \pi_{\mathcal{E}_0}^*}^{\mathcal{E}_0}(s, a) \\
&\leq \alpha_{\mathcal{E}_0}^2 \cdot H. \tag{A.22}
\end{aligned}$$

Thus, we have

$$\lim_{t \rightarrow \infty} \mathbb{E}[\mathcal{T}^t] = \lim_{t \rightarrow \infty} \mathbb{E}[\mathbb{E}[\mathcal{T}^t | \mathcal{E}_0]] = \mathbb{E}[\lim_{t \rightarrow \infty} \mathbb{E}[\mathcal{T}^t | \mathcal{E}_0]] \leq \mathbb{E}[\alpha_{\mathcal{E}_0}^2 H] = \alpha^2 H. \tag{A.23}$$

**Step 5:** By Eqn. (A.23), we derive that there exists  $T_0 > 0$  such that  $\mathbb{E}[\mathcal{T}^t] \leq 2\alpha^2 H$  for  $t > T_0$ . Plugging into Eqn. (A.15) and then taking  $\lambda = \sqrt{\alpha^2 TH / \log(K(\epsilon))}$ , we obtain

$$\begin{aligned}
BR_T(\pi_{\text{r-IDS}}) &\leq \frac{T\alpha^2 H}{2\lambda} + \frac{\lambda}{2} \log(K(\epsilon)) + T\epsilon + T_0 \\
&\leq \alpha\sqrt{TH \log(K(\epsilon))} + T\epsilon + T_0, \tag{A.24}
\end{aligned}$$

which finishes the proof of Theorem 4.11.  $\square$

## A.2 Proof of Theorem 4.15

$$\begin{aligned}
& V_{1,\pi}^{\mathcal{E}}(s_1^\ell) - V_{1,\pi}^{\mathcal{E}'}(s_1^\ell) \\
& \stackrel{(a)}{=} \sum_{h=1}^H \mathbb{E}_{\pi}^{\mathcal{E}'} \left[ \mathbb{E}_{s' \sim P_h^{\mathcal{E}}(\cdot | s_h, a_h)} [V_{h+1,\pi}^{\mathcal{E}}(s')] - \mathbb{E}_{s' \sim P_h^{\mathcal{E}'}(\cdot | s_h, a_h)} [V_{h+1,\pi}^{\mathcal{E}'}(s')] \right] \\
& \quad + \sum_{h=1}^H \mathbb{E}_{\pi}^{\mathcal{E}'} \left[ r_h^{\mathcal{E}}(s_h, a_h) - r_h^{\mathcal{E}'}(s_h, a_h) \right] \\
& \stackrel{(b)}{=} \sum_{h=1}^H \mathbb{E}_{\pi}^{\mathcal{E}'} \left[ P_h^{\mathcal{E}}(\cdot | s_h, a_h)^T V_{h+1,\pi}^{\mathcal{E}}(\cdot) - P_h^{\mathcal{E}'}(\cdot | s_h, a_h)^T V_{h+1,\pi}^{\mathcal{E}'}(\cdot) \right] \\
& \quad + \sum_{h=1}^H \mathbb{E}_{\pi}^{\mathcal{E}'} \left[ P_h^{\mathcal{E}'}(\cdot | s_h, a_h)^T (V_{h+1,\pi}^{\mathcal{E}'}(\cdot) - V_{h+1,\pi}^{\mathcal{E}}(\cdot)) \right] + \sum_{h=1}^H \mathbb{E}_{\pi}^{\mathcal{E}'} \left[ r_h^{\mathcal{E}}(s_h, a_h) - r_h^{\mathcal{E}'}(s_h, a_h) \right] \\
& \stackrel{(c)}{\leq} \sum_{h=1}^H \mathbb{E}_{\pi}^{\mathcal{E}'} \left[ P_h^{\mathcal{E}}(\cdot | s_h, a_h)^T V_{h+1,\pi}^{\mathcal{E}}(\cdot) - P_h^{\mathcal{E}'}(\cdot | s_h, a_h)^T V_{h+1,\pi}^{\mathcal{E}'}(\cdot) \right] \\
& \quad + \sum_{h=1}^H \mathbb{E}_{\pi}^{\mathcal{E}'} \left[ \|V_{h+1,\pi}^{\mathcal{E}'}(\cdot) - V_{h+1,\pi}^{\mathcal{E}}(\cdot)\|_2 \right] + \sum_{h=1}^H \mathbb{E}_{\pi}^{\mathcal{E}'} \left[ r_h^{\mathcal{E}}(s_h, a_h) - r_h^{\mathcal{E}'}(s_h, a_h) \right], \tag{A.25}
\end{aligned}$$

where (a) uses Lemma C.1, (b) adds and subtracts the term  $P_h^{\mathcal{E}'}(\cdot | s_h, a_h)^T V_{h+1,\pi}^{\mathcal{E}'}(\cdot)$ , (c) uses the Cauchy-schwartz inequality.

Since  $P_h^{\mathcal{E}}(\cdot | s_h, a_h)^T V_{h+1,\pi}^{\mathcal{E}}(\cdot) \in [0, H]$ , we can divide the value range  $[0, H]$  evenly into  $\frac{3H^2}{\epsilon}$  parts. For each  $(s, a, h)$ , we construct a covering set  $\{\mathcal{I}_{sah}^1, \dots, \mathcal{I}_{sah}^m\}$  for  $[0, H]$  where  $m = \frac{3H^2}{\epsilon}$ . Each set is of length  $\frac{\epsilon}{3H}$ . Since  $\|V_{h+1,\pi}^{\mathcal{E}}(\cdot)\|_2 \in [0, H\sqrt{S}]$ , we construct a covering set  $\{\mathcal{J}_h^1, \dots, \mathcal{J}_h^{m'}\}$  for  $[0, H\sqrt{S}]$  where  $m' = \frac{6H^2\sqrt{S}}{\epsilon}$ . For reward function, we divide the value range  $[0, 1]$  evenly into  $\frac{2H}{\epsilon}$  parts for all  $(s, a) \in \mathcal{S} \times \mathcal{A}, h \in [H]$ . The covering set is  $\{\mathcal{C}_1, \dots, \mathcal{C}_n\}$  where  $n = \frac{3H}{\epsilon}$ . Then, we construct the partition  $\{\Theta_k\}_{k=1}^K$  that  $\mathcal{E} \in \Theta_k$  if for any  $s, a, h$ ,

$$P_h^{\mathcal{E}}(\cdot | s, a)^T V_{h+1,\pi}^{\mathcal{E}}(\cdot) \in \mathcal{I}_{sah}^{k_1}, \quad \|V_{h+1,\pi}^{\mathcal{E}}(\cdot)\|_2 \in \mathcal{J}_h^{k_2}, r_h^{\mathcal{E}}(s, a) \in \mathcal{C}_{k_3},$$

where  $k_1 \in [m], k_2 \in [m'], k_3 \in [n]$ .

Therefore,  $\{\Theta_k\}_{k=1}^K$  is a partition of  $\Theta$ . For any  $k \in [K]$  and  $\mathcal{E}, \mathcal{E}' \in \Theta_k$ , the following holds for any  $s, a, h$ ,

$$P_h^{\mathcal{E}}(\cdot | s, a)^T V_{h+1,\pi}^{\mathcal{E}}(\cdot) - P_h^{\mathcal{E}'}(\cdot | s, a)^T V_{h+1,\pi}^{\mathcal{E}'}(\cdot) \leq \frac{\epsilon}{3H},$$

$$\|V_{h+1,\pi}^{\mathcal{E}}(\cdot) - V_{h+1,\pi}^{\mathcal{E}'}(\cdot)\|_2 \leq \frac{\epsilon}{3H},$$

$$r_h^{\mathcal{E}}(s, a) - r_h^{\mathcal{E}'}(s, a) \leq \frac{\epsilon}{3H}.$$

Then, we have

$$V_{1,\pi_{\mathcal{E}}^*}^{\mathcal{E}}(s_1^t) - V_{1,\pi_{\mathcal{E}}^*}^{\mathcal{E}'}(s_1^t) \leq \epsilon.$$

Since

$$K(\epsilon) \leq \left(\frac{3H^2}{\epsilon}\right)^{SAH} \cdot \left(\frac{6H^2\sqrt{S}}{\epsilon}\right)^H \cdot \left(\frac{3H}{\epsilon}\right)^{SAH},$$

we have

$$\log(K(\epsilon)) \leq SAH \log\left(\frac{3H^2}{\epsilon}\right) + H \log\left(\frac{6H^2\sqrt{S}}{\epsilon}\right) + SAH \log\left(\frac{3H}{\epsilon}\right) \leq 3SAH \log\left(\frac{6H^2\sqrt{S}}{\epsilon}\right).$$

From Theorem 4.11, we have

$$BR_T(\pi_{\text{r-IDS}}) \leq \alpha \sqrt{3SATH^2 \log\left(\frac{6H^2\sqrt{S}}{\epsilon}\right)} + T\epsilon + T_0.$$

### A.3 Proof of Theorem 4.16

Recall that  $\mathcal{F}$  is the compact feature space of  $(\psi_h^P)_i$  and  $(\psi_h^R)_i$ . From the compactness of  $\mathcal{F}$ , there exists a finite  $\epsilon$ -covering number of  $\mathcal{F}$ . Let  $M \triangleq \sup_{i,s} \max\{(\psi_h^P(s))_i (\psi_h^R(s))_i\}$ . Denote the  $\frac{\epsilon}{dMH^2}$ -covering number of  $\mathcal{F}$  as  $K_{\mathcal{F}}(\epsilon)$ . We have  $\mathcal{F} \subset \mathcal{K}_1 \cup \dots \cup \mathcal{K}_{K_{\mathcal{F}}(\epsilon)}$  and for any  $f, f' \in \mathcal{K}_i$ ,

$$\ell_g(f, f') = \int_s |\log \frac{f(s)}{f'(s)}| \leq \frac{\epsilon}{dMH^2}.$$

Then we construct the partition of  $\Theta$  as following:  $\mathcal{E}$  and  $\mathcal{E}'$  belong to the same partition if and only if  $(\psi_h^{P,\mathcal{E}})_i$  and  $(\psi_h^{P,\mathcal{E}'}))_i$  belong to the same partition of  $\mathcal{F}$ ,  $\forall i \in [d]$ . Then, we have

$$\begin{aligned} & V_{1,\pi_{\mathcal{E}}^*}(s_1) - V_{1,\pi_{\mathcal{E}}^*}^{(\mathcal{E}')}(s_1) \\ & \stackrel{(a)}{=} \sum_{h=1}^H \mathbb{E}_{\pi_{\mathcal{E}}^*}^{(\mathcal{E}')} \left[ \mathbb{E}_{s' \sim P_h^{\mathcal{E}}(\cdot | s_h, a_h)} \left[ V_{h+1,\pi_{\mathcal{E}}^*}^{\mathcal{E}}(s') \right] - \mathbb{E}_{s' \sim P_h^{\mathcal{E}'}(\cdot | s_h, a_h)} \left[ V_{h+1,\pi_{\mathcal{E}}^*}^{\mathcal{E}}(s') \right] \right] + \sum_{h=1}^H \mathbb{E}_{\pi_{\mathcal{E}}^*}^{(\mathcal{E}')} \left[ R_h^{\mathcal{E}}(s_h, a_h) - R_h^{\mathcal{E}'}(s_h, a_h) \right] \\ & \leq \sum_{h=1}^H \mathbb{E}_{\pi_{\mathcal{E}}^*}^{(\mathcal{E}')} \left[ \int_{\mathcal{S}} \left| P_h^{\mathcal{E}}(s' | s_h, a_h) - P_h^{\mathcal{E}'}(s' | s_h, a_h) \right| \cdot V_{h+1,\pi_{\mathcal{E}}^*}^{\mathcal{E}}(s') d\mu_{\mathcal{S}} + \int_{[0,1]} \left| x \left( R_h^{\mathcal{E}}(x | s_h, a_h) - R_h^{\mathcal{E}'}(x | s_h, a_h) \right) \right| dx \right] \\ & \stackrel{(c)}{\leq} H \sum_{h=1}^H \ell_1(P_h^{\mathcal{E}}, P_h^{\mathcal{E}'}) + \sum_{h=1}^H \ell_1(R_h^{\mathcal{E}}, R_h^{\mathcal{E}'}), \end{aligned} \tag{A.26}$$

where (a) uses Lemma C.1, (b) follows from  $V_{h+1,\pi_{\mathcal{E}}^*}^{\mathcal{E}}(s') \leq H$ . Next, we use the coverage of  $\mathcal{F}$  under  $\ell_g$  to bound  $\ell_1(P_h^{\mathcal{E}}, P_h^{\mathcal{E}'})$  and  $\ell_1(R_h^{\mathcal{E}}, R_h^{\mathcal{E}'})$ .

First notice that

$$\|\phi_h^P(s, a)\|_2 \leq 1 \Rightarrow |\phi_h^P(s, a)_i| \leq 1, \forall i \in [d].$$

For any  $\mathcal{E}, \mathcal{E}'$  that belong to the same partition, we have

$$\begin{aligned}
\ell_1(P_h^{\mathcal{E}}, P_h^{\mathcal{E}'}) &= \sup_{s,a} \int_{s'} |P_h^{\mathcal{E}}(s'|s, a) - P_h^{\mathcal{E}'}(s'|s, a)| \\
&= \sup_{s,a} \int_{s'} |\langle \phi_h^P(s, a), \psi_h^{P,\mathcal{E}}(s') - \psi_h^{P,\mathcal{E}'}(s') \rangle| \\
&\leq \int_{s'} \sum_{i=1}^d |(\psi_h^{P,\mathcal{E}}(s') - \psi_h^{P,\mathcal{E}'}(s'))_i| \\
&\leq M \cdot \sum_{i=1}^d \ell_g((\psi_h^{P,\mathcal{E}})_i, (\psi_h^{P,\mathcal{E}'}))_i \\
&\leq \frac{\epsilon}{H^2}.
\end{aligned} \tag{A.27}$$

The penultimate inequality uses the fact that  $|a - b| \leq M|\log \frac{a}{b}|$  for any  $a, b \in (0, M)$ . By analogy, it can be concluded that  $\ell_1(R_h^{\mathcal{E}}, R_h^{\mathcal{E}'}) \leq \frac{\epsilon}{H^2}$ . Thus,

$$V_{1,\pi_{\mathcal{E}}^*}(s_1) - V_{1,\pi_{\mathcal{E}}^*}(s_1) \leq 2\epsilon.$$

Therefore, we get an  $\epsilon$ -value partition of  $\Theta$ . Since  $\psi_h^{P,\mathcal{E}} = ((\psi_h^{P,\mathcal{E}})_1, (\psi_h^{P,\mathcal{E}})_2, \dots, (\psi_h^{P,\mathcal{E}})_d)$  and each  $(\psi_h^{P,\mathcal{E}})_i$  belongs to one of these  $K_{\mathcal{F}}(\epsilon)$  sets ( $\mathcal{K}_1, \dots, \mathcal{K}_{K_{\mathcal{F}}(\epsilon)}$ ), the number of this  $\epsilon$ -value partition can be bounded by  $(K_{\mathcal{F}}(\epsilon))^{dH}$ . Thus, we have

$$BR_T(\pi_{\text{IDS}}) \leq \alpha H \sqrt{dT \log(K_{\mathcal{F}}(\epsilon))} + T\epsilon + T_0.$$

#### A.4 Proof of Proposition 5.1

**Proposition.** Define

$$r'_h(s, a) = r_h(s, a) + \frac{\lambda}{2} \mathbb{E}_t \left[ D_{\text{KL}} \left( (P_h^{\tilde{\mathcal{E}}_t^*} \otimes R_h^{\tilde{\mathcal{E}}_t^*})(\cdot|s_h, a_h) \middle\| (P_h^{\tilde{\mathcal{E}}_t} \otimes R_h^{\tilde{\mathcal{E}}_t})(\cdot|s_h, a_h) \right) \right]. \tag{A.28}$$

Then, for any policy  $\pi$ , we have

$$\left| \mathbb{E}_{\pi}^{\tilde{\mathcal{E}}_t} \left[ \sum_{h=1}^H r'_h(s_h, a_h) \right] - \mathbb{E}_{\pi}^{\tilde{\mathcal{E}}_t} \left[ \sum_{h=1}^H \bar{r}_h(s_h, a_h) \right] \right| \leq \frac{\lambda}{2} \cdot \epsilon \cdot \left( 1 + \log \frac{B}{\beta} \right). \tag{A.29}$$

*Proof.* We begin our proof by calculating the difference of the two KL divergence terms. Recall that

$$\bar{r}_h(s_h, a_h) = r_h(s, a) + \frac{\lambda}{2} \mathbb{E}_t \left[ D_{\text{KL}} \left( (P_h^{\mathcal{E}} \otimes R_h^{\mathcal{E}})(\cdot|s_h, a_h) \middle\| (P_h^{\tilde{\mathcal{E}}_t} \otimes R_h^{\tilde{\mathcal{E}}_t})(\cdot|s_h, a_h) \right) \right].$$

First, by triangle inequality, we have

$$\begin{aligned}
&\left| D_{\text{KL}} \left( (P_h^{\tilde{\mathcal{E}}_t^*} \otimes R_h^{\tilde{\mathcal{E}}_t^*})(\cdot|s_h, a_h) \middle\| (P_h^{\tilde{\mathcal{E}}_t} \otimes R_h^{\tilde{\mathcal{E}}_t})(\cdot|s_h, a_h) \right) - D_{\text{KL}} \left( (P_h^{\mathcal{E}} \otimes R_h^{\mathcal{E}})(\cdot|s_h, a_h) \middle\| (P_h^{\tilde{\mathcal{E}}_t} \otimes R_h^{\tilde{\mathcal{E}}_t})(\cdot|s_h, a_h) \right) \right| \\
&\leq \left| \int_{\mathcal{S}} P_h^{\tilde{\mathcal{E}}_t^*}(x|s_h, a_h) \log \frac{P_h^{\tilde{\mathcal{E}}_t^*}(x|s_h, a_h)}{P_h^{\tilde{\mathcal{E}}_t}(x|s_h, a_h)} dx - \int_{\mathcal{S}} P_h^{\mathcal{E}}(x|s_h, a_h) \log \frac{P_h^{\mathcal{E}}(x|s_h, a_h)}{P_h^{\tilde{\mathcal{E}}_t}(x|s_h, a_h)} dx \right|
\end{aligned}$$

$$+ \left| \int_{[0,1]} R_h^{\tilde{\mathcal{E}}_t^*}(x|s_h, a_h) \log \frac{R_h^{\tilde{\mathcal{E}}_t^*}(x|s_h, a_h)}{R_h^{\tilde{\mathcal{E}}_t}(x|s_h, a_h)} dx - \int_{[0,1]} R_h^{\mathcal{E}}(x|s_h, a_h) \log \frac{R_h^{\mathcal{E}}(x|s_h, a_h)}{R_h^{\tilde{\mathcal{E}}_t}(x|s_h, a_h)} dx \right|. \quad (\text{A.30})$$

Let  $o = (s_h, a_h)$ . For the first term in Eqn. (A.30), we have the following bound

$$\begin{aligned} & \left| \int_{\mathcal{S}} P_h^{\tilde{\mathcal{E}}_t^*}(x|o) \log \frac{P_h^{\tilde{\mathcal{E}}_t^*}(x|o)}{P_h^{\tilde{\mathcal{E}}_t}(x|o)} - P_h^{\mathcal{E}}(x|o) \log \frac{P_h^{\mathcal{E}}(x|o)}{P_h^{\tilde{\mathcal{E}}_t}(x|o)} dx \right| \\ & \leq \left| \int_{\mathcal{S}} P_h^{\tilde{\mathcal{E}}_t^*}(x|o) \log \frac{P_h^{\tilde{\mathcal{E}}_t^*}(x|o)}{P_h^{\tilde{\mathcal{E}}_t}(x|o)} - P_h^{\mathcal{E}}(x|o) \log \frac{P_h^{\tilde{\mathcal{E}}_t^*}(x|o)}{P_h^{\tilde{\mathcal{E}}_t}(x|o)} dx \right| \\ & \quad + \left| \int_{\mathcal{S}} P_h^{\mathcal{E}}(x|o) \log \frac{P_h^{\mathcal{E}}(x|o)}{P_h^{\tilde{\mathcal{E}}_t}(x|o)} - P_h^{\mathcal{E}}(x|o) \log \frac{P_h^{\mathcal{E}}(x|o)}{P_h^{\tilde{\mathcal{E}}_t}(x|o)} dx \right| \\ & \leq \int_{\mathcal{S}} \left| P_h^{\tilde{\mathcal{E}}_t^*}(x|o) - P_h^{\mathcal{E}}(x|o) \right| \cdot \left| \log \frac{P_h^{\tilde{\mathcal{E}}_t^*}(x|o)}{P_h^{\tilde{\mathcal{E}}_t}(x|o)} \right| dx + \int_{\mathcal{S}} B \cdot \left| \log \frac{P_h^{\tilde{\mathcal{E}}_t^*}(x|o)}{P_h^{\mathcal{E}}(x|o)} \right| dx \\ & \leq B \cdot \left( 1 + \log \frac{B}{\beta} \right) \int_{\mathcal{S}} \left| \log \frac{P_h^{\tilde{\mathcal{E}}_t^*}(x|o)}{P_h^{\mathcal{E}}(x|o)} \right| dx \\ & \leq \frac{(1 + \log(B/\beta))\epsilon}{2H^2}, \end{aligned} \quad (\text{A.31})$$

where the first inequality is due to triangle inequality; the second inequality again uses triangle inequality, and the fact that  $P_h^{\mathcal{E}}(x|o) \leq B$ ; the third inequality is due to the fact that  $|a - b| \leq B \cdot |\log \frac{a}{b}|$  for any  $a, b \in (0, B)$  and  $\left| \log \frac{P_h^{\tilde{\mathcal{E}}_t^*}(\cdot|o)}{P_h^{\tilde{\mathcal{E}}_t}(\cdot|o)} \right| \leq \log \frac{B}{\beta}$ . Note that when  $P_h^{\tilde{\mathcal{E}}_t}(\cdot|o)$  equals zero, since  $\tilde{\mathcal{E}}_t$  is the mean MDP, it turns out that  $P_h^{\tilde{\mathcal{E}}_t^*}(\cdot|o)$  also equals zero, yielding  $\left| \log \frac{P_h^{\tilde{\mathcal{E}}_t^*}(\cdot|o)}{P_h^{\tilde{\mathcal{E}}_t}(\cdot|o)} \right| = 0$ . Finally, the last inequality is due to Eqn. (4.10).

If we employed the KL divergence or  $\ell_1$  metric, the derivation of third inequality would become invalid.

For the second term in Eqn. (A.30), adopting a similar approach, we have

$$\begin{aligned} & \left| \int_{[0,1]} R_h^{\tilde{\mathcal{E}}_t^*}(x|o) \log \frac{R_h^{\tilde{\mathcal{E}}_t^*}(x|o)}{R_h^{\tilde{\mathcal{E}}_t}(x|o)} - R_h^{\mathcal{E}}(x|o) \log \frac{R_h^{\mathcal{E}}(x|o)}{R_h^{\tilde{\mathcal{E}}_t}(x|o)} dx \right| \\ & \leq \left| \int_{[0,1]} R_h^{\tilde{\mathcal{E}}_t^*}(x|o) \log \frac{R_h^{\tilde{\mathcal{E}}_t^*}(x|o)}{R_h^{\tilde{\mathcal{E}}_t}(x|o)} - R_h^{\mathcal{E}}(x|o) \log \frac{R_h^{\tilde{\mathcal{E}}_t^*}(x|o)}{R_h^{\tilde{\mathcal{E}}_t}(x|o)} dx \right| \\ & \quad + \left| \int_{[0,1]} R_h^{\mathcal{E}}(x|o) \log \frac{R_h^{\tilde{\mathcal{E}}_t^*}(x|o)}{R_h^{\tilde{\mathcal{E}}_t}(x|o)} - R_h^{\mathcal{E}}(x|o) \log \frac{R_h^{\mathcal{E}}(x|o)}{R_h^{\tilde{\mathcal{E}}_t}(x|o)} dx \right| \\ & \leq \int_{[0,1]} \left| R_h^{\tilde{\mathcal{E}}_t^*}(x|o) - R_h^{\mathcal{E}}(x|o) \right| \cdot \left| \log \frac{R_h^{\tilde{\mathcal{E}}_t^*}(x|o)}{R_h^{\tilde{\mathcal{E}}_t}(x|o)} \right| dx + \int_{[0,1]} B \cdot \left| \log \frac{R_h^{\tilde{\mathcal{E}}_t^*}(x|o)}{R_h^{\mathcal{E}}(x|o)} \right| dx \\ & \leq B \cdot \left( 1 + \log \frac{B}{\beta} \right) \int_{[0,1]} \left| \log \frac{R_h^{\tilde{\mathcal{E}}_t^*}(x|o)}{R_h^{\mathcal{E}}(x|o)} \right| dx \end{aligned}$$

$$\leq \frac{(1 + \log(B/\beta))\epsilon}{2H}. \quad (\text{A.32})$$

Hence, adding up Eqn. (A.31) and Eqn. (A.32), we obtain

$$|r'_h(s_h, a_h) - \bar{r}_h(s_h, a_h)| \leq \frac{\lambda}{2} \cdot \frac{(1 + \log(B/\beta))\epsilon}{H}.$$

Finally, summing over  $h \in [H]$ , we have

$$\left| \mathbb{E}_{\pi}^{\tilde{\mathcal{E}}_t} \left[ \sum_{h=1}^H r'_h(s_h, a_h) \right] - \mathbb{E}_{\pi}^{\tilde{\mathcal{E}}_t} \left[ \sum_{h=1}^H \bar{r}_h(s_h, a_h) \right] \right| \leq \frac{\lambda}{2} \cdot \epsilon \cdot (1 + \log \frac{B}{\beta}) \quad (\text{A.33})$$

The proof is finished.  $\square$

## A.5 Proof of Proposition 5.4

The proof follows essentially the same structure as that of Proposition 1, with the only difference lying in the third inequality of Eq. (A.31). By applying the Mean Value Theorem, we have

$$\int_{\mathcal{S}} \left| \log \frac{P_h^{\tilde{\mathcal{E}}_t^*}(x|o)}{P_h^{\mathcal{E}}(x|o)} \right| dx \leq \frac{1}{\beta} \int_{\mathcal{S}} \left| P_h^{\tilde{\mathcal{E}}_t^*}(x|o) - P_h^{\mathcal{E}}(x|o) \right| dx = \frac{1}{\beta} \ell_1(P_h^{\tilde{\mathcal{E}}_t^*}, P_h^{\mathcal{E}}) \leq \frac{1}{\beta} \cdot \text{constant} \cdot \epsilon.$$

We can derive the third inequality in Eq. (A.31) as

$$B \cdot \left( 1 + \frac{B}{\beta} \right) \int_{\mathcal{S}} \left| \log \frac{P_h^{\tilde{\mathcal{E}}_t^*}(x|o)}{P_h^{\mathcal{E}}(x|o)} \right| dx \leq \frac{1}{\beta} \cdot \text{constant} \cdot \epsilon.$$

In the conclusion of Proposition 1, the term involving  $\beta$  has been modified from  $\log \frac{1}{\beta}$  to  $\frac{1}{\beta}$ . Therefore, with a suitably chosen partition radius, we have

$$\left| \mathbb{E}_{\pi}^{\tilde{\mathcal{E}}_t} \left[ \sum_{h=1}^H r'_h(s_h, a_h) \right] - \mathbb{E}_{\pi}^{\tilde{\mathcal{E}}_t} \left[ \sum_{h=1}^H \bar{r}_h(s_h, a_h) \right] \right| \leq \frac{\lambda}{2} \cdot \epsilon \cdot (1 + \frac{B}{\beta}) \quad (\text{A.34})$$

## B Basic Properties of the Measure $\ell_g$

The following result deals with the problem of convexity. Let  $B(\mathcal{C}, \epsilon)$  be an  $\epsilon$ -ball with its center at  $\mathcal{C}$ . Note that  $B(\mathcal{C}, \epsilon)$  is not essentially convex under  $\ell_g$ : for  $P, Q \in B(\mathcal{C}, \epsilon)$  and  $\lambda \in (0, 1)$ , it does not hold that  $\lambda P + (1 - \lambda)Q \in B(\mathcal{C}, \epsilon)$ . However, using Lemma 4.3, we have the following result:

**Lemma B.1.** For any  $P, Q \in B(\mathcal{C}, \epsilon)$  and  $\lambda \in [0, 1]$ ,  $\lambda P + (1 - \lambda)Q$  lies in the  $(2\epsilon)$ -ball at  $\mathcal{C}$ , i.e.,

$$\ell_g(\lambda P + (1 - \lambda)Q, \mathcal{C}) \leq 2\epsilon.$$

*Proof.* By definition of  $\ell_g$ , we have

$$\begin{aligned} \ell_g(\lambda P + (1 - \lambda)Q, \mathcal{C}) &= \sup_o \int_x \left| \log \frac{\lambda P(x|o) + (1 - \lambda)Q(x|o)}{\mathcal{C}(x|o)} \right| \\ &\leq \sup_o \int_x \left| \log \frac{P(x|o)}{\mathcal{C}(x|o)} + \log \frac{Q(x|o)}{\mathcal{C}(x|o)} \right| \end{aligned}$$

$$\leq 2\epsilon,$$

where the first inequality uses the fact that  $|\log(\lambda a + (1 - \lambda)b)| \leq |\log a| + |\log b|$  for any  $a, b > 0$  and  $\lambda \in [0, 1]$ ; the second inequality is due to  $P, Q \in B(\mathcal{C}, \epsilon)$ . This finishes the proof of Lemma B.1.  $\square$

## B.1 A Counterexample to Demonstrate Non-convex

We present the following counterexample to demonstrate that the unit ball under  $\ell_g$  is not convex. Let  $X$  be the set of positive and Lebesgue integrable functions defined on  $[0, 1]$ . We show that the ball centered at  $f_0 \in X$

$$B(f_0, r) = \{f \in X : \ell_g(f, f_0) \leq r\}$$

is not convex. Let

$$f(x) = \begin{cases} 5 & 0 \leq x \leq 0.1, \\ \frac{5}{9} & 0.1 < x \leq 1, \end{cases} \quad g(x) = \begin{cases} \frac{5}{9} & 0 \leq x \leq 0.9, \\ 5 & 0.9 < x \leq 1, \end{cases} \quad f_0(x) = 1.$$

It can be verified:

$$\int_0^1 f(x) dx = 0.1 \cdot 5 + 0.9 \cdot \frac{5}{9} = 1,$$

and the same holds for  $g$ . Hence  $f, g$  can be seen as the two probability measures. We also have

$$\ell_g(f, f_0) = 0.1 \log 5 + 0.9 \log \left( \frac{9}{5} \right) \approx 0.68995.$$

The same holds for  $\ell_g(g, f_0)$ . And

$$\ell_g\left(\frac{f+g}{2}, f_0\right) = \ell_g\left(\frac{25}{9}, 1\right) = |\log \frac{25}{9}| \approx 1.02165.$$

Let  $r = 1$ , then  $\ell_g(f, f_0) = \ell_g(g, f_0) \approx 0.68995 < r$ ,  $\ell_g\left(\frac{f+g}{2}, f_0\right) \approx 1.02165 > r$ .

So,  $f, g \in B(f_0, r)$  but  $\frac{f+g}{2} \notin B(f_0, r)$ .

## C Technical Lemmas

### C.1 Value Function and Mutual Information

We cite the following lemma from [Moradipari et al. \(2023\)](#). Similar results can be found in [Osband et al. \(2013\)](#); [Foster et al. \(2021\)](#).

**Lemma C.1.** For any two environments  $\mathcal{E}, \mathcal{E}'$  with potentially different transition and reward functions, and any policy  $\pi$ , we have

$$\begin{aligned} V_{1,\pi}^{\mathcal{E}}(s_1) - V_{1,\pi}^{\mathcal{E}'}(s_1) &= \sum_{h=1}^H \mathbb{E}_{\pi}^{\mathcal{E}'} \left[ \mathbb{E}_{(s', r') \sim (P_h^{\mathcal{E}} \otimes R_h^{\mathcal{E}})(\cdot | s_h, a_h)} [r' + V_{h+1,\pi}^{\mathcal{E}}(s')] \right. \\ &\quad \left. - \mathbb{E}_{(s', r') \sim (P_h^{\mathcal{E}'} \otimes R_h^{\mathcal{E}'})(\cdot | s_h, a_h)} [r' + V_{h+1,\pi}^{\mathcal{E}'}(s')] \right]. \end{aligned} \tag{C.1}$$

**Lemma C.2.** The mutual information  $\mathbb{I}_t^{\pi_{\text{TS}}^t} \left( \tilde{\mathcal{E}}_t^*; (\mathcal{H}_t, \mathcal{R}_{t,H}) \right)$  can be lower bounded as

$$\mathbb{I}_t^{\pi_{\text{TS}}^t} \left( \tilde{\mathcal{E}}_t^*; (\mathcal{H}_t, \mathcal{R}_{t,H}) \right) \geq \sum_{h=1}^H \mathbb{E}_t \left[ \mathbb{E}_{\pi_{\text{TS}}^t} \left[ D_{\text{KL}} \left( (P_h^{\tilde{\mathcal{E}}_t^*} \otimes R_h^{\tilde{\mathcal{E}}_t^*})(\cdot | s_h, a_h) \middle\| (P_h^{\tilde{\mathcal{E}}_t} \otimes R_h^{\tilde{\mathcal{E}}_t})(\cdot | s_h, a_h) \right) \right] \right]. \quad (\text{C.2})$$

*Proof of Lemma C.2.* Using the chain rule of mutual information,

$$\begin{aligned} & \mathbb{I}_t^{\pi_{\text{TS}}^t} \left( \tilde{\mathcal{E}}_t^*; (\mathcal{H}_t, \mathcal{R}_{t,H}) \right) \\ &= \sum_{h=1}^H \mathbb{I}_t^{\pi_{\text{TS}}^t} \left( \tilde{\mathcal{E}}_t^*; (s_h^{t,1}, a_h^{t,1}, r_h^{t,1}, s_h^{t,0}, a_h^{t,0}, r_h^{t,0}) \mid (\mathcal{H}_{t,h-1}, \mathcal{R}_{t,h-1}) \right) + \mathbb{I}_t^{\pi_{\text{TS}}^t} \left( \tilde{\mathcal{E}}_t^*; o_t \mid (\mathcal{H}_{t,H}, \mathcal{R}_{t,H}) \right) \\ &= \sum_{h=1}^H \mathbb{I}_t^{\pi_{\text{TS}}^t} \left( \tilde{\mathcal{E}}_t^*; s_h^{t,1} \mid \mathcal{H}_{t,h-1}, \mathcal{R}_{t,h-1} \right) + \sum_{h=1}^H \mathbb{I}_t^{\pi_{\text{TS}}^t} \left( \tilde{\mathcal{E}}_t^*; a_h^{t,1} \mid s_h^{t,1}, \mathcal{H}_{t,h-1}, \mathcal{R}_{t,h-1} \right) \\ &\quad + \sum_{h=1}^H \mathbb{I}_t^{\pi_{\text{TS}}^t} \left( \tilde{\mathcal{E}}_t^*; r_h^{t,1} \mid s_h^{t,1}, s_h^{t,1}, \mathcal{H}_{t,h-1}, \mathcal{R}_{t,h-1} \right) + \sum_{h=1}^H \mathbb{I}_t^{\pi_{\text{TS}}^t} \left( \tilde{\mathcal{E}}_t^*; s_h^{t,0} \mid s_h^{t,1}, a_h^{t,1}, r_h^{t,1}, \mathcal{H}_{t,h-1}, \mathcal{R}_{t,h-1} \right) \quad (\text{C.3}) \\ &\quad + \sum_{h=1}^H \mathbb{I}_t^{\pi_{\text{TS}}^t} \left( \tilde{\mathcal{E}}_t^*; a_h^{t,0} \mid s_h^{t,1}, a_h^{t,1}, r_h^{t,1}, s_h^{t,0}, \mathcal{H}_{t,h-1}, \mathcal{R}_{t,h-1} \right) \\ &\quad + \sum_{h=1}^H \mathbb{I}_t^{\pi_{\text{TS}}^t} \left( \tilde{\mathcal{E}}_t^*; r_h^{t,0} \mid s_h^{t,1}, a_h^{t,1}, r_h^{t,1}, s_h^{t,0}, a_h^{t,0}, \mathcal{H}_{t,h-1}, \mathcal{R}_{t,h-1} \right) + \mathbb{I}_t^{\pi_{\text{TS}}^t} \left( \tilde{\mathcal{E}}_t^*; o_t \mid (\mathcal{H}_{t,H}, \mathcal{R}_{t,H}) \right). \end{aligned}$$

From [Moradipari et al. \(2023\)](#), the first three terms on the right side of the above equation are equal to

$$\sum_{h=1}^H \mathbb{E}_t \left[ \mathbb{E}_{\pi_{\text{TS}}^t} \left[ D_{\text{KL}} \left( (P_h^{\tilde{\mathcal{E}}_t^*} \otimes r_h^{\tilde{\mathcal{E}}_t^*})(\cdot | s_h, a_h) \middle\| (P_h^{\tilde{\mathcal{E}}_t} \otimes r_h^{\tilde{\mathcal{E}}_t})(\cdot | s_h, a_h) \right) \right] \right]. \quad (\text{C.4})$$

Based on the non-negativity of mutual information, we obtain the conclusion of the lemma.  $\square$

## C.2 Posterior Consistency

**Lemma C.3.** Assume that there exists a strongly consistent estimator of the true environment given the history. Let  $\Pi$  be some measure. For any  $\Pi$ -integrable function  $f : \Theta \rightarrow \mathbb{R}$  and almost every  $\mathcal{D}_\infty$  sampled from the true environment  $\mathcal{E}_0$ , we have

$$\lim_{t \rightarrow \infty} \mathbb{E}_t [f(\mathcal{E})] = f(\mathcal{E}_0).$$

And if  $f : \Theta \times \Theta \rightarrow \mathbb{R}$  is bounded and  $(\Pi \times \Pi)$ -integrable, for almost every  $\mathcal{D}_\infty$  sampled from the true environment  $\mathcal{E}_0$ , we have

$$\lim_{t \rightarrow \infty} \mathbb{E}_t [f(\mathcal{E}, \mathcal{E}')] = f(\mathcal{E}_0, \mathcal{E}_0),$$

where the expectation is taken over all  $\mathcal{E}$  and  $\mathcal{E}'$ .

We refer the readers to Theorem 6.9 in [Ghosal and van der Vaart \(2017\)](#) or Appendix K in [Moradipari et al. \(2023\)](#) for the definition of a strongly consistent estimator and for more details of the proof.