Offline and Online KL-Regularized RLHF under Differential Privacy

Yulian Wu
King Abdullah University of Science and Technology
Thuwal, Saudi Arabia, 23955-6900
yulian.wu@kaust.edu.sa

Rushil Thareja
Mohamed bin Zayed University of Artificial Intelligence
Masdar City, Abu Dhabi, United Arab Emirates
rushil.thareja@mbzuai.ac.ae

Praneeth Vepakomma
Mohamed bin Zayed University of Artificial Intelligence
Massachusetts Institute of Technology
vepakom@mit.edu

Francesco Orabona
King Abdullah University of Science and Technology
Thuwal, Saudi Arabia, 23955-6900
francesco@orabona.com

Abstract

In this paper, we study the offline and online settings of reinforcement learning from human feedback (RLHF) with KL-regularization—a widely used objective function in large language model alignment—under the ϵ local differential privacy (ϵ-LDP) model on the label of the human preference. In the offline setting, we design an algorithm based on the principle of pessimism and derive a new suboptimality gap of $\tilde{O}(1/[(e^\epsilon - 1)^2 n])$ on the KL-regularized objective under single-policy concentrability. We also prove its optimality by providing a matching lower bound where n is the sample size. In the online setting, we are the first one to theoretically investigate the problem of KL-regularized RLHF with LDP. We design an optimism-based algorithm and derive a logarithmic regret bound of $O(d_{\mathcal{F}} \log(N_{\mathcal{F}} \cdot T)/(e^\epsilon - 1)^2)$, where T is the total time step, $N_{\mathcal{F}}$ is cardinality of the reward function space $\mathcal{F}$ and $d_{\mathcal{F}}$ is a variant of eluder dimension for RLHF. As a by-product of our analysis, our results also imply the first analysis for online KL-regularized RLHF without privacy. Finally, we implement our algorithm in the offline setting on real data to verify our theoretical results.

1 INTRODUCTION

The alignment of Large Language Models (LLMs) with human preferences, often achieved through Reinforcement Learning from Human Feedback (RLHF), has become a central area of research. A key technique in this process is the Kullback-Leibler (KL) regularization, which is widely used to prevent

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Socially Responsible and Trustworthy Foundation Models (ResponsibleFM).

the model from deviating too far from its original behavior and to avoid overfitting [37, 2, 40, 30]. Mathematically, this objective function encourages the maximization of a reward model while forcing the learned policy π to stay close to a base policy π_{ref} for a given state s (prompt) and action a (response):

$$J(\pi) := \mathbb{E}_{(s,a)\sim d_0\times\pi}\left[r^*(s,a) - \beta^{-1}\log\frac{\pi(a \mid s)}{\pi_{\text{ref}}(a \mid s)}\right], \tag{1}$$

where $r^*(s,a)$ represents the ground truth reward and $\beta > 0$ is the inverse temperature parameter. The performance of algorithms is measured by the suboptimality gap in the offline setting, defined as

$$\text{SubOpt}(\pi) := J(\pi^*) - J(\pi), \tag{2}$$

where π^* is the optimal policy $\pi^* := \arg\max_\pi J(\pi)$. In the online setting, performance is measured by regret:

$$\text{Reg}(\pi_{1:T}) := \sum_{t=1}^{T}(J(\pi^*) - J(\pi_t)) \,. \tag{3}$$

While RLHF is effective, significant privacy concerns arise because the preference data used for alignment may contain personal or sensitive information [36, 24]. The standard framework for quantifying and mitigating privacy leakage is Differential Privacy (DP) [9]. By introducing calibrated randomness, DP ensures that the output of an algorithm is not overly sensitive to any single individual's data, thereby protecting their privacy. In the context of learning from human feedback, a key challenge is to preserve the privacy of the potentially sensitive preference labels provided by users. This has motivated recent work on applying DP specifically to preference-based learning, often referred to as label differential privacy (label DP) [12]. Label differential privacy in KL-regularized RLHF for the offline setting is studied in [35] under a central differential privacy model in which the learner can access the raw information of human labels. However, in some applications, individual labelers may be unwilling—or legally unable—to share raw feedback with the learner. These considerations motivate studying a local model for label differential privacy, where each human preference is privatized before disclosure.

Several recent works consider privacy issues on preference labels and study the problem by adopting differential privacy. However, the intersection of these two areas—KL-regularized RLHF and local model label differential privacy—remains unexplored. In particular, it is unknown whether applying label LDP to KL-regularized RLHF can yield strong theoretical guarantees on suboptimality and regret. Motivated by this gap, we are interested in our first question:

1. In the offline setting, can we achieve an optimal rate for KL-regularized RLHF under the label-LDP setting?

A primary challenge in offline RLHF is the distribution shift, which occurs when the data distribution used to train the reward model mismatches the response distribution of the optimized policy. This can lead to out-of-distribution errors, reward over-optimization, and degraded performance. While many recent works on theoretical offline RLHF derive rates that depend on notions of data coverage, one effective method to mitigate distribution shift is to use an online version of RLHF. For instance, [38] achieves logarithmic regret for online KL-regularized RL, depending on the eluder dimension. However, no existing work has studied the privacy problem in online KL-regularized RLHF, which leads us to our second question:

2. In the online setting, can we provide a logarithmic regret bound for KL-regularized RLHF under a local differential privacy mechanism?

We answer both of these questions affirmatively and summarize our contributions as follows:

- For the problem of private KL-regularized RLHF in the offline setting, we propose the PPKL-RLHF algorithm (Algorithm 1), which uses a Random Response (RR) mechanism to achieve label ϵ-LDP. Using these privatized preference labels for a private Maximum Likelihood Estimation (MLE), we obtain a conservative reward estimation via the principle of pessimism, which is then used for policy optimization with Gibbs sampling. We derive a suboptimality gap upper bound of $\widetilde{O}\left(1/[(e^\epsilon-1)^2 n]\right)$ (Equation (2)), with sample size n and under single policy concentrability. To demonstrate optimality, we also establish a matching lower bound.

- For the online setting, we design the POKL-RLHF algorithm (Algorithm 2), which uses RR to locally privatize human feedback. With the privatized labels and historical data, we design an exploitation agent using private least squares estimation and strategically design exploration via optimism for reward estimation. This exploration strategy yields a logarithmic regret bound for the exploration agent (Equation 3). To the best of our knowledge, we are the first to study the private online KL-regularized RLHF problem.
- As a by-product, our analysis provides insights into the non-private online KL-regularized RLHF setting. In particular, we establish the first logarithmic regret bound for online KL-regularized RLHF using a new variant of the eluder dimension. This result outperforms the sublinear regret bound for online RLHF in [30, 28] and sheds light on future research directions, such as online f-regularized RLHF or analyzing online KL-regularized RLHF from a Markov decision process perspective.
- Finally, we also run some experiments on a real dataset by implementing our algorithm design for the offline setting.

2 RELATED WORK

Given the large literature on trustworthy LLM alignment, this is necessarily a short review of the most related theory work. We refer the reader to [17] for a more comprehensive survey of this topic.

Non-Private Offline KL-regularized RLHF Offline RLHF suffers from a distribution shift problem, since the model is trained on a fixed dataset. Coverage conditions are used to measure the ability of the training-data distribution to cover the test-data distribution. With sample size n in KL-regularized RLHF, [30] derives a suboptimality gap of[1] $\widetilde{O}(1/\sqrt{n})$ under single-policy coverage. [37] achieves a suboptimality gap of $\widetilde{O}(1/n)$ but under their all-policy concentrability, which is a strong condition that requires the sample distribution to cover all possible distributions. [39] first establishes the suboptimality gap of $\widetilde{\Theta}(1/n)$ under single-policy coverage. Building on these, we derive the optimal convergence of $\widetilde{\Theta}(1/[(e^\epsilon-1)^2 n]$ with single-policy concentrability for the private offline KL-regularized RLHF under ϵ-LDP.

Non-private Online KL-regularized RLHF Online methods are a promising approach to overcome the out-of-distribution problems in offline RLHF. [30, 31] show the benefits of the online exploration agent and provides regret of $\widetilde{O}(\sqrt{T})$ for online KL-regularized RLHF with an eluder-type condition. [33] investigate the online KL-regularized RLHF problem via a Nash equilibrium reformulation. [28] study online KL-regularized RLHF via adding an exploration term on their loss function based on optimism in the face of uncertainty, and establishes regret of $\widetilde{O}(\sqrt{T})$ under their trajectory-level coverability coefficient. Our result improves has a better regret, but for a different objective function. In fact, taking the privacy parameter $\epsilon \to +\infty$, our results imply the first logarithmic regret bound of $\widetilde{O}(\log T)$ depending on the eluder dimension.

Locally Private RLHF [42, 43] achieve sub-optimality gap of $\widetilde{O}(1/[(e^\epsilon-1)\sqrt{n}])$ for locally private RLHF on the unregularized suboptimality gap as the performance measure for policy in the offline setting. We adopt a KL-regulized objective function to evaluate progress on the same function the algorithm optimizes, which avoids evaluation–training mismatch. With KL-regularized performance measure, we can improve the sub-optimality gap to $\widetilde{\Theta}(1/(1/[(e^\epsilon-1)^2 n])$ for the offline setting and achieve $\widetilde{O}(\log T/(e^\epsilon-1)^2)$ with eluder dimension for the online setting, due to the strongly convexity of the KL-regularized objection function. [7] considers label DP in both local and central models in offline RLHF, but they focus on the estimation error of the parameter, not suboptimality gaps.

3 PRELIMINARY

In this section, we introduce the necessary background of KL-regularized RLHF via the contextual bandits view, for both offline and online settings, as well as the basic knowledge of privacy in human feedback. We refer the readers to [16] for a unified view of RLHF via contextual bandits.

[1]We use $\widetilde{O}(\cdot), \widetilde{\Omega}(\cdot), \widetilde{\Theta}(\cdot)$ to hide polylog factors.

3.1 Offline and Online KL-regularized RLHF

KL-regularized RLHF seeks to optimize a target policy π by using human preferences to learn a reward function $r(s,a)$, while constraining the policy update to stay close to a reference policy π_{ref}. Without loss of generality, we will assume $r(s,a)$ in $[0,B]$ (e.g., via clipping in [14] or normalization). This leads to the following objective function:

$$\max_{\pi} \mathbb{E}_{s\sim d_0,\, a\sim\pi(\cdot|x)}[r(s,a)] - \frac{1}{\beta}\mathrm{KL}(\pi(\cdot \mid s) \,\|\, \pi_{\text{ref}}(\cdot \mid s)), \tag{4}$$

where π_{ref} is often a reference policy (e.g., SFT model). It is easy to see that the optimal solution of (4) is the Gibbs distribution, that is

$$\pi_r^*(a \mid s) = \frac{1}{Z_r(s)}\pi_{\text{ref}}(a \mid s)\exp(\beta \cdot r(s,a)), \tag{5}$$

where $Z_r(s)$ is the normalization constant.

Offline KL-regularized RLHF In the offline case, the learning agent aims to learn a good policy from a pre-collected dataset $\mathcal{D} = \{(s_i, a_i^1, a_i^2, y_i)\}_{i=1}^n$, where $y_i \in \{-1,1\}$ denotes the human's preference between two candidate responses a_i^1, a_i^2 generated from the reference policy π_{ref} given a prompt s_i sampled from d_0. The binary label $y_i \in \{-1,1\}$ indicates whether $a_i^1 \succ a_i^2$ ($y_i = 1$) or $a_i^2 \succ a_i^1$ ($y_i = -1$), that is, which response is preferred.

Remark 3.1. We use $y \in \{-1,1\}$ here, which is also adopted in [43], not in $\{0,1\}$ as in most of the RLHF literature, since this will help us simplify the math. The analysis under either convention can be translated back and forth without loss of generality.

We will need some definitions to quantify the "concentrability" of π_{ref}, that is, its ability to generate a diverse set of actions.

Definition 3.2 (40)**.** Given a class of functions $\mathcal{F} \subset (\mathcal{S}\times\mathcal{A} \to [0,B])$ and some policy π, let $\mathcal{B} = (\mathcal{S} \to [-B,B])$ be the function class of biases, and define $D^2_{\mathcal{F}}((s,a);\pi)$ as

$$\sup_{g,h\in\mathcal{F}} \inf_{b\in\mathcal{B}} \frac{(g(s,a) - h(s,a) - b(s))^2}{\mathbb{E}_{s'\sim d_0}\mathrm{Var}_{a'\sim\pi(\cdot|s')}\left[g\left(s',a'\right) - h\left(s',a'\right)\right]}\,.$$

Definition 3.3 (Single-policy Concentrability [40])**.** $D^2_{\pi^*} := \mathbb{E}_{(s,a)\sim d_0\times\pi^*} D^2_{\mathcal{F}}\left((s,a);\pi_{\text{ref}}\right) < \infty$

Definition 3.4 (Density-ratio-based concentrability)**.** For policy class Π and reference policy π_{ref}, the density-ratio-based all-policy concentrability C^{Π} is $C^{\Pi} := \sup_{\pi\in\Pi, s\in\mathcal{S}, a\in\mathcal{A}} \pi(a \mid s)/\pi_{\text{ref}}(a \mid s)$. The single-policy counterpart under the optimal policy π^* is $C^{\pi^*} := \sup_{s\in\mathcal{S},a\in\mathcal{A}} \pi^*(a \mid s)/\pi_{\text{ref}}(a \mid s)$.

Online KL-regularized RLHF Online KL-regularized RLHF updates the policy π_t over rounds. At each step t, a context s_t is drawn, two actions $a_t^1 \sim \pi_t^1$ and $a_t^2 \sim \pi_t^2$ are sampled (possibly asymmetrically), and human feedback $y_t \in \{-1,1\}$ is queried. The second policy π_t^2 is used to facilitate exploration. Based on accumulated feedback $\mathcal{D}_t = \{(s_i, a_i^1, a_i^2, y_i)\}_{i=1}^t$, the reward is re-estimated to get $\hat{r}_t$, and the next policy is updated via (5):

$$\pi_{t+1}^1(a \mid s) \propto \pi_{\text{ref}}(a \mid s)\cdot\exp\left(\beta\cdot\hat{r}_t(s,a)\right)\,.$$

Definition 3.5 (Uncertainty and pair eluder dimension)**.** For any sequence $\mathcal{D}_{t-1} = \left\{\left(s_i, a_i^1, a_i^2\right)\right\}_{i=1}^{t-1}$, we define $U_{\mathcal{F}}\left(\lambda, s, a; \mathcal{D}_t; \pi_{t+1}\right)$, the uncertainty of (s,a) with respect to $\mathcal{F}$, as

$$\sup_{r_1,r_2\in\mathcal{F}} \frac{\left|r_1(s,a) - r_2(s,a) - \mathbb{E}_{b\sim\pi_{t+1}}[r_1(s,b) - r_2(s,b)]\right|}{\sqrt{\lambda + \sum_{i=1}^t \left(r_1\left(s_i,a_i^1\right) - r_1\left(s_i,a_i^2\right) - \left[r_2\left(s_i,a_i^1\right) - r_2\left(s_i,a_i^2\right)\right]\right)^2}}.$$

The pair eluder dimension is given by $d_{\mathcal{F}} := \sup_{s_{1:T}, a_{1:T}^2} \sum_{t=1}^T \min\left(1, \left[U_{\mathcal{F}_t}\left(\lambda, s_t, a_t^2; \mathcal{D}_t; \pi_{t+1}^1\right)\right]^2\right)$.

Remark 3.6. The eluder dimension definition was first proposed by [22] for multi arm bandits problem to measure the efficacy with which observed data support inference about the values of unobserved actions and then widely used in RL problem [20, 38, 25, 27, 32, 1, 41] and preference-based RL [26, 6, 33]. Our definition is a variant of the eluder dimension for the design of the exploration strategy based on the exploitation agent.

For both offline and online setting, we adopt the standard Bradley-Terry (BT) model for the preference model and we will assume realizability.

Assumption 3.7 (Bradley-Terry Preference Model)**.** Given a context s and two actions a_1, a_2, we assume the preference label y is sampled according to the the ground truth reward function r^* difference between the two actions:

$$\mathbb{P}[y = 1 \mid s, a^1, a^2] = \sigma(r^*(s, a^1) - r^*(s, a^2)), \tag{6}$$

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

Assumption 3.8 (Realizability of reward function)**.** We assume that $r^* \in \mathcal{F} \subset (\mathcal{S} \times \mathcal{A} \to [0, B])$.

To derive uniform theoretical guarantees when $|\mathcal{F}|$ is infinite, we approximate it by a finite subset that is sufficiently dense with respect to an appropriate metric. This allows us to apply analysis to the finite subset and then transfer the bound to the entire class via a discretization argument. The complexity of $\mathcal{F}$ in this sense is captured by the covering number, which measures how many elements are required to approximate every function in $\mathcal{F}$ within a prescribed tolerance. We recall the formal definition below.

Definition 3.9 (Net and covering number)**.** Given a function class $\mathcal{F} \subset (\mathcal{S} \times \mathcal{A} \to [0, B])$ and $\tau \in (0, 1)$, a finite set $\mathcal{F}(\tau) \subset \mathcal{F}$ is a τ-net of $\mathcal{F}$ w.r.t. $\|\cdot\|_\infty$, if for any $f \in \mathcal{F}$, there exists $f' \in \mathcal{F}(\tau)$ such that $\|f - f'\|_\infty \leq \tau$. The τ-covering number is the smallest cardinality $\mathcal{N}_{\mathcal{F}}(\tau)$ of such $\mathcal{F}(\tau)$.

3.2 Privacy in Human Feedback

Here, we formally introduce the Label Differential Privacy in the local model.

Definition 3.10 (ε-Pure Local Label DP [7])**.** If each label is first privatized by a local randomizer $\mathcal{R}$, which satisfies for any y, y' and any subset S in the range of $\mathcal{R}$, it holds that for $\varepsilon > 0$,

$$\mathbb{P}[\mathcal{R}(y) \in S] \leq e^{\varepsilon} \cdot \mathbb{P}\left[\mathcal{R}\left(y'\right) \in S\right],$$

then, we say that $\mathcal{R}$ is an ε-pure label differentially private local randomizer, where $\varepsilon > 0$ is the privacy parameter. Smaller values of ε provide stronger privacy guarantees, but introduce more noise.

Instead of directly observing the true binary preference $y \in \{-1, 1\}$ at each round, the learning agent receives a privatized label $z \in \{-1, 1\}$ obtained via randomized response (RR):

$$\begin{aligned} \mathbb{P}(z = y) &= \alpha := \frac{e^{\varepsilon}}{e^{\varepsilon} + 1} \in (0.5, 1), \\ \mathbb{P}(z \neq y) &= 1 - \alpha = \frac{1}{e^{\varepsilon} + 1}\,. \end{aligned} \tag{7}$$

The above randomized response mechanism satisfies ε-pure local label DP [9].

4 OFFLINE PRIVATE KL-REGULARIZED RLHF WITH PESSIMISM

In this section, we will study the locally private KL-regularized RLHF in the offline setting. We will first provide the algorithm for the problem and derive its suboptimality upper bound. In order to show the optimality of the theoretical guarantee, we will also present the lower bound under the same assumptions.

4.1 Algorithm and Upper Bound

The main idea of Algorithm 1 is that we first take the precollected data set $\widetilde{\mathcal{D}} = \{(s_i, a_i^1, a_i^2, z_i)\}_{i=1}^n$, where $z_i \in \{-1, +1\}$ are the privatized version of the true (unobserved) preference label y_i through the randomized response mechanism in (7) with flip probability $1 - \alpha$. For each sample (s, a^1, a^2, z), the probability of private label z given s, a^1, a^2 is

$$\widetilde{P}_{r^*}(z \mid s, a^1, a^2) := \mathbb{P}(z|s, a^1, a^2) = \alpha \cdot \sigma(z \cdot \Delta_{r^*}(s, a^1, a^2)) + (1-\alpha) \cdot \sigma(-z \cdot \Delta_{r^*}(s, a^1, a^2)), \tag{8}$$

where $\Delta_{r^*}(s, a^1, a^2) := r^*(s, a^1) - r^*(s, a^2)$ and $\sigma(x) = (1 + e^{-x})^{-1}$ is the sigmoid function. Building on the probability function

$$\widetilde{P}_r(z \mid s, a^1, a^2) = \alpha \cdot \sigma(z \cdot \Delta_r(s, a^1, a^2)) + (1-\alpha) \cdot \sigma(-z \cdot \Delta_r(s, a^1, a^2)) \tag{9}$$

of z as a function of the reward r, we can estimate the reward by the Maximum Likelihood Estimation (MLE) on $\widetilde{P}_r(z \mid s, a^1, a^2)$ in step 1 of the algorithm. After we get the estimation of the reward $\bar{r}$, we construct a pessimistic estimator $\hat{r}$ in step 2 with the following value of the bonus $\Gamma_n(s, a)$:

$$\sqrt{D^2_{\mathcal{F}}((s,a); \pi_{\text{ref}}) \frac{c \cdot e^B}{(2\alpha-1)^2} \left(\frac{\log(\mathcal{N}_{\mathcal{F}}(\tau)/\delta)}{n} + \tau \right)}, \tag{10}$$

where c is a constant. Finally, we get the policy output by Gibbs distribution from (5) based on $\hat{r}$.

Remark 4.1. The pessimism principle is well-known in offline RL [15] and offline RLHF [37]. It consists in adopting the lower confidence bound of the reward estimation, since the conservative estimate helps the distributional shift challenge in the offline setting. In our local DP case, the main difference compared with the non-private case is that the effective sample size changes from n to $(2\alpha-1)^2 \cdot n = [(e^\epsilon - 1)/(e^\epsilon + 1)]^2 \cdot n < n$ due to the randomness from the privacy-preserving mechanism.

We now provide the theoretical guarantee of the suboptimality gap for the output policy in Algorithm 1. We defer its detailed proof in Appendix B.

Theorem 4.2 (Sub-optimality gap upper bound in offline setting)**.** *Under Assumptions 3.7 and 3.8, Definitions 3.2, 3.3 3.4, and 3.9, for $\epsilon > 0, \beta > 0$ and a sufficiently small $\tau \in (0, 1)$, with probability at least $1 - \delta$, we have that the suboptimality gap of the output of Algorithm 1, SubOpt$(\hat{\pi})$ is of the order of*

$$O\left(\beta D^2_{\pi^*} \frac{e^B}{(2\alpha-1)^2} \left(\frac{\log(\mathcal{N}_{\mathcal{F}}(\tau)/\delta)}{n} + \tau \right) \right). \tag{11}$$

Remark 4.3 (Discussion of the parameters in the upper bound). In the above results, β is a hyperparameter in the regularized objective function (1) to trade off the reward maximization and how close the target policy is to π_{ref}. e^B comes from the sigmoid function in BT preference model and it is common in the RLHF literature [42, 30, 37, 39].

Remark 4.4 (Comparision with prior work for upper bound). Compared with the unregularized suboptimality upper bound of $\widetilde{O}(1/[(2\alpha-1)\sqrt{n}])$ in [42] with their single-policy relative condition number, our result with KL-regularization of $\widetilde{O}(1/[(2\alpha-1)^2 n])$ is tighter when the sample size n is large enough, but on a different objective function. When $\epsilon \in (0, 1]$, which means a strong privacy guarantee, we obtain $\widetilde{O}(1/[(2\alpha-1)^2 n]) = \widetilde{O}(1/[(e^\epsilon - 1)^2 n])$ that matches the lower bound we prove in the following. Note that when $\epsilon \to +\infty$, i.e., $\alpha = 1$, we recover the non-private case in [40].

4.2 Lower Bound Analysis

We verify the optimality of the above bound by proving the following lower bound and defer the complete proof to Appendix B.

Theorem 4.5 (Sub-optimality gap lower bound in offline setting)**.** *For reward function class $\mathcal{F} \subset (\mathcal{S} \times \mathcal{A} \to [0, B])$, $\tau \in (0, 1)$ small enough, $\beta > 0$, $S = \log \mathcal{N}_{\mathcal{F}}(\tau)$, $C^* \in (2, e^{(\beta B)/2} + 1)$,*

Algorithm 1 Private Pessimistic KL-Regularized RLHF (PPKL-RLHF) for Offline Setting

---

Require: Regularization parameter β, reference policy π_{ref}, function class $\mathcal{F}$, offline dataset $\widetilde{\mathcal{D}} = \{(s_i, a_i^1, a_i^2, z_i)\}_{i=1}^n$

1: Compute the private MLE estimation of the reward function:

$$\bar{r} \in \arg\max_{r \in \mathcal{F}} \sum_{i=1}^n \log \widetilde{P}_r(z_i \mid s_i, a_i^1, a_i^2)$$

2: Use pessimism: $\hat{r}(s, a) \leftarrow \bar{r}(s, a) - \Gamma_n(s, a), \forall (s, a)$, where Γ_n is the bonus term in (10)

3: **return** $\hat{\pi}(a \mid s) \propto \pi_{\text{ref}}(a \mid s) \exp(\beta \cdot \hat{r}(s, a))$

---

algorithm set Π*,* $C^{\pi^*} \leq C^*$*, and KL-regularized RLHF instance set* $\mathcal{I}$*, the minimax suboptimality gap* $\inf_{\hat{\pi}\in\Pi}\sup_{I\in\mathcal{I}}$ *SubOpt*$(\hat{\pi}, I)$ *under* ϵ*-LDP mechanism for labels is*

$$\Omega\left(\min\left\{\frac{\beta C^* \log \mathcal{N}_{\mathcal{F}}(\tau)}{(e^\epsilon-1)^2 n}, \frac{\sqrt{\log \mathcal{N}_{\mathcal{F}}(\tau) C^*}}{(e^\epsilon-1)\sqrt{n}}\right\}\right). \tag{12}$$

Remark 4.6 (Comparision with prior work for lower bound). A lower bound for the *parameter estimation* for RLHF under label LDP is provided in [7]. In particular, they show a lower bound of $\Omega(\frac{1}{e^\epsilon-1}\sqrt{\frac{d}{n}})$ for the estimation error bound of the parameter in a linear reward model in $\mathbb{R}^d$. As far as we know, we are the first ones to provide the lower bound for the *suboptimality gap* for this problem of RLHF under LDP, matching the same effective sample size of $(e^\epsilon-1)^2 n \approx \epsilon^2 n$ when $\epsilon \in (0,1)$ as [7]. Taking $\mathcal{N}_{\mathcal{F}}(\tau) = (1/\tau)^d$ in the linear model, we can imply the suboptimality gap of $\widetilde{\Omega}\left(\min\left\{\frac{\beta C^* d}{(e^\epsilon-1)^2 n}, \frac{\sqrt{dC^*}}{(e^\epsilon-1)\sqrt{n}}\right\}\right)$ for private KL-regularized RLHF which also demonstrates the importance of β and C^* in this problem.

Remark 4.7 (Discussion of the parameters in the lower bound). From the above lower bound and the upper bound of the suboptimality gap in Theorem 4.2, we obtain that the single-policy coverage C^{π^*} is necessary due to the distribution shift between the behavior policy and optimal in the private RLHF problem. In fact, [11] showed that in the non-private RLHF setting the single policy coverage coefficient is also unavoidable. Motivated by this, in the next section we study the problem of private KL-regularized RLHF under an online setting, which will help remove the dependence on the coverage condition.

5 ONLINE PRIVATE KL-REGULARIZED RLHF WITH OPTIMISM

In this section, we turn our attention to KL-Regularized RLHF with LDP on labels in the online setting. Compared with the online RL problem, the main challenge of online RLHF comes from the imperfect information on the reward. That is, the reward can be observed in RL and used to estimate the reward model. However, in online RLHF, given a context, we need to sample two actions and receive human labels to train the reward model. This raises another problem: How to sample two actions?

The sampling methods of two actions in online RLHF are mainly divided into two classes: symmetric and non-symmetric.

- In the symmetric class, we sample two actions from the same policy, e.g., the one got from the last iteration as in [4, 13]. However, [28, Proposition 2.1] shows that this strategy can suffer from a constant lower bound on the suboptimality gap. Hence, some kind of exploration is necessary in online RLHF.
- In the non-symmetric class, some algorithms sample actions from different polices—one policy from exploitation and another one for exploration based on the first one—for KL regularized RLHF [31, 30]. [28, 5] sample an action from the last iteration policy and another from the reference policy for KL regularized RLHF, but adds a bias term in the loss function for exploration.

Inspired by the above works, we adopt the optimism principle for our exploration policy, which is a principle widely used in online RL [29, 19, 18, 38]. We develop the Private Optimistic KL-Regularized RLHF (POKL-RLHF) algorithm (see Algorithm 2). In each time step $t \in \{1, \ldots, T\}$, after the learner observes the context s_t (the prompt in the large language model) sampled from a fixed distribution d_0, two actions (two answers from the LLM) are compared. In our LDP model, only the private label z_i privatized by the RR mechanism in (7) is available to the learner, instead of the true label y_i. With these historical data till time step t, we update the reward model by the private least squares estimation at Step 7. Then, we update the exploitation policy π^1_{t+1} based on the reward estimation by the solution of the KL-regularized objective function in (5). Given π^1_{t+1}, we design the exploration policy by using an exploration bonus. In particular, we construct a confidence set that will shrink with time:

$$\mathcal{F}_t = \left\{r \in \mathcal{F} : \sum_{i=1}^{t} \left(\Delta_i^r - \Delta_i^{\bar{r}_t}\right)^2 + \lambda \leq \Gamma_T^2\right\},$$

Algorithm 2 Private Optimistic KL-Regularized RLHF (POKL-RLHF) for Online Setting

Require: KL coefficient β, reward function class $\mathcal{F}$, exploration scale λ, reference policy π_{ref}, DP parameter ε

Initialize: $\mathcal{D}_0 = \emptyset$; $\pi_1^1, \pi_1^2 = \pi_{\text{ref}}$

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

  Observe context $s_t \sim d_0$

  Sample $a_t^1 \sim \pi_t^1(\cdot \mid s_t)$ and $a_t^2 \sim \pi_t^2(\cdot \mid s_t)$

  Observe private preference label $z_t \in \{-1, 1\}$ via randomized response in (7)

  Update $\mathcal{D}_t \leftarrow \mathcal{D}_{t-1} \cup \{(s_t, a_t^1, a_t^2, z_t)\}$

  Estimate reward from private least square:

$$\bar{r}_t = \arg\min_{r \in \mathcal{F}} \sum_{\mathcal{D}_t} \left[(2\sigma(\Delta_i^r) - 1)(2\alpha - 1) - z_i\right]^2,$$

  where $\Delta_i^r := r(s_i, a_i^1) - r(s_i, a_i^2)$

  Update exploitation policy: $\pi_{t+1}^1(a \mid s) \propto \pi_{\text{ref}}(a \mid s) \cdot \exp(\beta \cdot \bar{r}_t(s, a))$

  Set exploration policy: $\pi_{t+1}^2(a \mid s) \propto \pi_{t+1}^1(a \mid s) \cdot \exp(\beta \cdot b_t(s, a))$ with bonus b_t defined in (13)

end for

where

$$\Gamma_T = \frac{ce^B \sqrt{\log\left(T \cdot N_{\mathcal{F}}/\delta\right)}}{2\alpha - 1}$$

and c is a constant. Then, the exploration bonus b_t is defined through the uncertainty in Definition 3.5:

$$b_t(s, a) = \min\left\{1, \Gamma_T \, U_{\mathcal{F}_t}\left(\lambda, s, a; \mathcal{D}_t; \pi_{t+1}^1\right)\right\}. \tag{13}$$

Remark 5.1. As in [14, 38], we assume that the reward function space $\mathcal{F}$ is finite. The infinite case can be solved easily by an ϵ-net and uniform convergence argument (refer to Lemma C.1 in [38] and Lemma C.2 in [37]), similarly to our offline case.

Based on the optimism principle for exploration policy, we derive the following theoretical guarantee.

Theorem 5.2 (Regret Bound)**.** *Under Assumptions 3.7 and 3.8, for $\delta \in (0, 1), \epsilon > 0$ and $\lambda \leq \frac{1}{2}\Gamma_T^2$ with probability at least $1 - \delta$, Algorithm 2 satisfies*

$$\sum_{t=1}^{T} (J(\pi^*) - J(\pi_t^2)) = O\left(\frac{\beta \cdot d_{\mathcal{F}} \cdot e^{2B}}{(2\alpha - 1)^2} \log(N_{\mathcal{F}} \cdot T/\delta)\right),$$

where $d_{\mathcal{F}}$ is the pair eluder dimension in Definition 3.5, β is the hyperparameter in (1), $N_{\mathcal{F}}$ is the cardinality of reward function space.

Remark 5.3. In the context of online RL/RLHF, bounds in terms of the eluder dimension characterize the statistical learnability of exploration strategies. However, it is important to note that such guarantees are information-theoretic rather than computational: While they demonstrate that learning is possible with a finite number of iterations, the corresponding algorithms are often computationally intractable when the function class is large. We leave how to find a computationally efficient method with logarithmic regret for online RLHF as an open problem.

Remark 5.4. In the above results, e^{2B} comes from the sigmoid function for the preference model. The effect of LDP is a factor of $\frac{1}{(2\alpha-1)^2} = (\frac{e^\epsilon+1}{e^\epsilon-1})^2 > 1$ due to the randomness from the differential privacy mechanism. As a by-product, taking $\epsilon \to +\infty$, i.e., $\alpha = 1$ in the algorithm analysis, the result implies a bound for the corresponding non-private case.

Corollary 5.5. *Under Assumptions 3.7 and 3.8, for $\alpha = 1$, $\delta \in (0, 1)$, with probability at least $1 - \delta$, Algorithm 2 satisfies*

$$\sum_{t=1}^{T} (J(\pi^*) - J(\pi_t^2)) = O\left(\beta \cdot d_{\mathcal{F}} \cdot e^{2B} \log(N_{\mathcal{F}} \cdot T/\delta)\right).$$

Remark 5.6. Online RLHF is also studied in [30, Section 4], and from their proofs a sublinear regret bound of $\tilde{O}(\sqrt{T})$ for the exploration policy can be implied. Compared with their results, we are the first ones to achieve a logarithmic regret bound with the eluder dimension.

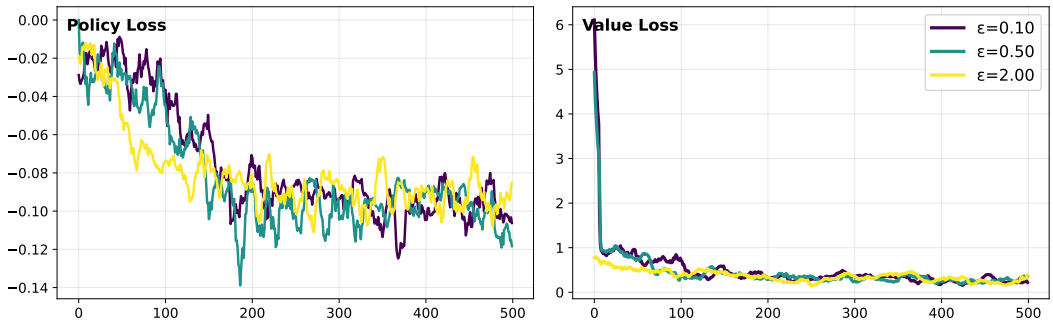

Figure 1: Training metrics for our Private KL-Regularized RLHF over iterations for different ϵ vals.

6 EXPERIMENTAL RESULTS

As noted in Remark 5.3, the online algorithm based on the eluder dimension is computationally intractable in practice. Thus, we choose to only experiment in the offline case to empirically verify our theoretical findings about the effect of the ϵ-LDP model and release our open source code at:https://github.com/rushil-thareja/PPKL-RLHF-Official.

Dataset and Compute For all experiments, we use the helpful assistant preference corpus[2] tailored for RLHF [3]. The dataset consists of two complementary components: (i) Supervised Fine-Tuning (SFT) dialogues, where each sample contains a user query and a preferred assistant response; and (ii) preference pairs, where each sample provides a prompt together with one chosen and one rejected response. The SFT corpus contains $38{,}821$ training examples and $4{,}413$ validation examples. Preference pairs are split into $38{,}821$ training, $2{,}100$ validation, and $2{,}313$ held-out test examples. We used a single AMD MI-200 GPU equipped with 64 GB of VRAM.

SFT training and Baseline We use the Llama-3.2-1B-Instruct model[3] as the backbone for all experiments. To obtain the baseline policy π_0, we performed SFT on the dialogue part of the dataset, with standard next-token prediction.

We also use Direct Preference Optimization (DPO) [21] as a baseline, training the policy relative to the frozen SFT reference π_0 on the preference pairs. The objective is optimized for $\beta = 0.1$ with AdamW, linear warmup, gradient accumulation, and validation every 500 steps, and the best checkpoint is selected by validation loss after a few thousand iterations. This baseline is non-private and without KL regularization.

Implementation of PPKL-RLHF To implement this setup, we first train a privatized reward model (Algorithm 1) that adds a scalar linear head with EOS pooling on top of the Llama-3.2-1B-Instruct backbone, clipped to $[-5, 5]$. The reward model is optimized in two phases: first warming up by training only the head, then fine-tuning the full backbone for 5 epochs. The policy is optimized with PPO [23] against the corrected rewards and a KL penalty to the SFT baseline, using $\beta = 0.1$. Training runs for 500 iterations with 16 rollouts per iteration; each update applies 3 PPO epochs with minibatch size 4, generation length capped at 64 tokens (prompts up to 256, temperature 1.0, top-p 0.9), and standard PPO hyperparameters (clip $\epsilon_c = 0.2$, policy lr 1×10^{-6}, value lr 5×10^{-6}, value loss weight 0.5, entropy coefficient 0.01, max grad norm 1.0).

Training Performance In Figure 1, we track two core metrics of training: policy loss and value loss. As showcased in Figure 1, the policy loss decreases steadily and converges to a low plateau, while the value loss drops sharply before stabilizing. As privacy is relaxed, both metrics improve. At $\epsilon = 0.10$, the policy loss remains relatively high and the value loss bottoms out at 0.072. At $\epsilon = 0.50$, both show stronger improvement, with the value loss converging to a lower value. At $\epsilon = 2.00$, training achieves the best utility: policy loss decreases most rapidly and value loss reaches its lowest point (0.062). These results confirm that higher ϵ (weaker privacy) yields stronger learning signals and more effective optimization, showing the expected trade-off between performance and privacy.

[2] https://huggingface.co/datasets/Anthropic/hh-rlhf
[3] https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct

Table 1: Win rates of different methods evaluated on the preference test set. PPKL-RLHF uses $\beta = 0.10$.

Method	Setting	Win rate
SFT (π_0)	–	0.538
DPO(non-private)	$\beta = 0.1$	0.704
PPKL-RLHF	$\epsilon = 0.1$	0.530
PPKL-RLHF	$\epsilon = 0.5$	0.554
PPKL-RLHF	$\epsilon = 2.0$	0.607

Results and Baseline Comparison The final results of our evaluation are presented in Table 1 where we use the win rate as our performance metric, as in [21, 42]. At stronger privacy (ϵ=0.1) performance is close to SFT, while at ϵ=0.5 it surpasses the SFT baseline (0.554 vs. 0.538). The best setting reaches around 0.607 at ϵ=2.0, indicating utility gains with weaker theoretical privacy. These results highlight that even with noisy privatized labels, training a reward model followed by our PPKL-RLHF procedure retains competitive utility and offers tunable privacy–utility trade-offs. However, PPKL-RLHF's win-rate remains behind DPO (0.704), likely because label privatization and the pessimistic KL correction restrict the effective learning signal compared to the non-private baseline.

7 CONCLUSION

In this paper, we investigated the KL-regularized RLHF problem in both offline and online settings. We designed algorithms based on pessimistic and optimistic principles for the offline and online settings, respectively, and provided theoretical guarantees for both cases. We established the optimal sub-optimality gaps for the offline setting and a logarithmic regret bound for the online setting while preserving privacy. Finally, we also showed some experimental results to verify our theoretical findings.

Potential avenues for future research include extending private KL-regularized RLHF to more general preference models, such as the Plackett–Luce model. Another promising direction is to investigate alternative differential privacy mechanisms that could provide improved privacy–utility trade-offs, for instance, through shuffle differential privacy or related variants.

Acknowledgement

We thank Wei Xiong, Xingyu Zhou, and Yuhui Wang for insightful discussions. We would like to acknowledge the MBZUAI SU Fund and MIT–MBZUAI Collaborative Research Program for supporting this work.

References

[1] Alekh Agarwal, Yujia Jin, and Tong Zhang. Vo q l: Towards optimal regret in model-free rl with nonlinear function approximation. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 987–1063. PMLR, 2023.

[2] Gholamali Aminian, Amir R Asadi, Idan Shenfeld, and Youssef Mroueh. Theoretical analysis of kl-regularized rlhf with multiple reference models. *arXiv preprint arXiv:2502.01203*, 2025.

[3] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[4] Shicong Cen, Jincheng Mei, Katayoon Goshvadi, Hanjun Dai, Tong Yang, Sherry Yang, Dale Schuurmans, Yuejie Chi, and Bo Dai. Value-incentivized preference optimization: A unified approach to online and offline rlhf. *arXiv preprint arXiv:2405.19320*, 2024.

[5] Fan Chen, Zeyu Jia, Alexander Rakhlin, and Tengyang Xie. Outcome-based online reinforcement learning: Algorithms and fundamental limits. *arXiv preprint arXiv:2505.20268*, 2025.

[6] Xiaoyu Chen, Han Zhong, Zhuoran Yang, Zhaoran Wang, and Liwei Wang. Human-in-the-loop: Provably efficient preference-based reinforcement learning with general function approximation. In *International Conference on Machine Learning*, pages 3773–3793. PMLR, 2022.

[7] Sayak Ray Chowdhury, Xingyu Zhou, and Nagarajan Natarajan. Differentially private reward estimation with preference feedback. In *International Conference on Artificial Intelligence and Statistics*, pages 4843–4851. PMLR, 2024.

[8] John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy and statistical minimax rates. In *2013 IEEE 54th annual symposium on foundations of computer science*, pages 429–438. IEEE, 2013.

[9] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

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

[11] Dylan J Foster, Zakaria Mhammedi, and Dhruv Rohatgi. Is a good foundation necessary for efficient reinforcement learning? the computational role of the base model in exploration. *arXiv preprint arXiv:2503.07453*, 2025.

[12] Badih Ghazi, Noah Golowich, Ravi Kumar, Pasin Manurangsi, and Chiyuan Zhang. Deep learning with label differential privacy. *Advances in neural information processing systems*, 34:27131–27145, 2021.

[13] Shangmin Guo, Biao Zhang, Tianlin Liu, Tianqi Liu, Misha Khalman, Felipe Llinares, Alexandre Rame, Thomas Mesnard, Yao Zhao, Bilal Piot, et al. Direct language model alignment from online ai feedback. *arXiv preprint arXiv:2402.04792*, 2024.

[14] Audrey Huang, Wenhao Zhan, Tengyang Xie, Jason D. Lee, Wen Sun, Akshay Krishnamurthy, and Dylan J. Foster. Correcting the mythos of kl-regularization: Direct alignment without overoptimization via chi-squared preference optimization, 2025.

[15] Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl?, 2022.

[16] Long-Fei Li, Yu-Yang Qian, Peng Zhao, and Zhi-Hua Zhou. Provably efficient rlhf pipeline: A unified view from contextual bandits. *ArXiv preprint*, 2502, 2025.

[17] Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*, 2023.

[18] Antoine Moulin and Gergely Neu. Optimistic planning by regularized dynamic programming. In *International Conference on Machine Learning*, pages 25337–25357. PMLR, 2023.

[19] Antoine Moulin, Gergely Neu, and Luca Viano. Optimistically optimistic exploration for provably efficient infinite-horizon reinforcement and imitation learning. *arXiv preprint arXiv:2502.13900*, 2025.

[20] Ian Osband and Benjamin Van Roy. Model-based reinforcement learning and the eluder dimension. *Advances in neural information processing systems*, 27, 2014.

[21] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

[22] Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. *Advances in Neural Information Processing Systems*, 26, 2013.

[23] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017.

[24] Weijie Su. Do large language models (really) need statistical foundations?, 2025.

[25] Ruosong Wang, Russ R Salakhutdinov, and Lin Yang. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. *Advances in Neural Information Processing Systems*, 33:6123–6135, 2020.

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

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

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

[29] Wei Xiong. A sufficient condition of sample-efficient reinforcement learning with general function approximation. *The Hong Kong University of Science and Technology*, 2023.

[30] Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: Bridging theory and practice for rlhf under kl-constraint. *arXiv preprint arXiv:2312.11456*, 2023.

[31] Wei Xiong, Chengshuai Shi, Jiaming Shen, Aviv Rosenberg, Zhen Qin, Daniele Calandriello, Misha Khalman, Rishabh Joshi, Bilal Piot, Mohammad Saleh, et al. Building math agents with multi-turn iterative preference learning. *arXiv preprint arXiv:2409.02392*, 2024.

[32] Chenlu Ye, Wei Xiong, Quanquan Gu, and Tong Zhang. Corruption-robust algorithms with uncertainty weighting for nonlinear contextual bandits and markov decision processes. In *International Conference on Machine Learning*, pages 39834–39863. PMLR, 2023.

[33] Chenlu Ye, Wei Xiong, Yuheng Zhang, Hanze Dong, Nan Jiang, and Tong Zhang. Online iterative reinforcement learning from human feedback with general preference model. *Advances in Neural Information Processing Systems*, 37:81773–81807, 2024.

[34] Tong Zhang. *Mathematical analysis of machine learning algorithms*. Cambridge University Press, 2023.

[35] Yizhou Zhang, Kishan Panaganti, Laixi Shi, Juba Ziani, and Adam Wierman. Kl-regularization itself is differentially private in bandits and rlhf, 2025.

[36] Zhexin Zhang, Yuhao Sun, Junxiao Yang, Shiyao Cui, Hongning Wang, and Minlie Huang. Be careful when fine-tuning on open-source llms: Your fine-tuning data could be secretly stolen!, 2025.

[37] Heyang Zhao, Chenlu Ye, Quanquan Gu, and Tong Zhang. Sharp analysis for kl-regularized contextual bandits and rlhf. *arXiv preprint arXiv:2411.04625*, 2024.

[38] Heyang Zhao, Chenlu Ye, Wei Xiong, Quanquan Gu, and Tong Zhang. Logarithmic regret for online kl-regularized reinforcement learning. *arXiv preprint arXiv:2502.07460*, 2025.

[39] Qingyue Zhao, Kaixuan Ji, Heyang Zhao, Tong Zhang, and Quanquan Gu. Nearly optimal sample complexity of offline kl-regularized contextual bandits under single-policy concentrability. *arXiv preprint arXiv:2502.06051*, 2025.

[40] Qingyue Zhao, Kaixuan Ji, Heyang Zhao, Tong Zhang, and Quanquan Gu. Towards a sharp analysis of offline policy learning for f-divergence-regularized contextual bandits, 2025.

[41] Han Zhong, Wei Xiong, Sirui Zheng, Liwei Wang, Zhaoran Wang, Zhuoran Yang, and Tong Zhang. Gec: A unified framework for interactive decision making in mdp, pomdp, and beyond. *arXiv preprint arXiv:2211.01962*, 2022.

[42] Xingyu Zhou, Yulian Wu, and Francesco Orabona. A unified theoretical analysis of private and robust offline alignment: from rlhf to dpo, 2025.

[43] Xingyu Zhou, Yulian Wu, Wenqian Weng, and Francesco Orabona. Square χpo: Differentially private and robust χ^2-preference optimization in offline direct alignment. *arXiv preprint arXiv:2505.21395*, 2025.

A Useful Lemmas

Lemma A.1 (10)**.** *For any sequence of real-valued random variables $(X_t)_{t\le T}$ adapted to a filtration $(\mathcal{F}_t)_{t\le T}$, it holds that with probability at least $1-\delta$, for all $T' \le T$,*

$$\sum_{t=1}^{T'} X_t \le \sum_{t=1}^{T'} \log \mathbb{E}_{t-1}\left[e^{X_t}\right] + \log\frac{1}{\delta}\,.$$

Lemma A.2. *Let*

$$f(x) = \log\big(\alpha\sigma(x) + (1-\alpha)\big(1-\sigma(x)\big)\big)\,, \quad \sigma(x) = \frac{1}{1+e^{-x}},$$

where $\alpha \in (0.5, 1)$ and $x \in [-B, B]$. Then for any $a, b \in [-B, B]$, we have

$$|f(a) - f(b)| \le \sigma(B)\,|a-b|\,.$$

Proof. First, observe that

$$\alpha\sigma(x) + (1-\alpha)(1-\sigma(x)) = 1-\alpha+(2\alpha-1)\sigma(x)\,.$$

So, we have

$$f'(x) = \frac{(2\alpha-1)\sigma(x)(1-\sigma(x))}{1-\alpha+(2\alpha-1)\sigma(x)} \le 1-\sigma(x),$$

where the inequality due to the fact that $1-\alpha \ge 0$.

Maximizing over $x \in [-B, B]$, we obtain

$$\sup_{x\in[-B,B],\ \alpha\in(0.5,1)} f'(x) \le \sup_{x\in[-B,B]} 1-\sigma(x) = 1-\sigma(-B) = \sigma(B)\,.$$

Finally, by the Mean Value Theorem, for any $a, b \in [-B, B]$ there exists c between a and b such that

$$|f(a) - f(b)| = |f'(c)|\,|a-b| \le \sigma(B)\,|a-b|\,. \qquad \square$$

Lemma A.3 (Freedman's Inequality)**.** *Let $\delta \in (0,1)$. Let $M, v > 0$ be fixed constants. Let $\{X_i\}_{i=1}^n$ be a stochastic process, $\{\mathcal{G}_i\}_i$ be a sequence of σ-fields, and X_i be $\mathcal{G}_i$-measurable, while almost surely*

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

Then, with probability at least $1-\delta$, it holds that

$$\sum_{i=1}^n X_i \le \sqrt{2v\log\frac{1}{\delta}} + \frac{2}{3}M\log\frac{1}{\delta}\,.$$

Lemma A.4 ([37])**.** *Suppose $a, b \ge 0$. If $x^2 \le a + b\cdot x$, then $x^2 \le b^2 + 2a$.*

Lemma A.5 (Theorem 1 in [8])**.** *For any $\epsilon \ge 0$, let Q be a conditional distribution that guarantees ϵ-local differential privacy. Then for any pair of distributions P_1 and P_2, the induced marginals M_1 and M_2 where $M_j(S) = \int_{\mathcal{X}} Q(S \mid x) dP_j(x)$ for $j = 1, 2$ satisfy the bound*

$$D_{\mathrm{kl}}\left(M_1 \| M_2\right) + D_{\mathrm{kl}}\left(M_2 \| M_1\right) \le \min\left\{4, e^{2\epsilon}\right\}\left(e^\epsilon - 1\right)^2 \|P_1 - P_2\|_{\mathrm{TV}}^2\,.$$

Lemma A.6 (Assouad's Lemma)**.** *Let $\mathcal{I}$ be the set of instances, Π be the set of estimators, $\Theta := \{\pm 1\}^S$ for some $S > 0$, and $\{L_j\}_{j=1}^S$ be S functions from $\Pi \times \mathcal{I}$ to $\mathbb{R}_+$. Suppose $\{I_\theta\}_{\theta\in\Theta} \subset \mathcal{I}$ and the loss function is*

$$L(\pi, I) := \sum_{j=1}^{S} L_j(\pi, I), \forall(\pi, I) \in \Pi \times \mathcal{I} .$$

We denote $\theta \sim_j \theta'$ if they differ only in the j-th coordinate. Further, assume that

$$\theta \sim_j \theta' \Rightarrow \inf_{\pi\in\Pi} L_j(\pi, J_\theta) + L_j(\pi, J_{\theta'}) \geq c,$$

for some $c > 0$. Then, we have

$$\inf_{\pi\in\Pi} \sup_{I\in\mathcal{I}} L(\pi, I) \geq S \cdot \frac{c}{4} \min_{\exists j:\theta\sim_j\theta'} \exp\left(-\mathrm{KL}\left(P_{I_\theta} \| P_{I_{\theta'}}\right)\right),$$

where P_I denotes the distribution of the dataset given $I \in \mathcal{I}$.

Lemma A.7 (40)**.** *Let $b(s) : \mathcal{S} \to \mathbb{R}$ be some bias function, then for all $r(s, a) \in \mathcal{F}$ we have $J(\pi_r) = J(\pi_{r-b})$ since $\pi_r = \pi_{r-b}$ where $\pi_r = \frac{\pi_{ref}\exp(\beta r)}{\sum_{a\in\mathcal{A}}\pi_{ref}\exp(\beta r)}$, where $(r - b)(s, a) = r(s, a) - b(s)$.*

Lemma A.8. *Let $\sigma(x) = \frac{1}{1+e^{-x}}$ be sigmoid function and $f(x) = (2\sigma(x) - 1)(2\alpha - 1)$ for a fixed $\alpha \in (0.5, 1]$. For any $B \geq 0$ and any $x, x' \in [-B, B]$,*

$$|x - x'| \leq \frac{e^{-B} + 2 + e^{B}}{2(2\alpha - 1)} |f(x) - f(x')| .$$

Proof. First we have $f'(x) = 2(2\alpha - 1)\sigma'(x)$ with

$$\sigma'(x) = \frac{e^{-x}}{(1 + e^{-x})^2} = \frac{1}{e^x + 2 + e^{-x}} .$$

On $[-B, B]$, σ' attains its minimum at $\pm B$:

$$\min_{|x|\leq B} \sigma'(x) = \frac{1}{e^B + 2 + e^{-B}} .$$

Hence

$$m := \inf_{|x|\leq B} |f'(x)| = \frac{2(2\alpha - 1)}{e^B + 2 + e^{-B}} .$$

By the Mean Value Theorem there exists ξ between x and x' such that

$$|f(x) - f(x')| = |f'(\xi)| \, |x - x'| \geq m \, |x - x'|,$$

which gives the stated inequality. □

B Proofs of Section 4

In Algorithm 1, we estimate the reward function via MLE. Thus, we extend the approach in [37] to establish the generalization error bound of reward difference the MLE, taking into account that here the MLE is on the private probabilities.

Lemma B.1. *For an arbitrary policy π, and a set of offline data $\{(s_i, a_i^1, a_i^2, z_i)\}_{i=1}^n$ generated i.i.d from the BT model and π, and privatized by RR. Suppose that $\bar{r}$ is the result of the private MLE in step 1 of Algorithm 1, then there exists a function $b(s) : \mathcal{S} \to [-B, B]$ such that with probability at least $1 - 2\delta$ and for all values of τ small enough, we have*

$$\mathbb{E}_{s\sim d_0, a\sim\pi(\cdot|s)}[\bar{r}(s, a) - r^*(s, a) - b(s)]^2 = O\left(\frac{e^B}{(2\alpha - 1)^2} \cdot \left(\frac{\log(\mathcal{N}_{\mathcal{F}}(\tau)/\delta)}{n} + \tau\right)\right) . \quad (14)$$

From the proof of the lemma, define $b(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}[\bar{r}(s, a) - r^*(s, a)]$, then $\mathbb{E}_{s\sim d_0} \mathrm{Var}_{a\sim\pi(\cdot|s)}[\bar{r}(s, a) - r^*(s, a)] = \mathbb{E}_{(s,a)\sim d_0\times\pi}[(\bar{r}(s, a) - r^*(s, a) - b(s))^2]$. Note that, in the offline setting, the actions are sampled from π_{ref}.

Proof of Lemma B.1. **Step 1: Connect private MLE and the reward difference.** Since we estimate the reward function by private MLE, let

$$\widetilde{L}(r|s_i,a_i^1,a_i^2)=\log\left[\alpha\cdot\sigma(z_i\cdot\Delta_r(s_i,a_i^1,a_i^2))+(1-\alpha)\cdot\sigma(-z_i\cdot\Delta_r(s_i,a_i^1,a_i^2))\right]\ .$$

We first use Lemma A.1 on the sequence

$$\left\{\frac{1}{2}\widetilde{L}(r|s_i,a_i^1,a_i^2)-\frac{1}{2}\widetilde{L}(r^*|s_i,a_i^1,a_i^2)\right\}_{i=1}^n=\left\{\frac{1}{2}\log\frac{\widetilde{P}_r(z_i|s_i,a_i^1,a_i^2)}{\widetilde{P}_{r^*}(z_i|s_i,a_i^1,a_i^2)}\right\}_{i=1}^n,$$

for any $r\in\mathcal{F}$ where $\widetilde{P}_r$ is defined in (9). Then, for $s\le n$, we have with probability at least $1-\delta$ that

$$\begin{aligned}
&\frac{1}{2}\sum_{i=1}^s\left[\widetilde{L}(r|s_i,a_i^1,a_i^2)-\widetilde{L}(r^*|s_i,a_i^1,a_i^2)\right]\\
&\le\sum_{i=1}^s\log\mathbb{E}\left[\sqrt{\frac{\widetilde{P}_r(z_i|s_i,a_i^1,a_i^2)}{\widetilde{P}_{r^*}(z_i|s_i,a_i^1,a_i^2)}}\right]+\log\frac{1}{\delta}\\
&=\sum_{i=1}^s\log\left[\sqrt{\widetilde{P}_r(z_i=-1|s_i,a_i^1,a_i^2)\widetilde{P}_{r^*}(z_i=-1|s_i,a_i^1,a_i^2)}\right.\\
&\quad\left.+\sqrt{\widetilde{P}_r(z_i=+1|s_i,a_i^1,a_i^2)\widetilde{P}_{r^*}(z_i=+1|s_i,a_i^1,a_i^2)}\right]+\log\frac{1}{\delta}\\
&\overset{(a)}{\le}\sum_{i=1}^s\left[\sqrt{\widetilde{P}_r(z_i=-1|s_i,a_i^1,a_i^2)\widetilde{P}_{r^*}(z_i=-1|s_i,a_i^1,a_i^2)}\right.\\
&\quad\left.+\sqrt{\widetilde{P}_r(z_i=+1|s_i,a_i^1,a_i^2)\widetilde{P}_{r^*}(z_i=+1|s_i,a_i^1,a_i^2)}-1\right]+\log\frac{1}{\delta}\\
&=\log\frac{1}{\delta}-\frac{1}{2}\sum_{i=1}^s\left(\sqrt{\widetilde{P}_{r^*}(z_i=+1|s_i,a_i^1,a_i^2)}-\sqrt{\widetilde{P}_r(z_i=+1|s_i,a_i^1,a_i^2)}\right)^2\\
&\quad-\frac{1}{2}\sum_{i=1}^s\left(\sqrt{\widetilde{P}_{r^*}(z_i=-1|s_i,a_i^1,a_i^2)}-\sqrt{\widetilde{P}_r(z_i=-1|s_i,a_i^1,a_i^2)}\right)^2\\
&\overset{(b)}{\le}\log\frac{1}{\delta}-\frac{1}{8}\sum_{i=1}^s\left(\widetilde{P}_{r^*}(z_i=+1|s_i,a_i^1,a_i^2)-\widetilde{P}_r(z_i=+1|s_i,a_i^1,a_i^2)\right)^2\\
&=\log\frac{1}{\delta}-\frac{1}{8}\sum_{i=1}^s(2\alpha-1)^2\cdot[\sigma(\Delta_{r^*}(s_i,a_i^1,a_i^2))-\sigma(\Delta_r(s_i,a_i^1,a_i^2))]^2\\
&\le\log\frac{1}{\delta}-\frac{(2\alpha-1)^2\cdot e^B}{8(1+e^B)^2}\sum_{i=1}^s[\Delta_{r^*}(s_i,a_i^1,a_i^2)-\Delta_r(s_i,a_i^1,a_i^2)]^2,
\end{aligned}\tag{15}$$

where (a) is from $\log x\le x-1$ for $x>0$, (b) is from $(\sqrt{a}-\sqrt{b})^2\ge\frac{1}{4}(a-b)^2$ for $a,b\in[0,1]$ since $(\sqrt{a}-\sqrt{b})^2=\frac{(a-b)^2}{(\sqrt{a}+\sqrt{b})^2}\ge\frac{1}{4}(a-b)^2,\ a,b\in[0,1]$ and the last inequality is from $\sigma'(x)\ge\frac{e^B}{(1+e^B)^2}$ for $x\in[-B,B]$.

Step 2: private likelihood function class well-covered by τ-net of reward function. For any $\tau>0$, define $\mathcal{F}_\tau$ as a τ-net for the reward function class $\mathcal{F}$ with covering number $\mathcal{N}_{\mathcal{F}}(\tau)$ in Definition 3.9. Then, for any $s\in\mathcal{S},a^1,a^2\in\mathcal{A},z\in\{-1,+1\}$ and $r\in\mathcal{F}$, there exists $r'\in\mathcal{F}_\tau$ such that

$$|\widetilde{L}(r|s,a^1,a^2)-\widetilde{L}(r'|s,a^1,a^2)|\le\sigma(B)|\Delta_r(s,a^1,a^2)-\Delta_{r'}(s,a^1,a^2)|\le2\sigma(B)\tau,\tag{16}$$

where the first inequality is from Lemma A.2 by taking $x=z\cdot\Delta_r(s,a^1,a^2)$ and $\sigma(-x)=1-\sigma(x)$. This yields

$$\sum_{i=1}^s\widetilde{L}(r|s_i,a_i^1,a_i^2)\le\sum_{i=1}^s\widetilde{L}(r'|s_i,a_i^1,a_i^2)+2\sigma(B)\tau s\ .\tag{17}$$

Step 3: confidence bound for the private MLE estimator. Based on (15) and the union bound, for all $r' \in \mathcal{F}_\tau$ we obtain

$$\frac{1}{2}\sum_{i=1}^{n}\left[\widetilde{L}(r'|s_i,a_i^1,a_i^2)-\widetilde{L}(r^*|s_i,a_i^1,a_i^2)\right] \le \log\frac{\mathcal{N}_{\mathcal{F}}(\tau)}{\delta}-\frac{(2\alpha-1)^2\cdot e^B}{8(1+e^B)^2}\sum_{i=1}^{n}[\Delta_{r^*}(s_i,a_i^1,a_i^2)-\Delta_{r'}(s_i,a_i^1,a_i^2)]^2 .$$

Building on the above inequality and (17), we have with probability at least $1-\delta$, for any $r \in \mathcal{F}$, there exists $r' \in \mathcal{F}_\tau$ such that

$$\begin{aligned}
&\frac{1}{2}\sum_{i=1}^{n}\left\{\widetilde{L}(r|s_i,a_i^1,a_i^2)-\widetilde{L}(r^*|s_i,a_i^1,a_i^2)\right\}\\
\le&\frac{1}{2}\sum_{i=1}^{n}\left\{\widetilde{L}(r|s_i,a_i^1,a_i^2)-\widetilde{L}(r'|s_i,a_i^1,a_i^2)+\widetilde{L}(r'|s_i,a_i^1,a_i^2)-\widetilde{L}(r^*|s_i,a_i^1,a_i^2)\right\}\\
\overset{(a)}{\le}&\log\frac{\mathcal{N}_{\mathcal{F}}(\tau)}{\delta}-\frac{(2\alpha-1)^2\cdot e^B}{8(1+e^B)^2}\sum_{i=1}^{n}[\Delta_{r^*}(s_i,a_i^1,a_i^2)-\Delta_{r'}(s_i,a_i^1,a_i^2)]^2+\sigma(B)\tau n\\
=&\log\frac{\mathcal{N}_{\mathcal{F}}(\tau)}{\delta}-\frac{(2\alpha-1)^2\cdot e^B}{8(1+e^B)^2}\sum_{i=1}^{n}[\Delta_{r^*}(s_i,a_i^1,a_i^2)-\Delta_{r}(s_i,a_i^1,a_i^2)+\Delta_{r}(s_i,a_i^1,a_i^2)-\Delta_{r'}(s_i,a_i^1,a_i^2)]^2+\sigma(B)\tau n\\
\overset{(b)}{\le}&\log\frac{\mathcal{N}_{\mathcal{F}}(\tau)}{\delta}-\frac{(2\alpha-1)^2\cdot e^B}{4(1+e^B)^2}\sum_{i=1}^{n}[\Delta_{r^*}(s_i,a_i^1,a_i^2)-\Delta_{r}(s_i,a_i^1,a_i^2)]^2+\frac{(2\alpha-1)^2\cdot e^B}{(1+e^B)^2}\tau^2 n+\sigma(B)\tau n\\
\le&\log\frac{\mathcal{N}_{\mathcal{F}}(\tau)}{\delta}-\frac{(2\alpha-1)^2\cdot e^B}{4(1+e^B)^2}\sum_{i=1}^{n}[\Delta_{r^*}(s_i,a_i^1,a_i^2)-\Delta_{r}(s_i,a_i^1,a_i^2)]^2+2\tau n,
\end{aligned} \tag{18}$$

where (a) is from the union bound over $\mathcal{F}_\tau$, (b) is from $(a+b)^2 \le 2a^2+2b^2$ and the definition of τ-net for the reward functions, and the last inequality is from the small value of τ.

Since $\bar{r}$ is the private MLE estimator, by the realizability of the reward function, we have $\sum_{i=1}^{n}\{\widetilde{L}(\bar{r}|s_i,a_i^1,a_i^2)-\widetilde{L}(r^*|s_i,a_i^1,a_i^2)\} \ge 0$. So, we get

$$0 \le \log\frac{\mathcal{N}_{\mathcal{F}}(\tau)}{\delta}-\frac{(2\alpha-1)^2\cdot e^B}{4(1+e^B)^2}\sum_{i=1}^{n}[\Delta_{r^*}(s_i,a_i^1,a_i^2)-\Delta_{\bar{r}}(s_i,a_i^1,a_i^2)]^2+2\tau n .$$

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

$$\sum_{i=1}^{n}[\Delta_{r^*}(s_i,a_i^1,a_i^2)-\Delta_{\bar{r}}(s_i,a_i^1,a_i^2)]^2 \le \frac{4(1+e^B)^2}{(2\alpha-1)^2\cdot e^B}\left(\log\frac{\mathcal{N}_{\mathcal{F}}(\tau)}{\delta}+2\tau n\right) . \tag{19}$$

Step 4: On-policy error bound of reward difference function. We first get the bound on the finite reward function set $\mathcal{F}_\tau$, then derive it for an infinite set $\mathcal{F}$. We now use Lemma A.3 by taking $X_i = \mathbb{E}[Y_i]-Y_i$ as zero mean r.v. where $Y_i = [\Delta_{r'}(s_i,a_i^1,a_i^2)-\Delta_{r^*}(s_i,a_i^1,a_i^2)]^2 \in [0,4B^2]$, thus, $|X_i| \le 4B^2$ and $\mathbb{E}X_i^2 = \mathbb{E}[Y_i^2]-[\mathbb{E}Y_i]^2 \le \mathbb{E}Y_i^2 \le 4B^2\mathbb{E}Y_i$. Hence, by the union bound, with probability at least $1-\delta$ we have for all $r' \in \mathcal{F}_\tau$ that

$$\begin{aligned}
&n\mathbb{E}_{s\sim d_0,a^1,a^2\sim\pi}[\Delta_{r'}(s,a^1,a^2)-\Delta_{r^*}(s,a^1,a^2)]^2-\sum_{i=1}^{n}[\Delta_{r'}(s_i,a_i^1,a_i^2)-\Delta_{r^*}(s_i,a_i^1,a_i^2)]^2\\
&\le\sqrt{4nB^2\log\frac{\mathcal{N}_{\mathcal{F}}(\tau)}{\delta}\mathbb{E}_{s\sim d_0,a^1,a^2\sim\pi}[\Delta_{r'}(s,a^1,a^2)-\Delta_{r^*}(s,a^1,a^2)]^2}+\frac{8}{3}B^2\log\frac{\mathcal{N}_{\mathcal{F}}(\tau)}{\delta} .
\end{aligned} \tag{20}$$

From the above inequality and by taking $x = \sqrt{n\mathbb{E}_{s\sim d_0,a^1,a^2\sim\pi}[\Delta_{r'}(s,a^1,a^2)-\Delta_{r^*}(s,a^1,a^2)]^2}, b = 2B, a = \frac{8}{3}B^2\log(\mathcal{N}_{\mathcal{F}}(\tau)/\delta)+\sum_{i=1}^{n}[\Delta_{r'}(s_i,a_i^1,a_i^2)-\Delta_{r^*}(s_i,a_i^1,a_i^2)]^2$ in Lemma A.4, we get

$$n\mathbb{E}_{s\sim d_0,a^1,a^2\sim\pi}[\Delta_{r'}(s,a^1,a^2)-\Delta_{r^*}(s,a^1,a^2)]^2 = O\left(B^2\log\frac{\mathcal{N}_{\mathcal{F}}(\tau)}{\delta}\right)+\sum_{i=1}^{n}[\Delta_{r'}(s_i,a_i^1,a_i^2)-\Delta_{r^*}(s_i,a_i^1,a_i^2)]^2 .$$

By the definition of τ-net in Definition 3.9, we have for the private MLE estimator $\bar{r}$, there exists a $r' \in \mathcal{F}_\tau$, such that, for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, we have $|r'(s,a) - \bar{r}(s,a)| \leq \tau$ from which and the result in step 3 we can further derive with probability at least $1-2\delta$

$$\begin{aligned}
&\mathbb{E}_{s\sim d_0, a^1, a^2 \sim \pi}[\Delta_{\bar{r}}(s,a^1,a^2) - \Delta_{r^*}(s,a^1,a^2)]^2 \\
&= O\left(\frac{B^2 \log(\mathcal{N}_{\mathcal{F}}(\tau)/\delta)}{n}\right) + \frac{1}{n}\sum_{i=1}^{n}[\Delta_{r'}(s_i,a_i^1,a_i^2) - \Delta_{r^*}(s_i,a_i^1,a_i^2)]^2 + 8\tau^2 \\
&= \frac{4(1+e^B)^2}{(2\alpha-1)^2 \cdot e^B} \cdot \left(\frac{\log(\mathcal{N}_{\mathcal{F}}(\tau)/\delta)}{n} + 2\tau\right) + O\left(\frac{B^2 \log(\mathcal{N}_{\mathcal{F}}(\tau)/\delta)}{n}\right) + 8\tau^2 \\
&= O\left(\frac{e^B}{(2\alpha-1)^2} \cdot \left(\frac{\log(\mathcal{N}_{\mathcal{F}}(\tau)/\delta)}{n} + \tau\right)\right),
\end{aligned}$$

for all values of τ small enough. Then, we get the result by taking $b(s) = \mathbb{E}_{a^2\sim\pi}[\bar{r}(s,a^2) - r^*(s,a^2)]$. □

Lemma B.2. *From Lemma 2.16 in [40] and Lemma E.2 in [40], if pessimistic event $(g - r^*)(s,a) \leq 0$ holds, we have*

$$J(\pi^*) - J(\pi_g) \leq \beta \mathbb{E}_{(s,a)\sim\rho\times\pi^*}\left[(g-r^*)^2(s,a)\right].$$

We state the details of the proof here.

Proof of Theorem 4.2. Similar to Lemma E.1 in [40], it is easy to get with probability at least $1-\delta$, the event $\mathcal{E}(\delta) := \{\exists b : \mathcal{S} \to [-B,B], \forall (s,a) \in \mathcal{S}\times\mathcal{A}, |\bar{r}(s,a) - b(s) - r^*(s,a)| \leq \Gamma_n(s,a)\}$ holds for $\delta \in (0,1)$.

From the result of Lemma B.1, we have with probability at least $1-\delta$,

$$\mathbb{E}_{s'\sim d_0}\mathrm{Var}_{a'\sim\pi_{\mathrm{ref}}}[\bar{r}(s',a') - r^*(s',a')] \leq O\left(\frac{e^B}{(2\alpha-1)^2} \cdot \left(\frac{\log(\mathcal{N}_{\mathcal{F}}(\tau)/\delta)}{n} + \tau\right)\right).$$

Then we have

$$\begin{aligned}
&\inf_b (\bar{r}(s,a) - b(s) - r^*(s,a))^2 \\
&= \inf_b \frac{(\bar{r}(s,a) - b(s) - r^*(s,a))^2}{\mathbb{E}_{s'\sim d_0}\mathrm{Var}_{a'\sim\pi_{\mathrm{ref}}}[\bar{r}(s',a') - r^*(s',a')]} \mathbb{E}_{s'\sim d_0}\mathrm{Var}_{a'\sim\pi_{\mathrm{ref}}}[\bar{r}(s',a') - r^*(s',a')] \\
&\leq D_{\mathcal{F}}^2((s,a),\pi_{\mathrm{ref}})\mathbb{E}_{s'\sim d_0}\mathrm{Var}_{a'\sim\pi_{\mathrm{ref}}}[\bar{r}(s',a') - r^*(s',a')] \\
&\leq D_{\mathcal{F}}^2((s,a),\pi_{\mathrm{ref}}) O\left(\frac{e^B}{(2\alpha-1)^2} \cdot \left(\frac{\log(\mathcal{N}_{\mathcal{F}}(\tau)/\delta)}{n} + \tau\right)\right).
\end{aligned}$$

Thus, we get $\mathcal{E}(\delta)$ holds with probability at least $1-\delta$.

Under event $\mathcal{E}(\delta)$, we have $\hat{r}(s,a) - b(s) \leq r^*(s,a)$,

$$J(\pi^*) - J(\pi_{\hat{r}}) = J(\pi^*) - J(\pi_{\hat{r}-b}) \leq \beta \cdot \mathbb{E}_{(s,a)\sim d_0\times\pi^*}[(\hat{r}(s,a) - b(s) - r^*(s,a))^2],$$

where $\hat{r}(s,a) = \bar{r}(s,a) - \Gamma_n(s,a)$ in Step 2 of Algorithm 1, the equation is from Lemma A.7 and the inequality is from Lemma B.2. Therefore, we obtain

$$\begin{aligned}
J(\pi^*) - J(\pi_{\hat{r}}) &\leq \beta \cdot \mathbb{E}_{(s,a)\sim d_0\times\pi^*}[(\hat{r}(s,a) - b(s) - r^*(s,a))^2] \\
&= \beta \cdot \mathbb{E}_{(s,a)\sim d_0\times\pi^*}[(\bar{r}(s,a) - \Gamma_n(s,a) - b(s) - r^*(s,a))^2] \\
&\leq \beta\left(2\mathbb{E}_{(s,a)\sim d_0\times\pi^*}[\Gamma_n(s,a)]^2 + 2\mathbb{E}_{(s,a)\sim d_0\times\pi^*}[(\bar{r}(s,a) - b(s) - r^*(s,a))^2]\right) \\
&\leq 4\beta\mathbb{E}_{(s,a)\sim d_0\times\pi^*}[\Gamma_n(s,a)]^2 \\
&= 4\beta D_{\pi^*}^2 \cdot O\left(\frac{e^B}{(2\alpha-1)^2} \cdot \left(\frac{\log(\mathcal{N}_{\mathcal{F}}(\tau)/\delta)}{n} + \tau\right)\right) \\
&= O\left(\beta D_{\pi^*}^2 \frac{e^B}{(2\alpha-1)^2} \cdot \left(\frac{\log(\mathcal{N}_{\mathcal{F}}(\tau)/\delta)}{n} + \tau\right)\right),
\end{aligned}$$

□

Proof of Theorem 4.5. Consider the set of private RLHF instances

$$\mathcal{I} = \{(\mathcal{S}, \mathcal{A}, r, \pi_{ref}, \beta, \mathcal{R})\},$$

where $\mathcal{R}$ is the LDP randomizer. We aim to construct a specific instance in the set to get the minimax lower bound.

Step 1: Construct the instance. Inspired by [40], we consider the following instance for the private RLHF problem via the contextual dueling bandits view: the state space $\mathcal{S} = [S]$ where $S \geq 1$, binary action space $\mathcal{A} = \{-1, +1\}$, $d_0 = Unif(\mathcal{S})$ is a uniform distribution, the reward function in some function class $\mathcal{F} \subseteq \mathcal{S} \times \mathcal{A} \to [0, B]$ and the reference policy for any $s \in \mathcal{S}$ to be

$$\pi_{ref}(-1|s) = 1/C, \quad \pi_{ref}(+1|s) = 1 - 1/C,$$

where $C \geq 1$ is a parameter to be decided later. We consider collections of distributions indexed using the Boolean hypercube $\mathcal{V} = \{-1, +1\}^S$. In particular, for any $\mathbf{v} = (v_1, v_2, \ldots, v_S) \in \mathcal{V}$, the mean function of the reward indexed by $\mathbf{v}$ is defined as

$$r_{\mathbf{v}}(s, -1) = B/2 + v_s \cdot a, \quad r_{\mathbf{v}}(s, +1) = B/2 - b,$$

for any state $s \in \mathcal{S}$, where $a, b \in (0, B/2)$ will be specified later. With this reward function, from Definition 5, the optimal policy $\pi^*_{\mathbf{v}}$ for the KL-regularized RLHF is for any $s \in \mathcal{S}$,

$$\pi^*_{\mathbf{v}}(-1|s) = \frac{\pi_{ref}(-1|s)\exp\left(\beta \cdot r_{\mathbf{v}}(s,-1)\right)}{\pi_{ref}(-1|s)\exp\left(\beta \cdot r_{\mathbf{v}}(s,-1)\right) + \pi_{ref}(+1|s)\exp\left(\beta \cdot r_{\mathbf{v}}(s,+1)\right)} = \frac{\exp(\beta(b + v_s a))}{\exp(\beta(b + v_s a)) + C - 1},$$
$$\pi^*_{\mathbf{v}}(+1|s) = \frac{C-1}{\exp(\beta(b + v_s a)) + C - 1}.$$

Step 2: Verify the single policy concentrability. Following [40], we state the verification for concentrability here for completeness. Set $C^* \geq 2$, $C = C^*$ and $b = \beta^{-1}\log(C-1)$, then for any $s \in \mathcal{S}$,

$$\frac{\pi^*_{\mathbf{v}}(-1|s)}{\pi_{ref}(-1|s)} = C \cdot \frac{\exp(\beta(b + v_s a))}{\exp(\beta(b + v_s a)) + C - 1} = C \cdot \frac{\exp(\beta v_s a)}{1 + \exp(\beta v_s a)} \leq C = C^*,$$
$$\frac{\pi^*_{\mathbf{v}}(+1|s)}{\pi_{ref}(+1|s)} = \frac{C}{C-1} \cdot \frac{1}{1 + \exp(\beta v_s a)} \leq C = C^*.$$

Therefore, we get $\max_{\mathbf{v} \in \mathcal{V}} C^{\pi^*_{\mathbf{v}}} \leq C^*$.

Step 3: Construction of hard-to-distinguish pair for Sub-optimality gap. In order to get the minimax lower bound, since $d_0 = Unif(\mathcal{S})$, we define

$$\text{SubOpt}(\hat{\pi}, \mathbf{v}) = \frac{1}{S}\sum_{s=1}^{S} \text{SubOpt}_s(\hat{\pi}, \mathbf{v}),$$

and, simpler than the analysis in [40], we have the following derivation from sub-optimality gap to the KL divergence between estimated policy and optimal policy:

$$\begin{aligned}
\text{SubOpt}_s(\hat{\pi}, \mathbf{v}) &= \left\langle \pi^*_{\mathbf{v}}(\cdot \mid s), r_{\mathbf{v}}(s,\cdot) - \beta^{-1}\log\frac{\pi^*_{\mathbf{v}}(\cdot \mid s)}{\pi_{\text{ref}}(\cdot \mid s)} \right\rangle - \left\langle \hat{\pi}(\cdot \mid s), r_{\mathbf{v}}(s,\cdot) - \beta^{-1}\log\frac{\hat{\pi}(\cdot \mid s)}{\pi_{\text{ref}}(\cdot \mid s)} \right\rangle \\
&= \frac{1}{\beta}\mathbb{E}_{a \sim \pi^*_{\mathbf{v}}(\cdot|s)}\left[\log\frac{\pi_{\text{ref}}(a|s)\cdot\exp(\beta r_{\mathbf{v}}(s,a))}{\pi^*_{\mathbf{v}}(a \mid s)}\right] - \frac{1}{\beta}\mathbb{E}_{a \sim \hat{\pi}(\cdot|s)}\left[\log\frac{\pi_{\text{ref}}(a|s)\cdot\exp(\beta r_{\mathbf{v}}(s,a))}{\hat{\pi}(a \mid s)}\right] \\
&\overset{(a)}{=} \frac{1}{\beta}\log Z(s) - \frac{1}{\beta}\mathbb{E}_{a \sim \hat{\pi}(\cdot|s)}\left[\log\frac{\pi_{\text{ref}}(a|s)\cdot\exp(\beta r_{\mathbf{v}}(s,a))}{\pi^*_{\mathbf{v}}(a \mid s)} \cdot \frac{\pi^*_{\mathbf{v}}(a \mid s)}{\hat{\pi}(a \mid s)}\right] \\
&\overset{(b)}{=} \frac{1}{\beta}\log Z(s) - \frac{1}{\beta}\log Z(s) + \frac{1}{\beta}\mathbb{E}_{a \sim \hat{\pi}(\cdot|s)}\left[\log\frac{\hat{\pi}(a \mid s)}{\pi^*_{\mathbf{v}}(a \mid s)}\right] \\
&= \frac{1}{\beta}\text{KL}(\hat{\pi} \| \pi^*_{\mathbf{v}}),
\end{aligned}$$

where $(a), (b)$ is from the definition of $\pi^*_{\mathbf{v}}(\cdot|s) = \frac{\pi_{\text{ref}}(\cdot|s)\cdot\exp(\beta r_{\mathbf{v}}(s,\cdot))}{Z(s)}$ and $\mathbb{E}_{a \sim \pi^*_{\mathbf{v}}(\cdot|s)} Z(s) = Z(s) = \mathbb{E}_{a \sim \hat{\pi}(\cdot|s)} Z(s)$ is the normalization constant.

We denote $\mathbf{v} \sim_s \mathbf{v}'$ if $\mathbf{v}, \mathbf{v}' \in \mathcal{V} = \{-1, +1\}^S$ only differ in the s-th element and $\mathbf{v} \sim \mathbf{v}'$ means there exists $s \in \mathcal{S}, \mathbf{v} \sim_s \mathbf{v}'$. By following the equations of (B.10) and (B.11) in Appendix B.4 of [40] and taking $C - 1 = \exp(\beta b)$, for any $s \in \mathcal{S}$, we consider $\mathbf{v} \sim_s \mathbf{v}'$ and obtain

$$\text{SubOpt}_s(\hat{\pi}, \mathbf{v}) + \text{SubOpt}_s(\hat{\pi}, \mathbf{v}') \geq \min\left\{\frac{\beta a^2}{8}, \frac{3a}{10}\right\} .$$

Step 4: LDP mechanism on labels. Let P_r be the distribution of $\left(s, a^1, a^2, z\right)$ for $s \sim d_0, a^1 = -1, a^2 = +1 \overset{\text{i.i.d.}}{\sim} \pi_{\text{ref}}(\cdot \mid s)$, $z = \mathcal{R}(y)$ with LDP randomizer $\mathcal{R}$ and $y \sim \text{Bern}\left(\sigma\left(r\left(s, a^1\right) - r\left(s, a^2\right)\right)\right)$. Note that for the value of the KL divergence the $\{-1, +1\}$ labels are the same as $\{0, 1\}$ labels. Then for $\mathbf{v} \sim \mathbf{v}'$ with $v_s = -v_s'$,

$$\begin{aligned}
&\text{KL}\left(P_{r_{\mathbf{v}}} \| P_{r_{\mathbf{v}'}}\right) \\
&\leq \frac{(C-1)}{SC^2} \sum_{s', a^1, a^2} \left[\text{KL}(\mathcal{R}(y_{\mathbf{v}}) \| \mathcal{R}(y_{\mathbf{v}'})) + \text{KL}(\mathcal{R}(y_{\mathbf{v}'}) \| \mathcal{R}(y_{\mathbf{v}}))\right] \\
&\leq \frac{4(e^\epsilon - 1)^2 (C-1)}{SC^2} \sum_{s', a^1, a^2} \text{TV}^2\left(\text{Bern}\left(\sigma\left(r_{\mathbf{v}}\left(s', a^1\right) - r_{\mathbf{v}}\left(s', a^2\right)\right)\right) \| \, \text{Bern}\left(\sigma\left(r_{\mathbf{v}'}\left(s', a^1\right) - r_{\mathbf{v}'}\left(s', a^2\right)\right)\right)\right) \\
&= \frac{4(e^\epsilon - 1)^2 (C-1)}{SC^2} \text{TV}^2(\text{Bern}(\sigma(b+a)) \| \, \text{Bern}(\sigma(b-a))) \\
&= \frac{4(e^\epsilon - 1)^2 (C-1)}{SC^2} \left(\frac{1}{1 + e^{-(a+b)}} - \frac{1}{1 + e^{a-b}}\right)^2 \\
&\overset{(a)}{\leq} \frac{(e^\epsilon - 1)^2 a^2}{SC},
\end{aligned}$$

where the second inequality is from Lemma A.5 since the offline setting is non-interactive and (a) is from mean-value theorem

$$\left|\sigma(b+a) - \sigma(b-a)\right| \leq \sup_{t \in [b-a, b+a]} \left|\sigma'(t)\right| \cdot \left|(b+a) - (b-a)\right| \leq \frac{1}{4} \cdot 2|a| = \frac{|a|}{2} .$$

Step 5: Minimax lower bound. We evaluate procedures through the minimax suboptimality, which means among all algorithms, pick the one that achieves the smallest possible worst-case suboptimality. From Assouad's lemma in Lemma A.6 and by taking $a = \frac{\sqrt{SC}}{(e^\epsilon - 1)\sqrt{n}}$, $S = \log \mathcal{N}_{\mathcal{F}}(\tau)$, and $C = C^*$, we get

$$\begin{aligned}
\inf_{\hat{\pi} \in \Pi} \sup_{I \in \mathcal{I}} \text{SubOpt}(\hat{\pi}, I) &\geq \frac{1}{4} S \cdot \frac{1}{S} \min\left\{\frac{\beta a^2}{8}, \frac{3a}{10}\right\} \min_{\mathbf{v} \sim \mathbf{v}'} \exp\left(-\text{KL}\left(P_{r_{\mathbf{v}}}^n \| P_{r_{\mathbf{v}'}}^n\right)\right) \\
&= \frac{1}{4} \min\left\{\frac{\beta a^2}{8}, \frac{3a}{10}\right\} \exp\left(-n\text{KL}\left(P_{r_{\mathbf{v}}} \| P_{r_{\mathbf{v}'}}\right)\right) \\
&= \Omega\left(\min\left\{\frac{\beta C S}{(e^\epsilon - 1)^2 n}, \frac{\sqrt{SC}}{(e^\epsilon - 1)\sqrt{n}}\right\}\right) \\
&= \Omega\left(\min\left\{\frac{\beta C^* \log \mathcal{N}_{\mathcal{F}}(\tau)}{(e^\epsilon - 1)^2 n}, \frac{\sqrt{C^* \log \mathcal{N}_{\mathcal{F}}(\tau)}}{(e^\epsilon - 1)\sqrt{n}}\right\}\right) . \quad \square
\end{aligned}$$

C Proofs of Section 5

By direct calculation, it is easy to get the following lemma that will be used in our follow-up analysis.

Lemma C.1. *From the Bernoulli distribution of y in Bradley-Terry model (Assumption 6), we denote $\mathbb{E}_r[y|s, a^1, a^2] = h^*(s, a^1, a^2) = 2\sigma(\Delta_{r^*}(s, a^1, a^2)) - 1$, then based on the randomness of random response, $\mathbb{E}_{RR}[z|s, a^1, a^2] = \widetilde{h}^*(s, a^1, a^2) = (2\alpha - 1) \cdot h^*(s, a^1, a^2)$.*

Proof of Lemma C.1. First, $\mathbb{E}_r[y|s,a^1,a^2] = (+1)\cdot\sigma(\Delta_{r^*}(s,a^1,a^2)) + (-1)\cdot(1-\sigma(\Delta_{r^*}(s,a^1,a^2))) = 2\sigma(\Delta_{r^*}(s,a^1,a^2)) - 1 = h^*(s,a^1,a^2)$. Then, $\mathbb{E}_{RR}[z|s,a^1,a^2] = 1\cdot\mathbb{P}(z=+1|s,a^1,a^2)+(-1)\cdot\mathbb{P}(z=-1|s,a^1,a^2) = \alpha\mathbb{P}(y=+1|s,a^1,a^2)+(1-\alpha)\mathbb{P}(y=-1|s,a^1,a^2) - \alpha\mathbb{P}(y=-1|s,a^1,a^2) - (1-\alpha)\mathbb{P}(y=+1|s,a^1,a^2) = (2\alpha-1)h^*(s,a^1,a^2)$. □

Lemma C.2 (In-sample error of ERM [38, 34, 32])**.** *Consider a function space $\mathcal{H}: \mathcal{Z}\to\mathbb{R}$ and a filtered sequence $\{x_t,\epsilon_t\}\in\mathcal{X}\times\mathbb{R}$ so that ϵ_t is conditional zero-mean σ-sub-Gaussian noise. Suppose that $\mathcal{H}$ is a finite space with cardinality $N_{\mathcal{H}}$. For $h^*(\cdot):\mathcal{Z}\to\mathbb{R}$, suppose that $z_t = h^*(x_t)+\epsilon_t$. If $\widehat{f}_t$ is an ERM solution:*

$$\hat{h}_t = \operatorname*{argmin}_{h\in\mathcal{H}} \sum_{i=1}^{t} \left(h(x_i) - z_i\right)^2,$$

with probability at least $1-\delta$, we have for all $t\in[T]$,

$$\sum_{i=1}^{t}\left(\widehat{h}_t(x_i) - h^*(x_i)\right)^2 \le 8\sigma^2\log\frac{T\cdot N_{\mathcal{F}}}{\delta}.$$

Lemma C.3 (In sample error bound of reward difference)**.** *Under Assumption 3.7, finite reward space $\mathcal{F}$ with cardinality $N_{\mathcal{F}}$, the reward $\bar{r}$ estimated by step 7 in Algorithm 2 satisfies w ith probability at least $1-\delta$, for all $t\in[T]$,*

$$\sum_{i=1}^{t}\left(r^*\left(s_i,a_i^1\right) - r^*\left(s_i,a_i^2\right) - \left[\bar{r}_t\left(s_i,a_i^1\right) - \bar{r}_t\left(s_i,a_i^2\right)\right]\right)^2 \le \frac{8(e^{-B}+2+e^B)^2}{(2\alpha-1)^2}\log\frac{T\cdot N_{\mathcal{F}}}{\delta}.$$

Proof. By the mean value theorem from Lemma C.2 and Lemma A.8 where the noise is from random response with zero-mean 2-sub-Gaussian noise based on Lemma C.1, with probability at least $1-\delta$, we have for all $t\in[T]$

$$\begin{aligned}
\sum_{i=1}^{t}\left(r^*\left(s_i,a_i^1\right) - r^*\left(s_i,a_i^2\right) - \left[\bar{r}_t\left(s_i,a_i^1\right) - \bar{r}_t\left(s_i,a_i^2\right)\right]\right)^2 &\le \frac{(e^{-B}+2+e^B)^2}{4(2\alpha-1)^2}\sum_i(\hat{\widetilde{h}}_t - \widetilde{h}^*)^2 \\
&\le \frac{8(e^{-B}+2+e^B)^2}{(2\alpha-1)^2}\log\frac{T\cdot N_{\mathcal{H}}}{\delta} \\
&\le \frac{8(e^{-B}+2+e^B)^2}{(2\alpha-1)^2}\log\frac{T\cdot N_{\mathcal{F}}}{\delta} \\
&= \frac{1}{2}\Gamma_T^2,
\end{aligned}$$

where the last inequality is since $N_{\mathcal{H}}\le N_{\mathcal{F}}$. □

Lemma C.4. *Under Algorithm 2 and Assumption 3.7, the noises of the random response on labels $\{-1,+1\}$ are zero mean 2-sub-Gaussian, we have with probability $1-\delta$, the optimism event that $\mathcal{E}_t = \{\bar{r}_t(s,a)+b_t(s,a)+c_t(s) - r^*(s,a)\ge 0\}$ holds for any $(s,a)\in\mathcal{S}\times\mathcal{A}$ for all $t\in[T]$ uniformly where $c_t(s) = \mathbb{E}_{b\sim\pi^1_{t+1}}[r^*(s,b)-\bar{r}_t(s,b)]$.*

Proof. For any $(s,a) \in \mathcal{S} \times \mathcal{A}$, we have

$$
\begin{aligned}
&|r^*(s,a) - \bar{r}_t(s,a) - c_t(s)| \\
&\leq \frac{|r^*(s,a) - \bar{r}_t(s,a) - c_t(s)|}{\sqrt{\lambda + \sum_{i=1}^{t} \left(r^*\left(s_i, a_i^1\right) - r^*\left(s_i, a_i^2\right) - \left[\bar{r}_t\left(s_i, a_i^1\right) - \bar{r}_t\left(s_i, a_i^2\right)\right]\right)^2}} \\
&\quad \cdot \sqrt{\lambda + \sum_{i=1}^{t} \left(r^*\left(s_i, a_i^1\right) - r^*\left(s_i, a_i^2\right) - \left[\bar{r}_t\left(s_i, a_i^1\right) - \bar{r}_t\left(s_i, a_i^2\right)\right]\right)^2} \\
&\leq \sup_{r_1, r_2 \in \mathcal{F}_t} \frac{\left|r_1(s,a) - r_2(s,a) - \mathbb{E}_{b \sim \pi_{t+1}^1}[r_1(s,b) - r_2(s,b)]\right|}{\sqrt{\lambda + \sum_{i=1}^{t} \left(r_1\left(s_i, a_i^1\right) - r_1\left(s_i, a_i^2\right) - \left[r_2\left(s_i, a_i^1\right) - r_2\left(s_i, a_i^2\right)\right]\right)^2}} \\
&\quad \cdot \sqrt{\lambda + \sum_{i=1}^{t} \left(r^*\left(s_i, a_i^1\right) - r^*\left(s_i, a_i^2\right) - \left[\bar{r}_t\left(s_i, a_i^1\right) - \bar{r}_t\left(s_i, a_i^2\right)\right]\right)^2} \\
&= U_{\mathcal{F}_t}\left(\lambda, s, a; \mathcal{D}_t; \pi_{t+1}^1\right) \cdot \sqrt{\lambda + \sum_{i=1}^{t} \left(r^*\left(s_i, a_i^1\right) - r^*\left(s_i, a_i^2\right) - \left[\bar{r}_t\left(s_i, a_i^1\right) - \bar{r}_t\left(s_i, a_i^2\right)\right]\right)^2} \\
&\leq U_{\mathcal{F}_t}\left(\lambda, s, a; \mathcal{D}_t; \pi_{t+1}^1\right) \cdot \sqrt{\lambda + \frac{1}{2}\Gamma_T^2} \\
&\leq U_{\mathcal{F}_t}\left(\lambda, s, a; \mathcal{D}_t; \pi_{t+1}^1\right) \cdot \Gamma_T \\
&= b_t(s,a),
\end{aligned}
$$

where the last inequality is from taking $\lambda \leq \frac{1}{2}\Gamma_T^2$. □

Lemma C.5 (Objective Decomposition, Lemma A.1 in [38])**.** *For any $t \in [T]$, conditioning on the uniform optimism event that $\mathcal{E}_t = \{\bar{r}_t(x,a) + b_t(x,a) - r^*(x,a) \geq 0, \forall (x,a) \in \mathcal{X} \times \mathcal{A}\}$ holds, we have*

$$
J\left(\pi^*\right) - J\left(\pi_t\right) \leq \beta \mathbb{E}_{x \sim d_0} \mathbb{E}_{a \sim \pi_t}\left[\left(\bar{r}_{t-1}(s,a) + b_{t-1}(s,a) - r^*(s,a)\right)^2\right].
$$

where $\pi_t = \pi_{(\bar{r}_{t-1} + b_{t-1})(s,a)}$.

Proof of Theorem 5.2. Based on the uniform event that $\cup_{t \in [T]} \mathcal{E}_t$ holds with probability at least $1 - \delta$, and denoting $c_{t-1}(s) = \mathbb{E}_{b \sim \pi_t^1}[r^*(s,b) - \bar{r}_{t-1}(s,b)]$, from Lemma A.7, we have

$$
J(\pi^*) - J(\pi_t^2) = J(\pi^*) - J(\pi_{\bar{r}_{t-1} + b_{t-1}}) = J(\pi^*) - J(\pi_{(\bar{r}_{t-1} + b_{t-1})(s,a) + c_{t-1}(s)}).
$$

From Lemma C.5 for objective decomposition, under the event $\mathcal{E}_t$, we have

$$
J(\pi^*) - J(\pi_t^2) \leq \beta \mathbb{E}_{s \sim d_0} \mathbb{E}_{a \sim \pi_t^2}[(\bar{r}_{t-1}(s,a) + b_{t-1}(s,a) + c_{t-1}(s) - r^*(s,a))^2] \leq 4\beta \mathbb{E}_{s \sim d_0} \mathbb{E}_{a \sim \pi_t^2}[b_{t-1}(s,a)]^2 \,..
$$

where the last inequality is from Lemma C.4.

Thus, we get the cumulative regret bound is

$$
\sum_{t=1}^{T} (J(\pi^*) - J(\pi_t^2)) \leq \sum_{t=1}^{T} 4\beta \mathbb{E}_{s \sim d_0} \mathbb{E}_{a \sim \pi_t^2}[b_{t-1}(s,a)]^2 .
$$

By plugging in $b_t(s,a) = U_{\mathcal{F}_t}\left(\lambda, s, a; \mathcal{D}_t; \pi_{t+1}^1\right) \cdot \Gamma_T$, we get the final result. □

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: We state our contributions in the abstract and in Section 1 of the introduction. We consider private KL-regularized RLHF in both offline and online settings, design algorithms, and derive theoretical guarantees. We also run some experiments to verify our findings.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We assume the preference model is the Bradley-Terry model in Assumption 3.7. And we also discuss that the future work can extend to be a more general preference model in Section 7. For our online case, we discuss in Remark 5.3 in Section 5 that our algorithm is not computationally efficient because of the design of the Euler dimension.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

   Justification: We number and cross-reference all the theorems, lemmas, and corollaries. And we provide complete proofs for the offline setting in Section 4 in Appendix B and for the online setting in Section 5 in Appendix C.

   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: [Yes]

   Justification: Our experimental section (Section 6) specifies all components needed for reproducibility, including dataset sources, preprocessing steps, model architectures, hyperparameter settings, and evaluation metrics. These details ensure that other researchers can reproduce our reported results and verify the main claims without relying on additional resources.

   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: [Yes]

   Justification: The paper relies exclusively on publicly available datasets. We provide the information on the dataset in Section 6 and the link to the dataset in footnote 2.And the implementation code (including training scripts and hyperparameter configurations) will be released in an open-access repository upon acceptance.

   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, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper provides comprehensive details for all experimental settings, including dataset descriptions, data splits, model architectures, optimizer types, hyperparameter choices, and training schedules in Section 6.

   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: [Yes]

Justification: The main experimental results are reported in terms of win rate in Table 1 of Section 6, which is commonly used in RLHF evaluation.

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: [Yes]

   Justification: We used a single AMD MI-200 GPU equipped with 64 GB of VRAM and the information of computer resources is clarified in Section 6.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The research adheres to the NeurIPS Code of Ethics. All experiments are conducted on publicly available datasets and do not involve human subjects or private data collection. Moreover, the proposed framework explicitly incorporates differential privacy mechanisms to enhance data protection and promote responsible AI development. No ethical risks or conflicts of interest are associated with this work.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

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

    Answer: [NA]

    Justification: The paper focuses on theoretical and algorithmic aspects of private reinforcement learning from human feedback and does not involve direct societal deployment or user interaction. Therefore, it does not have immediate societal impacts beyond standard considerations of responsible AI research.

    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: The paper does not involve such risks.

    Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: The paper builds on publicly available preference datasets (HH-RLHF) and a base model (Llama-3.2-1B-Instruct model). We put the link in footnote 2 and footnote 3 in Section 6.

Guidelines:

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

13. **New assets**

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

Answer:[NA]

Justification: The paper does not introduce any new datasets, pretrained models, or other reusable assets. All experiments are conducted using publicly available resources, and no new assets requiring additional documentation are released.

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 research does not involve crowdsourcing, user studies, or any experiments 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 study does not involve any human subjects or crowdsourced participants.

    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 rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not956 involve LLMs as any important, original, or non-standard components.

    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.