

---

# Sharp Analysis for KL-Regularized Contextual Bandits and RLHF

---

Anonymous Author(s)

Affiliation

Address

email

## Abstract

*Reverse-Kullback-Leibler* (KL) regularization has emerged to be a predominant technique to enhance policy optimization in reinforcement learning (RL) and reinforcement learning from human feedback (RLHF), which forces the learned policy to stay close to a reference policy. While the effectiveness of KL-regularization has been empirically demonstrated in various practical scenarios, current theoretical analyses of KL-regularized RLHF still yield the same  $\mathcal{O}(1/\epsilon^2)$  sample complexity as ones without KL-regularization. To understand the fundamental distinction between objectives with KL-regularization and ones without KL-regularization, we are the first to theoretically demonstrate the power of KL-regularization by providing a sharp analysis for KL-regularized contextual bandits and RLHF, revealing an  $\mathcal{O}(1/\epsilon)$  sample complexity when  $\epsilon$  is sufficiently small. We also prove matching lower bounds for both settings. More specifically, we study how the coverage of the reference policy affects the sample complexity of KL-regularized online contextual bandits and RLHF. We show that with sufficient coverage from the reference policy, a simple two-stage mixed sampling algorithm can achieve an  $\mathcal{O}(1/\epsilon)$  sample complexity with only an additive dependence on the coverage coefficient, thus proving the benefits of online data even without explicit exploration. Our results provide a comprehensive understanding of the roles of KL-regularization and data coverage in online decision making, shedding light on the design of more efficient algorithms.

## 1 Introduction

The KL-regularized contextual bandit problem (Langford and Zhang, 2007; Xiong et al., 2024a) has raised tremendous interest recently because of the significant development of the post-training stage in large language models (LLMs) and diffusion models from preference feedback (Christiano et al., 2017; Ziegler et al., 2019; Ouyang et al., 2022; Bai et al., 2022; Rafailov et al., 2024), which is called *Reinforcement Learning from Human Feedback* (RLHF). RLHF aims to optimize the policy by aligning it with human feedback, exhibiting impressive capabilities in applications such as Chatgpt (Achiam et al., 2023), Claude (Anthropic, 2023), Gemini (Team et al., 2023), and LLaMA-3 (Meta, 2024).

In RLHF, we treat the language model as a policy that takes a prompt  $x$  and produces a response  $a$  conditioned on  $x$ , optimizing the policy by aligning it with human feedback. There are mainly two kinds of feedback: absolute rating and preference comparison. For absolute rating, the collection typically involves human annotators to provide rating scores like 1 to 5 (Wang et al., 2024a,b) for responses or hard 0-1 scores for math reasoning tasks since the reasoning tasks often have gold standard answers (Cobbe et al., 2021; Hendrycks et al., 2021; Xiong et al., 2024b). Although discrete feedback is believed to be more intuitive for human users and easier to collect, it also poses more challenges for the RLHF algorithms to effectively leverage the feedback signals since the reward signals are not directly observed. In practice, the learning process typically involves (a) constructing a reward model based on the maximum likelihood estimation (MLE) of *Bradley-Terry* (BT) (Bradley and Terry, 1952a) model from the preference feedback; (b) applying RL algorithms like

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

PPO (Schulman et al., 2017b) to train the language model so that it maximizes the reward signals with KL regularization (Ouyang et al., 2022; Bai et al., 2022; Touvron et al., 2023).

Since the human feedback data only covers a tiny fraction of possible interactions, if we optimize the model purely for the reward without constraints, it might learn behaviors that work well for the training feedback but fail catastrophically on slightly different inputs. For example, the policy may generate disproportionate bold words or emoji to please the learned reward (Zhang et al., 2024). Hence, the KL-regularization between the learned policy and a reference policy (the pre-trained model after supervised fine-tuning) plays a fundamental role in RLHF to avoid overfitting. There is a line of theoretical RLHF work that modeled the problem as a reverse-KL regularized contextual bandit (Xiong et al., 2024a; Ye et al., 2024a; Zhong et al., 2024; Wu et al., 2024; Xie et al., 2024). However, they adopt the techniques from contextual bandits and neglect the power of reverse-KL-regularization, thus obtaining almost the same  $\mathcal{O}(1/\epsilon^2)$ <sup>1</sup> sample complexity as learning objectives without KL-regularization. Therefore, the question of

*whether there exists a fundamental distinction between bandit learning objectives with and without KL-regularization*

is still largely under-explored.

Additionally, an emerging line of offline RLHF literature highlights the coverage of the reference policy  $\pi_0$ . The coverage of  $\pi_0$  refers to the ability of the model to generate diverse responses for a wide range of prompts. A model with good coverage can generalize well to unseen contexts and actions, which is essential for the learned reward function to also generalize well. In practice, this is evidenced by the fact that the simple best-of- $n$  sampling based on  $\pi_0$  is competitive with the well-tuned PPO algorithm for general open-ended conversation tasks (Dong et al., 2023), and the fact that the  $\pi_0$  can solve a majority of the math problems with multiple responses (Shao et al., 2024; Nakano et al., 2021). While the coverage of  $\pi_0$  is recognized as a key factor in offline RLHF, its impact on the sample complexity of online RLHF is still largely unknown. Thus, it is natural to ask:

*If online RLHF is theoretically more efficient than offline RLHF under strong coverage of  $\pi_0$ ?*

In this paper, we answer the above questions by (1) providing a novel fine-grained decomposition for the suboptimality of objective functions, which adapts to the strongly convex optimization landscape of the reverse-KL regularization and obtains a sharper sample complexity than the existing results, and (2) proposing an efficient 2-stage mixed sampling strategy for online RLHF with good coverage of  $\pi_0$ , which achieves sample complexity with only an additive dependence on the coverage coefficient. In contrast, the existing RLHF algorithms typically require a multiplicative dependence on the coverage coefficient.

## 1.1 Our Contributions

In this work, we make a first attempt to illustrate the statistical benefits of KL-regularization for contextual bandits and RLHF. Our main contributions are summarized as follows:

- In Section 2, we study the contextual bandit problem with KL-regularization, which also serves as a mathematical formulation for RLHF with absolute-rating feedback. We provide a lower bound for the KL-regularized contextual bandit problem, which indicates that the sample complexity of the problem is  $\Omega(\eta \log N_{\mathcal{R}}(\epsilon)/\epsilon)$  when  $\epsilon$  is sufficiently small, where  $N_{\mathcal{R}}(\epsilon)$  is the covering number of the reward function class and  $\eta$  is the KL-regularization coefficient.
- We provide a novel analysis to upper bound the suboptimality gap of the KL-regularized objective in contextual bandits, and propose a simple two-stage mixed sampling strategy to achieve a sample complexity of  $\mathcal{O}(\max(\eta^2 D^2, \eta/\epsilon) \log N_{\mathcal{R}}(\epsilon/\delta))$  when the reward scale is a constant, where  $D$  is the coverage coefficient of the reference policy  $\pi_0$  and  $\delta$  is the confidence parameter. To the best of our knowledge, this is the first work to provide an  $\mathcal{O}(1/\epsilon)$  sample complexity for KL-regularized contextual bandits.
- In Section 3, we extend our analysis to RLHF. We rigorously demonstrate that KL-regularization is essential for more efficient policy learning in RLHF with preference data. We further propose a two-stage mixed sampling strategy for online RLHF with good coverage of  $\pi_0$ , which achieves a sample complexity of  $\mathcal{O}(\max(\eta^2 D^2, \eta/\epsilon) \log N_{\mathcal{R}}(\epsilon/\delta))$  when the reward scale is a constant.
- In Appendix B, we consider the setting where the coverage condition is not satisfied. We show that an active querying subroutine can be used to improve the coverage of the collected data in the first

<sup>1</sup>For simplicity, we omit here the dependencies on quantities other than  $\epsilon$ .

stage, which can be combined with our two-stage mixed sampling strategy to achieve a sample complexity of  $\tilde{O}(d \log N_{\mathcal{R}}(\epsilon/\delta) \eta^3/\epsilon)$ , where  $d$  is the eluder dimension (Russo and Van Roy, 2013) of the reward function class.

## 1.2 Previous Understanding of KL Regularization in RL

Our analysis of KL-regularized contextual bandits and RLHF also contributes to the theoretical understanding of the impact of KL-regularization in RL since contextual bandits can be viewed as a simplified version of Markov decision processes (MDPs). In RL, KL-regularization has been widely used to stabilize the learning process and prevent the policy from deviating too far from the reference policy. Here, we provide a brief overview of the existing understanding of KL-regularization in decision-making problems. From the perspective of policy optimization, KL-regularization captures entropy regularization as a special case<sup>2</sup>, which is also an extensively used technique in RL literature (Sutton, 2018; Szepesvári, 2022). There is a large body of literature that has explored the benefits of entropy regularization or KL-regularization in RL (Schulman et al., 2015; Fox et al., 2016; Schulman et al., 2017a; Haarnoja et al., 2017, 2018; Ahmed et al., 2019). Most related to our work, Ahmed et al. (2019) provided a comprehensive understanding of the role of entropy regularization in RL, showing that entropy regularization can improve the training efficiency and stability of the policy optimization process by changing the optimization landscape through experiments on continuous control tasks (Brockman, 2016). Theoretically, Neu et al. (2017) provided a unified view of entropy regularization as approximate variants of Mirror Descent or Dual Averaging, and left the statistical justification for using entropy regularization in RL as an open question. Geist et al. (2019) provided a framework for analyzing the error propagation in regularized MDPs, which also focused on the proof of the convergence for the policy optimization methods with regularization and lacked a sharp sample complexity analysis.

## 2 KL-Regularized Contextual Bandits

In this section, we formally define the KL-regularized contextual bandit problem and provide a lower bound for the sample complexity of the problem. We then propose a novel two-stage mixed sampling strategy for online regularized bandits with good coverage of the reference policy  $\pi_0$ .

### 2.1 Problem Setup

In the contextual bandit setting, in each round  $t$ , the agent observes a context  $x_t \in \mathcal{X}$  generated from a distribution  $d_0$  and chooses an action  $a_t \in \mathcal{A}$ . The agent receives a stochastic reward  $r_t \in \mathbb{R}$  depending on the context  $x_t$  and the action  $a_t$ . The goal is to maximize the expected cumulative reward over  $T$  rounds.

The learner has access to a family of reward functions  $R(\theta, x, a)$  parameterized by  $\theta \in \Theta$ , such that there exists  $\theta_* \in \Theta$  satisfying  $\mathbb{E}[r_t | x_{1:t}, a_{1:t}] = R(\theta_*, x_t, a_t)$ . WLOG, we assume that the reward feedback  $r_t$  at all rounds is a non-negative real number bounded by  $B$ . We consider a KL-regularized objective as follows:

$$Q(\pi) = \mathbb{E}_{x \sim d_0} \mathbb{E}_{a \sim \pi(\cdot|x)} \left[ R(\theta_*, x, a) - \eta^{-1} \log \frac{\pi(a|x)}{\pi_0(a|x)} \right], \quad (2.1)$$

where  $\pi_0$  is a known fixed policy, and  $\eta > 0$  is a hyperparameter that controls the trade-off between maximizing rewards and staying close to the reference policy  $\pi_0$ .

**Remark 2.1.** In RLHF with the absolute-rating feedback, we can directly measure the quality of the responses by querying absolute reward value. For instance, in the NVIDIA Helpsteer project (Wang et al., 2023b, 2024c), human labelers are required to provide absolute score in five attributes: helpfulness, correctness, coherence, complexity, and verbosity.

The absolute-rating feedback is directly modeled as the stochastic reward in the contextual bandit setting (Wang et al., 2024a; Xiong et al., 2024b). Under the online RLHF setting, in each round  $t$ , the learner observes a prompt  $x_t$  (modeled as the context) and chooses a response  $a_t$  (modeled as the action). The learner then updates the model (policy) based on the absolute-rating feedback.

**Remark 2.2.** It is worth noting that entropy or Kullback-Leibler (KL) regularization is also widely used in contextual bandits (Berthet and Perchet, 2017; Wu et al., 2016) and deep RL algorithms

<sup>2</sup>We can regard the entropy regularization as a special case of KL-regularization by setting the reference policy as the uniform distribution.

(Schulman et al., 2015; Fox et al., 2016; Schulman et al., 2017a; Haarnoja et al., 2017, 2018), where KL-divergence regularization is a popular technique for preventing drastic updates to the policy. Algorithms such as Trust Region Policy Optimization (TRPO) (Schulman et al., 2015) explicitly incorporate KL-regularization to limit the policy updates during optimization, ensuring that the updated policy does not deviate too much from the current policy. This constraint promotes stable and reliable learning, particularly in high-dimensional state-action spaces. Additionally, KL-regularization is central to Proximal Policy Optimization (PPO) (Schulman et al., 2017a), where a penalty term involving KL-divergence ensures updates remain within “trust region”.

**Reward function class** We consider a function class  $\mathcal{R} = \{R(\theta, \cdot, \cdot) | \theta \in \Theta\}$  and for the realizability, we assume that the ground truth reward function  $R(\theta_*, x, a)$  is in the function class  $\mathcal{R}$ . Then, we define the covering number of  $\mathcal{R}$  as follows.

**Definition 2.3** ( $\epsilon$ -cover and covering number). Given a function class  $\mathcal{F}$ , for each  $\epsilon > 0$ , an  $\epsilon$ -cover of  $\mathcal{F}$  with respect to  $\|\cdot\|_\infty$ , denoted by  $\mathcal{C}(\mathcal{F}, \epsilon)$ , satisfies that for any  $f \in \mathcal{F}$ , we can find  $f' \in \mathcal{C}(\mathcal{F}, \epsilon)$  such that  $\|f - f'\|_\infty \leq \epsilon$ . The  $\epsilon$ -covering number, denoted as  $N_{\mathcal{F}}(\epsilon)$ , is the smallest cardinality of such  $\mathcal{C}(\mathcal{F}, \epsilon)$ .

**Planning oracle** Given a reward model, we can learn the policy by optimizing the KL-regularized objective in (2.1). To simplify the analysis, we assume that there exists a planning oracle, which in empirical can be efficiently approximated by rejection sampling (Liu et al., 2023), Gibbs sampling (Xiong et al., 2024a), and iterative preference learning with a known reward (Dong et al., 2024).

**Definition 2.4** (Policy Improvement Oracle). For a reward function  $R(\theta, \cdot, \cdot) \in \mathcal{R}$  and a reference policy  $\pi_0$ , for any prompt  $x \sim d_0$ , we can compute:

$$\pi_\theta^\eta(\cdot|x) := \operatorname{argmax}_{\pi(\cdot|x) \in \Delta(\mathcal{A})} \mathbb{E}_{a \sim \pi(\cdot|x)} \left[ R(\theta, x, a) - \eta^{-1} \log \frac{\pi(a|x)}{\pi_0(a|x)} \right] \propto \pi_0(\cdot|x) \cdot \exp(\eta R(\theta, x, \cdot)).$$

Hence, the comparator policy is the solution to the oracle given the true reward function  $R(\theta_*, \cdot, \cdot)$ :  $\pi^*(\cdot|x) \propto \pi_0(\cdot|x) \cdot \exp(\eta R(\theta_*, x, \cdot))$ . The **goal** is to minimize the sub-optimality of our learned policy  $\hat{\pi}$  with respect to  $\pi^*$ :  $Q(\pi^*) - Q(\hat{\pi})$ .

**Coverage conditions** It is crucial to assume that our data-collector policy  $\pi_0$  possesses good coverage, which can ensure that the learned reward function can generalize well to unseen contexts (prompts) and actions (responses), and thus can enable us to approximate the optimal policy.

**Definition 2.5** (Data Coverage). Given a reference policy  $\pi_0$ ,  $D^2$  is the minimum positive real number satisfying  $\forall (x, a) \in \mathcal{X} \times \mathcal{A}$ ,  $\pi(a|x) > 0$ , we have for any pair of  $\theta, \theta' \in \Theta$ ,

$$\frac{[R(\theta', x, a) - R(\theta, x, a)]^2}{\mathbb{E}_{x' \sim d_0, a' \sim \pi_0(\cdot|x')} [(R(\theta', x', a') - R(\theta, x', a'))^2]} \leq D^2.$$

The coverage coefficient  $D$  measures how well the in-sample error induced by distribution  $d_0 \times \pi_0$  can characterize the out-of-sample error. This concept is adapted from the F-design for online RL under general function approximation (Agarwal et al., 2024), and follows the coverage coefficient for offline RL (Di et al., 2023; Ye et al., 2024b), and the eluder dimension (Wang et al., 2020; Ye et al., 2023; Agarwal et al., 2023; Zhao et al., 2023a) for online RL. Take the linear model as an example, where the reward function is embedded into a  $d$ -dimensional vector space:  $R(\theta, x, a) = \theta^\top \phi(x, a)$  for  $\theta \in \mathbb{R}^d$ . Let the covariance matrix  $\Sigma = \mathbb{E}_{x \sim d_0} \mathbb{E}_{a \sim \pi_0(\cdot|x)} \phi(x, a) \phi(x, a)^\top$ . Then, the coverage condition turns into  $\sup_{\theta, \theta' \in \Theta} \frac{|(\theta' - \theta)^\top \phi(x, a)|^2}{(\theta' - \theta)^\top \Sigma (\theta' - \theta)} \leq \|\phi(x, a)\|_{\Sigma^{-1}}^2 \leq D^2$ , where the first inequality uses the Cauchy-Schwarz inequality. Hence, this quantity measures how much does the reference policy covers all directions of the feature space, and we can show that there exists  $\pi_0$  such that  $D^2 = O(d)$  through G-optimal design (Zhang, 2023; Lattimore and Szepesvári, 2020).

## 2.2 Lower Bound

In this subsection, we provide a lower bound for the KL-regularized contextual bandit problem.

**Theorem 2.6.** For any  $\epsilon \in (0, 1/256)$ ,  $\eta > 4$ , and any algorithm  $A$ , there exists a KL-regularized contextual bandit problem with reward function class  $\mathcal{R}$  and  $O(N_{\mathcal{R}}(\epsilon))$  data coverage coefficient (as defined in Definition 2.5) such that  $A$  requires at least  $\Omega\left(\min\left(\frac{\eta \log N_{\mathcal{R}}(\epsilon)}{\epsilon}, \frac{\log N_{\mathcal{R}}(\epsilon)}{\epsilon^2}\right)\right)$  rounds to achieve a suboptimality gap of  $\epsilon$ .

---

**Algorithm 1** Two-stage Mixed-Policy Sampling (TMPS)

---

- 1: **Input:**  $\eta, \epsilon, \pi_0, \Theta$ .  
▷ **Stage 1: Use policy  $\pi_0$  to achieve sufficient data coverage**
- 2: **for**  $i = 1, \dots, m$  **do**
- 3:     Sample context  $x_i^0 \sim d_0$  and action  $a_i^0 \sim \pi_0(\cdot | x_i^0)$ .
- 4:     Observe reward  $r_i^0 = R(\theta_*, x_i^0, a_i^0) + \epsilon_i^0$ , where  $\epsilon_i^0$  is the random noise.
- 5: **end for**
- 6: Compute the least square estimate of the reward function based on  $D_0 = \{(x_i^0, a_i^0, r_i^0)\}_{i=1}^m$ :

$$\hat{\theta}_0 \leftarrow \operatorname{argmin}_{\theta \in \Theta} \sum_{i=1}^m (R(\theta, x_i^0, a_i^0) - r_i^0)^2.$$

- 7: Apply the planning oracle to compute  $\pi_{\hat{\theta}_0}^\eta(\cdot | \cdot) \propto \pi_0(\cdot | \cdot) \exp(\eta R(\hat{\theta}_0, \cdot, \cdot))$ .  
▷ **Stage 2: Use policy  $\pi_{\hat{\theta}_0}^\eta$  to sample new responses**
- 8:
- 9: **for**  $i = 1, \dots, n$  **do**
- 10:     Sample context  $x_i \sim d_0$  and action  $a_i \sim \pi_{\hat{\theta}_0}^\eta(\cdot | x_i)$ .
- 11:     Observe reward  $r_i = R(\theta_*, x_i, a_i) + \epsilon_i$ , where  $\epsilon_i$  is the random noise.
- 12: **end for**
- 13: Compute the least square estimate of the reward function using  $\{(x_i, a_i, r_i)\}_{i=1}^n$  together with  $D_0$ :

$$\hat{\theta} \leftarrow \operatorname{argmin}_{\theta \in \Theta} \sum_{i=1}^m (R(\theta, x_i^0, a_i^0) - r_i^0)^2 + \sum_{i=1}^n (R(\theta, x_i, a_i) - r_i)^2.$$

- 14: **Output**  $\pi_{\hat{\theta}}^\eta(\cdot | \cdot) \propto \pi_0(\cdot | \cdot) \exp(\eta R(\hat{\theta}, \cdot, \cdot))$ .
- 

**Remark 2.7.** The lower bound in Theorem 2.6 indicates that the sample complexity of the KL-regularized contextual bandit problem is  $\Omega(\eta \log N_{\mathcal{R}}(\epsilon)/\epsilon)$  when  $\epsilon$  is sufficiently small. In our proof, the KL-regularization term shifts the local landscape of the objective function, which prevents us to directly apply the standard bandit analysis, and thus requires a novel analysis to derive the new lower bound. This  $\Omega(\eta \log N_{\mathcal{R}}(\epsilon)/\epsilon)$  lower bound suggests that the KL-regularized contextual bandit problem potentially enjoy a lower sample complexity compared to the standard contextual bandit.

### 2.3 The Proposed Algorithm

We present our algorithm in Algorithm 1 for the KL-regularized contextual bandit problem, which serves as a theoretical model for online RLHF with absolute-rating feedback. The algorithm consists of two stages:

- In the first stage, we sample  $m$  contexts (prompts) and actions (answers) from the foundation model  $\pi_0$  and observe the corresponding rewards (absolute ratings). These ratings can be regarded as noisy observations of the underlying reward function  $R(\theta_*, x, a)$ . In line 6, we compute an estimate of the reward function  $\hat{\theta}_0$  using least squares regression based on the collected data. In line 7, we apply the planning oracle to obtain the policy  $\pi_{\hat{\theta}_0}^\eta$  which maximizes the following KL-regularized estimated objective in Definition 2.4 with reward function  $R(\theta, \cdot, \cdot) = R(\hat{\theta}_0, \cdot, \cdot)$ .
- In the second stage, we utilize the trained policy  $\pi_{\hat{\theta}_0}^\eta$  to sample  $n$  contexts (prompts) and actions (responses). With the intermediate policy  $\pi_{\hat{\theta}_0}^\eta$ , we can collect new data  $\{(x_i, a_i, r_i)\}_{i=1}^n$  which is more aligned with the data distribution induced by the optimal policy  $\pi_*$ . In line 13, the algorithm combines data from both stages  $\{(x_i, a_i, r_i)\}_{i=1}^n$  and  $\{(x_i^0, a_i^0, r_i^0)\}_{i=1}^m$  to compute a refined least squares estimate  $\hat{\theta}$  of the reward function, minimizing the sum of squared errors across both datasets. By aggregating the two datasets together, there is an overlap between the data to compute  $\hat{\theta}$  and  $\hat{\theta}_0$ , so that the output policy  $\pi_{\hat{\theta}}^\eta$  is well covered by the intermediate policy  $\pi_{\hat{\theta}_0}^\eta$ .



## 2.4 Theoretical Guarantees

**Review of previous analysis** The previous analysis (e.g., Xiong et al., 2024a) basically follows the techniques of bandits and neglects the significance of KL-regularization. For simplicity, We use short-hand notation  $R(\theta, x, \pi) = \mathbb{E}_{a \sim \pi(\cdot|x)} R(\theta, x, a)$  and denote  $\text{KL}(\pi(\cdot|x) \parallel \pi'(\cdot|x))$  by  $\text{KL}(\pi \parallel \pi')$  when there is no confusion. We make the estimation on a dataset  $\{(x_i, a_i, r_i) : x_i \sim d_0, a_i \sim \pi_0(\cdot|x_i)\}_{i=1}^n$ :  $\pi_\theta^\eta = \arg\max_{\pi \in \Pi} \mathbb{E}_{x \sim d_0} [R(\hat{\theta}, x, \pi) - \eta^{-1} \text{KL}(\pi \parallel \pi_0)]$ , and has a small in-sample-error:  $\mathbb{E}_{x \sim d_0} \mathbb{E}_{a \sim \pi_\theta^\eta(\cdot|x)} [(R(\hat{\theta}, x, a) - R(\theta_*, x, a))^2] = O(1/n)$ . The sub-optimality is decomposed as:

$$\begin{aligned} Q(\pi^*) - Q(\pi_\theta^\eta) &= \mathbb{E}_{x \sim d_0} [R(\theta_*, x, \pi^*) - R(\hat{\theta}, x, \pi^*)] + \mathbb{E}_{x \sim d_0} [R(\hat{\theta}, x, \pi_\theta^\eta) - R(\theta_*, x, \pi_\theta^\eta)] \\ &\quad + \mathbb{E}_{x \sim d_0} [R(\hat{\theta}, x, \pi^*) - \eta^{-1} \text{KL}(\pi^* \parallel \pi_0)] - \mathbb{E}_{x \sim d_0} [R(\hat{\theta}, x, \pi_\theta^\eta) - \eta^{-1} \text{KL}(\pi_\theta^\eta \parallel \pi_0)] \\ &\leq \mathbb{E}_{x \sim d_0} [R(\theta_*, x, \pi^*) - R(\hat{\theta}, x, \pi^*) + R(\hat{\theta}, x, \pi_\theta^\eta) - R(\theta_*, x, \pi_\theta^\eta)], \end{aligned}$$

where the inequality holds since  $\pi_\theta^\eta$  is the maximum.

Then, the suboptimality can be further bounded by using the coverage condition (Definition 2.10) and concentration inequalities:

$$\begin{aligned} Q(\pi^*) - Q(\pi_\theta^\eta) &\leq 2C_{\text{GL}} \mathbb{E}_{x \sim d_0} \mathbb{E}_{a \sim \pi_0(\cdot|x)} [|R(\theta_*, x, a) - R(\hat{\theta}, x, a)|] \\ &\leq 2C_{\text{GL}} \sqrt{\mathbb{E}_{a \sim \pi_0(\cdot|x)} [(R(\theta_*, x, a) - R(\hat{\theta}, x, a))^2]} = O(C_{\text{GL}}/\sqrt{n}). \end{aligned}$$

Hence, they need  $\Theta(C_{\text{GL}}^2/\epsilon^2)$  sample complexity to ensure  $O(\epsilon)$  sub-optimality.

### Sharper results and analysis

**Theorem 2.8.** Suppose that Assumption 2.5 holds. For any  $\delta \in (0, 1/5)$ ,  $\epsilon > 0$  and constant  $c_{m,n} > 0$ , if we set  $m = \Theta(\eta^2 D^2 \cdot B^2 \log(2N_{\mathcal{R}}(\epsilon_c)/\delta))$  and  $n = \Theta(\eta/\epsilon \cdot B^2 \log(N_{\mathcal{R}}(\epsilon_c)/\delta))$  and  $\epsilon_c = \min\{\frac{\epsilon}{2(1+c_{m,n})B}, \frac{1}{8(1+c_{m,n})B\eta^2 D^2}\}$ , then with probability at least  $1 - 5\delta$  the output policy of Algorithm 1  $\pi_\theta^\eta$  is  $O(\epsilon)$  optimal.

Theorem 2.8 shows that the sample complexity of Algorithm 1 is  $O(\eta/\epsilon \log N_{\mathcal{R}}(\epsilon/\delta))$  when the reward scale is a constant and  $\epsilon$  is sufficiently small. The result indicates that the proposed two-stage mixed sampling strategy can achieve a suboptimality gap of  $\epsilon$  with only an additive dependence on the coverage coefficient  $D^2$ .

To illustrate the novel techniques to obtain the sharper bound, we highlight the crucial points in the sequel and defer the detailed proof to Appendix E.2.

**Part I: Decomposition of the suboptimality gap** The most challenging part is how to proceed with the suboptimality gap based on the strong convexity of the objective  $Q$  with the KL-regularization. Given the closed-form solution of  $\pi^*(a|x) = \pi_0(a|x) \exp(\eta R(\theta_*, x, a))/Z_{\theta_*}^\eta(x)$  and  $\pi_\theta^\eta(a|x) = \pi_0(a|x) \exp(\eta R(\hat{\theta}, x, a))/Z_\theta^\eta(x)$ , where  $Z_\theta^\eta(x) = \sum_{a \in \mathcal{A}} \pi_0(a|x) \cdot \exp(\eta R(\theta, x, a))$  denotes the normalization constant, we can write the suboptimality as

$$\begin{aligned} &\mathbb{E}_{\pi^*} \left[ R(\theta_*, x, a) - \frac{1}{\eta} \log \frac{\pi^*(a|x)}{\pi_0(a|x)} \right] - \mathbb{E}_{\pi_\theta^\eta} \left[ R^*(x, a) - \frac{1}{\eta} \log \frac{\pi_\theta^\eta(a|x)}{\pi_0(a|x)} \right] \\ &= \frac{1}{\eta} \mathbb{E}_{\pi^*} \left[ \log \frac{\pi_0(a|x) \exp(\eta R(\theta_*, x, a))}{\pi^*(a|x)} \right] - \frac{1}{\eta} \mathbb{E}_{\pi_\theta^\eta} \left[ \log \frac{\pi_0(a|x) \exp(\eta R(\theta_*, x, a))}{\pi_\theta^\eta(a|x)} \right] \\ &= -\frac{1}{\eta} (J(x; \hat{\theta}) - J(x; \theta_*)), \end{aligned}$$

where we define  $J(x; \theta) = \log Z_\theta^\eta(x) - \eta \mathbb{E}_{\pi_\theta^\eta} [R(\theta, x, a) - R(\theta_*, x, a)]$ , and the last equation is deduced by taking the distribution of  $\pi^*$  and  $\pi_\theta^\eta$  in the terms.

Thus, the suboptimality is expressed by the gap between  $\hat{\theta}$  and  $\theta_*$  with respect to the function  $J$ . By taking the first-order Taylor expansion with respect to  $\{\Delta(x, a) = R(\hat{\theta}, x, a) - R(\theta_*, x, a) : a \in \mathcal{A}\}$ , we can prove the following lemma.

243 **Lemma 2.9.** For any estimator  $\hat{\theta} \in \Theta$ , and the policy  $\pi_{\hat{\theta}}^{\eta}$  satisfying Definiton 2.4, we have

$$\begin{aligned} Q(\pi^*) - Q(\pi_{\hat{\theta}}^{\eta}) &= \eta \mathbb{E}_{x \sim d_0} \left[ \sum_{a \in \mathcal{A}} \pi_f^{\eta}(a|x) \Delta^2(x, a) - \sum_{a_1, a_2 \in \mathcal{A}} \pi_f^{\eta}(a_1|x) \pi_f^{\eta}(a_2|x) \Delta(x, a_1) \Delta(x, a_2) \right] \\ &\leq \eta \mathbb{E}_{x \sim d_0} \left[ \sum_{a \in \mathcal{A}} \pi_f^{\eta}(a|x) \Delta^2(x, a) \right], \end{aligned} \quad (2.2)$$

244 where  $f(\cdot, \cdot) = \gamma R(\hat{\theta}, \cdot, \cdot) + (1 - \gamma) R(\theta_*, \cdot, \cdot)$  ( $\gamma \in (0, 1)$ ) the inequality uses the fact that second  
245 term on the right-hand side of the equality is  $(\sum_{a \in \mathcal{A}} \pi_f^{\eta}(a|x) \Delta(x, a))^2 \geq 0$ .

246 The proof of this lemma is provided in Appendix E.2.

247 **Part II: Utilizing the coverage condition** In Algorithm 1, with the coverage condition (Defini-  
248 tion 2.5) and the concentration inequalities, if the sample size  $m = \Theta(\eta^2 D^2)$ , we can prove that  
249  $\|R(\hat{\theta}, \cdot, \cdot) - R(\theta_*, \cdot, \cdot)\|_{\infty} \leq \eta^{-1}$  and  $\|R(\hat{\theta}_0, \cdot, \cdot) - R(\theta_*, \cdot, \cdot)\|_{\infty} \leq \eta^{-1}$ , which implies the whole-  
250 policy coverage condition for all contexts:  $\|\pi_f^{\eta}(\cdot|x)/\pi_{\hat{\theta}_0}^{\eta}(\cdot|x)\|_{\infty} \leq e^4$ . Therefore, substituting it back  
251 into (2.2) leads to

$$Q(\pi^*) - Q(\pi_{\hat{\theta}}^{\eta}) \lesssim \eta \mathbb{E}_{x \sim d_0} \mathbb{E}_{a \sim \pi_{\hat{\theta}_0}^{\eta}} \left[ (R(\hat{\theta}, x, a) - R(\theta_*, x, a))^2 \right].$$

252 Note that  $\hat{\theta}$  in the RHS is computed using the data sampled from  $\pi_{\hat{\theta}_0}^{\eta}$ . By setting  $n = \Theta(\eta/\epsilon)$ , we  
253 obtain that  $\pi_{\hat{\theta}}^{\eta}$  is  $O(\epsilon)$  optimal.

## 254 2.5 Result under Local-Coverage Condition

255 In this subsection, we consider another coverage conditions appearing in previous work as described  
256 in Definition 2.11.

257 **Definition 2.10** (Global-Policy Coverage). Given a reference policy  $\pi_0$ ,  $C_{\text{GL}}$  is the minimum posi-  
258 tive real number satisfying that for any  $\pi : \mathcal{X} \rightarrow \mathcal{A}$

$$\sup_{x \sim d_0, a \in \mathcal{A}} \frac{\pi(a|x)}{\pi_0(a|x)} \leq C_{\text{GL}}.$$

259 **Definition 2.11** (Local KL-ball Coverage, Song et al. 2024). Given a reference policy  $\pi_0$ , for a  
260 positive constant  $\rho_{\text{KL}} < \infty$ , and all policy satisfying that  $\mathbb{E}_{x \sim d_0} [\text{KL}(\pi, \pi_0)] \leq \rho_{\text{KL}}$ , we define

$$\sup_{x \sim d_0, a \in \mathcal{A}} \frac{\pi(a|x)}{\pi_0(a|x)} := C_{\rho_{\text{KL}}}.$$

261 **Remark 2.12** (Relation between Local and Global Coverage Conditions). The local-coverage con-  
262 dition (Definition 2.11) is more precise because compared to the global conditions targeting all  
263 possible policies, it only constrains the coverage to a KL-ball. In Song et al. (2024), because  
264 of the specific form of the oracle (Definition 2.4), the considered policy class is  $\Pi = \{\pi(\cdot|x) \propto$   
265  $\pi_0(\cdot|x) \exp(\eta R(\theta, \cdot, \cdot)) : R(\theta, \cdot, \cdot) \in \mathcal{R}\}$ . Thus, they only need to assume that the condition hold for  
266  $\rho = 2\eta B$ , indicating that  $C_{\rho_{\text{KL}}} \leq C_{\text{GL}}$ . On the other hand, the data coverage condition (Definition  
267 2.5) is measured on the level of reward functions instead of policies. In this sense, the data coverage  
268 condition and local-coverage condition do not encompass each other.

269 **Corollary 2.13.** Let  $C_{\rho_{\text{KL}}}$  be in Definition 2.11 where  $\rho_{\text{KL}} = 2\eta B$ . For any  $\delta \in (0, 1/6)$  and  
270  $\epsilon > 0$ , if we set  $n = c_{m,n} m = \Theta(C_{\rho_{\text{KL}}} \eta / \epsilon \cdot B \log(N_{\mathcal{R}}(\epsilon_c) / \delta))$  (where constant  $c_{m,n} > 0$ ,  
271  $\epsilon_c = \epsilon / (2(1 + c_{m,n}^{-1})B)$ ) then with probability at least  $1 - 6\delta$  the output policy of Algorithm 3  $\pi_{\hat{\theta}}^{\eta}$  is  
272  $O(\epsilon)$  optimal.

273 In comparison with the sample complexity  $\Theta(\eta^2 D^2 + \eta/\epsilon)$  under data coverage in Theorem 2.8,  
274 the order  $\Theta(C_{\rho_{\text{KL}}} \eta / \epsilon)$  depends on a local coverage coefficient  $C_{\rho_{\text{KL}}}$ , but has a multiplicative depen-  
275 dence on the coverage coefficient instead of additive dependence. whether the additive dependence  
276 can be achieved under the local-coverage condition is left as future work. Moreover, we compare  
277 this result with Theorem 4.2 in Song et al. (2024) and suppose that the in-sample-error  $\epsilon_{\text{reward}}$  of  
278 Song et al. (2024) is  $O(1/n)$ , their sample complexity is  $\Theta(C_{\rho_{\text{KL}}}^2 / \epsilon^2)$ , which is looser than ours  
279  $\Theta(C_{\rho_{\text{KL}}} \eta / \epsilon)$  when  $\eta = o(C_{\rho_{\text{KL}}} / \epsilon)$ .

### 3 Reinforcement Learning from Preference Feedback

In this section, we consider the problem of aligning the language model with preference feedback.

#### 3.1 Problem Setup

In each round, we can sample a pair of actions (responses)  $a_1, a_2$  and query a preference oracle to get the preference label  $y \in \{0, 1\}$ , where  $y = 1$  means that the user prefers  $a_1$  over  $a_2$ . Specifically, when receiving a prompt  $x \in \mathcal{X}$ , and two actions (responses)  $a^1, a^2 \in \mathcal{A}$  from some LLM policy  $\pi(\cdot|x)$ , a preference oracle will give feedback  $y$  defined as follows:

**Definition 3.1** (Preference Oracle). A Preference Oracle is a function  $\mathbb{P} : \mathcal{X} \times \mathcal{A} \times \mathcal{A} \rightarrow \{0, 1\}$ . Given a context  $x \in \mathcal{X}$  and two actions  $a_1, a_2 \in \mathcal{A}$ , the oracle can be queried to obtain a preference signal  $y \sim \text{Bernoulli}(\mathbb{P}(x, a_1, a_2))$ , where  $y = 1$  indicates that  $a_1$  is preferred to  $a_2$  in the context  $x$ , and  $y = 0$  indicates the opposite.

To learn the preference, we follow [Ouyang et al. \(2022\)](#); [Zhu et al. \(2023\)](#); [Rafailov et al. \(2024\)](#); [Liu et al. \(2023\)](#); [Xiong et al. \(2024a\)](#) and assume that the preference oracle is measured by the difference of ground-truth reward functions  $R(\theta_*, x, a)$ , which is named the Bradley-Terry (BT) model ([Bradley and Terry, 1952b](#)).

**Definition 3.2** (Bradley-Terry Model). Given a context  $x \in \mathcal{X}$  and two actions  $a_1, a_2 \in \mathcal{A}$ , the probability of  $a_1$  being preferred to  $a_2$  is modeled as

$$\mathbb{P}(x, a_1, a_2) = \frac{\exp(R(\theta_*, x, a_1))}{\exp(R(\theta_*, x, a_1)) + \exp(R(\theta_*, x, a_2))} = \sigma(R(\theta_*, x, a_1) - R(\theta_*, x, a_2)), \quad (3.1)$$

where  $\sigma(u) = (1 + e^{-u})^{-1}$  is the sigmoid function.

The RLHF training always follows the fine-tuning process, which yields a reference policy  $\pi_0$ . When performing RLHF on specific tasks, to avoid overfitting, we impose KL-regularization to the learned reward model when optimizing the policy. Hence, our objective function is also (2.1).

To learn the reward function, we introduce the following assumption to ensure the existence of an MLE estimation oracle that can globally maximize the likelihood of the BT model over all possible reward functions.

**Definition 3.3** (MLE estimation oracle). Given a set of context-action pairs  $\{(x_i, a_i^1, a_i^2, y_i)\}_{i=1}^n$  generated from the BT model, can output the parameter  $\hat{\theta}$  such that

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}} \sum_{i=1}^n \underbrace{y_i \cdot \log \sigma(R(\theta, x_i, a_i^1) - R(\theta, x_i, a_i^2)) + (1 - y_i) \cdot \log \sigma(R(\theta, x_i, a_i^2) - R(\theta, x_i, a_i^1))}_{\mathcal{L}(\theta|x_i, a_i^1, a_i^2, y_i)}.$$

Following the previous analysis for RLHF ([Xiong et al., 2024a](#)), we assume the existence of a policy improvement oracle (Definition 2.4) that can compute the optimal policy  $\pi_\theta^\eta$  based on the reward function  $\hat{\theta}$ .

**Remark 3.4.** We learn the reward function since we can always control the reward (like clipping and normalization) to ensure that the reward function is always bounded by  $B$ . The bounded assumption does not apply for direct preference learning like DPO ([Rafailov et al., 2024](#)) since there is no intrinsic policy function class encompassing the soundness ([Song et al., 2024](#)), thus increasing the cases of overfitting.

In RLHF setting, we cannot directly observe or estimate absolute reward values. Consequently, the most intuitive estimation approach is to focus on relative rewards: for any context  $x$  and actions  $a^1, a^2$ , our estimated difference  $f(s, a^1) - f(s, a^2)$  should closely approximate the true reward difference  $r(s, a^1) - r(s, a^2)$ . Therefore, we extend the data coverage condition to the RLHF setting as follows.

**Definition 3.5** (Data Coverage). Given a reference policy  $\pi_0$ ,  $D^2$  is the minimum positive real number satisfying  $\forall (x, a) \in \mathcal{X} \times \mathcal{A}$ ,  $\pi(a|x) > 0$ , we have for any pair of  $\theta, \theta' \in \Theta$ , there exists  $b : \mathcal{X} \rightarrow [-B, B]$  such that

$$\frac{|R(\theta', x, a) - R(\theta, x, a) - b(x)|^2}{\mathbb{E}_{x' \sim d_0} \operatorname{Var}_{a' \sim \pi_0(\cdot|x')} [R(\theta', x', a') - R(\theta, x', a')]} \leq D^2.$$



### 3.2 Theoretical Guarantees

**Lower bound** We provide a lower bound for the RLHF problem with preference feedback. The lower bound is derived by constructing a hard instance where the reward function is difficult to estimate from the preference feedback.

**Theorem 3.6.** For any  $\epsilon \in (0, 1/256)$ ,  $\eta > 4$ , and any algorithm  $A$ , there exists a KL-regularized preference learning problem with  $O(N_{\mathcal{R}}(\epsilon))$  coverage coefficient and reward function class  $\mathcal{R}$  such that  $A$  requires at least  $\Omega\left(\min\left(\frac{\eta \log N_{\mathcal{R}}(\epsilon)}{\epsilon}, \frac{\log N_{\mathcal{R}}(\epsilon)}{\epsilon^2}\right)\right)$  samples to achieve a suboptimality gap of  $\epsilon$ .

We present a two-stage mixed-policy sampling algorithm for RLHF, which can be seen as an extension of Algorithm 1. Due to space limit, we defer it to Algorithm 3 in Appendix C.

**Upper bound** We provide the theoretical guarantees for Algorithm 3 in the following theorem.

**Theorem 3.7.** Suppose that Assumption 2.5 holds. For any  $\delta \in (0, 1/6)$  and  $\epsilon > 0$ , if we set

$$m = \Theta(\eta^2 D^2 \cdot e^B \log(N_{\mathcal{R}}(\epsilon_c)/\delta)) \text{ and } n = \Theta(\eta/\epsilon \cdot e^B \log(N_{\mathcal{R}}(\epsilon_c)/\delta))$$

where  $\epsilon_c = \min\left\{\frac{\epsilon}{2(1+c_{m,n}^{-1})e^B}, \frac{1}{8(1+c_{m,n})e^B \eta^2 D^2}\right\}$  then with probability at least  $1 - 6\delta$  the output policy of Algorithm 3  $\pi_{\theta}^{\eta}$  is  $O(\epsilon)$  optimal.

**Remark 3.8** (Comparison with Hybrid Framework). We compare our two-stage mixed sampling method with hybrid RL. From the algorithmic perspective, a hybrid RL algorithm first learns from an offline dataset and then requires sufficient online iterations to ensure the performance (Xiong et al., 2024a). For example, for a finite action space with  $A$  actions, the number of online iterations should be  $\Theta(A)$ . In contrast, our method only requires two iterations of sampling from mixed policy and interacting with the environment. Moreover, the results of hybrid literature depend on both the coverage coefficient and the structure complexity of the function class (like the dimension for a linear function class or eluder dimension (Russo and Van Roy, 2013)). Our result only needs the coverage condition of the reference policy. More importantly, we obtain a sharper bound on the sample complexity and derive the additive dependence on the coverage coefficient.

**Remark 3.9.** Although the coefficient  $e^B$  appearing in sample size  $m, n$  can be exponentially large, this term is caused by the non-linearity of the link function for the preference model, and is common in RLHF literature (Zhu et al., 2023; Xiong et al., 2024a; Ye et al., 2024a; Song et al., 2024).

Theorem 3.7 shows that the sample complexity of Algorithm 3 is  $\mathcal{O}(\eta/\epsilon \log N_{\mathcal{R}}(\epsilon/\delta))$  when the reward scale is a constant and  $\epsilon$  is sufficiently small. The result indicates that the proposed two-stage mixed sampling strategy can achieve a suboptimality gap of  $\epsilon$  with only an additive dependence on the coverage coefficient  $D^2$ .

Besides, the algorithm only requires sampling from the reference policy  $\pi_0$  and the intermediate policy  $\pi_{\theta_0}^{\eta}$ , which is more aligned with the practical scenarios where the preference feedback is collected from the human users and it is expensive to collect the data while the language model is being updated. Our result implies that we may achieve a near-optimal sample complexity by simply leveraging an intermediate policy to collect more data, and the training process of the reward model and the policy (language model) can be highly decoupled.

**Upper bound for local coverage** We also show the result under the local-coverage assumption (Definition 2.11) as follows.

**Corollary 3.10.** Let  $C_{\rho_{\text{KL}}}$  be in Definition 2.11 where  $\rho = 2\eta B$ . For any  $\delta \in (0, 1/6)$  and  $\epsilon > 0$ , if we set  $n = c_{m,n} m = \Theta(C_{\rho_{\text{KL}}} \eta/\epsilon \cdot e^B \log(N_{\mathcal{R}}(\epsilon_c)/\delta))$  (where constant  $c_{m,n} > 0$ ,  $\epsilon_c = \frac{\epsilon}{2(1+c_{m,n}^{-1})e^B}$ ) then with probability at least  $1 - 6\delta$  the output policy of Algorithm 3  $\pi_{\theta}^{\eta}$  is  $O(\epsilon)$  optimal.

## 4 Conclusions and Future Work

We have presented a comprehensive theoretical analysis of the role of reverse-KL regularization in decision-making models including contextual bandits and reinforcement learning from preference feedback, highlighting its significance in terms of sample complexity. Our results provide new insights into the power of regularization extending beyond its traditional role of mitigating errors from the current critic (or reward) model. Additionally, we examined the role of data coverage in both contextual bandits and RLHF. Our analysis shows that with sufficient coverage from the reference

policy, a mixed sampling strategy can achieve a sample complexity that exhibits only an additive dependence on the coverage coefficient without the need for explicit exploration or additional structural assumptions. For future directions, it is interesting to study if the sharp bound exists for the KL-regularized Markov Decision Process (MDP) problem. Additionally, it is also worthwhile to explore how to achieve a sample complexity bound with only additive dependence on the global and local coverage conditions in Definitions 2.10 and 2.11.

## 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*.
- AGARWAL, A., JIN, Y. and ZHANG, T. (2023). Vo  $q$  l: Towards optimal regret in model-free rl with nonlinear function approximation. In *The Thirty Sixth Annual Conference on Learning Theory*. PMLR.
- AGARWAL, A., KAKADE, S. M., LEE, J. D. and MAHAJAN, G. (2020). Optimality and approximation with policy gradient methods in markov decision processes. In *Conference on Learning Theory*. PMLR.
- AGARWAL, A., KAKADE, S. M., LEE, J. D. and MAHAJAN, G. (2021). On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research* **22** 1–76.
- AGARWAL, A., QIAN, J., RAKHLIN, A. and ZHANG, T. (2024). The non-linear  $f$ -design and applications to interactive learning. In *Forty-first International Conference on Machine Learning*.
- AHMED, Z., LE ROUX, N., NOROUZI, M. and SCHUURMANS, D. (2019). Understanding the impact of entropy on policy optimization. In *International conference on machine learning*. PMLR.
- ANTHROPIC, A. (2023). Introducing claude.
- AZAR, M. G., GUO, Z. D., PIOT, B., MUNOS, R., ROWLAND, M., VALKO, M. and CALANDRIELLO, D. (2024). A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*. PMLR.
- BAI, Y., JONES, A., NDOUSSE, K., ASKELL, A., CHEN, A., DASSARMA, N., DRAIN, D., FORT, S., GANGULI, D., HENIGHAN, T. ET AL. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.
- BERTHET, Q. and PERCHET, V. (2017). Fast rates for bandit optimization with upper-confidence frank-wolfe. *Advances in Neural Information Processing Systems* **30**.
- BRADLEY, R. A. and TERRY, M. E. (1952a). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika* **39** 324.
- BRADLEY, R. A. and TERRY, M. E. (1952b). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika* **39** 324–345.
- BROCKMAN, G. (2016). Openai gym. *arXiv preprint arXiv:1606.01540*.
- CALANDRIELLO, D., GUO, D., MUNOS, R., ROWLAND, M., TANG, Y., PIRES, B. A., RICHEMOND, P. H., LAN, C. L., VALKO, M., LIU, T. ET AL. (2024). Human alignment of large language models through online preference optimisation. *arXiv preprint arXiv:2403.08635*.
- CHEN, Z., DENG, Y., YUAN, H., JI, K. and GU, Q. (2024). Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*.
- CHRISTIANO, P. F., LEIKE, J., BROWN, T., MARTIC, M., LEGG, S. and AMODEI, D. (2017). Deep reinforcement learning from human preferences. *Advances in neural information processing systems* **30**.

COBBE, K., KOSARAJU, V., BAVARIAN, M., CHEN, M., JUN, H., KAISER, L., PLAPPERT, M.,  
TWAREK, J., HILTON, J., NAKANO, R. ET AL. (2021). Training verifiers to solve math word  
problems. *arXiv preprint arXiv:2110.14168* .

DI, Q., ZHAO, H., HE, J. and GU, Q. (2023). Pessimistic nonlinear least-squares value iteration  
for offline reinforcement learning. *arXiv preprint arXiv:2310.01380* .

DONG, H., XIONG, W., GOYAL, D., ZHANG, Y., CHOW, W., PAN, R., DIAO, S., ZHANG, J.,  
SHUM, K. and ZHANG, T. (2023). Raft: Reward ranked finetuning for generative foundation  
model alignment. *arXiv preprint arXiv:2304.06767* .

DONG, H., XIONG, W., PANG, B., WANG, H., ZHAO, H., ZHOU, Y., JIANG, N., SAHOO, D.,  
XIONG, C. and ZHANG, T. (2024). Rlhf workflow: From reward modeling to online rlhf. *arXiv  
preprint arXiv:2405.07863* .

FOSTER, D. J., KAKADE, S. M., QIAN, J. and RAKHLIN, A. (2021). The statistical complexity of  
interactive decision making. *arXiv preprint arXiv:2112.13487* .

FOX, R., PAKMAN, A. and TISHBY, N. (2016). Taming the noise in reinforcement learning via soft  
updates. In *32nd Conference on Uncertainty in Artificial Intelligence 2016, UAI 2016*. Association  
For Uncertainty in Artificial Intelligence (AUAI).

GEIST, M., SCHERRER, B. and PIETQUIN, O. (2019). A theory of regularized markov decision  
processes. In *International Conference on Machine Learning*. PMLR.

GOU, Z., SHAO, Z., GONG, Y., YELONG SHEN, YANG, Y., HUANG, M., DUAN, N. and CHEN,  
W. (2024). ToRA: A tool-integrated reasoning agent for mathematical problem solving. In *The  
Twelfth International Conference on Learning Representations*.  
URL <https://openreview.net/forum?id=Ep0TtjVoap>

GUI, L., GÂRBACEA, C. and VEITCH, V. (2024). Bonbon alignment for large language models  
and the sweetness of best-of-n sampling. *arXiv preprint arXiv:2406.00832* .

GULCEHRE, C., PAINE, T. L., SRINIVASAN, S., KONYUSHKOVA, K., WEERTS, L., SHARMA,  
A., SIDDHANT, A., AHERN, A., WANG, M., GU, C. ET AL. (2023). Reinforced self-training  
(rest) for language modeling. *arXiv preprint arXiv:2308.08998* .

GUO, S., ZHANG, B., LIU, T., LIU, T., KHALMAN, M., LLINARES, F., RAME, A., MESNARD,  
T., ZHAO, Y., PIOT, B. ET AL. (2024). Direct language model alignment from online ai feedback.  
*arXiv preprint arXiv:2402.04792* .

HAARNOJA, T., TANG, H., ABBEEL, P. and LEVINE, S. (2017). Reinforcement learning with deep  
energy-based policies. In *International conference on machine learning*. PMLR.

HAARNOJA, T., ZHOU, A., ABBEEL, P. and LEVINE, S. (2018). Soft actor-critic: Off-policy max-  
imum entropy deep reinforcement learning with a stochastic actor. In *International conference on  
machine learning*. PMLR.

HENDRYCKS, D., BURNS, C., KADAVATH, S., ARORA, A., BASART, S., TANG, E., SONG, D.  
and STEINHARDT, J. (2021). Measuring mathematical problem solving with the math dataset.  
*arXiv preprint arXiv:2103.03874* .

LAN, G. (2023). Policy mirror descent for reinforcement learning: Linear convergence, new sam-  
pling complexity, and generalized problem classes. *Mathematical programming* **198** 1059–1106.

LANGFORD, J. and ZHANG, T. (2007). The epoch-greedy algorithm for multi-armed bandits with  
side information. *Advances in neural information processing systems* **20**.

LATTIMORE, T. and SZEPESVÁRI, C. (2020). *Bandit algorithms*. Cambridge University Press.

LIU, T., ZHAO, Y., JOSHI, R., KHALMAN, M., SALEH, M., LIU, P. J. and LIU, J. (2023). Sta-  
tistical rejection sampling improves preference optimization. *arXiv preprint arXiv:2309.06657*  
.

462 MEI, J., XIAO, C., SZEPESVARI, C. and SCHUURMANS, D. (2020). On the global convergence  
463 rates of softmax policy gradient methods. In *International conference on machine learning*.  
464 PMLR.

465 META, A. (2024). Introducing meta llama 3: The most capable openly available llm to date. *Meta*  
466 *AI* .

467 NAKANO, R., HILTON, J., BALAJI, S., WU, J., OUYANG, L., KIM, C., HESSE, C., JAIN, S.,  
468 KOSARAJU, V., SAUNDERS, W. ET AL. (2021). Webgpt: Browser-assisted question-answering  
469 with human feedback. *arXiv preprint arXiv:2112.09332* .

470 NEU, G., JONSSON, A. and GÓMEZ, V. (2017). A unified view of entropy-regularized markov  
471 decision processes. *arXiv preprint arXiv:1705.07798* .

472 OPENAI (2023). Gpt-4 technical report. *ArXiv* **abs/2303.08774**.

473 OUYANG, L., WU, J., JIANG, X., ALMEIDA, D., WAINWRIGHT, C., MISHKIN, P., ZHANG, C.,  
474 AGARWAL, S., SLAMA, K., RAY, A. ET AL. (2022). Training language models to follow in-  
475 structions with human feedback. *Advances in Neural Information Processing Systems* **35** 27730–  
476 27744.

477 RAFAILOV, R., SHARMA, A., MITCHELL, E., MANNING, C. D., ERMON, S. and FINN, C. (2024).  
478 Direct preference optimization: Your language model is secretly a reward model. *Advances in*  
479 *Neural Information Processing Systems* **36**.

480 RUSSO, D. and VAN ROY, B. (2013). Eluder dimension and the sample complexity of optimistic  
481 exploration. *Advances in Neural Information Processing Systems* **26**.

482 SCHULMAN, J., CHEN, X. and ABBEEL, P. (2017a). Equivalence between policy gradients and  
483 soft q-learning. *arXiv preprint arXiv:1704.06440* .

484 SCHULMAN, J., LEVINE, S., ABBEEL, P., JORDAN, M. and MORITZ, P. (2015). Trust region  
485 policy optimization. In *International Conference on Machine Learning*. PMLR.

486 SCHULMAN, J., WOLSKI, F., DHARIWAL, P., RADFORD, A. and KLIMOV, O. (2017b). Proximal  
487 policy optimization algorithms. *arXiv preprint arXiv:1707.06347* .

488 SHANI, L., EFRONI, Y. and MANNOR, S. (2020). Adaptive trust region policy optimization: Global  
489 convergence and faster rates for regularized mdps. In *Proceedings of the AAAI Conference on*  
490 *Artificial Intelligence*, vol. 34.

491 SHAO, Z., WANG, P., ZHU, Q., XU, R., SONG, J., ZHANG, M., LI, Y., WU, Y. and GUO, D.  
492 (2024). Deepseekmath: Pushing the limits of mathematical reasoning in open language models.  
493 *arXiv preprint arXiv:2402.03300* .

494 SONG, Y., SWAMY, G., SINGH, A., BAGNELL, D. and SUN, W. (2024). The importance of online  
495 data: Understanding preference fine-tuning via coverage. In *ICML 2024 Workshop: Aligning*  
496 *Reinforcement Learning Experimentalists and Theorists*.

497 SONG, Y., ZHOU, Y., SEKHARI, A., BAGNELL, D., KRISHNAMURTHY, A. and SUN, W. (2023).  
498 Hybrid rl: Using both offline and online data can make rl efficient. In *The Eleventh International*  
499 *Conference on Learning Representations*.

500 SUTTON, R. S. (2018). Reinforcement learning: An introduction. *A Bradford Book* .

501 SZEPESVÁRI, C. (2022). *Algorithms for reinforcement learning*. Springer nature.

502 TEAM, G., ANIL, R., BORGEAUD, S., WU, Y., ALAYRAC, J.-B., YU, J., SORICUT, R., SCHALK-  
503 WYK, J., DAI, A. M., HAUTH, A. ET AL. (2023). Gemini: a family of highly capable multimodal  
504 models. *arXiv preprint arXiv:2312.11805* .

505 TONG, Y., ZHANG, X., WANG, R., WU, R. and HE, J. (2024). Dart-math: Difficulty-aware  
506 rejection tuning for mathematical problem-solving. *arXiv preprint arXiv:2407.13690* .

TOUVRON, H., MARTIN, L., STONE, K., ALBERT, P., ALMAHAIRI, A., BABAEI, Y., BASH-  
LYKOV, N., BATRA, S., BHARGAVA, P., BHOSALE, S. ET AL. (2023). Llama 2: Open foundation  
and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

WANG, H., LIN, Y., XIONG, W., YANG, R., DIAO, S., QIU, S., ZHAO, H. and ZHANG, T. (2024a).  
Arithmetic control of llms for diverse user preferences: Directional preference alignment with  
multi-objective rewards. *arXiv preprint arXiv:2402.18571*.

WANG, H., XIONG, W., XIE, T., ZHAO, H. and ZHANG, T. (2024b). Interpretable preferences via  
multi-objective reward modeling and mixture-of-experts. *arXiv preprint arXiv:2406.12845*.

WANG, R., SALAKHUTDINOV, R. R. and YANG, L. (2020). Reinforcement learning with gen-  
eral value function approximation: Provably efficient approach via bounded eluder dimension.  
*Advances in Neural Information Processing Systems* **33** 6123–6135.

WANG, Y., LIU, Q. and JIN, C. (2023a). Is rlhf more difficult than standard rl? a theoretical  
perspective. *Advances in Neural Information Processing Systems* **36** 76006–76032.

WANG, Z., DONG, Y., DELALLEAU, O., ZENG, J., SHEN, G., EGERT, D., ZHANG, J. J., SREED-  
HAR, M. N. and KUCHAIEV, O. (2024c). Helpsteer2: Open-source dataset for training top-  
performing reward models. *arXiv preprint arXiv:2406.08673*.

WANG, Z., DONG, Y., ZENG, J., ADAMS, V., SREEDHAR, M. N., EGERT, D., DELALLEAU,  
O., SCOWCROFT, J. P., KANT, N., SWOPE, A. ET AL. (2023b). Helpsteer: Multi-attribute  
helpfulness dataset for steerlm. *arXiv preprint arXiv:2311.09528*.

WU, Y., SHARIFF, R., LATTIMORE, T. and SZEPESVÁRI, C. (2016). Conservative bandits. In  
*International Conference on Machine Learning*. PMLR.

WU, Y., SUN, Z., YUAN, H., JI, K., YANG, Y. and GU, Q. (2024). Self-play preference optimiza-  
tion for language model alignment. *arXiv preprint arXiv:2405.00675*.

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*.

XIONG, W., DONG, H., YE, C., WANG, Z., ZHONG, H., JI, H., JIANG, N. and ZHANG, T.  
(2024a). Iterative preference learning from human feedback: Bridging theory and practice for  
rlhf under kl-constraint. In *Forty-first International Conference on Machine Learning*.

XIONG, W., SHI, C., SHEN, J., ROSENBERG, A., QIN, Z., CALANDRIELLO, D., KHALMAN, M.,  
JOSHI, R., PIOT, B., SALEH, M. ET AL. (2024b). Building math agents with multi-turn iterative  
preference learning. *arXiv preprint arXiv:2409.02392*.

YE, C., XIONG, W., GU, Q. and ZHANG, T. (2023). Corruption-robust algorithms with uncertainty  
weighting for nonlinear contextual bandits and markov decision processes. In *International Con-  
ference on Machine Learning*. PMLR.

YE, C., XIONG, W., ZHANG, Y., JIANG, N. and ZHANG, T. (2024a). A theoretical analysis  
of nash learning from human feedback under general kl-regularized preference. *arXiv preprint  
arXiv:2402.07314*.

YE, C., YANG, R., GU, Q. and ZHANG, T. (2024b). Corruption-robust offline reinforcement learn-  
ing with general function approximation. *Advances in Neural Information Processing Systems*  
**36**.

YUE, Y., BRODER, J., KLEINBERG, R. and JOACHIMS, T. (2012). The k-armed dueling bandits  
problem. *Journal of Computer and System Sciences* **78** 1538–1556.

ZHANG, T. (2023). *Mathematical analysis of machine learning algorithms*. Cambridge University  
Press.

ZHANG, X., XIONG, W., CHEN, L., ZHOU, T., HUANG, H. and ZHANG, T. (2024). From lists to  
emojis: How format bias affects model alignment. *arXiv preprint arXiv:2409.11704*.



554 ZHANG, Z., ZHOU, Y. and JI, X. (2020). Almost optimal model-free reinforcement learning via  
555 reference-advantage decomposition. *Advances in Neural Information Processing Systems* **33**  
556 15198–15207.

557 ZHAO, H., HE, J. and GU, Q. (2023a). A nearly optimal and low-switching algorithm for rein-  
558 forcement learning with general function approximation. *arXiv preprint arXiv:2311.15238* .

559 ZHAO, Y., JOSHI, R., LIU, T., KHALMAN, M., SALEH, M. and LIU, P. J. (2023b). Slic-hf:  
560 Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425* .

561 ZHONG, H., FENG, G., XIONG, W., ZHAO, L., HE, D., BIAN, J. and WANG, L. (2024). Dpo  
562 meets ppo: Reinforced token optimization for rlhf. *arXiv preprint arXiv:2404.18922* .

563 ZHU, B., JIAO, J. and JORDAN, M. I. (2023). Principled reinforcement learning with human  
564 feedback from pairwise or  $k$ -wise comparisons. *arXiv preprint arXiv:2301.11270* .

565 ZIEGLER, D. M., STIENNON, N., WU, J., BROWN, T. B., RADFORD, A., AMODEI, D., CHRIS-  
566 TIANO, P. and IRVING, G. (2019). Fine-tuning language models from human preferences. *arXiv*  
567 *preprint arXiv:1909.08593* .

## 568 **NeurIPS Paper Checklist**

### 569 **1. Claims**

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

572 Answer: [\[Yes\]](#)

573 Justification: We clearly state the contributions of the paper in the abstract and introduct-  
574 tion. The main claims are that we provide a theoretical analysis of the role of reverse-KL  
575 regularization in decision-making models including contextual bandits and reinforcement  
576 learning from preference feedback, highlighting its significance in terms of sample com-  
577 plexity. We also show that with sufficient coverage from the reference policy, a mixed  
578 sampling strategy can achieve a sample complexity that exhibits only an additive depen-  
579 dence on the coverage coefficient without the need for explicit exploration or additional  
580 structural assumptions.

581 Guidelines:

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

### 591 **2. Limitations**

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

593 Answer: [\[Yes\]](#)

594 Justification: We detail in our conclusion the assumptions that are crucial for our analysis  
595 but may be relaxed in future work.

596 Guidelines:

- 597 • The answer NA means that the paper has no limitation while the answer No means  
598 that the paper has limitations, but those are not discussed in the paper.
- 599 • The authors are encouraged to create a separate ”Limitations” section in their paper.

- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory assumptions and proofs

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

Answer: [\[Yes\]](#)

Justification: Every theorem and lemma in this paper has a complete and rigorous proof, either in the main text or deferred to appendix.

Guidelines:

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

### 4. Experimental result reproducibility

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

Answer: [\[NA\]](#)

Justification: This paper is mostly theoretical and provides necessary details to reproduce the synthetic experiments. The paper does not include real-world experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.

- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

## 5. Open access to data and code

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

Answer: [NA]

Justification: This paper is purely theoretical with only synthetic experiments.

Guidelines:

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

## 6. Experimental setting/details

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

Answer: [NA]

Justification: This paper is purely theoretical with only synthetic experiments.

Guidelines:

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

## 7. Experiment statistical significance

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

Answer: [NA]

Justification: This paper is purely theoretical.

Guidelines:

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

## 8. Experiments compute resources

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

Answer: [NA]

Justification: This paper is purely theoretical.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

## 9. Code of ethics

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

Answer: [Yes]

Justification: This theoretical research is fully in line with the NeurIPS Code of Ethics.

Guidelines:

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

## 10. Broader impacts

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

Answer: [No]

Justification: The purpose of this research is to advance the theory of reinforcement learning. It may have downstream impacts, none of which is worth a specific discussion.

Guidelines:

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

## 11. Safeguards

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

Answer: [NA]

Justification: We believe that our theoretical analysis poses no such risk.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.



- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

## 12. Licenses for existing assets

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

Answer: [NA]

Justification: We do not use existing assets.

Guidelines:

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

## 13. New assets

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

Answer: [NA]

Justification: We do not release new assets.

Guidelines:

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

## 14. Crowdsourcing and research with human subjects

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

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

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

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

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

#### 16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [NA]

Justification: All the proof techniques in this theoretical research are proposed by human.

Guidelines:

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

## A Additional Related Work

**Analyses of Policy Optimization with Regularization** While it is previously unknown whether regularization can improve the sample complexity of policy optimization without additional assumptions, there are some works that provided a sharp convergence rate in the presence of regularization (Mei et al., 2020; Shani et al., 2020; Agarwal et al., 2020, 2021; Lan, 2023). However, these works either assumed the access of exact or unbiased policy gradient or required uniform value function approximation error, which are not the standard case in general sample-based RL setting. For instance, Lan (2023) provided a sharp convergence rate for policy optimization with KL-regularization, assuming the access to an unbiased value function estimator (Condition 4.1) and the bounded infinity norm on the error (Conditions 4.2, 4.3), which is standard in the literature of optimization. However, RL algorithms usually make biased estimation to balance exploration and exploitation. Instead of focusing on the influence of regularization on the optimization, our work aims to understand how the KL-regularization affects the exploration and exploitation trade-off in the bandit and RLHF settings through a novel analysis on the optimal sample complexity.

**RLHF Algorithms** There are mainly three types of RLHF algorithms: offline, online and hybrid. The most well-known offline algorithms are Slic (Zhao et al., 2023b), Direct Preference Optimization (DPO) (Rafailov et al., 2024), Identity-PO (IPO) (Azar et al., 2024) and (SPIN) (Chen et al., 2024). They aim to approximate the closed-form solution of the optimization problem on a fixed offline dataset. For the online algorithms, the most representative one is Proximal Policy Optimization (PPO) (Schulman et al., 2017b). PPO has been used in the Chat-GPT (OpenAI, 2023), Gemini (Team et al., 2023), and Claude (Bai et al., 2022). However, the deep RL method PPO is known to be sample inefficient and unstable, making its success hard to reproduce for the open-source community. In response to this, there have been many efforts to propose alternative algorithms to the PPO algorithm. The Reward ranked fine-tuning (RAFT) (also known as rejection sampling finetuning) (Dong et al., 2023; Touvron et al., 2023; Gulcehre et al., 2023; Gui et al., 2024) is a stable framework requiring minimal hyper-parameter tuning, which iteratively learns from the best-of-n policy (Nakano et al., 2021). This framework proves to be particularly effective in the reasoning task such as (Gou et al., 2024; Tong et al., 2024). However, the RAFT-like algorithms only use the positive signal by imitating the best-of-n sampling. To further improve the efficiency, there is an emerging body of literature that proposes online direct preference optimization by extending DPO or IPO to an online iterative framework (Xiong et al., 2024a; Guo et al., 2024; Wu et al., 2024; Calandriello et al., 2024; Xiong et al., 2024b). Finally, for the third type, the common point of hybrid and online algorithms is that they both require further interaction with the preference oracle and on-policy data collection. The difference is that hybrid algorithms start with a pre-collected dataset (Xiong et al., 2024a; Song et al., 2024; Touvron et al., 2023), while the online algorithms learn from scratch.

**RLHF Theory** The theoretical study of RLHF can date back to the dueling bandits (Yue et al., 2012) and follow-up work on MDPs (Wang et al., 2023a; Zhu et al., 2023). However, these works deviate from the practice because they do not realize the significance of KL-regularization and still choose the greedy policy that simply maximizes the reward. After this line of work, Xiong et al. (2024a); Ye et al. (2024a); Song et al. (2024) highlight the KL-regularization theoretically and incorporate the KL term into the learning objective. However, they circumvent the special advantages of KL-regularization and still follow the techniques in bandit analysis, thus obtaining loose bounds. In our paper, we establish a new lower bound and a sharper upper bound for the KL-regularized framework, thus validating the empirical advantage of KL-regularization. There are also some works extending KL-regularized RLHF from bandit problems to the Markov decision process (MDP) problems (Zhong et al., 2024; Xiong et al., 2024b). We expect that our techniques can also be extended to the MDP setting, which we leave for future work.

## B Alternative Approach with Active Query

In this section, we remove the convergence assumption and show that by actively querying the reward signal, we can still achieve the sharper sample complexity bound. The key idea is to actively query the context-action pairs that are most informative for the current reward model in stage 1, and then follow the same steps as in stage 2 of Algorithm 1.

In the following definition, we introduce the uncertainty measure  $\mathcal{D}_{\mathcal{R}}(x, a; D)$  which characterizes the informativeness of the context-action pair  $(x, a)$  with respect to a dataset  $D$ .

---

**Algorithm 2** Active Querying for KL-Regularized Bandits

---

```

1: Input:  $\eta, \epsilon, \Theta$ .
2: for  $i = 1, \dots, m$  do
3:   Sample context  $x \sim d_0$ .
4:   if there exists  $a \in \mathcal{A}$  such that  $\mathcal{D}_{\mathcal{R}}(x, a; D_0) \geq u$  then
5:     Query the reward signal  $r$  for  $(x, a)$  and update the dataset  $D_0 \leftarrow D_0 \cup (x, a, r)$ .
6:   end if
7: end for
8: Compute the least square estimate  $\hat{\theta}_0$  of the reward function based on  $D_0$ .

```

---

952 **Definition B.1** (Uncertainty Measure). For a given dataset  $D = \{x_i, a_i\}_{i=1}^m$ , we define the uncer-  
953 tainty measure  $\mathcal{D}_{\mathcal{R}}(x, a; D)$  as follows:

$$\mathcal{D}_{\mathcal{R}}(x, a; D) := \sup_{R_1, R_2 \in \mathcal{R}} \frac{|R_1(x, a) - R_2(x, a)|}{\sqrt{1 + \sum_{i=1}^m |R_1(x_i, a_i) - R_2(x_i, a_i)|^2}}.$$

954 **Definition B.2** (Definition of eluder dimension, [Russo and Van Roy 2013](#)). The eluder dimension  
955 of a function class  $\mathcal{G}$  with domain  $\mathcal{Z}$  is defined as follows:

- 956 • A point  $z \in \mathcal{Z}$  is  $\epsilon$ -dependent on  $z_1, z_2, \dots, z_k \in \mathcal{Z}$  with respect to  $\mathcal{G}$ , if for all  $g_1, g_2 \in \mathcal{G}$  such  
957 that  $\sum_{i=1}^k [g_1(z_i) - g_2(z_i)]^2 \leq \epsilon$ , it holds that  $|g_1(z) - g_2(z)| \leq \epsilon$ .
- 958 • Further  $z$  is said to be  $\epsilon$ -independent of  $z_1, z_2, \dots, z_k$  with respect to  $\mathcal{G}$ , if  $z$  is not dependent on  
959  $z_1, z_2, \dots, z_k$ .
- 960 • The eluder dimension of  $\mathcal{G}$ , denoted by  $\dim_E(\mathcal{G}, \epsilon)$ , is the length of the longest sequence of  
961 elements in  $\mathcal{Z}$  such that every element is  $\epsilon'$ -independent of its predecessors for some  $\epsilon' \geq \epsilon$ .

962 **Lemma B.3.** Let  $\dim_E(\mathcal{R})$  be the eluder dimension of the reward function class  $\mathcal{R}$  as defined in  
963 Definition B.2. In Algorithm 2, if we set  $m = \tilde{O}(\dim_E(\mathcal{R})\eta/\epsilon u^2)$ , then with probability at least  
964  $1 - \delta$ ,  $\mathbb{P}_{x \sim d_0} \{\max_{a \in \mathcal{A}} \mathcal{D}_{\mathcal{R}}(x, a; D_0) \geq u\} \leq \epsilon/\eta$ .

965 *Proof.* We denote the  $m$  sampled context by  $x_1, \dots, x_m$  numbered in the order of operations,  
966 and denote by  $D_{0,i}$  the dataset  $D_0$  at the beginning of the  $i$ -th iteration. Also, we define  $\mu_i :=$   
967  $\mathbb{P}_{x \sim d_0} \{\max_{a \in \mathcal{A}} \mathcal{D}_{\mathcal{R}}(x, a; D_{0,i}) \geq u\}$ .

968 Applying freedman's inequality (Lemma G.6), we have with probability at least  $1 - \delta$ :

$$\begin{aligned} \sum_{i=1}^m [\mu_i - \mathbb{1}\{\max_{a \in \mathcal{A}} \mathcal{D}_{\mathcal{R}}(x_i, a; D_{0,i}) \geq u\}] &\leq 2 \sqrt{\sum_{i=1}^m \mu_i \cdot \log((2m+2)/\delta)} + 2 \sqrt{\log((2m+2)/\delta)} \\ &\quad + 2 \log((2m+2)/\delta) \\ &\leq \frac{1}{2} \sum_{i=1}^m \mu_i + 4 \log((2m+2)/\delta) + 2 \sqrt{\log((2m+2)/\delta)} \end{aligned}$$

969 where the last inequality follows from Young's inequality.

970 Rearranging the terms, we obtain:

$$\begin{aligned} \sum_{i=1}^m \mu_i &\leq 2|D_0| + 8 \log((2m+2)/\delta) + 4 \sqrt{\log((2m+2)/\delta)} \\ &\leq 2 \dim_E(\mathcal{R})/u^2 + 8 \log((2m+2)/\delta) + 4 \sqrt{\log((2m+2)/\delta)} \end{aligned}$$

971 where the last inequality holds due to the definition of eluder dimension.<sup>3</sup>

972 Hence, by setting  $m = \tilde{O}(\dim_E(\mathcal{R})\eta/\epsilon u^2)$ , we can ensure that  $\frac{1}{m} \sum_{i=1}^m \mu_i \leq \frac{\epsilon}{\eta}$ .  $\square$

973 **Theorem B.4.** Suppose we alternatively execute Algorithm 2 for stage 1 in Algorithm 1, and sup-  
974 pose that the reward signals lies in  $[0, 1]$ . If we set  $u = \frac{1}{\eta \sqrt{16 \log(2N_{\mathcal{R}}(\epsilon_c)/\delta) + 5}}$ , set  $m$  as in Lemma

---

<sup>3</sup>Here we use the reduction from generalized eluder dimension to eluder dimension as discussed in Theorem B.3 of [Zhao et al. \(2023a\)](#)

975 **B.3**, and set  $n = \Theta(\eta \cdot \log(N_{\mathcal{R}}(\epsilon_c)/\delta)/\epsilon)$ , then with probability at least  $1 - 6\delta$ , the output policy  
 976  $\pi_{\hat{\theta}}^\eta$  is  $O(\epsilon)$  optimal.

977 *Proof.* By Lemma **B.3**, we have with probability at least  $1 - \delta$ ,  $\mathbb{P}_{x \sim d_0} \{\max_{a \in \mathcal{A}} \mathcal{D}_{\mathcal{R}}(x, a; D_0) \geq$   
 978  $u\} \leq \epsilon/\eta$ .

979 Let  $\mathcal{X}_0 := \{x \in \mathcal{X} : \max_{a \in \mathcal{A}} \mathcal{D}_{\mathcal{R}}(x, a; D_0) < u\}$ . Similar to the proof of Lemma **E.4**, by Lemma  
 980 **E.2** and the definition of  $u$ , we have with probability at least  $1 - 4\delta$ , for all  $x \in \mathcal{X}_0$ ,

$$\eta |R(\hat{\theta}_0, x, a) - R(\theta_*, x, a)| \leq 1, \eta |R(\hat{\theta}, x, a) - R(\theta_*, x, a)| \leq 1. \quad (\text{B.1})$$

981 By Lemma **2.9**, we have

$$Q(\pi^*) - Q(\pi_{\hat{\theta}}^\eta) \leq \eta \mathbb{E}_{x \sim d_0} \left[ \sum_{a \in \mathcal{A}} \pi_f^\eta(a|x) (R(\hat{\theta}, x, a) - R(\theta_*, x, a))^2 \right].$$

From (B.1), we have for any  $x \in \mathcal{X}_0$ ,  $a \in \mathcal{A}$ ,

$$\frac{\pi_f^\eta(a|x)}{\pi_{\theta_0}^\eta(a|x)} \leq e^4.$$

982 Therefore, we can proceed as in the proof of Theorem **2.8** to obtain

$$\begin{aligned} Q(\pi^*) - Q(\pi_{\hat{\theta}}^\eta) &\leq \eta e^4 \mathbb{E}_{x \sim d_0} \left[ \mathbb{1}(x \in \mathcal{X}_0) \sum_{a \in \mathcal{A}} \pi_{\hat{\theta}_0}^\eta(a|x) (R(\hat{\theta}, x, a) - R(\theta_*, x, a))^2 \right] \\ &\quad + \eta \cdot \mathbb{P}_{x \sim d_0} \{\max_{a \in \mathcal{A}} \mathcal{D}_{\mathcal{R}}(x, a; D_0) \geq u\}. \end{aligned}$$

983 Then we can complete the proof by Lemma **E.3** and the union bound.  $\square$

984 **Remark B.5.** Theorem **B.4** shows that we can achieve a sample complexity of  
 985  $\tilde{O}(d \log N_{\mathcal{R}}(\epsilon/\delta) \eta^3/\epsilon)$ , where  $d$  is the eluder dimension (Russo and Van Roy, 2013) of the  
 986 reward function class  $\mathcal{R}$ . This is a significant completion to the results in Theorem **2.8**, which  
 987 requires the coverage assumption but do not rely on the structural assumption of the reward function  
 988 class. Compared to the coverage condition (Assumption **2.5**), the eluder dimension is more common  
 989 in the online setting (Russo and Van Roy, 2013; Agarwal et al., 2023; Zhao et al., 2023a). It is worth  
 990 noting that for the hybrid learning setting, which our proposed algorithm is designed for, coverage  
 991 condition and structural assumption on the function class were both required in the previous works  
 992 (Song et al., 2023; Xiong et al., 2024a).

## 993 C KL-Regularized Algorithm for RLHF

994 In this section, we present a two-stage mixed-policy sampling algorithm for RLHF in Algorithm **3**,  
 995 which can be seen as an extension of Algorithm **1**. There are two stages in the algorithm.

996 In the first stage, we sample  $m$  context-action pairs  $\{(\tilde{x}_i, \tilde{a}_i^1, \tilde{a}_i^2, \tilde{y}_i)\}_{i=1}^m$  from the BT model and  
 997 call the preference oracle to get the preference labels. We then compute the MLE estimator of the  
 998 reward function  $\hat{\theta}_0$  based on the preference feedback in line **6**. Afterwards, we apply the planning  
 999 oracle to compute the optimal policy  $\pi_{\hat{\theta}_0}^\eta$  based on the reward function  $\hat{\theta}_0$  in line **7**. Line **6** and line **7**  
 1000 correspond to the practical implementation of RLHF (Ouyang et al., 2022; Bai et al., 2022; Touvron  
 1001 et al., 2023) given a dataset of preference feedback.

1002 In the second stage, we sample  $n$  context-action pairs  $\{(x_i, a_i^1, a_i^2, y_i)\}_{i=1}^n$  using the intermediate  
 1003 policy  $\pi_{\hat{\theta}_0}^\eta$  and call the preference oracle to get the preference labels. We then compute the MLE  
 1004 estimator of the reward function  $\hat{\theta}$  based on the preference feedback from both stages. Finally, we  
 1005 apply the planning oracle to compute the optimal policy  $\pi_{\hat{\theta}}^\eta$  based on the reward function  $\hat{\theta}$ .

## 1006 D Experimental Results

1007 In this section, we conduct experiments with synthetic data to investigate the benefit of mixed-policy  
 1008 sampling and the effect of KL-regularization coefficient on the sample complexity of the problem.  
 1009 We plot the experimental results for RL from preference feedback in Figure **2** and the results for



---

**Algorithm 3** Two-stage Mixed-Policy Sampling from Preference Feedback (TMPS-PF)

---

- 1: **Input:**  $\eta, \epsilon, \pi_0, \Theta$ .  
 $\triangleright$  Use policy  $\pi_0$  to achieve sufficient data coverage
  - 2: **for**  $i = 1, \dots, m$  **do**
  - 3:     Sample context  $\tilde{x}_i \sim d_0$  and 2 actions  $\tilde{a}_i^1, \tilde{a}_i^2 \sim \pi_0(\cdot | \tilde{x}_i)$ .
  - 4:     Observe preference label  $\tilde{y}_i \in \{0, 1\}$  from the preference oracle defined in Definition 3.1.
  - 5: **end for**
  - 6: Compute the MLE estimator of the reward function based on  $\{(\tilde{x}_i, \tilde{a}_i^1, \tilde{a}_i^2, \tilde{y}_i)\}_{i=1}^m$ :  

$$\hat{\theta}_0 \leftarrow \operatorname{argmax}_{\theta} \sum_{i=1}^m \tilde{y}_i \cdot \log \sigma(R(\theta, \tilde{x}_i, \tilde{a}_i^1) - R(\theta, \tilde{x}_i, \tilde{a}_i^2)) + (1 - \tilde{y}_i) \cdot \log \sigma(R(\theta, \tilde{x}_i, \tilde{a}_i^2) - R(\theta, \tilde{x}_i, \tilde{a}_i^1)).$$
  - 7: Apply the planning oracle to compute  $\pi_{\hat{\theta}_0}^{\eta}(\cdot | \cdot) \propto \pi_0(\cdot | \cdot) \exp(\eta R(\hat{\theta}_0, \cdot, \cdot))$ .  
 $\triangleright$  Use policy  $\pi_{\hat{\theta}_0}^{\eta}$  to sample new responses
  - 8: **for**  $i = 1, \dots, n$  **do**
  - 9:     Sample context  $x_i \sim d_0$  and 2 actions  $a_i^1, a_i^2 \sim \pi_{\hat{\theta}_0}^{\eta}(\cdot | x_i)$ .
  - 10:     Observe preference label  $y_i \in \{0, 1\}$  from the preference oracle defined in Definition 3.1.
  - 11: **end for**
  - 12: Compute the MLE estimator of the reward function using  $\{(x_i, a_i^1, a_i^2, y_i)\}_{i=1}^n$  together with  $\{(\tilde{x}_i, \tilde{a}_i^1, \tilde{a}_i^2, \tilde{y}_i)\}_{i=1}^m$ :  

$$\hat{\theta} \leftarrow \operatorname{argmax}_{\theta} \sum_{i=1}^m \tilde{y}_i \cdot \log \sigma(R(\theta, \tilde{x}_i, \tilde{a}_i^1) - R(\theta, \tilde{x}_i, \tilde{a}_i^2)) + (1 - \tilde{y}_i) \cdot \log \sigma(R(\theta, \tilde{x}_i, \tilde{a}_i^2) - R(\theta, \tilde{x}_i, \tilde{a}_i^1))$$

$$+ \sum_{i=1}^n y_i \cdot \log \sigma(R(\theta, x_i, a_i^1) - R(\theta, x_i, a_i^2)) + (1 - y_i) \cdot \log \sigma(R(\theta, x_i, a_i^2) - R(\theta, x_i, a_i^1))$$
  - 13: **Output**  $\pi_{\hat{\theta}}^{\eta}(\cdot | \cdot) \propto \pi_0(\cdot | \cdot) \exp(\eta R(\hat{\theta}, \cdot, \cdot))$ .
- 

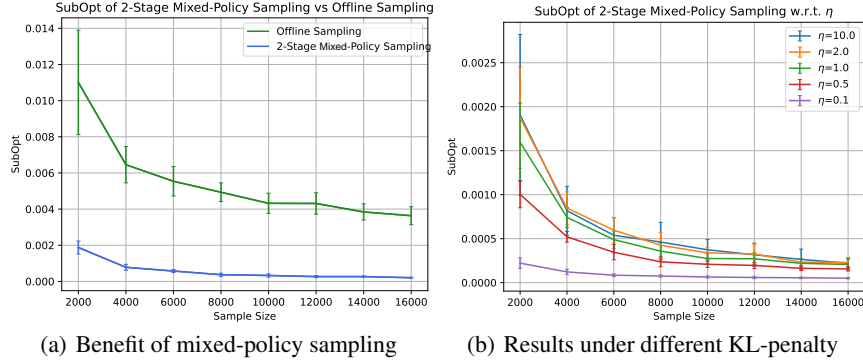


Figure 1: Suboptimality gap for KL-regularized contextual bandits.

1010 KL-regularized contextual bandits in Figure 1. All the trials are repeated for 10 times and plotted  
 1011 with the standard variation.

1012 We consider the case where context distribution  $d_0$  is a projected Gaussian distribution over  
 1013 the unit sphere and  $\mathcal{A}$  is a discrete set with  $|\mathcal{A}| = 5$ . We construct the reward functions as  
 1014  $R(\phi, x, a) = \langle x, \phi(a) \rangle$ , parameterized by a mapping  $\phi$  from  $\mathcal{A}$  to  $\mathbb{R}^{10}$ , and set the reference policy  
 1015  $\pi_0$  to be the uniform random policy. To generate  $\phi_*$ , we sample  $\phi_*(a)$  independently for each  $a \in \mathcal{A}$   
 1016 according to another projected gaussian distribution over the sphere with radius equal to 5. In Figure  
 1017 2(a), we compare the suboptimality gaps of mixed-policy sampling with  $m = n$  to those of offline  
 1018 sampling using  $\pi_0$  under the same sample sizes. The result indicates that the usage of mixed-policy  
 1019 sampling reduces the suboptimality gap by a large margin. In Figure 2(b), it is shown that the sample  
 1020 complexity is remarkably affected by the KL-regularization term, corroborating our sharp analysis  
 1021 for regularized RLHF.

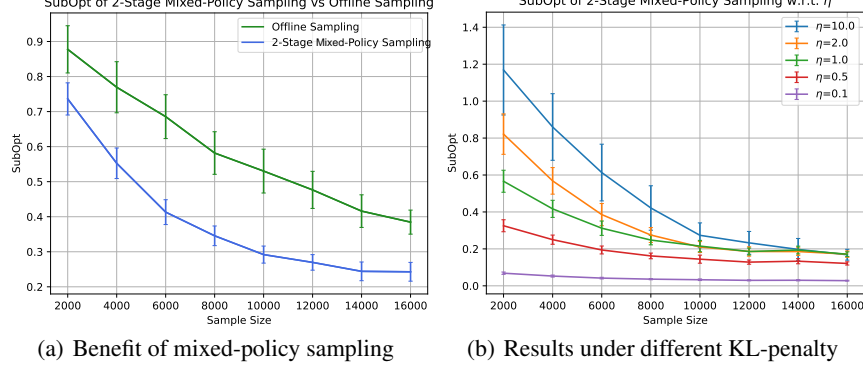


Figure 2: Suboptimality gap for reinforcement learning from preference feedback.

## E Proofs from Section 2

### E.1 Proof of Theorem 2.6

*Proof.* Consider a simple case when  $|\mathcal{X}| = M$  and  $|\mathcal{A}| = 2$ . We suppose that the context  $x$  is drawn uniformly from  $\mathcal{X}$  at the beginning of each round. Let  $\Theta$  be the set consisting of mappings from  $\mathcal{X}$  to  $\mathcal{A} = \{0, 1\}$ . For each  $\theta \in \Theta$ , we have  $R(\theta, x, a) = \begin{cases} 1/2 + c & \text{if } a = \theta(x), \\ 1/2 & \text{if } a \neq \theta(x), \end{cases}$  where  $c \in (0, 1/4)$

is a constant, and  $\theta(x)$  is the optimal action under context  $x$  when the model is  $\theta$ .

For any  $(\theta, x, a) \in \Theta \times \mathcal{X} \times \mathcal{A}$ , we assume the reward feedback  $r \sim \text{Bernoulli}(R(\theta, x, a))$  when the model is  $\theta$  and  $a$  is chosen under context  $x$ .

We pick a pair of model  $\theta_1, \theta_2$  in  $\Theta$ , such that  $\theta_1(x) = \begin{cases} \theta_2(x) & \text{if } x \neq x_0, \\ 1 - \theta_2(x) & \text{if } x = x_0. \end{cases}$

We denote by  $\mathbb{P}_\theta, \mathbb{E}_\theta$  the probability measure and expectation under the model  $\theta$ . Let  $N(x)$  be the number of times the context  $x$  is observed in the first  $T$  rounds for an  $x \in \mathcal{X}$ .

For two Bernoulli random variables  $X$  and  $Y$  with parameters  $1/2 - c$  and  $1/2 + c$ , we have

$$\begin{aligned} \text{KL}(X \| Y) &= (1/2 - c) \log \frac{1/2 - c}{1/2 + c} + (1/2 + c) \log \frac{1/2 + c}{1/2 - c} \\ &= 2c \cdot \log \frac{1 + 2c}{1 - 2c} \leq 16c^2 \end{aligned}$$

where the inequality follows from the fact that  $\log(1 + x) \leq x$  for  $x \geq 0$  and  $c \in (0, 1/4)$ .

Applying Pinsker's inequality (Lemma G.3), we have for all event  $A$  measurable with respect to the filtration generated by the observations,

$$|\mathbb{P}_{\theta_1}(A) - \mathbb{P}_{\theta_2}(A)| \leq \sqrt{8c^2 \mathbb{E}_{\theta_1}[N(x_0)]} = \sqrt{8c^2 T/M},$$

where the first inequality follows from the chain rule of KL divergence, and the fact that  $\mathbb{E}_{\theta_1}[N(x_0)] = T/M$ .

Set  $A$  to be the event that  $\pi_{\text{out}}(\theta_1(x_0)|x_0) > 1/2$ . Then we have

$$\mathbb{P}_{\theta_1}(\pi_{\text{out}}(\theta_1(x_0)|x_0) \leq 1/2) + \mathbb{P}_{\theta_2}(\pi_{\text{out}}(\theta_2(x_0)|x_0) \leq 1/2) \geq 1 - |\mathbb{P}_{\theta_1}(A) - \mathbb{P}_{\theta_2}(A)| \geq 1 - \sqrt{8c^2 T/M}.$$

If the model  $\theta$  is uniformly drawn from  $\Theta$ , then we have

$$\mathbb{E}_{\theta \sim \text{Unif}(\Theta)} \mathbb{P}_\theta(\pi_{\text{out}}(\theta(x_0)) \leq 1/2) \geq \frac{1}{2} - \sqrt{2c^2 T/M}$$

for an arbitrary  $x_0$ .

Then we consider the following suboptimality gap:

$$\begin{aligned} &\mathbb{E}_{\pi_{\theta_*}^\eta} \left[ R(\theta_*, x, a) - \frac{1}{\eta} \log \frac{\pi_{\theta_*}^\eta(a|x)}{\pi_0(a|x)} \right] - \mathbb{E}_{\pi_{\text{out}}} \left[ R(\theta_*, x, a) - \frac{1}{\eta} \log \frac{\pi_{\text{out}}(a|x)}{\pi_0(a|x)} \right] \\ &= \frac{1}{\eta} \mathbb{E}_{\pi_{\theta_*}^\eta} \left[ \log \frac{\pi_0(a|x) \cdot \exp(\eta R(\theta_*, x, a))}{\pi_{\theta_*}^\eta(a|x)} \right] - \frac{1}{\eta} \mathbb{E}_{\pi_{\text{out}}} \left[ \log \frac{\pi_0(a|x) \cdot \exp(\eta R(\theta_*, x, a))}{\pi_{\text{out}}(a|x)} \right] \end{aligned}$$

$$= \frac{1}{\eta} \mathbb{E}_{\pi_{\text{out}}} \left[ \log \frac{\pi_{\text{out}}(a|x)}{\pi^*(a|x)} \right],$$

1043 where the last equality follows from the fact that  $\pi_{\theta_*}^\eta \propto \pi_0(a|x) \cdot \exp(\eta R(\theta_*, x, a))$ . Note that  
 1044 we handle the difference in reward and the KL-divergence term together, which is distinct from the  
 1045 standard analysis of the lower bound for contextual bandits.

1046 To bound the suboptimality gap, we further have

$$\begin{aligned} & \mathbb{E}_{\theta \sim \text{Unif}(\Theta)} \mathbb{E}_{\pi_{\text{out}}} \left[ \log \frac{\pi_{\text{out}}(a|x)}{\pi^*(a|x)} \right] \\ &= \mathbb{E}_{\theta \sim \text{Unif}(\Theta)} \frac{1}{M} \sum_{x \in \mathcal{X}} \mathbb{E}_{a \sim \pi_{\text{out}}(\cdot|x)} \left[ \log \frac{\pi_{\text{out}}(a|x)}{\pi^*(a|x)} \right] \\ &\geq \mathbb{E}_{\theta \sim \text{Unif}(\Theta)} \frac{1}{M} \sum_{x \in \mathcal{X}} \mathbb{P}_{\theta}(\pi_{\text{out}}(\theta(x)) \leq 1/2) \cdot \left[ \frac{1}{2} \log \frac{1 + \exp(-\eta c)}{2} + \frac{1}{2} \log \frac{1 + \exp(\eta c)}{2} \right] \\ &\geq \left( \frac{1}{2} - \sqrt{2c^2 T/M} \right) \left[ \frac{1}{2} \log \frac{1 + \exp(-\eta c)}{2} + \frac{1}{2} \log \frac{1 + \exp(\eta c)}{2} \right], \end{aligned} \quad (\text{E.1})$$

1047 where the first inequality follows from the fact that  $\text{KL}(\pi_{\text{out}}(\cdot|x) \parallel \pi^*(\cdot|x)) \geq$   
 1048  $\text{KL}(\pi_{\text{unif}}(\cdot|x) \parallel \pi^*(\cdot|x))$  if  $\pi_{\text{out}}(\theta(x)) \leq 1/2$ . Here  $\pi_{\text{unif}}$  is the uniform distribution over  $\mathcal{A}$ .  
 1049 Note that

$$\begin{aligned} & \frac{d}{du} \left[ \frac{1}{2} \log \frac{1 + e^{-u}}{2} + \frac{1}{2} \log \frac{1 + e^u}{2} \right] \Big|_{u=0} = \frac{1}{2} \left[ \frac{1}{1 + \exp(-u)} - \frac{1}{1 + \exp(u)} \right] \Big|_{u=0} = 0, \\ & \frac{d^2}{du^2} \left[ \frac{1}{2} \log \frac{1 + e^{-u}}{2} + \frac{1}{2} \log \frac{1 + e^u}{2} \right] = \frac{\exp(u)}{[1 + \exp(u)]^2}. \end{aligned}$$

1050 Thus, applying Taylor's expansion on the right-hand side of (E.1), we have

$$\mathbb{E}_{\theta \sim \text{Unif}(\Theta)} \mathbb{E}_{\pi_{\text{out}}} \left[ \log \frac{\pi_{\text{out}}(a|x)}{\pi^*(a|x)} \right] \geq \frac{1}{2} \cdot \left( \frac{1}{2} - \sqrt{2c^2 T/M} \right) \eta^2 c^2 \cdot \frac{1}{3 + \exp(\eta c)},$$

1051 which follows from the Taylor's expansion where  $f(x) = f(0) + f'(0)x + \frac{1}{2}f''(z)x^2$  where  $z \in$   
 1052  $[0, \eta c]$ , and the fact that  $\frac{1}{2} \frac{e^z}{(1+e^z)^2} = \frac{1}{2} \frac{1}{e^{-z} + e^z + 2} \leq \frac{1}{2} \frac{1}{3 + e^{\eta c}}$ .

1053 When  $\epsilon < 1/64\eta$ , we can set  $c = 8\sqrt{\epsilon/\eta}$ . To achieve a suboptimality gap of  $\epsilon$ , we need to satisfy:

$$\frac{1}{2} \cdot \left( \frac{1}{2} - \sqrt{2c^2 T/M} \right) \eta^2 c^2 \cdot \frac{1}{3 + \exp(\eta c)} \leq \eta \epsilon,$$

1054 indicating that  $T \geq \frac{\eta M}{2048\epsilon} = \Omega(\frac{\eta M}{\epsilon})$ .

1055 When  $\epsilon \geq 1/64\eta$ , or equivalently,  $\eta \geq 1/64\epsilon$ , we employ a different lower bound for (E.1) as  
 1056 follows:

$$\begin{aligned} \frac{1}{2} \log \frac{1 + \exp(-\eta c)}{2} + \frac{1}{2} \log \frac{1 + \exp(\eta c)}{2} &= \frac{1}{2} \log \frac{2 + \exp(\eta c) + \exp(-\eta c)}{4} \\ &\geq \frac{1}{2} \cdot \frac{1}{2} \left( \log \frac{\exp(\eta c) + \exp(-\eta c)}{2} \right) \\ &\geq \frac{1}{4} (\eta c - \log 2), \end{aligned} \quad (\text{E.2})$$

1057 where the first inequality follows from Jensen's inequality.

1058 Substituting (E.2) into (E.1), we have

$$\epsilon \geq \frac{1}{\eta} \mathbb{E}_{\theta \sim \text{Unif}(\Theta)} \mathbb{E}_{\pi_{\text{out}}} \left[ \log \frac{\pi_{\text{out}}(a|x)}{\pi^*(a|x)} \right] \geq \frac{1}{4} \cdot \left( \frac{1}{2} - \sqrt{c^2 T/2M} \right) (\eta c - \log 2) \cdot \frac{1}{\eta}.$$

1059 Set  $c = 64\epsilon$ . Then we have  $T = \Omega(M/\epsilon^2)$ .

1060 Note that for the given  $\epsilon$ ,  $N_{\mathcal{R}}(\epsilon) = M$ , and since  $d_0$  is the uniform distribution over  $\mathcal{X}$ , we have  
 1061  $D^2 = O(M)$ , which completes the proof.

1062 □

## 1063 E.2 Proof of Theorem 2.8

1064 We start with the following lemma, which provides an on-policy generalization bound for the reward  
 1065 function. Due to the on-policy nature of the algorithm (i.e., the usage of intermediate  $\pi_{\theta_0}^\eta$ ), we can  
 1066 leverage the covering number of the reward function class  $\mathcal{R}$  to derive the generalization error. Since  
 1067 we are using a fixed policy  $\pi_{\theta_0}^\eta$  to sample in the second stage, we can derive the generalization error  
 1068 of the reward function as follows:

1069 **Lemma E.1** (Generalization error of reward function). For an arbitrary policy  $\pi$ , a set of context-  
 1070 action pairs  $\{(x_i, a_i)\}_{i=1}^n$  generated i.i.d. from  $\pi$ , and a distance threshold  $0 < \epsilon_c \leq B$ , we have  
 1071 with probability at least  $1 - \delta$ , for any pair of parameters  $\theta_1$  and  $\theta_2$ ,

$$\begin{aligned} & \mathbb{E}_\pi |R(\theta_1, x, a) - R(\theta_2, x, a)|^2 \\ & \leq \frac{2}{n} \sum_{i=1}^n |R(\theta_1, x_i, a_i) - R(\theta_2, x_i, a_i)|^2 + \frac{32B^2}{3n} \log(2N_{\mathcal{R}}(\epsilon_c)/\delta) + 10\epsilon_c B. \end{aligned}$$

1072 *Proof.* We first consider an  $\epsilon_c$ -net  $\mathcal{R}^c$  of the reward function class  $\mathcal{R}$  where  $\mathcal{R}^c = \{R(\theta, \cdot, \cdot) | \theta \in$   
 1073  $\Theta^c\}$  with size  $N_{\mathcal{R}}(\epsilon_c)$ . For any  $R(\theta, \cdot, \cdot) \in \mathcal{R}$ , there exists  $\theta^c$  such that  $\|R(\theta, \cdot, \cdot) - R(\theta^c, \cdot, \cdot)\|_\infty \leq$   
 1074  $\epsilon_c$ .

1075 By Lemma G.1, for each pair of  $\theta_1^c, \theta_2^c \in \Theta^c$  (corresponding to  $\theta_1, \theta_2$ ), we have with probability at  
 1076 least  $1 - \delta$ ,

$$\begin{aligned} & \left| \frac{1}{n} \sum_{i=1}^n (R(\theta_1^c, x_i, a_i) - R(\theta_2^c, x_i, a_i))^2 - \mathbb{E}_\pi |R(\theta_1^c, x, a) - R(\theta_2^c, x, a)|^2 \right| \\ & \leq \sqrt{\frac{2\text{Var}_\pi |R(\theta_1^c, x, a) - R(\theta_2^c, x, a)|^2}{n} \log(2/\delta)} + \frac{2}{3n} B^2 \log(2/\delta) \\ & \leq \sqrt{\frac{2B^2 \mathbb{E}_\pi |R(\theta_1^c, x, a) - R(\theta_2^c, x, a)|^2}{n} \log(2/\delta)} + \frac{2}{3n} B^2 \log(2/\delta) \end{aligned}$$

1077 where the second inequality follows from the fact that  $R(\theta_1^c, x, a), R(\theta_2^c, x, a) \leq B$ .

1078 Using union bound over all  $\theta_1^c, \theta_2^c \in \Theta^c$ , we have with probability at least  $1 - \delta$ , for all  $\theta_1^c, \theta_2^c \in \Theta^c$ ,

$$\begin{aligned} & \mathbb{E}_\pi |R(\theta_1^c, x, a) - R(\theta_2^c, x, a)|^2 - \frac{1}{n} \sum_{i=1}^n (R(\theta_1^c, x_i, a_i) - R(\theta_2^c, x_i, a_i))^2 \\ & \leq \sqrt{\frac{4B^2 \mathbb{E}_\pi |R(\theta_1^c, x, a) - R(\theta_2^c, x, a)|^2}{n} \log(2N_{\mathcal{R}}(\epsilon_c)/\delta)} + \frac{4B^2}{3n} \log(2N_{\mathcal{R}}(\epsilon_c)/\delta), \end{aligned}$$

1079 from which we further obtain the following inequality by Lemma G.2,

$$\mathbb{E}_\pi |R(\theta_1^c, x, a) - R(\theta_2^c, x, a)|^2 \leq \frac{2}{n} \sum_{i=1}^n (R(\theta_1^c, x_i, a_i) - R(\theta_2^c, x_i, a_i))^2 + \frac{32B^2}{3n} \log(2N_{\mathcal{R}}(\epsilon_c)/\delta). \quad (\text{E.3})$$

1080 Then we can complete the proof by the definition of  $\epsilon$ -net.  $\square$

1081 Next, we provide the following lemma, which gives an upper bound on the cumulative square error  
 1082 of the learned reward function.

1083 **Lemma E.2** (Confidence bound for reward function). For an arbitrary policy  $\pi$ , and a set of data  
 1084  $\{(x_i, a_i, r_i)\}_{i=1}^n$  generated i.i.d. from  $\pi$ , suppose that  $\hat{\theta}$  is the least squares estimator of  $\theta_*$ , i.e.,  
 1085  $\hat{\theta} = \arg \min_{\theta \in \Theta} \sum_{i=1}^n (R(\theta, x_i, a_i) - r_i)^2$ . Then for any threshold  $\epsilon_c > 0$ , with probability at least  
 1086  $1 - \delta$ , it holds that

$$\sum_{i=1}^n (R(\hat{\theta}, x_i, a_i) - R(\theta_*, x_i, a_i))^2 \leq 16B^2 \log(2N_{\mathcal{R}}(\epsilon_c)/\delta) + 4\epsilon_c nB.$$

1087 *Proof.* We have the following inequality for  $\sum_{i=1}^n (R(\hat{\theta}, x_i, a_i) - R(\theta_*, x_i, a_i))^2$ ,

$$\begin{aligned} & \sum_{i=1}^n (R(\hat{\theta}, x_i, a_i) - R(\theta_*, x_i, a_i))^2 \\ &= \sum_{i=1}^n (R(\hat{\theta}, x_i, a_i) - r_i)^2 - \sum_{i=1}^n (R(\theta_*, x_i, a_i) - r_i)^2 \\ & \quad + 2 \sum_{i=1}^n (R(\hat{\theta}, x_i, a_i) - R(\theta_*, x_i, a_i))(r_i - R(\theta_*, x_i, a_i)) \\ & \leq 2 \sum_{i=1}^n (R(\hat{\theta}, x_i, a_i) - R(\theta_*, x_i, a_i))(r_i - R(\theta_*, x_i, a_i)), \end{aligned}$$

1088 where the last inequality follows from the fact that  $\sum_{i=1}^n (R(\hat{\theta}, x_i, a_i) - r_i)^2 \leq$   
 1089  $\sum_{i=1}^n (R(\theta_*, x_i, a_i) - r_i)^2$ .

1090 We then consider an  $\epsilon_c$ -net  $\mathcal{R}^c$  of the reward function class  $\mathcal{R}$  where  $\mathcal{R}^c = \{R(\theta, \cdot, \cdot) | \theta \in \Theta^c\}$  with  
 1091 size  $N_{\mathcal{R}}(\epsilon_c)$ . For any  $R(\theta, \cdot, \cdot) \in \mathcal{R}$ , there exists  $\theta^c$  such that  $\|R(\theta, x, a) - R(\theta^c, x, a)\|_{\infty} \leq \epsilon_c$ .

1092 From Azuma-Hoeffding inequality, with probability at least  $1 - \delta$ , it holds for all  $\theta \in \Theta^c$  that

$$\begin{aligned} & \sum_{i=1}^n (R(\theta, x_i, a_i) - R(\theta_*, x_i, a_i))(r_i - R(\theta_*, x_i, a_i)) \\ & \leq \sqrt{2B^2 \sum_{i=1}^n (R(\theta, x_i, a_i) - R(\theta_*, x_i, a_i))^2 \log(2N_{\mathcal{R}}(\epsilon_c)/\delta)}. \end{aligned}$$

1093 Then we further have with probability at least  $1 - \delta$ , there exists  $\|R(\theta^c, \cdot, \cdot) - R(\hat{\theta}, \cdot, \cdot)\| \leq \epsilon_c$  such  
 1094 that

$$\begin{aligned} & \sum_{i=1}^n (R(\hat{\theta}, x_i, a_i) - R(\theta_*, x_i, a_i))(r_i - R(\theta_*, x_i, a_i)) \\ & \leq \sqrt{2B^2 \sum_{i=1}^n (R(\theta, x_i, a_i) - R(\theta_*, x_i, a_i))^2 \log(2N_{\mathcal{R}}(\epsilon_c)/\delta) + 2\epsilon_c nB}, \end{aligned}$$

1095 which implies that

$$\sum_{i=1}^n (R(\hat{\theta}, x_i, a_i) - R(\theta_*, x_i, a_i))^2 \leq 16B^2 \log(2N_{\mathcal{R}}(\epsilon_c)/\delta) + 4\epsilon_c nB \quad (\text{E.4})$$

1096 by Lemma G.2. □

1097 With the above lemmas, we are now ready to prove the following lemma that bounds the estimation  
 1098 error of the reward function  $R(\hat{\theta}, \cdot, \cdot)$  under the sampled policy  $\pi_{\hat{\theta}_0}^{\eta}$ .

1099 **Lemma E.3.** Let  $\hat{\theta}_0$  be the least squares estimator of the reward function based on the data  
 1100  $\{(x_i^0, a_i^0, r_i^0)\}_{i=1}^m$  generated from  $\pi_0$  as defined in Algorithm 1. Then for any threshold  $\epsilon_c > 0$ ,  
 1101 with probability at least  $1 - 2\delta$ , we have

$$\mathbb{E}_{\pi_{\hat{\theta}_0}^{\eta}} |R(\hat{\theta}, x, a) - R(\theta_*, x, a)|^2 \leq \frac{43B^2}{n} \log(2N_{\mathcal{R}}(\epsilon_c)/\delta) + 10\epsilon_c(1 + m/n)B.$$

1102 *Proof.* By Lemma E.1, we have with probability at least  $1 - \delta$ , the following upper bound holds for  
 1103  $\mathbb{E}_{\pi_{\hat{\theta}_0}^{\eta}} |R(\theta_1, x, a) - R(\theta_2, x, a)|^2$ ,

$$\begin{aligned} & \mathbb{E}_{\pi_{\hat{\theta}_0}^{\eta}} |R(\theta_1, x, a) - R(\theta_2, x, a)|^2 \\ & \leq \frac{2}{n} \sum_{i=1}^n |R(\theta_1, x_i, a_i) - R(\theta_2, x_i, a_i)|^2 + \frac{32B^2}{3n} \log(2N_{\mathcal{R}}(\epsilon_c)/\delta) + 10\epsilon_c B. \end{aligned} \quad (\text{E.5})$$

1104 By Lemma E.2, with probability at least  $1 - \delta$

$$\sum_{i=1}^n |R(\theta_*, x_i, a_i) - R(\hat{\theta}, x_i, a_i)|^2 \leq 16B^2 \log(2N_{\mathcal{R}}(\epsilon_c)/\delta) + 4\epsilon_c(n+m)B. \quad (\text{E.6})$$

1105 Then we can complete the proof using a union bound and substituting (E.6) into (E.5).  $\square$

1106 **Lemma E.4.** If  $m \geq 128\eta^2 D^2 B^2 \cdot \log(2N_{\mathcal{R}}(\epsilon_c)/\delta)$ , and there exists a positive constant  $c_{m,n} > 0$   
 1107 such that  $n = c_{m,n}m$  in Algorithm 1 and Assumption 2.5 holds, then by taking  $\epsilon_c \leq \min\{B, (8(1 +$   
 1108  $c_{m,n})B\eta^2 D^2)^{-1}\}$ , with probability at least  $1 - 3\delta$ , we have

$$\eta |R(\hat{\theta}_0, x, a) - R(\theta_*, x, a)| \leq 1, \quad \eta |R(\hat{\theta}, x, a) - R(\theta_*, x, a)| \leq 1$$

1109 for any pair  $(x, a) \in \mathcal{X} \times \mathcal{A}$  such that  $\pi_0(a|x) > 0$ .

1110 *Proof.* By Lemma E.1, with probability at least  $1 - \delta$ , for all  $\theta_1, \theta_2 \in \Theta$ , we have

$$\mathbb{E}_{\pi_0} |R(\theta_1, x, a) - R(\theta_2, x, a)|^2 \leq \frac{2}{m} \sum_{i=1}^m |R(\theta_1, x_i^0, a_i^0) - R(\theta_2, x_i^0, a_i^0)|^2 + \frac{32B^2}{3m} \log(2N_{\mathcal{R}}(\epsilon_c)/\delta).$$

1111 By Lemma E.2, with probability at least  $1 - \delta$ , we have

$$\sum_{i=1}^m |R(\hat{\theta}_0, x_i^0, a_i^0) - R(\theta_*, x_i^0, a_i^0)|^2 \leq 16B^2 \log(2N_{\mathcal{R}}(\epsilon_c)/\delta) + 4\epsilon_c m.$$

1112 Also, with probability at least  $1 - \delta$ , we have

$$\sum_{i=1}^m |R(\theta_*, x_i^0, a_i^0) - R(\hat{\theta}, x_i^0, a_i^0)|^2 \leq 16B^2 \log(2N_{\mathcal{R}}(\epsilon_c)/\delta) + 4\epsilon_c(m+n)B.$$

1113 Similar to the proof of Lemma E.3, we have if  $m \geq 128\eta^2 D^2 B^2 \cdot \log(2N_{\mathcal{R}}(\epsilon_c)/\delta)$ ,  $n = c_{m,n}m$ ,  
 1114 then with probability at least  $1 - 3\delta$ ,

$$\mathbb{E}_{\pi_0} |R(\theta_*, x, a) - R(\hat{\theta}_0, x, a)|^2 \leq 1/\eta^2 D^2, \quad \mathbb{E}_{\pi_0} |R(\theta_*, x, a) - R(\hat{\theta}, x, a)|^2 \leq 1/\eta^2 D^2.$$

1115 which implies that  $\eta |R(\hat{\theta}_0, x, a) - R(\theta_*, x, a)| \leq 1$  and  $\eta |R(\hat{\theta}, x, a) - R(\theta_*, x, a)| \leq 1$  for all  
 1116  $(x, a) \in \mathcal{X} \times \mathcal{A}$  such that  $\pi_0(a|x) > 0$ .  $\square$

1117 **Lemma E.5** (Restatement of Lemma 2.9). For any estimator  $\hat{\theta} \in \Theta$ , and the policy  $\pi_{\hat{\theta}}^{\eta}$  satisfying  
 1118 Definiton 2.4, we have

$$\begin{aligned} Q(\pi^*) - Q(\pi_{\hat{\theta}}^{\eta}) &= \eta \mathbb{E}_{x \sim d_0} \left[ \sum_{a \in \mathcal{A}} \pi_f^{\eta}(a|x) \Delta^2(x, a) - \sum_{a_1, a_2 \in \mathcal{A}} \pi_f^{\eta}(a_1|x) \pi_f^{\eta}(a_2|x) \Delta(x, a_1) \Delta(x, a_2) \right] \\ &\leq \eta \mathbb{E}_{x \sim d_0} \left[ \sum_{a \in \mathcal{A}} \pi_f^{\eta}(a|x) \Delta^2(x, a) \right], \end{aligned}$$

1119 where  $\Delta(x, a) = R(\hat{\theta}, x, a) - R(\theta_*, x, a)$ ,  $f(\cdot, \cdot) = \gamma R(\hat{\theta}, \cdot, \cdot) + (1 - \gamma)R(\theta_*, \cdot, \cdot)$  ( $\gamma \in$   
 1120  $(0, 1)$ ) the inequality uses the fact that second term on the right-hand side of the equality is  
 1121  $(\sum_{a \in \mathcal{A}} \pi_f^{\eta}(a|x) \Delta(x, a))^2 \geq 0$ .

1122 *Proof of Lemma 2.9.* We have

$$\begin{aligned} &\mathbb{E}_{\pi_{\theta_*}^{\eta}} \left[ R(\theta_*, x, a) - \frac{1}{\eta} \log \frac{\pi_{\theta_*}^{\eta}(a|x)}{\pi_0(a|x)} \right] - \mathbb{E}_{\pi_{\hat{\theta}}^{\eta}} \left[ R(\theta_*, x, a) - \frac{1}{\eta} \log \frac{\pi_{\hat{\theta}}^{\eta}(a|x)}{\pi_0(a|x)} \right] \\ &= \frac{1}{\eta} \mathbb{E}_{\pi_{\theta_*}^{\eta}} \left[ \log \frac{\pi_0(a|x) \cdot \exp(\eta R(\theta_*, x, a))}{\pi_{\theta_*}^{\eta}(a|x)} \right] - \frac{1}{\eta} \mathbb{E}_{\pi_{\hat{\theta}}^{\eta}} \left[ \log \frac{\pi_0(a|x) \cdot \exp(\eta R(\theta_*, x, a))}{\pi_{\hat{\theta}}^{\eta}(a|x)} \right] \\ &= \frac{1}{\eta} \mathbb{E}_{x \sim d_0} [\log Z_{\theta_*}^{\eta}(x)] - \frac{1}{\eta} \mathbb{E}_{x \sim d_0} [\log Z_{\hat{\theta}}^{\eta}(x)] - \mathbb{E}_{x \sim d_0} \left[ \sum_{a \in \mathcal{A}} \pi_{\hat{\theta}}^{\eta}(a|x) \cdot (R(\theta_*, x, a) - R(\hat{\theta}, x, a)) \right], \end{aligned}$$



1123 where the first equality follows from the definition of the KL-divergence, the second equality follows  
 1124 from Lemma G.5.

1125 For an arbitrary reward function  $f : \mathcal{X} \times \mathcal{A} \rightarrow \mathbb{R}$ , let  $\Delta_f(x, a) = f(x, a) - R(\theta_*, x, a)$ . Consider  
 1126 the following first derivative of  $J(f) = \log Z_f^\eta(x) - \eta \sum_{a \in \mathcal{A}} \pi_f^\eta(a|x) \cdot \Delta_f(x, a)$ , where  $Z_f^\eta(x) =$   
 1127  $\sum_{a \in \mathcal{A}} \pi_0(a|x) \cdot \exp(\eta \cdot f(x, a))$  and  $\pi_f^\eta(a|x) \propto \pi_0(a|x) \cdot \exp(\eta \cdot f(x, a))$ .

$$\begin{aligned} & \frac{\partial}{\partial \Delta_f(x, a)} \left[ \log Z_f^\eta(x) - \eta \sum_{a \in \mathcal{A}} \pi_f^\eta(a|x) \cdot \Delta_f(x, a) \right] \\ &= \frac{1}{Z_f^\eta(x)} \cdot \pi_0(a|x) \exp(\eta \cdot f(x, a)) \cdot \eta - \eta \cdot \pi_f^\eta(a|x) \\ & \quad - \eta \cdot \Delta_f(x, a) \cdot \frac{\pi_0(a|x) \cdot \exp(\eta \cdot f(x, a))}{Z_f^\eta(x)} \cdot \eta + \eta \cdot \Delta_f(x, a) \cdot \frac{[\pi_0(a|x) \cdot \exp(\eta \cdot f(x, a))]^2}{[Z_f^\eta(x)]^2} \cdot \eta \\ & \quad + \eta \sum_{a' \in \mathcal{A} \setminus \{a\}} \frac{\pi_0(a'|x) \cdot \exp(\eta \cdot f(x, a'))}{Z_f^\eta(x)} \cdot \eta \cdot \Delta_f(x, a') \cdot \frac{\pi_0(a|x) \cdot \exp(\eta \cdot f(x, a))}{Z_f^\eta(x)} \\ &= -\eta^2 \pi_f^\eta(a|x) \Delta_f(x, a) + \eta^2 [\pi_f^\eta(a|x)]^2 \cdot \Delta_f(x, a) + \eta^2 \sum_{a' \in \mathcal{A} \setminus \{a\}} \pi_f^\eta(a'|x) \pi_f^\eta(a|x) \Delta_f(x, a') \end{aligned}$$

1128 where the first equality is derived by taking the derivative of  $\log Z_f^\eta(x)$  and the second term with  
 1129 respect to  $\Delta_f$ . Therefore, by the Mean Value Theorem, there exists an  $f(\cdot, \cdot) = \gamma R(\hat{\theta}, \cdot, \cdot) + (1 -$   
 1130  $\gamma) R(\theta_*, \cdot, \cdot)$  for some  $\gamma \in [0, 1]$  such that

$$\begin{aligned} & \mathbb{E}_{x \sim d_0} [J(R(\hat{\theta}, \cdot, \cdot)) - J(R(\theta_*, \cdot, \cdot))] = \frac{1}{\eta} \mathbb{E}_{x \sim d_0} \left[ -\eta^2 \sum_{a \in \mathcal{A}} \pi_f^\eta(a|x) \cdot \gamma \cdot (R(\hat{\theta}, x, a) - R(\theta_*, x, a))^2 \right] \\ & \quad + \frac{1}{\eta} \mathbb{E}_{x \sim d_0} \left[ \gamma \eta^2 \sum_{a_1 \in \mathcal{A}} \sum_{a_2 \in \mathcal{A}} \pi_f^\eta(a_1|x) \pi_f^\eta(a_2|x) (R(\hat{\theta}, x, a_1) - R(\theta_*, x, a_1)) (R(\hat{\theta}, x, a_2) - R(\theta_*, x, a_2)) \right] \\ & \geq -\eta \cdot \mathbb{E}_{\pi_f^\eta} [(R(\hat{\theta}, x, a) - R(\theta_*, x, a))^2] \end{aligned}$$

1131 where the last inequality holds since

$$\begin{aligned} & \sum_{a_1 \in \mathcal{A}} \sum_{a_2 \in \mathcal{A}} \pi_f^\eta(a_1|x) \pi_f^\eta(a_2|x) (R(\hat{\theta}, x, a_1) - R(\theta_*, x, a_1)) (R(\hat{\theta}, x, a_2) - R(\theta_*, x, a_2)) \\ &= [\mathbb{E}_{a \sim \pi_f^\eta(\cdot|x)} [R(\hat{\theta}, x, a) - R(\theta_*, x, a)]]^2 \geq 0. \end{aligned}$$

1132

□

1133 Now, we are ready to prove the theorem.

1134 *Proof of Theorem 2.8.* By Lemma 2.9, we have

$$Q(\pi^*) - Q(\pi_\theta^\eta) \leq \eta \mathbb{E}_{x \sim d_0} \left[ \sum_{a \in \mathcal{A}} \pi_f^\eta(a|x) (R(\hat{\theta}, x, a) - R(\theta_*, x, a))^2 \right].$$

1135 By Lemma E.4, if  $m \geq 128\eta^2 D^2 B^2 \cdot \log(2N_{\mathcal{R}}(\epsilon_c)/\delta)$ , for any  $(x, a) \in \mathcal{X} \times \mathcal{A}$  such that  $\pi_0(a|x) >$   
 1136 0, it holds that

$$\eta |R(\hat{\theta}_0, x, a) - R(\theta_*, x, a)| \leq 1, \quad \eta |R(\hat{\theta}, x, a) - R(\theta_*, x, a)| \leq 1,$$

which means that for any  $(x, a) \in \mathcal{X} \times \mathcal{A}$

$$\frac{\pi_f^\eta(a|x)}{\pi_{\theta_0}^\eta(a|x)} \leq e^4.$$

1137 Let  $\epsilon_c = \min\{\frac{\epsilon}{(1+c_{m,n}^{-1})B}, \frac{1}{8(1+c_{m,n})B\eta^2 D^2}, B\}$ . By Lemma E.3, if  $m \geq 128\eta^2 D^2 B^2 \cdot$   
 1138  $\log(2N_{\mathcal{R}}(\epsilon_c)/\delta)$  and  $n \geq \eta/\epsilon \cdot B^2 \log(N_{\mathcal{R}}(\epsilon_c)/\delta)$  and  $n = c_{m,n} m$  then with high probability  
 1139 the output policy  $\pi_\theta^\eta$  is  $O(\epsilon)$  optimal. □

### 1140 E.3 Proof of Corollary 2.13

1141 *Proof of Corollary 2.13.* The proof follows the same lines as Theorem 3.7 by replacing the data  
1142 coverage condition with the local-coverage condition. It still holds that

$$Q(\pi^*) - Q(\pi_{\hat{\theta}_0}^\eta) \leq \eta \cdot \mathbb{E}_{\pi_f^\eta} [(R(\hat{\theta}_0, x, a) - R(\theta_*, x, a))^2]$$

1143 where  $\pi_f^\eta(a|x) \propto \pi_0(a|x) \cdot \exp(\eta \cdot f(x, a))$  and  $f(\cdot, \cdot) = \gamma R(\hat{\theta}_0, \cdot, \cdot) + (1 - \gamma)R(\theta_*, \cdot, \cdot)$  for some  
1144  $\gamma \in (0, 1)$ . Thus, We have  $\text{KL}(\pi_f^\eta(a|x) \parallel \pi_0) \leq 2\eta B$ , which further implies that

$$Q(\pi^*) - Q(\pi_{\hat{\theta}}^\eta) \leq \eta \cdot C_{\rho_{\text{KL}}} \cdot O\left(\frac{1}{n} B \log(N_{\mathcal{R}}(\epsilon_c)/\delta) + B(1 + c_{m,n}^{-1})\epsilon_c\right)$$

1145 by Lemma F.4. Then we can conclude by substituting the value of  $m$  into the suboptimality gap.  $\square$

## 1146 F Proofs from Section 3

### 1147 F.1 Proof of Theorem 3.6

1148 *Proof of Theorem 3.6.* The proof follows a similar construction as the one for Theorem 2.6. Con-  
1149 sider a simple case when  $|\mathcal{X}| = M$  and  $|\mathcal{A}| = 2$ . We suppose that the context  $x$  is drawn uni-  
1150 formly from  $\mathcal{X}$  at the beginning of each round. Let  $\Theta$  be the set consisting of mappings from  $\mathcal{X}$   
1151 to  $\mathcal{A} = \{0, 1\}$ . For each  $\theta \in \Theta$ , we have  $R(\theta, x, a) = \begin{cases} c & \text{if } a = \theta(x), \\ 0 & \text{if } a \neq \theta(x), \end{cases}$  where  $c \in (0, 1/4)$  is a  
1152 constant, and  $\theta(x)$  is the optimal action under context  $x$  when the model is  $\theta$ .

1153 We pick a pair of model  $\theta_1, \theta_2$  in  $\Theta$ , such that  $\theta_1(x) = \begin{cases} \theta_2(x) & \text{if } x \neq x_0, \\ 1 - \theta_2(x) & \text{if } x = x_0. \end{cases}$

1154 We denote by  $\mathbb{P}_\theta, \mathbb{E}_\theta$  the probability measure and expectation under the model  $\theta$ .

1155 We have the following upper bound for two Bernoulli distribution  $y_1 \sim \text{Bernoulli}(\sigma(c))$  and  $y_2 \sim$   
1156  $\text{Bernoulli}(\sigma(-c))$  with  $\sigma(x) = 1/(1 + \exp(-x))$ :

$$\begin{aligned} \sigma(c) \log \frac{\sigma(c)}{\sigma(-c)} + \sigma(-c) \log \frac{\sigma(-c)}{\sigma(c)} &= 2 \cdot \left(\frac{1}{2} - \sigma(-c)\right) \log \frac{\sigma(c)}{\sigma(-c)} \\ &= \frac{1 - e^{-c}}{1 + e^{-c}} \cdot \log \frac{1 + e^c}{1 + e^{-c}} \\ &\leq \frac{1 - e^{-c}}{1 + e^{-c}} \cdot \frac{e^c - e^{-c}}{1 + e^{-c}} \\ &\leq \frac{c \cdot (c + e^{\frac{1}{4}}c)}{[(1 + e^{-\frac{1}{4}})]^2} \leq c^2, \end{aligned}$$

1157 where the first equality follows from the fact that  $\sigma(-c) = 1 - \sigma(c)$ , the first inequality holds since  
1158  $\log x \leq x - 1$ , and the second inequality holds since  $c \leq 1/4$ .

1159 Applying Pinsker's inequality (Lemma G.3), we have for all event  $A$  measurable with respect to the  
1160 filtration generated by the observations,

$$|\mathbb{P}_{\theta_1}(A) - \mathbb{P}_{\theta_2}(A)| \leq \sqrt{c^2 \mathbb{E}_{\theta_1}[N(x_0)]} = \sqrt{c^2 T/M},$$

1161 where the first inequality follows from the chain rule of KL divergence, and the fact that  
1162  $\mathbb{E}_{\theta_1}[N(x_0)] = T/M$ .

1163 Set  $A$  to be the event that  $\pi_{\text{out}}(\theta_1(x_0)|x_0) > 1/2$ . Then we have

$$\mathbb{P}_{\theta_1}(\pi_{\text{out}}(\theta_1(x_0)|x_0) \leq 1/2) + \mathbb{P}_{\theta_2}(\pi_{\text{out}}(\theta_2(x_0)|x_0) \leq 1/2) \geq 1 - |\mathbb{P}_{\theta_1}(A) - \mathbb{P}_{\theta_2}(A)| \geq 1 - \sqrt{c^2 T/M}.$$

1164 If the model  $\theta$  is uniformly drawn from  $\Theta$ , then we have

$$\mathbb{E}_{\theta \sim \text{Unif}(\Theta)} \mathbb{P}_\theta(\pi_{\text{out}}(\theta(x_0)) \leq 1/2) \geq \frac{1}{2} - \sqrt{c^2 T/4M}$$

1165 for an arbitrary  $x_0$ .

1166 Then we consider the following suboptimality gap:

$$\begin{aligned}
& \mathbb{E}_{\pi_{\theta_*}^\eta} \left[ R(\theta_*, x, a) - \frac{1}{\eta} \log \frac{\pi_{\theta_*}^\eta(a|x)}{\pi_0(a|x)} \right] - \mathbb{E}_{\pi_{\text{out}}} \left[ R(\theta_*, x, a) - \frac{1}{\eta} \log \frac{\pi_{\text{out}}(a|x)}{\pi_0(a|x)} \right] \\
&= \frac{1}{\eta} \mathbb{E}_{\pi_{\theta_*}^\eta} \left[ \log \frac{\pi_0(a|x) \cdot \exp(\eta R(\theta_*, x, a))}{\pi_{\theta_*}^\eta(a|x)} \right] - \frac{1}{\eta} \mathbb{E}_{\pi_{\text{out}}} \left[ \log \frac{\pi_0(a|x) \cdot \exp(\eta R(\theta_*, x, a))}{\pi_{\text{out}}(a|x)} \right] \\
&= \frac{1}{\eta} \mathbb{E}_{\pi_{\text{out}}} \left[ \log \frac{\pi_{\text{out}}(a|x)}{\pi^*(a|x)} \right],
\end{aligned}$$

1167 where the last equality follows from the fact that  $\pi_{\theta_*}^\eta \propto \pi_0(a|x) \cdot \exp(\eta R(\theta_*, x, a))$ . To bound the  
1168 suboptimality gap, we further have

$$\begin{aligned}
& \mathbb{E}_{\theta \sim \text{Unif}(\Theta)} \mathbb{E}_{\pi_{\text{out}}} \left[ \log \frac{\pi_{\text{out}}(a|x)}{\pi^*(a|x)} \right] \\
&= \mathbb{E}_{\theta \sim \text{Unif}(\Theta)} \frac{1}{M} \sum_{x \in \mathcal{X}} \mathbb{E}_{a \sim \pi_{\text{out}}(\cdot|x)} \left[ \log \frac{\pi_{\text{out}}(a|x)}{\pi^*(a|x)} \right] \\
&\geq \mathbb{E}_{\theta \sim \text{Unif}(\Theta)} \frac{1}{M} \sum_{x \in \mathcal{X}} \mathbb{P}_\theta(\pi_{\text{out}}(\theta(x)) \leq 1/2) \cdot \left[ \frac{1}{2} \log \frac{1 + \exp(-\eta c)}{2} + \frac{1}{2} \log \frac{1 + \exp(\eta c)}{2} \right] \\
&\geq \left( \frac{1}{2} - \sqrt{c^2 T / 4M} \right) \left[ \frac{1}{2} \log \frac{1 + \exp(-\eta c)}{2} + \frac{1}{2} \log \frac{1 + \exp(\eta c)}{2} \right] \tag{F.1}
\end{aligned}$$

1169 Note that

$$\begin{aligned}
& \frac{d}{du} \left[ \frac{1}{2} \log \frac{1 + e^{-u}}{2} + \frac{1}{2} \log \frac{1 + e^u}{2} \right] \Big|_{u=0} = \frac{1}{2} \left[ \frac{1}{1 + \exp(-u)} - \frac{1}{1 + \exp(u)} \right] \Big|_{u=0} = 0, \\
& \frac{d^2}{du^2} \left[ \frac{1}{2} \log \frac{1 + e^{-u}}{2} + \frac{1}{2} \log \frac{1 + e^u}{2} \right] = \frac{\exp(u)}{[1 + \exp(u)]^2}.
\end{aligned}$$

1170 Thus, applying Taylor's expansion on the right-hand side of (F.1), we have

$$\mathbb{E}_{\theta \sim \text{Unif}(\Theta)} \mathbb{E}_{\pi_{\text{out}}} \left[ \log \frac{\pi_{\text{out}}(a|x)}{\pi^*(a|x)} \right] \geq \frac{1}{2} \cdot \left( \frac{1}{2} - \sqrt{c^2 T / 4M} \right) \eta^2 c^2 \cdot \frac{1}{3 + \exp(\eta c)}$$

1171 When  $\epsilon < 1/64\eta$ , we can set  $c = 8\sqrt{\epsilon/\eta}$ . To achieve a suboptimality gap of  $\epsilon$ , we need to satisfy:

$$\frac{1}{2} \cdot \left( \frac{1}{2} - \sqrt{c^2 T / 4M} \right) \eta^2 c^2 \cdot \frac{1}{3 + \exp(\eta c)} \leq \eta \epsilon,$$

1172 indicating that  $T \geq \frac{\eta M}{512\epsilon} = \Omega\left(\frac{\eta M}{\epsilon}\right)$ .

1173 When  $\epsilon \geq 1/64\eta$ , or equivalently,  $\eta \geq 1/64\epsilon$ , we employ a different lower bound for (E.1) as  
1174 follows:

$$\begin{aligned}
\frac{1}{2} \log \frac{1 + \exp(-\eta c)}{2} + \frac{1}{2} \log \frac{1 + \exp(\eta c)}{2} &= \frac{1}{2} \log \frac{2 + \exp(\eta c) + \exp(-\eta c)}{4} \\
&\geq \frac{1}{2} \cdot \frac{1}{2} \left( \log \frac{\exp(\eta c) + \exp(-\eta c)}{2} \right) \\
&\geq \frac{1}{4} (\eta c - \log 2), \tag{F.2}
\end{aligned}$$

1175 where the first inequality follows from Jensen's inequality. Substituting (F.2) into (F.1), we have

$$\epsilon \geq \frac{1}{\eta} \mathbb{E}_{\theta \sim \text{Unif}(\Theta)} \mathbb{E}_{\pi_{\text{out}}} \left[ \log \frac{\pi_{\text{out}}(a|x)}{\pi^*(a|x)} \right] \geq \frac{1}{4} \cdot \left( \frac{1}{2} - \sqrt{c^2 T / 4M} \right) (\eta c - \log 2) \cdot \frac{1}{\eta}.$$

1176 Set  $c = 64\epsilon$ . Then we have  $T = \Omega(M/\epsilon^2)$ .

1177

□

1178 **F.2 Proof of Theorem 3.7**

1179 First, we provide the following lemma for the connection between the likelihood loss and the reward  
1180 difference, which is a key step to upper bound the reward difference between  $\hat{\theta}$  and  $\theta_*$ .

1181 **Lemma F.1.** For an arbitrary policy  $\pi$ , and a set of context-action pairs  $\{(x_i, a_i^1, a_i^2, y_i)\}_{i=1}^n$  gener-  
1182 ated i.i.d. from the BT model and  $\pi$ , we have with probability at least  $1 - \delta$ , for any  $s \leq n$ ,

$$\begin{aligned} & \frac{1}{2} \sum_{i=1}^s \mathcal{L}(\theta | x_i, a_i^1, a_i^2, y_i) - \mathcal{L}(\theta_* | x_i, a_i^1, a_i^2, y_i) \\ & \leq \log(1/\delta) - \frac{1}{8} e^{-B} \sum_{i=1}^s ([R(\theta, x_i, a_i^2) - R(\theta, x_i, a_i^1)] - [R(\theta_*, x_i, a_i^2) - R(\theta_*, x_i, a_i^1)])^2 \end{aligned}$$

*Proof.* Applying Lemma G.4 to the sequence

$$\left\{ -\frac{1}{2} y_i \cdot \log \frac{\sigma(R(\theta_*, x_i, a_i^1) - R(\theta_*, x_i, a_i^2))}{\sigma(R(\theta, x_i, a_i^1) - R(\theta, x_i, a_i^2))} - \frac{1}{2} (1 - y_i) \cdot \log \frac{\sigma(R(\theta_*, x_i, a_i^2) - R(\theta_*, x_i, a_i^1))}{\sigma(R(\theta, x_i, a_i^2) - R(\theta, x_i, a_i^1))} \right\}_{i=1}^n,$$

1183 We have with probability at least  $1 - \delta$ , for all  $s \leq n$ ,

$$\begin{aligned} & \frac{1}{2} \sum_{i=1}^s \mathcal{L}(\theta | x_i, a_i^1, a_i^2, y_i) - \mathcal{L}(\theta_* | x_i, a_i^1, a_i^2, y_i) \\ & \leq \log(1/\delta) + \sum_{i=1}^s \log \left( \sqrt{\sigma(R(\theta_*, x_i, a_i^2) - R(\theta_*, x_i, a_i^1)) \cdot \sigma(R(\theta, x_i, a_i^2) - R(\theta, x_i, a_i^1))} \right. \\ & \quad \left. + \sqrt{\sigma(R(\theta_*, x_i, a_i^1) - R(\theta_*, x_i, a_i^2)) \cdot \sigma(R(\theta, x_i, a_i^1) - R(\theta, x_i, a_i^2))} \right) \\ & \leq \log(1/\delta) + \sum_{i=1}^s \left( \sqrt{\sigma(R(\theta_*, x_i, a_i^2) - R(\theta_*, x_i, a_i^1)) \cdot \sigma(R(\theta, x_i, a_i^2) - R(\theta, x_i, a_i^1))} \right. \\ & \quad \left. + \sqrt{\sigma(R(\theta_*, x_i, a_i^1) - R(\theta_*, x_i, a_i^2)) \cdot \sigma(R(\theta, x_i, a_i^1) - R(\theta, x_i, a_i^2))} - 1 \right) \\ & = \log(1/\delta) - \frac{1}{2} \sum_{i=1}^s \left( \sqrt{\sigma(R(\theta_*, x_i, a_i^2) - R(\theta_*, x_i, a_i^1))} - \sqrt{\sigma(R(\theta, x_i, a_i^2) - R(\theta, x_i, a_i^1))} \right)^2 \\ & \quad - \frac{1}{2} \sum_{i=1}^s \left( \sqrt{\sigma(R(\theta_*, x_i, a_i^1) - R(\theta_*, x_i, a_i^2))} - \sqrt{\sigma(R(\theta, x_i, a_i^1) - R(\theta, x_i, a_i^2))} \right)^2 \\ & \leq \log(1/\delta) - \frac{1}{8} \sum_{i=1}^s (\sigma(R(\theta_*, x_i, a_i^2) - R(\theta_*, x_i, a_i^1)) - \sigma(R(\theta, x_i, a_i^2) - R(\theta, x_i, a_i^1)))^2 \\ & \leq \log(1/\delta) - \frac{1}{8} e^{-B} \sum_{i=1}^s ([R(\theta, x_i, a_i^2) - R(\theta, x_i, a_i^1)] - [R(\theta_*, x_i, a_i^2) - R(\theta_*, x_i, a_i^1)])^2, \end{aligned}$$

1184 where the second inequality holds due to  $\log(1 + r) \leq r$  for  $r > -1$ , the equality follows from the  
1185 fact that  $\sigma(r) + \sigma(-r) = 1$  and the last inequality holds since  $\sigma'(r) = \sigma(r) \cdot (1 - \sigma(r)) \geq e^{-B}$  for  
1186 all  $r \in [-B, B]$ .  $\square$

1187 To further control the error bound for the reward function with the help of Lemma F.1, we introduce  
1188 the following lemma to show that the likelihood function class  $\mathcal{L}$  can be well-covered by the  $\epsilon$ -net  
1189 of the reward function class  $\mathcal{R}$ .

1190 **Lemma F.2** (Covering number of  $\mathcal{L}$ ). For any  $\epsilon_c > 0$ , consider an  $\epsilon_c$ -net  $\mathcal{R}^c = \{R(\theta, \cdot, \cdot) | \theta \in \Theta^c\}$   
1191 for the reward function class  $\mathcal{R}$  with size  $N_{\mathcal{R}}(\epsilon_c)$ . Then for any  $\theta \in \Theta$ , there exists  $\theta^c \in \Theta^c$  such  
1192 that for any  $s \in [n]$ ,

$$\sum_{i=1}^s \mathcal{L}(\theta | x_i, a_i^1, a_i^2, y_i) \leq \sum_{i=1}^s \mathcal{L}(\theta^c | x_i, a_i^1, a_i^2, y_i) + 2s\epsilon_c.$$

1193 *Proof.* For any  $r \in \mathbb{R}$ , we have

$$\frac{d \log(\sigma(r))}{dr} = \frac{1}{\sigma(r)} \cdot \sigma(r) \cdot (1 - \sigma(r)) = 1 - \sigma(r) \in (0, 1).$$

1194 Thus, for any  $\theta \in \Theta$ ,  $x \in \mathcal{X}$ ,  $a^1, a^2 \in \mathcal{A}$  and  $y \in \{0, 1\}$ , there exists  $\theta^c \in \Theta^c$  such that

$$\begin{aligned} & |\mathcal{L}(\theta|x, a^1, a^2, y) - \mathcal{L}(\theta^c|x, a^1, a^2, y)| \\ & \leq |[R(\theta, x, a^1) - R(\theta, x, a^2)] - [R(\theta^c, x, a^1) - R(\theta^c, x, a^2)]| = 2\epsilon_c. \end{aligned}$$

1195

□

1196 With the above two lemmas, we are now ready to provide the confidence bound for the MLE esti-  
1197 mator of the reward function.

1198 **Lemma F.3.** Consider a set of context-action pairs  $\{(x_i, a_i^1, a_i^2, y_i)\}_{i=1}^n$  where labels  $\{y_i\}_{i=1}^n$  are  
1199 generated independently from the BT model. Suppose that  $\hat{\theta}$  is the MLE estimator as defined in  
1200 Definition 3.3. We have with probability at least  $1 - \delta$ ,

$$\sum_{i=1}^n ([R(\hat{\theta}, x_i, a_i^2) - R(\hat{\theta}, x_i, a_i^1)] - [R(\theta_*, x_i, a_i^2) - R(\theta_*, x_i, a_i^1)])^2 \leq O(e^B \log(N_{\mathcal{R}}(\epsilon_c)/\delta) + e^B n \epsilon_c).$$

1201 *Proof.* By Lemmas F.1 and F.2, we have with probability at least  $1 - \delta$ , for any  $\theta \in \Theta$ ,

$$\begin{aligned} & \frac{1}{2} \sum_{i=1}^n \mathcal{L}(\theta|x_i, a_i^1, a_i^2, y_i) - \mathcal{L}(\theta_*|x_i, a_i^1, a_i^2, y_i) \\ & \leq \log(N_{\mathcal{R}}(\epsilon_c)/\delta) - \frac{1}{8} e^{-B} \sum_{i=1}^n ([R(\theta, x_i, a_i^2) - R(\theta, x_i, a_i^1)] - [R(\theta_*, x_i, a_i^2) - R(\theta_*, x_i, a_i^1)])^2 + O(n \epsilon_c). \end{aligned}$$

1202 Since  $\hat{\theta}$  is the MLE estimator, we have  $\sum_{i=1}^n \mathcal{L}(\hat{\theta}|x_i, a_i^1, a_i^2, y_i) - \mathcal{L}(\theta_*|x_i, a_i^1, a_i^2, y_i) \geq 0$ , which  
1203 further implies

$$\begin{aligned} 0 & \leq \log(N_{\mathcal{R}}(\epsilon_c)/\delta) - \frac{1}{8} e^{-B} \sum_{i=1}^n ([R(\hat{\theta}, x_i, a_i^2) - R(\hat{\theta}, x_i, a_i^1)] - [R(\theta_*, x_i, a_i^2) - R(\theta_*, x_i, a_i^1)])^2 \\ & \quad + O(n \epsilon_c). \end{aligned}$$

1204 Then we have

$$\sum_{i=1}^n ([R(\hat{\theta}, x_i, a_i^2) - R(\hat{\theta}, x_i, a_i^1)] - [R(\theta_*, x_i, a_i^2) - R(\theta_*, x_i, a_i^1)])^2 \leq O(e^B \log(N_{\mathcal{R}}(\epsilon_c)/\delta) + e^B n \epsilon_c).$$

1205

□

1206 Finally, we provide the on-policy confidence bound for the squared reward difference between the  
1207 MLE estimator  $\hat{\theta}$  and the optimal reward function  $\theta_*$ .

1208 **Lemma F.4.** Consider an arbitrary policy  $\pi$ , and a set of context-action pairs  $\{(x_i, a_i^1, a_i^2, y_i)\}_{i=1}^n$   
1209 generated i.i.d. from the BT model and  $\pi$ . Suppose that  $\hat{\theta}$  is the MLE estimator. We have with  
1210 probability at least  $1 - 2\delta$ , there exists a mapping  $b : \mathcal{X} \rightarrow \mathbb{R}$  such that

$$\mathbb{E}_{\pi} [(R(\hat{\theta}, x, a) - R(\theta_*, x, a) - b(x))^2] \leq O\left(\frac{1}{n} e^B \log(N_{\mathcal{R}}(\epsilon_c)/\delta) + e^B \epsilon_c\right).$$

1211 *Proof.* By Lemma F.3, we have with probability at least  $1 - \delta$ ,

$$\sum_{i=1}^n ([R(\hat{\theta}, x_i, a_i^2) - R(\hat{\theta}, x_i, a_i^1)] - [R(\theta_*, x_i, a_i^2) - R(\theta_*, x_i, a_i^1)])^2 \leq O(e^B \log(N_{\mathcal{R}}(\epsilon_c)/\delta) + e^B n \epsilon_c).$$

1212 We consider an  $\epsilon_c$ -net  $\mathcal{R}^c = \{R(\theta, \cdot, \cdot) | \theta \in \Theta^c\}$  for the reward function class  $\mathcal{R}$  with size  $N_{\mathcal{R}}(\epsilon_c)$ .

1213 For any  $R(\theta, \cdot, \cdot)$ , there exists  $R(\theta^c, \cdot, \cdot)$  such that

$$|R(\theta, x, a) - R(\theta^c, x, a)| \leq O(\epsilon_c)$$

1214 for all  $x \in \mathcal{X}, a \in \mathcal{A}$ .

1215 Applying Lemma G.1, with probability at least  $1 - \delta$ , we have

$$\begin{aligned} & \sum_{i=1}^n ([R(\theta^c, x_i, a_i^2) - R(\theta^c, x_i, a_i^1)] - [R(\theta_*, x_i, a_i^2) - R(\theta_*, x_i, a_i^1)])^2 \\ & - n \mathbb{E}_{x \sim d_0} \mathbb{E}_{a^1, a^2 \sim \pi} [(R(\theta^c, x, a^1) - R(\theta_*, x, a^1) - R(\theta^c, x, a^2) + R(\theta_*, x, a^2))^2] \\ & \leq \sqrt{\sum_{i=1}^n 4B^2 \mathbb{E}_{x \sim d_0} \mathbb{E}_{a^1, a^2 \sim \pi} [(R(\theta^c, x, a^1) - R(\theta_*, x, a^1) - R(\theta^c, x, a^2) + R(\theta_*, x, a^2))^2] \log(N_{\mathcal{R}}(\epsilon_c)/\delta)} \\ & + \frac{8}{3} B^2 \log(N_{\mathcal{R}}(\epsilon_c)/\delta) \end{aligned}$$

1216 for all  $\theta^c \in \Theta^c$ . By Lemma G.2 and the definition of  $\Theta^c$ , we further have

$$\begin{aligned} & \mathbb{E}_{x \sim d_0} \mathbb{E}_{a^1, a^2 \sim \pi} [(R(\hat{\theta}, x, a^1) - R(\theta_*, x, a^1) - R(\hat{\theta}, x, a^2) + R(\theta_*, x, a^2))^2] \\ & \leq O\left(\frac{1}{n} B^2 \log(N_{\mathcal{R}}(\epsilon_c)/\delta) + \frac{1}{n} \sum_{i=1}^n ([R(\hat{\theta}, x_i, a_i^2) - R(\hat{\theta}, x_i, a_i^1)] - [R(\theta_*, x_i, a_i^2) - R(\theta_*, x_i, a_i^1)])^2 + B\epsilon_c\right), \end{aligned} \tag{F.3}$$

1217 from which we can further derive that

$$\begin{aligned} & \mathbb{E}_{x \sim d_0} \mathbb{E}_{a^1, a^2 \sim \pi} [(R(\hat{\theta}, x, a^1) - R(\theta_*, x, a^1) - R(\hat{\theta}, x, a^2) + R(\theta_*, x, a^2))^2] \\ & \leq O\left(\frac{1}{n} e^B \log(N_{\mathcal{R}}(\epsilon_c)/\delta) + e^B \epsilon_c\right) \end{aligned}$$

1218 with probability at least  $1 - 2\delta$  by Lemma F.3 and the union bound.

1219 We can then complete the proof by setting

$$b(x) = \mathbb{E}_{a^2 \sim \pi(\cdot|x)} [R(\hat{\theta}, x, a^2) - R(\theta_*, x, a^2)].$$

1220

□

1221 **Lemma F.5** (Coverage of  $\pi_*$  and  $\pi_{\hat{\theta}}$  by  $\pi_{\hat{\theta}_0}$ ). If  $m \geq 32\eta^2 D^2 e^B \log(N_{\mathcal{R}}(\epsilon_c))$ ,  $n = c_{m,n} m$  and  
 1222  $\epsilon_c \leq \frac{1}{(1+c_{m,n})e^B \eta^2 D^2}$  in Algorithm 3 and Assumption 2.5 holds, then with probability at least  
 1223  $1 - 4\delta$ , there exists  $b_1 : \mathcal{X} \rightarrow \mathbb{R}$  and  $b_2 : \mathcal{X} \rightarrow \mathbb{R}$  such that

$$\eta |R(\hat{\theta}_0, x, a) - R(\theta_*, x, a) - b_1(x)| \leq 1, \quad \eta |R(\hat{\theta}, x, a) - R(\theta_*, x, a) - b_2(x)| \leq 1$$

1224 for all  $x \in \mathcal{X}, a \in \mathcal{A}$  such that  $\pi_0(a|x) > 0$ .

1225 *Proof.* By Lemma F.3 and the union bound, we have with probability at least  $1 - \delta$ , it holds that

$$\begin{aligned} & \sum_{i=1}^m ([R(\hat{\theta}, \tilde{x}_i, \tilde{a}_i^2) - R(\hat{\theta}, \tilde{x}_i, \tilde{a}_i^1)] - [R(\theta_*, \tilde{x}_i, \tilde{a}_i^2) - R(\theta_*, \tilde{x}_i, \tilde{a}_i^1)])^2 \\ & + \sum_{i=1}^n ([R(\hat{\theta}, x_i, a_i^2) - R(\hat{\theta}, x_i, a_i^1)] - [R(\theta_*, x_i, a_i^2) - R(\theta_*, x_i, a_i^1)])^2 \\ & \leq O(e^B \log(N_{\mathcal{R}}(\epsilon_c)/\delta) + e^B (n + m) \epsilon_c). \end{aligned} \tag{F.4}$$

1226 Consider an  $\epsilon_c$ -net  $\mathcal{R}^c = \{R(\theta, \cdot, \cdot) | \theta \in \Theta^c\}$  for the reward function class  $\mathcal{R}$  with size  $N_{\mathcal{R}}(\epsilon_c)$ . For  
 1227 any  $R(\theta, \cdot, \cdot)$ , there exists  $R(\theta^c, \cdot, \cdot)$  such that

$$|R(\theta, x, a) - R(\theta^c, x, a)| \leq O(\epsilon_c)$$

1228 for all  $x \in \mathcal{X}, a \in \mathcal{A}$ .

1229 Applying Lemma G.1, with probability at least  $1 - \delta$ , we have

$$\sum_{i=1}^m ([R(\theta^c, \tilde{x}_i, \tilde{a}_i^2) - R(\theta^c, \tilde{x}_i, \tilde{a}_i^1)] - [R(\theta_*, x_i, a_i^2) - R(\theta_*, x_i, a_i^1)])^2$$



$$\begin{aligned}
& -m\mathbb{E}_{x \sim d_0} \mathbb{E}_{a^1, a^2 \sim \pi_0} [(R(\theta^c, x, a^1) - R(\theta_*, x, a^1) - R(\theta^c, x, a^2) + R(\theta_*, x, a^2))^2] \\
& \leq \sqrt{\sum_{i=1}^m 4B^2 \mathbb{E}_{x \sim d_0} \mathbb{E}_{a^1, a^2 \sim \pi_0} [(R(\theta^c, x, a^1) - R(\theta_*, x, a^1) - R(\theta^c, x, a^2) + R(\theta_*, x, a^2))^2] \log(N_{\mathcal{R}}(\epsilon_c)/\delta)} \\
& \quad + \frac{8}{3} B^2 \log(N_{\mathcal{R}}(\epsilon_c)/\delta)
\end{aligned}$$

1230 for all  $\theta^c \in \Theta^c$ . By Lemma G.2 and the definition of  $\Theta^c$ , we further have

$$\begin{aligned}
& \mathbb{E}_{x \sim d_0} \mathbb{E}_{a^1, a^2 \sim \pi} [(R(\hat{\theta}, x, a^1) - R(\theta_*, x, a^1) - R(\hat{\theta}, x, a^2) + R(\theta_*, x, a^2))^2] \\
& \leq O\left(\frac{1}{m} B^2 \log(N_{\mathcal{R}}(\epsilon_c)/\delta)\right) \\
& \quad + \frac{1}{m} \sum_{i=1}^n [(R(\hat{\theta}, \tilde{x}_i, \tilde{a}_i^2) - R(\hat{\theta}, \tilde{x}_i, \tilde{a}_i^1)) - [R(\theta_*, \tilde{x}_i, \tilde{a}_i^2) - R(\theta_*, \tilde{x}_i, \tilde{a}_i^1)]]^2 + B\epsilon_c. \quad (\text{F.5})
\end{aligned}$$

1231 Substituting (F.4) into (F.5), we have with probability at least  $1 - 2\delta$ ,

$$\begin{aligned}
& \mathbb{E}_{x \sim d_0} \mathbb{E}_{a^1, a^2 \sim \pi_0} [(R(\hat{\theta}, x, a^1) - R(\theta_*, x, a^1) - R(\hat{\theta}, x, a^2) + R(\theta_*, x, a^2))^2] \\
& \leq O\left(\frac{1}{m} e^B \log(N_{\mathcal{R}}(\epsilon_c)/\delta) + e^B \cdot \frac{n+m}{m} \cdot \epsilon_c\right).
\end{aligned}$$

1232 Therefore, there exists a mapping  $b_2 : \mathcal{X} \rightarrow \mathbb{R}$  such that

$$\mathbb{E}_{\pi_0} [(R(\hat{\theta}, x, a) - R(\theta_*, x, a) - b_2(x))^2] \leq O\left(\frac{1}{m} e^B \log(N_{\mathcal{R}}(\epsilon_c)/\delta) + e^B \cdot \frac{n+m}{m} \cdot \epsilon_c\right).$$

1233 By Lemma F.4, we have with probability at least  $1 - 2\delta$ , there exists a mapping  $b_1 : \mathcal{X} \rightarrow \mathbb{R}$  such  
1234 that

$$\mathbb{E}_{\pi_0} [(R(\hat{\theta}_0, x, a) - R(\theta_*, x, a) - b_1(x))^2] \leq O\left(\frac{1}{m} e^B \log(N_{\mathcal{R}}(\epsilon_c)/\delta) + e^B (1 + c_{m,n}) \epsilon_c\right).$$

1235 Hence, we can complete the proof by a union bound over the two events and Assumption 3.5.  $\square$

1236 Now we are ready to prove Theorem 3.7.

1237 *Proof of Theorem 3.7.* Let  $b$  be the mapping defined in Lemma F.4 for  $\hat{\theta}$ . We have

$$\begin{aligned}
& \mathbb{E}_{\pi_{\theta_*}^\eta} \left[ R(\theta_*, x, a) - \frac{1}{\eta} \log \frac{\pi_{\theta_*}^\eta(a|x)}{\pi_0(a|x)} \right] - \mathbb{E}_{\pi_{\hat{\theta}}^\eta} \left[ R(\theta_*, x, a) - \frac{1}{\eta} \log \frac{\pi_{\hat{\theta}}^\eta(a|x)}{\pi_0(a|x)} \right] \\
& = \frac{1}{\eta} \mathbb{E}_{\pi_{\theta_*}^\eta} \left[ \log \frac{\pi_0(a|x) \cdot \exp(\eta R(\theta_*, x, a))}{\pi_{\theta_*}^\eta(a|x)} \right] - \frac{1}{\eta} \mathbb{E}_{\pi_{\hat{\theta}}^\eta} \left[ \log \frac{\pi_0(a|x) \cdot \exp(\eta R(\theta_*, x, a))}{\pi_{\hat{\theta}}^\eta(a|x)} \right] \\
& = \frac{1}{\eta} \mathbb{E}_{x \sim d_0} [\log Z_{\theta_*}^\eta(x)] - \frac{1}{\eta} \mathbb{E}_{x \sim d_0} [\log Z_{\hat{\theta}}^\eta(x)] - \mathbb{E}_{x \sim d_0} \left[ \sum_{a \in \mathcal{A}} \pi_{\hat{\theta}}^\eta(a|x) \cdot (R(\theta_*, x, a) - R(\hat{\theta}, x, a)) \right].
\end{aligned}$$

1238 For an arbitrary reward function  $f : \mathcal{X} \times \mathcal{A} \rightarrow \mathbb{R}$ , let  $\Delta(x, a) = f(x, a) - R(\theta_*, x, a)$ . Consider  
1239 the following first derivative of  $J(f) = \log Z_f^\eta(x) - \eta \sum_{a \in \mathcal{A}} \pi_f^\eta(a|x) \cdot \Delta(x, a)$ , where  $Z_f^\eta(x) =$   
1240  $\sum_{a \in \mathcal{A}} \pi_0(a|x) \cdot \exp(\eta \cdot f(x, a))$  and  $\pi_f^\eta(a|x) \propto \pi_0(a|x) \cdot \exp(\eta \cdot f(x, a))$ .

1241 Similar to the proof of Theorem 2.8, we still have

$$\begin{aligned}
& \frac{\partial}{\partial \Delta(x, a)} \left[ \log Z_f^\eta(x) - \eta \sum_{a \in \mathcal{A}} \pi_f^\eta(a|x) \cdot \Delta(x, a) \right] \\
& = \frac{1}{Z_f^\eta(x)} \cdot \pi_0(a|x) \exp(\eta \cdot f(x, a)) \cdot \eta - \eta \cdot \pi_f^\eta(a|x) \\
& \quad - \eta \cdot \Delta(x, a) \cdot \frac{\pi_0(a|x) \cdot \exp(\eta \cdot f(x, a))}{Z_f^\eta(x)} \cdot \eta + \eta \cdot \Delta(x, a) \cdot \frac{[\pi_0(a|x) \cdot \exp(\eta \cdot f(x, a))]^2}{[Z_f^\eta(x)]^2} \cdot \eta
\end{aligned}$$

$$\begin{aligned}
& + \eta \sum_{a' \in \mathcal{A} \setminus \{a\}} \frac{\pi_0(a'|x) \cdot \exp(\eta \cdot f(x, a'))}{Z_f^\eta(x)} \cdot \eta \cdot \Delta(x, a') \cdot \frac{\pi_0(a|x) \cdot \exp(\eta \cdot f(x, a))}{Z_f^\eta(x)} \\
& = -\eta^2 \pi_f^\eta(a|x) \Delta(x, a) + \eta^2 [\pi_f^\eta(a|x)]^2 \cdot \Delta(x, a) + \eta^2 \sum_{a' \in \mathcal{A} \setminus \{a\}} \pi_f^\eta(a'|x) \pi_f^\eta(a|x) \Delta(x, a').
\end{aligned}$$

1242 Note that

$$\begin{aligned}
J(R(\hat{\theta}, x, \cdot)) &= \log Z_\theta^\eta(x) - \eta \sum_{a \in \mathcal{A}} \pi_\theta^\eta(a|x) \cdot (R(\hat{\theta}, x, a) - R(\theta_*, x, a)) \\
&= \log \sum_{a \in \mathcal{A}} \pi_0(a|x) \cdot \exp(\eta(R(\hat{\theta}, x, a) - b(x))) - \eta \sum_{a \in \mathcal{A}} \pi_\theta^\eta(a|x) \cdot (R(\hat{\theta}, x, a) - R(\theta_*, x, a) - b(x)) \\
&= J(R(\hat{\theta}, x, \cdot) - b(x)).
\end{aligned}$$

1243 Therefore, there exists  $f(\cdot, \cdot) = \gamma[R(\hat{\theta}, \cdot, \cdot) - b(\cdot)] + (1 - \gamma)R(\theta_*, \cdot, \cdot)$  such that  $(\gamma \in (0, 1))$

$$\begin{aligned}
& \mathbb{E}_{x \sim d_0} [J(R(\hat{\theta}, \cdot, \cdot)) - J(R(\theta_*, \cdot, \cdot))] \\
&= \frac{1}{\eta} \mathbb{E}_{x \sim d_0} \left[ -\eta^2 \sum_{a \in \mathcal{A}} \pi_f^\eta(a|x) \cdot \gamma \cdot (R(\hat{\theta}, x, a) - R(\theta_*, x, a) - b(x))^2 \right] \\
&+ \frac{1}{\eta} \mathbb{E}_{x \sim d_0} \left[ \gamma \eta^2 \sum_{a_1 \in \mathcal{A}} \sum_{a_2 \in \mathcal{A}} \pi_f^\eta(a_1|x) \pi_f^\eta(a_2|x) (R(\hat{\theta}, x, a_1) - R(\theta_*, x, a_1) - b(x)) \right. \\
&\quad \left. (R(\hat{\theta}, x, a_2) - R(\theta_*, x, a_2) - b(x)) \right] \\
&\geq -\eta \cdot \mathbb{E}_{\pi_f^\eta} [(R(\hat{\theta}, x, a) - R(\theta_*, x, a) - b(x))^2]
\end{aligned}$$

1244 By Lemma F.2, if  $m \geq 32\eta^2 D^2 e^B \cdot \log(2N_{\mathcal{R}}(\epsilon_c)/\delta)$ , for any  $(x, a) \in \mathcal{X} \times \mathcal{A}$  such that  $\pi_0(a|x) > 0$ ,  
1245 it holds that

$$\eta |R(\hat{\theta}_0, x, a) - R(\theta_*, x, a) - b_1(x)| \leq 1, \quad \eta |R(\hat{\theta}, x, a) - R(\theta_*, x, a) - b_2(x)| \leq 1,$$

which means that

$$\frac{\pi_f^\eta}{\pi_{\hat{\theta}_0}^\eta} \leq e^4.$$

1246 Let  $\epsilon_c = \min\{\frac{\epsilon}{2(1+c_{m,n}^{-1})e^B}, \frac{1}{(1+c_{m,n})e^B \eta^2 D^2}\}$ . By Lemma F.4, under the condition of the theorem,  
1247 with high probability the output policy  $\pi_\theta^\eta$  is  $O(\epsilon)$  optimal.  $\square$

### 1248 F.3 Proof of Corollary 3.10

1249 In this subsection, we also discuss our result under the local-coverage condition (Definition 2.11).

1250 *Proof of Corollary 3.10.* The proof follows the same lines as Theorem 3.7 by replacing the data  
1251 coverage condition with the local-coverage condition. It still holds that

$$Q(\pi^*) - Q(\pi_\theta^\eta) \leq \eta \cdot \mathbb{E}_{\pi_f^\eta} [(R(\hat{\theta}_0, x, a) - R(\theta_*, x, a) - b(x))^2],$$

1252 where  $\pi_f^\eta(a|x) \propto \pi_0(a|x) \cdot \exp(\eta \cdot f(x, a))$  and  $f(\cdot, \cdot) = \gamma[R(\hat{\theta}_0, \cdot, \cdot) - b(\cdot)] + (1 - \gamma)R(\theta_*, \cdot, \cdot)$   
1253 for some  $\gamma \in (0, 1)$ . Thus, We have  $\text{KL}(\pi_f^\eta(a|x) \parallel \pi_0) \leq 2\eta B$ , which further implies that

$$Q(\pi^*) - Q(\pi_\theta^\eta) \leq \eta \cdot C_{\rho_{\text{KL}}} \cdot O\left(\frac{1}{n} e^B \log(N_{\mathcal{R}}(\epsilon_c)/\delta) + e^B (1 + c_{m,n}^{-1}) \epsilon_c\right)$$

1254 by Lemma F.4. Then we can conclude by substituting the value of  $m$  into the suboptimality gap.  $\square$

## 1255 G Auxiliary Lemmas

1256 **Lemma G.1** (Freedman's Inequality). Let  $M, v > 0$  be fixed constants. Let  $\{X_i\}_{i=1}^n$  be a stochastic  
 1257 process,  $\{\mathcal{G}_i\}_i$  be a sequence of  $\sigma$ -fields, and  $X_i$  be  $\mathcal{G}_i$ -measurable, while almost surely

$$\mathbb{E}[X_i|\mathcal{G}_i] = 0, |X_i| \leq M, \text{ and } \sum_{i=1}^n \mathbb{E}[X_i^2|\mathcal{G}_{i-1}] \leq v.$$

1258 Then for any  $\delta > 0$ , with probability at least  $1 - \delta$ , it holds that

$$\sum_{i=1}^n X_i \leq \sqrt{2v \log(1/\delta)} + \frac{2}{3} M \log(1/\delta).$$

1259 **Lemma G.2.** Suppose  $a, b \geq 0$ . If  $x^2 \leq a + b \cdot x$ , then  $x^2 \leq 2b^2 + 2a$ .

1260 *Proof.* By solving the root of quadratic polynomial  $q(x) := x^2 - b \cdot x - a$ , we obtain  $\max\{x_1, x_2\} =$   
 1261  $(b + \sqrt{b^2 + 4a})/2$ . Hence, we have  $x \leq (b + \sqrt{b^2 + 4a})/2$  provided that  $q(x) \leq 0$ . Then we further  
 1262 have

$$x^2 \leq \frac{1}{4} \left( b + \sqrt{b^2 + 4a} \right)^2 \leq \frac{1}{4} \cdot 2 (b^2 + b^2 + 4a) \leq 2b^2 + 2a. \quad (\text{G.1})$$

1263 □

1264 **Lemma G.3** (Pinsker's Inequality). If  $\mathbb{P}_1, \mathbb{P}_2$  are two probability measures on a common measur-  
 1265 able space  $(\Omega, \mathcal{F})$ , then it holds that

$$\delta(\mathbb{P}_1, \mathbb{P}_2) \leq \sqrt{\frac{1}{2} \text{KL}(\mathbb{P}_1 \| \mathbb{P}_2)},$$

1266 where  $\delta(\cdot, \cdot)$  is the total variation distance and  $\text{KL}(\cdot \| \cdot)$  is the Kullback-Leibler divergence.

1267 **Lemma G.4** (Lemma A.4, Foster et al. 2021). For any sequence of real-valued random variables  
 1268  $(X_t)_{t \leq T}$  adapted to a filtration  $(\mathcal{F}_t)_{t \leq T}$ , it holds that with probability at least  $1 - \delta$ , for all  $T' \leq T$ ,

$$\sum_{t=1}^{T'} X_t \leq \sum_{t=1}^{T'} \log(\mathbb{E}_{t-1}[e^{X_t}]) + \log(1/\delta).$$

1269 **Lemma G.5** (Solution of KL-regularized Optimization (Proposition 7.16 of Zhang 2023)). For any  
 1270 fixed  $x \in \mathcal{X}$  and reward function  $R$ , we have

$$\begin{aligned} & \max_{\pi} \mathbb{E}_{a \sim \pi(\cdot|x)} \left[ R(x, a) - \eta^{-1} \text{KL}(\pi(\cdot|x) \| \pi_0(\cdot|x)) \right] \\ &= \frac{1}{\eta} \cdot \log \mathbb{E}_{a \sim \pi_0(\cdot|x)} \exp(\eta R(x, a)), \end{aligned}$$

1271 where  $Z_R(x)$  is the normalization constant and the minimizer of the loss functional is

$$\pi_R^\eta(a|x) = \frac{1}{Z_R(x)} \pi_0(a|x) \exp(\eta R(x, a)).$$

1272 **Lemma G.6** (Lemma 10, Zhang et al. 2020). Let  $(M_n)_{n \geq 0}$  be a martingale such that  $M_0 = 0$  and  
 1273  $|M_n - M_{n-1}| \leq c$  for some  $c > 0$  and any  $n \geq 1$ . Let  $\text{Var}_n = \sum_{k=1}^n \mathbb{E}[(M_k - M_{k-1})^2 | \mathcal{F}_{k-1}]$   
 1274 for  $n \geq 0$ , where  $\mathcal{F}_k = \sigma(M_1, M_2, \dots, M_k)$ . Then for any positive integer  $n$  and any  $\varepsilon, p > 0$ , we  
 1275 have that

$$\mathbb{P} \left( |M_n| \geq 2\sqrt{\text{Var}_n \log\left(\frac{1}{p}\right)} + 2\sqrt{\varepsilon \log\left(\frac{1}{p}\right)} + 2c \log\left(\frac{1}{p}\right) \right) \leq (2nc^2/\varepsilon + 2)p$$