Reinforcement Learning from Human Feedback with Active Queries

Anonymous Authors[1]

Abstract

Aligning large language models (LLM) with human preference plays a key role in building modern generative models and can be achieved by reinforcement learning from human feedback (RLHF). Despite their superior performance, current RLHF approaches often require a large amount of human-labelled preference data, which is expensive to collect. In this paper, inspired by the success of active learning, we address this problem by proposing query-efficient RLHF methods. We first formalize the alignment problem as a contextual dueling bandit problem and design an Proximal Policy Optimization with Active Queries (PPO-AQ) algorithm with an $\widetilde{O}(d^2/\Delta)$ regret bound and an $\widetilde{O}(d^2/\Delta^2)$ query complexity, where d is the dimension of feature space and Δ is the sub-optimality gap over all the contexts. We then propose DPO-AQ, a practical version of our algorithm based on direct preference optimization (DPO) and apply it to fine-tuning LLMs. Our experiments show that DPO-AQ, while only making about half of queries for human preference, matches the performance of the state-of-the-art DPO method.

1 Introduction

Recent breakthroughs in large language models (LLM) significantly enhance performance across a diverse range of tasks, including commonsense reasoning, world knowledge, reading comprehension, math, code, and popular aggregated results (Jiang et al., 2023; Touvron et al., 2023; Chiang et al., 2023; Tunstall et al., 2023). In addition to their amazing capabilities in traditional natural language tasks (Gao et al., 2023a; Yuan et al., 2023; Han et al., 2023; Wei et al., 2023), they also demonstrate great potential in responding to human queries (Ouyang et al., 2022). One key step towards building these models is aligning them with human preference, where reinforcement learning from human feedback (RLHF) (Casper et al., 2023; Ouyang et al., 2022; Ziegler et al., 2019; Christiano et al., 2017; Rafailov et al., 2023) is widely employed. Typically, the process of RLHF is described as follows: At each time, the human user prompts the LLM with an instruction. Subsequently, the model generates several candidate responses and queries the users for their preferences. A reward model is then trained on this preference data to mimic human evaluation. The reward model can be parameterized by the language model itself (Rafailov et al., 2023) or by other neural networks (Gao et al., 2023b; Munos et al., 2023). The language models are then updated using reinforcement learning (RL) algorithms such as Proximal Policy Optimization (PPO) (Schulman et al., 2017) to optimize responses that maximize the reward.

Despite the notable success of RLHF in aligning language models with human preferences, its practical implementation often necessitates significant amounts of human-labeled preference data. For instance, the fine-tuning process of `zephyr-7b-beta` through RLHF relies on the utilization of a sizable 62k UltraCat-binarized dataset (Ding et al., 2023). The collection of such a substantial volume of human preference data is both costly and inefficient. Therefore, there exists a pressing need to develop query-efficient RLHF methods for effectively aligning large language models with human preferences.

Following recent theoretical advancements in RLHF (Xiong et al., 2023; Zhu et al., 2023; Sekhari et al., 2023), we formulate the RLHF problem as a contextual dueling bandit problem (Yue et al., 2012; Wu & Liu, 2016; Saha, 2021; Saha & Krishnamurthy, 2022; Saha & Gaillard, 2022; Wu et al., 2023; Di et al., 2023). In this setting, the learner proposes a pair of actions and receives noisy feedback regarding the preference between the dueling pair for each round. While numerous studies address regret minimization in dueling bandits, only Sekhari et al. (2023) have considered query complexity. However, the regret incurred by their algorithm exhibits a linear dependency on the size of the action set $\mathcal{A}$, limiting the practical applicability of their method.

In this paper, we adopt the principles of active learning (Zhang & Oles, 2000; Hoi et al., 2006) to design a query-efficient algorithm, **P**roximal **P**olicy **O**ptimization with

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Active **Q**ueries (PPO-AQ) for linear contextual dueling bandit. In each round, PPO-AQ employs the maximum likelihood estimator (Di et al., 2023) to estimate the underlying parameter and constructs an optimistic estimator for the reward gap between different arms. Subsequently, PPO-AQ selects the best arms and estimates the uncertainty associated with the potential feedback. To reduce the query complexity, PPO-AQ selectively queries the dueling preference and updates the estimation only when the uncertainty of the observation exceeds a threshold.

We further extend PPO-AQ to direct preference optimization (DPO) (Rafailov et al., 2023) and introduce a novel query-efficient method, **DPO** with **A**ctive **Q**uery (DPO-AQ). Following the methodology of PPO-AQ, DPO-AQ selectively queries human preference only for data where the model exhibits high uncertainty about the observation. For data where the model is less uncertain about, we employ the pseudo label predicted by the model to fine-tune the training.

- We propose an active-learning based algorithm PPO-AQ for linear contextual dueling bandit with a global sub-optimal gap. Theoretical analysis shows that our algorithm enjoys a constant regret $\widetilde{O}(d^2/\Delta)$[1]. Meanwhile, our proposed algorithm only requires $\widetilde{O}(d^2/\Delta^2)$ queries in total T rounds, where d is the dimension of the feature mapping, and Δ is the sub-optimal gap. Compared with previous regret bound of $\widetilde{O}(\min\{\sqrt{AT}, A^2d/\Delta\})$ achieved by Sekhari et al. (2023), where A is the size of the action space and T is the total number of rounds, our regret bound is independent of the size of the action space A, which is more favorable in practice [2].
- We propose an active learning-based DPO method, DPO-AQ. We apply our method to train `zephyr-7b-beta` on UltraCat-binarized dataset (Ding et al., 2023). Our experiment shows while DPO-AQ only make about half numbers of queries, the model trained by DPO-AQ outperforms DPO on Open-LLM-Benchmark (Beeching et al., 2023) by a margin of 0.35%

Notation. We employ $[n]$ to denote the set $\{1, \ldots, n\}$. In this work, we use lowercase letters to represent scalars, and denote vectors and matrices by lower and uppercase boldface letters respectively. Given a vector $\mathbf{x} \in \mathbb{R}^d$, we denote the vector's L_2-norm by $\|\mathbf{x}\|_2$. We further define $\|\mathbf{x}\|_{\mathbf{\Sigma}} = \sqrt{\mathbf{x}^\top \mathbf{\Sigma} \mathbf{x}}$ given a positive semidefinite matrix $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$. We use standard asymptotic notations $O(\cdot), \Omega(\cdot), \Theta(\cdot)$, and further use $\widetilde{O}(\cdot)$ to hide polylogarithmic factors other than the number of rounds T. We use $\mathbb{1}\{\cdot\}$ denote the indicator function,

[1] we use notation $\widetilde{\mathcal{O}}(\cdot)$ to hide the log factor other than number of rounds T

[2] In our work, we only focused on the regret of one selected action, which slightly differs from the regret in Sekhari et al. (2023). Further discussion can be found in Section 3.

2 Related Work

2.1 Reinforcement Learning from Human Feedback

Learning from human preference data dates back to Wirth et al. (2017); Christiano et al. (2017) and is recently popularized by generative language models (Achiam et al., 2023; Touvron et al., 2023). This procedure usually takes place after supervised finetuning (SFT). The canonical procedure of aligning with human preference includes two stages: reward modeling and reward maximization (Ouyang et al., 2022; Bai et al., 2022; Munos et al., 2023). Another approach is direct preference optimization (DPO) (Rafailov et al., 2023), which treats the generative models directly as reward models and trains them on preference data. Compared with the first approach, DPO simplifies the aligning process while maintaining its effectiveness. However, both paradigms require a large amount of human preference data. In this work, we follow the DPO approach and study its query-efficient modification.

The empirical success of RLHF also prompts a series of theoretical works, with a predominant focus on the reward maximization stage, modeling this process as learning a dueling bandit (Zhu et al., 2023; Xiong et al., 2023; Sekhari et al., 2023). Among these works, Sekhari et al. (2023) stands out for considering query complexity in the process. However, their regret upper bound is $\widetilde{O}(\min\sqrt{AT\beta}, A^2\beta^2 d/\Delta)$, depending on the size of the action set A, thereby limiting the practical applicability of their algorithm.

2.2 Dueling Bandits

Dueling bandits represent a variant of the multi-armed bandit problem, incorporating preference feedback between two selected arms (Yue et al., 2012). Existing results in this domain generally fall into two categories, shaped by their assumptions about preference probability. The first category of work (Yue et al., 2012; Falahatgar et al., 2017; 2018; Ren et al., 2019; Wu et al., 2022; Lou et al., 2022) assumes a transitivity property for preference probability and focuses on identifying the optimal action. Our work also belongs to this category. The second category of work (Jamieson et al., 2015; Heckel et al., 2018; Saha, 2021; Wu et al., 2023; Dudík et al., 2015; Ramamohan et al., 2016; Balsubramani et al., 2016) focuses on general preferences with various criteria for optimal actions, such as Borda winner and Copeland winner.

Expanding beyond the standard dueling bandit problem, Dudík et al. (2015) was the first to incorporate contextual information into the dueling bandit framework. Subsequently, Saha (2021) studied the K-arm contextual dueling bandit problem and proposed an algorithm with a near-optimal regret guarantee. In order to addressing the challenge of a potentially large action space, Bengs et al. (2022) also considered linear function approximation and extended these

results to the contextual linear dueling bandit problem and obtained a regret guarantee of $\widetilde{O}(d\sqrt{T})$. Recently, Di et al. (2023) introduced a layered algorithm, improving the results to a variance-aware guarantee of $\widetilde{O}(d\sqrt{\sum \sigma_t^2})$, where σ_t^2 denotes the variance of the observed preference in round t.

2.3 Active Learning

To mitigate the curse of label complexity, active learning serves as a valuable approach in supervised learning. The first line of work is pool-based active learning (Zhang & Oles, 2000; Hoi et al., 2006; Gu et al., 2012; 2014; Citovsky et al., 2021). In pool-based active learning, instead of acquiring labels for the entire dataset, the learner strategically selects a batch of the most informative data at each step and exclusively queries labels for this selected data batch. The learner then employs this labeled data batch to update the model. Subsequently, guided by the enhanced model, the learner queries another mini-batch of labels and continues the training process. These steps are iteratively repeated until the model achieves the desired performance level. The strategic selection of informative data in active learning significantly reduces the label complexity for supervised learning. The label complexity of pool-based active learning has been extensively studied by Dasgupta (2005); Dasgupta et al. (2005); Balcan et al. (2006; 2007); Hanneke & Yang (2015); Gentile et al. (2022). On the other hand, selective sampling (a.k.a., online active learning) (Cesa-Bianchi et al., 2005; 2006; 2009; Hanneke & Yang, 2021) is a learning framework that integrates online learning and active learning. In this framework, the algorithm sequentially observes different examples and determines whether to collect the label for the observed example. minimizing the total regret with more collected data. On the application side, there are also lines of works apply selective sampling to specific circumstance of RLHF in robots (Lee et al., 2021b;a; Liang et al., 2022).

Another line of research (Schulze & Evans, 2018; Krueger et al., 2020; Tucker et al., 2023) focused on active reinforcement learning and directly integrates the query cost into the received reward. Krueger et al. (2020) laid the groundwork for active reinforcement learning by introducing a cost c associated with each reward observation and evaluated various heuristic algorithms for active reinforcement learning. Recently, Tucker et al. (2023) studied the multi-arm bandit problem with costly reward observation. Their work not only suggests empirical advantages but also proves an $O(T^{2/3})$ regret guarantee.

3 Preliminaries

In this work, we formulate the RLHF problem as a contextual dueling bandit problem (Saha, 2021; Di et al., 2023). We assume a context set $\mathcal{X}$, and at the beginning of each round, a contextual variable x_t is i.i.d generated from the context set $\mathcal{X}$ with the distribution $\mathcal{D}$. Based on the context x_t, the learner then chooses two actions y_t^1, y_t^2 from the action space $\mathcal{A}$ and determines whether to query the environment for preferences between these actions. If affirmative, the environment generates the preference feedback o_t with the following probability $\mathbb{P}(o_t = 1|x_t, y_t^1, y_t^2) = \sigma\big(r(x_t, y_t^1) - r(x_t, y_t^2)\big)$, where $\sigma(\cdot) : \mathbb{R} \to [0, 1]$ is the link function and $r(\cdot, \cdot)$ is the reward model.

In this work, we consider the linear reward model, e.g., $r(x, y) = \langle \boldsymbol{\theta}^*, \boldsymbol{\phi}(x, y)\rangle$, where $\boldsymbol{\theta}^* \in \mathbb{R}^d$ and $\boldsymbol{\phi} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$ is a known feature mapping. For the sake of simplicity, we use $\boldsymbol{\phi}_t^1, \boldsymbol{\phi}_t^2$ to denote $\boldsymbol{\phi}(x_t, y_t^1), \boldsymbol{\phi}(x_t, y_t^2)$. Additionally, we assume the norm of the feature mapping $\boldsymbol{\phi}$ and the underlying vector $\boldsymbol{\theta}^*$ are bounded.

Assumption 3.1. The linear contextual dueling bandit satisfies the following conditions:

- For any contextual $x \in \mathcal{X}$ and action $y \in \mathcal{A}$, we have $\|\boldsymbol{\phi}(x, y)\|_2 \le L/2$.
- For the unknown environment parameter $\boldsymbol{\theta}^*$, it satisfies $\|\boldsymbol{\theta}^*\|_2 \le B$.

For the link function σ, we make the following assumption, which is commonly employed in the study of generalized linear contextual bandits (Filippi et al., 2010; Di et al., 2023).

Assumption 3.2. The link function σ is differentiable and the corresponding first derivative satisfied

$$\kappa_\sigma \le \dot{\sigma}(\cdot),$$

where $\kappa_\sigma > 0$ is a known constant.

The learning objective is to minimize the cumulative regret defined as:

$$\text{Regret}(T) = \sum_{t=1}^{T} r^*(x_t) - r(x_t, y_t^1),$$

where $r^*(x_t) = r^*(x_t, y_t^*) = \max_{y \in \mathcal{A}} r^*(x_t, y)$ stands for the largest possible reward in context x_t. In comparison with prior studies (Saha, 2021; Di et al., 2023), our regret measure only focuses on the gap between action y_t^1 and the optimal action y_t^*. In the context of RLHF, the model generates multiple candidate responses, and users will choose the most preferable response from the available options. Under this situation, sub-optimality is only associated with the selected response and therefore we focus on the regret from action y_t^1.

To quantify the cost of collecting human-labeled data, we introduce the concept of query complexity $\text{Query}(T)$ for an algorithm, which is the total number of data pairs that require human feedback for preference across the first T rounds. It's worth noting that while some prior work (Tucker et al., 2023) incorporates direct costs c in the reward for

required human feedback and counts it along with the cost of action-taking, in our approach, we distinguish between regret and query complexity as two separate performance metrics for an algorithm.

In addition, we consider the minimal sub-optimality gap (Simchowitz & Jamieson, 2019; Yang et al., 2020; He et al., 2021), which characterizes the difficulty of the bandit problem.

Definition 3.3 (Minimal sub-optimality gap)**.** For each context $x \in \mathcal{X}$ and action $y \in \mathcal{A}$, the sub-optimality gap $\Delta(x, y)$ is defined as

$$\Delta(x, y) = r^*(x) - r(x, y),$$

and the minimal sub-optimality gap is defined as

$$\Delta = \min_{x \in \mathcal{X}, y \in \mathcal{A}} \big\{\Delta(x, y) : \Delta(x, y) \neq 0\big\}.$$

In general, a larger sub-optimality gap Δ between action y and the optimal action y^* implies that it is easier to distinguish between these actions and results in a lower cumulative regret. Conversely, a task with a lower gap Δ indicates that it is more challenging to make such a distinction, leading to a larger regret. In this paper, we assume the minimal sub-optimality gap is strictly positive.

Assumption 3.4. The minimal sub-optimality gap is strictly positive, i.e., $\Delta > 0$.

4 Algorithm

In this section, we introduce our proposed query-efficient method for aligning LLMs. The main algorithm is illustrated in Algorithm 1. At a high level, the algorithm leverages the uncertainty-aware query criterion (Zhang et al., 2023) to issue queries and employs Optimistic Proximal Policy Optimization (OPPO) (Cai et al., 2020; He et al., 2022b) for policy updates. In the sequel, we introduce the key parts of the proposed algorithm.

4.1 Regularized MLE Estimator

For each round $k \in [K]$, we construct the regularized MLE estimator (Filippi et al., 2010; Li et al., 2017) of parameter $\boldsymbol{\theta}^*$ by solving the following equation:

$$\lambda\kappa_\sigma\boldsymbol{\theta} + \sum_{\tau \in \mathcal{C}_{t-1}} \Big(o_\tau - \sigma\big(\langle\boldsymbol{\theta}, \boldsymbol{\phi}_\tau^1 - \boldsymbol{\phi}_\tau^2\rangle\big)\Big)(\boldsymbol{\phi}_\tau^1 - \boldsymbol{\phi}_\tau^2) = \mathbf{0}. \tag{4.1}$$

Compared with previous work on linear dueling bandit (Saha, 2021; Di et al., 2023), only partial of the data requires human-labelled preference and we construct the MLE estimator only with rounds $t \in \mathcal{C}_T$. In addition, the estimation error between $\widehat{\boldsymbol{\theta}}_t$ and $\boldsymbol{\theta}^*$ satisfied

$$\|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_t\|_{\Sigma_{t-1}} \leq \widetilde{O}\big(\sqrt{d \log |\mathcal{C}_T|}/\kappa_\sigma\big).$$

Algorithm 1 Proximal Policy Optimization with Active Queries (PPO-AQ)

Require: Regularization parameter $\lambda > 0$, and B, an upper bound on the ℓ_2-norm of $\boldsymbol{\theta}^*$, confidence radius β, uncertainty threshold Γ, learning rate η

1: Set initial policy $\pi_1(\cdot|\cdot)$ as uniform distribution over the action set $\mathcal{A}$, $\boldsymbol{\Sigma}_0 \leftarrow \lambda\mathbf{I}$, $\mathcal{C}_0 = \emptyset$
2: **for** $t = 1, \ldots, T$ **do**
3: Compute the MLE $\widehat{\boldsymbol{\theta}}_t$ as in (4.1) and observe $\mathcal{A}$, select $y_t^2 \sim \text{Uniform}(\mathcal{A})$
4: Let $\widehat{D}_t(x_t, y) = \min\big\{\langle\widehat{\boldsymbol{\theta}}_t, \boldsymbol{\phi}(x_t, y) - \boldsymbol{\phi}_t^2\rangle + \beta\|\boldsymbol{\phi}(x_t, y) - \boldsymbol{\phi}_t^2\|_{\boldsymbol{\Sigma}_{t-1}^{-1}}, 2\big\}$
5: Let $y_t^1 = \text{argmax}_y \widehat{D}_t(x_t, y)$
6: **if** $\|\boldsymbol{\phi}_t^1 - \boldsymbol{\phi}_t^2\|_{\boldsymbol{\Sigma}_{t-1}} \leq \Gamma$ **then**
7: play the duel (y_t^1, y_t^2) and keep $\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}_{t-1}$
8: $\pi_{t+1}(a|s) = \pi_t(a|s)$, $\mathcal{C}_t = \mathcal{C}_{t-1}$
9: **else**
10: Sample $y_t^1 \sim \pi_t(\cdot|s_t)$ and play the action pair (y_t^1, y_t^2), $\mathcal{C}_t = \mathcal{C}_{t-1} \cup \{t\}$
11: Query for the preference and observe o_t
12: Update $\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}_{t-1} + (\boldsymbol{\phi}_t^1 - \boldsymbol{\phi}_t^2)(\boldsymbol{\phi}_t^1 - \boldsymbol{\phi}_t^2)^\top$
13: Update $\pi_{t+1}(y|x) \propto \pi_t(y|x) \exp\big(\eta\widehat{D}_t(y, x)\big)$
14: **end if**
15: **end for**

After constructing the estimator $\widehat{\boldsymbol{\theta}}_t$, the agent first selects a baseline action y_t^2 and compares each action $y \in \mathcal{A}$ with the baseline action y_t^2. For simplicity, we denote $D_t(x_t, y) = \langle\boldsymbol{\theta}^*, \boldsymbol{\phi}(x_t, y) - \boldsymbol{\phi}_t^2\rangle$ as the reward gap between y and action y_t^2. Then, we construct an optimistic estimator $\widehat{D}_t$ for the reward gap with linear function approximation and Upper Confidence Bound (UCB) bonus, i.e.,

$$\widehat{D}_t(x_t, y) \approx \langle\widehat{\boldsymbol{\theta}}_t, \boldsymbol{\phi}(x_t, y) - \boldsymbol{\phi}_t^2\rangle + \beta\|\boldsymbol{\phi}(x_t, y) - \boldsymbol{\phi}_t^2\|_{\boldsymbol{\Sigma}_{t-1}^{-1}}.$$

With the help of UCB bonus, we can show that our estimated reward gap $\widehat{D}_t$ is an upper bound of the true reward gap D_t.

4.2 Uncertainty-Aware Query Criterion

To mitigate the expensive costs from collecting human feedback, we introduce the uncertainty-based criterion (Line 1) (Zhang et al., 2023) to decide whether a pair of action requires y_t^1 and y_t^2 requires human-labelled preference. Similar criterion has also been used in corruption-robust linear contextual bandits (He et al., 2022c) and achieving nearly minimax optimal regret in learning linear (mixture) Markov decision processes (Zhou & Gu, 2022; He et al., 2022a; Zhao et al., 2023). Intuitively speaking, the UCB bonus $\beta\|\boldsymbol{\phi}_t^1 - \boldsymbol{\phi}_t^2\|_{\boldsymbol{\Sigma}_{t-1}^{-1}}$ captures the uncertainty associated with the preference feedback o_t. For the action pair (y_t^1, y_t^2) with low uncertainty, where the observation is nearly known and provides minimal information, we select the action y_t^1, y_t^2 without querying human preference feedback. In this situation, the policy $\pi(\cdot|\cdot)$ remains unchanged as there is no

observation in this round. By employing the uncertainty-based data selection rule, we will later prove that the query complexity is effectively controlled by:

$$\textbf{Query}(T) = |\mathcal{C}_T| \leq \widetilde{O}(d^2/\Delta^2).$$

4.3 Proximal Policy Optimization

In cases where the action pair (y_t^1, y_t^2) exhibits high uncertainty and the uncertainty-aware query criterion is triggered, the agent resample the action y_t^1 from policy π_t and queries human feedback for the duel y_t^1, y_t^2. Upon observing the preference o_t, this round is then added to the dataset $\mathcal{C}_t$. Subsequently, the policy π_{t+1} is updated using the Optimistic Proximal Policy Optimization (PPO) method (Cai et al., 2020; He et al., 2022b), i.e.,

$$\pi_{t+1}(y|x) \propto \pi_t(y|x) \exp\big(\eta \widehat{D}_t(y, x)\big).$$

In an extreme case where the uncertainty threshold Γ is chosen to be 0, the uncertainty-aware query criterion will always be triggered. Under this situation, Algorithm 1 will query the human-labeled preference for each duel (y_t^1, y_t^2), and Algorithm 1 will degenerate to the dueling bandit version of OPPO (Cai et al., 2020). Under this situation, Algorithm 1 enjoys $\widetilde{O}(d\sqrt{T})$ regret while having a linear query complexity with respect to the number of rounds T.

5 Theoretical Analysis

In this section, we present our main theoretical results.

Theorem 5.1. Let Δ be the minimal sub-optimal gap in Assumption 3.4. If we set the parameters $\Gamma = \widetilde{O}(\Delta/\sqrt{d})$, $\lambda = B^{-2}$, $\eta = \widetilde{O}(\sqrt{\Gamma^2 \log \mathcal{A}/d})$, and $\beta = \widetilde{O}(\sqrt{d}/\kappa_\sigma)$ in Algorithm 1, then with with probability at least $1-\delta$, the regret for Algorithm 1 across the first T rounds is upper bounded by

$$\text{Regret}(T) = \widetilde{O}(d^2/\Delta).$$

In addition, the query complexity of Algorithm 1 is upper bounded by:

$$|\mathcal{C}_T| = \widetilde{O}(d^2/\Delta^2).$$

Remark 5.2. Theorem 5.1 suggests that our algorithm achieves a constant level of regret and query complexity respect to the number of rounds T. It is important to highlight that our algorithm requires a prior knowledge of the sub-optimal gap Δ. Under the situation where Δ is unknown, the learner can set the parameter Δ via grid search process.

Remark 5.3. In comparison to the regret $\widetilde{O}(A^2 d/\Delta)$ obtained by the AURORA algorithm (Sekhari et al., 2023)[3], our algorithm's regret eliminates the dependency of the action space A. Moreover, we achieve an improvement in the query complexity by a factor of A^3.

[3] In our work, we only focused on the regret of one selected action, which slightly differs from the regret in Sekhari et al. (2023).

Algorithm 2 Direct Preference Optimization with Active Queries (DPO-AQ)

Require: Regularization parameter β, uncertainty threshold Γ, learning rate η, initial model parameter $\boldsymbol{\theta}_0$, batch size H

Set $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}_0$
for $t = 1, \ldots, T$ **do**
  Receive batch of data $\mathcal{B}_t = \{x_i, y_i^1, y_i^2\}_{i=1}^H$
  for $i = 1, \ldots, H$ **do**
    Set the confidence $C_{\boldsymbol{\theta}}(x_i, y_i^1, y_i^2)$ as in (6.2)
    if $C_{\boldsymbol{\theta}}(x_i, y_i^1, y_i^2) \leq \Upsilon$ **then**
      Query for the human label and set o_i as the queried preference.
    else
      Set $o_i \leftarrow \text{sign}\big(r_{\boldsymbol{\theta}}(x, y^1) - r_{\boldsymbol{\theta}}(x, y^2)\big)$
    end if
  end for
  Update $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{B}_t}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}_0})$
end for

6 Practical Algorithm

In this section, we introduce the practical version of our proposed algorithm, namely direct preference optimization with active queries (DPO-AQ) and summarized it in Algorithm 2. Inspired from Algorithm 1 and direct preference optimization (DPO) (Rafailov et al., 2023), DPO-AQ is an active-learning-based direct preference optimization. In this work, we consider the Bradley-Terry-Luce model (Bradley & Terry, 1952), where the link function has the form $\sigma(x) = 1/(1+e^{-x})$. At a high level, our proposed method follows the basic idea of Algorithm 1 and sets an uncertainty threshold to filter out informative training samples. However, adapting our algorithm to neural network training requires several key modifications as below.

Direct Preference Optimization. We follow the framework of DPO (Rafailov et al., 2023) for alignment training. In DPO, the reward is directly parameterized by the language model in the following ways[4]:

$$r_{\boldsymbol{\theta}}(x, y) = \beta\big(\log \pi_{\boldsymbol{\theta}}(y|x) - \log \pi_{\text{ref}}(y|x)\big),$$

where β is a regularization parameter, y is a potential answer given the prompt x and π_{ref} is the reference model, which is usually the SFT checkpoint in practice. With the reward model parameterized above, we have the following DPO training objective:

$$\mathcal{L}_{\text{DPO}}(\pi_{\boldsymbol{\theta}}, \pi_{\text{ref}})$$

[4] This definition is an equivalent one of the definition in Rafailov et al. (2023)

$$= -\mathbb{E}_{(x,y^1,y^2,o)\sim\mathcal{D}}\Big[\log\sigma\Big(o\cdot\big(r_{\boldsymbol{\theta}}(x,y^1)-r_{\boldsymbol{\theta}}(x,y^2)\big)\Big)\Big],$$

where $\mathcal{D}$ is the data distribution, y^1 and y^2 are the two answers given the prompt x, and o is the human preference such that $o = 1$ indicates a preference for y^1, and $o = -1$ indicates a preference for y^2. This approach bypasses the reward-model training process in standard RLHF and, therefore, eliminates the reward noise introduced in it. We follow this approach and treat the language model as the reward model and simultaneously as the agent.

Empirical Confidence Estimator. Inspired by Algorithm 1, we require a confidence-based criteria to filter out those samples that are not that informative. However, in real applications, rewards are no longer necessarily parameterized by a linear function. Thus, the uncertainty estimator cannot be directly transferred to empirical cases. Since the model is essentially predicting the probability of human preference labels, that is:

$$\mathbb{P}(o=1|x,y^1,y^2)=\sigma(r_{\boldsymbol{\theta}}(x,y^1)-r_{\boldsymbol{\theta}}(x,y^2)), \quad (6.1)$$

where o stands for the preference label and $r_{\boldsymbol{\theta}}$ is the reward model. We can use the reward model's predicted probability as its uncertainty. Specifically, if the probability is near 0 or 1, the models are highly confident in the human's preference. Therefore, we define the following function $C_{\boldsymbol{\theta}}$:

$$C_{\boldsymbol{\theta}}(x,y^1,y^2)=|r_{\boldsymbol{\theta}}(x,y^1)-r_{\boldsymbol{\theta}}(x,y^2)|, \quad (6.2)$$

as the confidence level of the model.

Training Objectives. We borrow ideas from previous active learning literature. For given answer pairs, if the model is very confident in its preference label, we then use the preference label predicted by the model (i.e., pseudo label) for training. To be specific, given a prompt x and the corresponding answers y^1 and y^2, the predicted preference label can be defined as follows:

$$o_{\boldsymbol{\theta}}(x,y^1,y^2) = \begin{cases} o & \text{if } C_{\boldsymbol{\theta}}(x,y^1,y^2)\le\Upsilon \\ \text{sign}\big(r_{\boldsymbol{\theta}}(x,y^1)-r_{\boldsymbol{\theta}}(x,y^2)\big) & \text{if } C_{\boldsymbol{\theta}}(x,y^1,y^2)>\Upsilon \end{cases}, \quad (6.3)$$

where o is the human preference upon query, $\text{sign}(z)$ is the signal of z and Υ is the confidence threshold (close to the threshold Γ in PPO-AQ). With the predicted preference labels of given prompts and answers, now we can formulate our training objective as the follows:

$$\mathcal{L}_{\mathcal{D}}(\pi_{\boldsymbol{\theta}},\pi_{\text{ref}}) = -\mathbb{E}_{(x,y^1,y^2)\sim\mathcal{D}}\big[\log\sigma\big(\Lambda_{\boldsymbol{\theta}}(x,y^1,y^2)\big)\big], \quad (6.4)$$

where

$$\Lambda_{\boldsymbol{\theta}}(x,y^1,y^2)=o_{\boldsymbol{\theta}}(x,y^1,y^2)\cdot\big(r_{\boldsymbol{\theta}}(x,y^1)-r_{\boldsymbol{\theta}}(x,y^2)\big).$$

To make our approach more time efficient in practice, we follow the standard approach in DPO and use mini-batch gradient descent to update the parameters of our model. At each time step, we feed the model with a batch of data $\{(x_i,y_i^1,y_i^2)\}_{i=1}^H$. We then compute the pseudo labels and update the model parameters by one-step gradient descent.

7 Experiments

In this section, we conducted extensive experiments to verify the effectiveness of our proposed method. Our experiments reveal that DPO-AQ performs than DPO while requiring only half the queries. Additionally, our ablation studies show that involving pseudo-labels plays a key role in the training process.

7.1 Experimental Setup

Models and Dataset. We start from the base model `Zephry-7b-sft-full`[5], which is supervised-finetuned from Mistral-7B (Jiang et al., 2023) model on the SFT dataset Ultrachat-200k (Ding et al., 2023). It has not been trained to align with any human preference. We adopt the human-preference dataset Ultrachat-binarized (Ding et al., 2023) for alignment. Ultrachat-binarized contains about 62k prompts, and each prompt corresponds to two candidate answers with human preference labels.

Baseline Method and Hyper-parameters. We use the standard direct preference optimization as our baseline method and follow the implementation alignment handbook codebase[6]. We use LoRA (Hu et al., 2021) to fine-tune the model. For both DPO and our method, we select the learning rate to be 1e-5 and train for one epoch. We fix other hyper-parameters the same as the original implementation and choose the uncertainty threshold $\Gamma = 1.5$ for DPO-AQ.

Evaluation. We utilize the widely used Huggingface Open LLM Leaderboard (Beeching et al., 2023) for our evaluation. This benchmark contains a bunch of different datasets covering a wide range of tasks, offering a comprehensive assessment of different aspects of large language models. Specifically, the tasks involved in the benchmark include commonsense reasoning (Arc (Clark et al., 2018), HellaSwag (Zellers et al., 2019), Winogrande (Sakaguchi et al., 2021)), multi-task language understanding (MMLU(Hendrycks et al., 2020)), human falsehood mimic (TruthfulQA (Lin et al., 2021)) and math problem solving (GSM8k (Cobbe et al., 2021)). During the evaluation process, we follow the standard approach and provide the language models with a few in-context examples. Besides the scores for each dataset, we also provide the average

[5] https://huggingface.co/alignment-handbook/zephyr-7b-sft-full

[6] https://github.com/huggingface/alignment-handbook

Table 1. Detailed information of HuggingFace Open LLM Leaderboard. Here we use WG to denote Winogrande and HS for HellaSwag. For each evaluation dataset, we present the number of few-shot examples and metric adopted for evaluation.

Datasets	Arc	TruthfulQA	WG	GSM8k	HS	MMLU
# few-shot	25	0	5	5	10	5
Metric	`acc_norm`	`mc2`	`acc`	`acc`	`acc_norm`	`acc`

score across all datasets. Please refer to Table 1 for detailed information about the metrics and few-shot numbers we use.

7.2 Experimental Results

Performance of DPO-AQ. We count the queries made through the training process of DPO-AQ. Besides the model trained with the whole dataset (denoted by DPO), we picked the intermediate checkpoint at step 550 (denoted by "DPO (550)"). This checkpoint has made about 35.7k queries, the most similar to the 32k used by DPO-AQ. We compare the performance of the three models. The results are presented in Table 2. Examining the table, we observe that while both DPO and DPO-AQ improve the average score, DPO-AQ improves the model's average score from 57.29 to 59.01, which is much higher than 58.66 achieved by DPO and 58.60 by DPO (550). In terms of label usage, DPO-AQ requires only about 32k queries, which is only 52% of the usage by DPO. Specifically, on the Arc, TruthfulQA, and HellaSwag datasets, our method outperforms our baselines, achieving scores of 61.26, 45.52, and 83.21, respectively. However, we also observed that DPO-AQ shows inferior performances on Winogrande, GSM8k, and MMLU datasets. An interesting finding is that on all these datasets, DPO (550) outperforms DPO. We will discuss this phenomenon in detail in the next paragraph.

Query Efficiency. To further demonstrate the query efficiency for DPO-AQ, we plot the accuracy curve for DPO-AQ and the baseline as the number of queries increases across all datasets, along with the average score. The curves for ARC, HellaSwag, TruthfulQA, Winogrande, GSM8k, and the average score are depicted in Figure 1. The results show that, in terms of ARC, HellaSwag, TruthfulQA and the average score, the improvement of DPO slows down after training on 20k samples. After training on 40k samples, the overall performance even begins to stagnate. In contrast, the performance of DPO-AQ enjoys a faster improvement after training with the first 10k samples, and quickly reaches the peak with only 1/3 to 1/2 of the total queries. This suggests that DPO-AQ can effectively select the most informative data to query the preference labels. Recall that DPO-AQ cannot outperform DPO on Winogrande and MMLU. From Figure 1(d) and Figure 1(e), we observe an apparent training instability in both DPO and DPO-AQ. Therefore, the failure of our method on these two datasets can be attributed to the unstable training dynamic.

7.3 Ablation Study

In this section, we study some important parts that might play central roles in DPO-AQ. We first empirically study the effect of uncertainty threshold in the practice. We also conduct experiments to demonstrate the impact of involving pseudo labels in the training process.

Values of Confidence Threshold. We first study the impact of different confidence thresholds. We varies the value of Υ to 1.0, 1.3, 1.5, and 1.8. For each Υ, we count the preference label used by the models and evaluate the trained models on the Open LLM Benchmark. As shown in Table 3, when the confidence threshold is small, with more predicted labels, these models perform better on the TruthfulQA dataset. However, these model performs poorly on datasets like GSM8k. On the other hand, when the confidence threshold goes larger, the models are making more queries, and the performance patterns become closer to the DPO baseline. Another observation is that when $\Upsilon \geq 1.5$, DPO-AQ consistently outperforms the DPO baseline, which implies that DPO-AQ is not very sensitive to the uncertainty threshold and an coarse grid search of confidence threshold can introduce a fairly good performance.

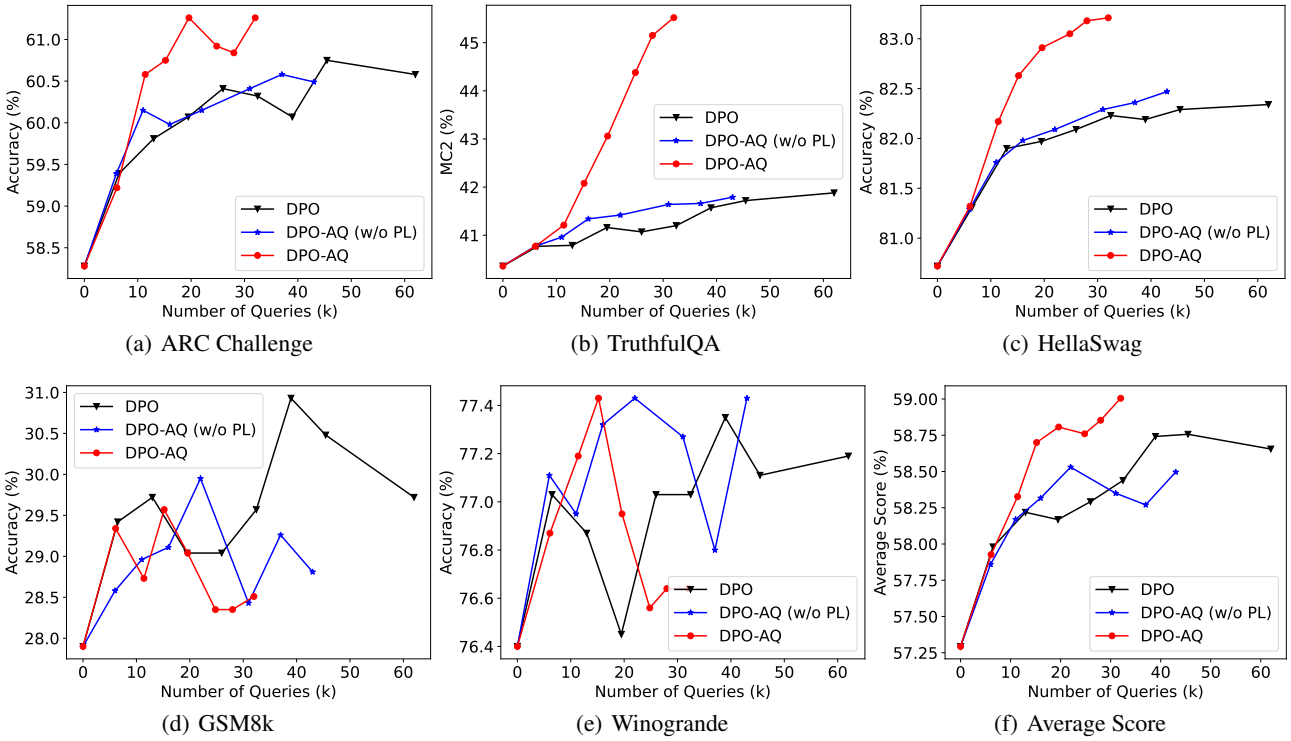
Pseudo Labels v.s. Without Pseudo Labels. An alternative to active learning is to directly follow algorithm 1 and simply neglect those training data with high confidence. Since neglected samples will not affect the loss and the corresponding gradient, we set the label to 0 so that they will not contribute to $\nabla_{\boldsymbol{\theta}}\mathcal{L}$ in Eq. (6.4) during the learning process. Therefore, we formally define the label $o'_{\boldsymbol{\theta}}$ as follows:

$$o'_{\boldsymbol{\theta}}(x, y^1, y^2) = \begin{cases} o & \text{if } C_{\boldsymbol{\theta}}(x, y^1, y^2) \leq \Upsilon \\ 0 & \text{if } C_{\boldsymbol{\theta}}(x, y^1, y^2) > \Upsilon \end{cases}.$$

We keep the remaining part of our method the same and denote this method as "DPO-AQ (w/o PL)". In this experiment, we also conduct a grid search for the confidence threshold Υ and finally pick Υ to be 0.8, 1.0, and 1.2. The performances of the trained models are shown in Table 4. We also plot the training curve in Figure 1. The results show that, without pseudo-labels, the performance suffers from a significant downgrade in average score compared to DPO-AQ. The training curves further indicate that, without pseudo labels, the training dynamics are much more similar to vanilla DPO. These results show the crucial role of pseudo-labels in the

Table 2. Result on Open LLM LeaderBoard. DPO-AQ significantly outperforms DPO on three of the datasets and the average score while makes 32k queries, which is about only half of the queries made by DPO. With the similar amount of queries, DPO-AQ also outperforms DPO (550).

Models	Arc	TruthfulQA	WG	GSM8k	HS	MMLU	Average Score	# Queries
SFT-full	58.28	40.36	76.4	27.9	80.72	60.1	57.29	0
DPO (550)	60.24	41.41	**77.27**	**30.17**	82.28	**60.23**	58.60	35.7k
DPO	60.58	41.88	77.19	29.72	82.34	60.22	58.66	62k
DPO-AQ	**61.26**	**45.52**	76.64	28.51	**83.21**	58.89	**59.01**	32k



(a) ARC Challenge (b) TruthfulQA (c) HellaSwag

(d) GSM8k (e) Winogrande (f) Average Score

Figure 1. The test accuracy curve of DPO, DPO-AQ (w/o PL) and DPO-AQ. The x-axis is the number of queries and the y-axis is the metric for corresponding dataset. Compared to DPO and DPO-AQ (w/o PL), DPO-AQ enjoys a faster performance improvement and a higher performance upper bound. We also observe an unstability in the training curve of GSM8k and Winogrande

Table 3. The effect of confidence threshold in practice setting. We vary the value of Υ and report the evaluation results. When Υ is increasing, DPO-AQ made more queries and the performance pattern is getting closer to DPO.

Method	Υ	Arc	TruthfulQA	WG	GSM8k	HS	MMLU	Average Score	# Queries
DPO	-	60.58	41.88	77.19	29.72	82.34	60.22	58.66	35.7k
DPO-550	-	60.24	41.41	77.27	30.17	82.28	60.23	58.60	62k
DPO-AQ	1.0	61.01	48.41	76.64	20.39	83.35	58.89	58.12	16k
	1.3	61.43	47.56	76.24	24.41	83.48	58.44	58.59	24k
	1.5	61.26	45.52	76.64	28.51	83.21	58.89	59.01	32k
	1.8	60.92	43.2	77.35	29.26	82.13	59.94	58.80	43k

active learning process.

8 Conclusion and Future Work

In this work, we considered query-efficient methods for aligning LLMs with human preference. We first formu-

Table 4. The effect of Pseudo-labels. The numbers in the parentheses stand for the corresponding confidence threshold Υ. DPO-AQ performs better than DPO-AQ (w/o PL) in terms of average scores.

Model	Arc	TruthfulQA	WG	GSM8k	HS	MMLU	Average Score	# Queries
DPO-AQ (w/o PL) (0.8)	60.49	41.39	77.35	**29.26**	82.23	60.01	58.46	38k
DPO-AQ (w/o PL) (1.0)	60.49	41.62	**77.43**	28.73	82.32	**60.02**	58.44	40k
DPO-AQ (w/o PL) (1.2)	60.49	41.79	**77.43**	28.81	82.47	59.99	58.50	43k
DPO-AQ	**61.26**	**45.52**	76.64	28.51	**83.21**	58.89	**59.01**	**32k**

late the problem as a contextual dueling bandit. Under linear reward and sub-optimal gap assumption, we proposed an active-learning based algorithm. Our theoretical analysis shows that our algorithm enjoys an constant regret upper bound and query complexity. We then adapt our algorithm to direct preference optimization and proposed a query efficient DPO method. Experiment results show that our method outperforms DPO with only a half demanding on human preference labels. Despite the good performance DPO-AQ achieves, since it use DPO as the framework of our practical method, our theoretical analysis of PPO-AQ cannot directly be applied to DPO-AQ. We leave the theoretical analysis of DPO-AQ as our future work.

Impact Statements

This paper studies aligning LLMs with human preference in a query-efficient manner. We believe that this topic has the following social impacts. First, LLM-based chatbots have shown great capacity as AI assistants and they are now relied by more and more people. The key step towards building helpful AI assistant is aligning LLMs with human ethics and preferences. Second, aligning LLMs requires a large number of human preference labels, which means huge human labor and material resources. In this paper, We propose a query-efficient method to align LLMs with human preference. Our experiments show that our method can better align LLMs with human preference with much fewer queries for human preferences. Therefore, we believe that our method can potentially address these problems.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Balcan, M.-F., Beygelzimer, A., and Langford, J. Agnostic active learning. In *Proceedings of the 23rd international conference on Machine learning*, pp. 65–72, 2006.

Balcan, M.-F., Broder, A., and Zhang, T. Margin based active learning. In *International Conference on Computational Learning Theory*, pp. 35–50. Springer, 2007.

Balsubramani, A., Karnin, Z., Schapire, R. E., and Zoghi, M. Instance-dependent regret bounds for dueling bandits. In *Conference on Learning Theory*, pp. 336–360. PMLR, 2016.

Beeching, E., Fourrier, C., Habib, N., Han, S., Lambert, N., Rajani, N., Sanseviero, O., Tunstall, L., and Wolf, T. Open llm leaderboard. https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard, 2023.

Bengs, V., Saha, A., and Hüllermeier, E. Stochastic contextual dueling bandits under linear stochastic transitivity models. In *International Conference on Machine Learning*, pp. 1764–1786. PMLR, 2022.

Bradley, R. A. and Terry, M. E. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

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

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, learning, and games*. Cambridge university press, 2006.

Cesa-Bianchi, N., Lugosi, G., and Stoltz, G. Minimizing regret with label efficient prediction. *IEEE Transactions on Information Theory*, 51(6):2152–2162, 2005.

Cesa-Bianchi, N., Gentile, C., Zaniboni, L., and Warmuth, M. Worst-case analysis of selective sampling for linear classification. *Journal of Machine Learning Research*, 7 (7), 2006.

Cesa-Bianchi, N., Gentile, C., and Orabona, F. Robust bounds for classification via selective sampling. In *Proceedings of the 26th annual international conference on machine learning*, pp. 121–128, 2009.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., and Xing, E. P. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Citovsky, G., DeSalvo, G., Gentile, C., Karydas, L., Rajagopalan, A., Rostamizadeh, A., and Kumar, S. Batch active learning at scale. *Advances in Neural Information Processing Systems*, 34:11933–11944, 2021.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafjord, O. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Dasgupta, S. Coarse sample complexity bounds for active learning. *Advances in neural information processing systems*, 18, 2005.

Dasgupta, S., Kalai, A. T., and Monteleoni, C. Analysis of perceptron-based active learning. In *International conference on computational learning theory*, pp. 249–263. Springer, 2005.

Di, Q., Jin, T., Wu, Y., Zhao, H., Farnoud, F., and Gu, Q. Variance-aware regret bounds for stochastic contextual dueling bandits. *arXiv preprint arXiv:2310.00968*, 2023.

Ding, N., Chen, Y., Xu, B., Qin, Y., Zheng, Z., Hu, S., Liu, Z., Sun, M., and Zhou, B. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*, 2023.

Dudík, M., Hofmann, K., Schapire, R. E., Slivkins, A., and Zoghi, M. Contextual dueling bandits. *ArXiv*, abs/1502.06362, 2015.

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

Falahatgar, M., Jain, A., Orlitsky, A., Pichapati, V., and Ravindrakumar, V. The limits of maxing, ranking, and preference learning. In *International conference on machine learning*, pp. 1427–1436. PMLR, 2018.

Filippi, S., Cappe, O., Garivier, A., and Szepesvári, C. Parametric bandits: The generalized linear case. *Advances in Neural Information Processing Systems*, 23, 2010.

Gao, J., Zhao, H., Yu, C., and Xu, R. Exploring the feasibility of chatgpt for event extraction. *arXiv preprint arXiv:2303.03836*, 2023a.

Gao, L., Schulman, J., and Hilton, J. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pp. 10835–10866. PMLR, 2023b.

Gentile, C., Wang, Z., and Zhang, T. Fast rates in pool-based batch active learning. *arXiv preprint arXiv:2202.05448*, 2022.

Gu, Q., Zhang, T., Han, J., and Ding, C. Selective labeling via error bound minimization. *Advances in neural information processing systems*, 25, 2012.

Gu, Q., Zhang, T., and Han, J. Batch-mode active learning via error bound minimization. In *UAI*, pp. 300–309, 2014.

Han, R., Peng, T., Yang, C., Wang, B., Liu, L., and Wan, X. Is information extraction solved by chatgpt? an analysis of performance, evaluation criteria, robustness and errors. *arXiv preprint arXiv:2305.14450*, 2023.

Hanneke, S. and Yang, L. Minimax analysis of active learning. *J. Mach. Learn. Res.*, 16(1):3487–3602, 2015.

Hanneke, S. and Yang, L. Toward a general theory of online selective sampling: Trading off mistakes and queries. In *International Conference on Artificial Intelligence and Statistics*, pp. 3997–4005. PMLR, 2021.

He, J., Zhou, D., and Gu, Q. Logarithmic regret for reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, pp. 4171–4180. PMLR, 2021.

He, J., Zhao, H., Zhou, D., and Gu, Q. Nearly minimax optimal reinforcement learning for linear markov decision processes. *arXiv preprint arXiv:2212.06132*, 2022a.

He, J., Zhou, D., and Gu, Q. Near-optimal policy optimization algorithms for learning adversarial linear mixture mdps. In *International Conference on Artificial Intelligence and Statistics*, pp. 4259–4280. PMLR, 2022b.

He, J., Zhou, D., Zhang, T., and Gu, Q. Nearly optimal algorithms for linear contextual bandits with adversarial corruptions. *Advances in Neural Information Processing Systems*, 35:34614–34625, 2022c.

Heckel, R., Simchowitz, M., Ramchandran, K., and Wainwright, M. Approximate ranking from pairwise comparisons. In *International Conference on Artificial Intelligence and Statistics*, pp. 1057–1066. PMLR, 2018.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Hoi, S. C., Jin, R., Zhu, J., and Lyu, M. R. Batch mode active learning and its application to medical image classification. In *Proceedings of the 23rd international conference on Machine learning*, pp. 417–424, 2006.

Hu, E. J., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021.

Jamieson, K., Katariya, S., Deshpande, A., and Nowak, R. Sparse dueling bandits. In *Artificial Intelligence and Statistics*, pp. 416–424. PMLR, 2015.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Krueger, D., Leike, J., Evans, O., and Salvatier, J. Active reinforcement learning: Observing rewards at a cost. *arXiv preprint arXiv:2011.06709*, 2020.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Lee, K., Smith, L., and Abbeel, P. Pebble: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training. *arXiv preprint arXiv:2106.05091*, 2021a.

Lee, K., Smith, L., Dragan, A., and Abbeel, P. B-pref: Benchmarking preference-based reinforcement learning. *arXiv preprint arXiv:2111.03026*, 2021b.

Li, L., Lu, Y., and Zhou, D. Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning*, pp. 2071–2080. PMLR, 2017.

Liang, X., Shu, K., Lee, K., and Abbeel, P. Reward uncertainty for exploration in preference-based reinforcement learning. *arXiv preprint arXiv:2205.12401*, 2022.

Lin, S., Hilton, J., and Evans, O. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.

Lou, H., Jin, T., Wu, Y., Xu, P., Gu, Q., and Farnoud, F. Active ranking without strong stochastic transitivity. *Advances in neural information processing systems*, 35:297–309, 2022.

Munos, R., Valko, M., Calandriello, D., Azar, M. G., Rowland, M., Guo, Z. D., Tang, Y., Geist, M., Mesnard, T., Michi, A., et al. Nash learning from human feedback. *arXiv preprint arXiv:2312.00886*, 2023.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions

with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023.

Ramamohan, S., Rajkumar, A., and Agarwal, S. Dueling bandits: Beyond condorcet winners to general tournament solutions. In *NIPS*, 2016.

Ren, W., Liu, J. K., and Shroff, N. On sample complexity upper and lower bounds for exact ranking from noisy comparisons. *Advances in Neural Information Processing Systems*, 32, 2019.

Saha, A. Optimal algorithms for stochastic contextual preference bandits. *Advances in Neural Information Processing Systems*, 34:30050–30062, 2021.

Saha, A. and Gaillard, P. Versatile dueling bandits: Best-of-both world analyses for learning from relative preferences. In *International Conference on Machine Learning*, pp. 19011–19026. PMLR, 2022.

Saha, A. and Krishnamurthy, A. Efficient and optimal algorithms for contextual dueling bandits under realizability. In *International Conference on Algorithmic Learning Theory*, pp. 968–994. PMLR, 2022.

Sakaguchi, K., Bras, R. L., Bhagavatula, C., and Choi, Y. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Schulze, S. and Evans, O. Active reinforcement learning with monte-carlo tree search. *arXiv preprint arXiv:1803.04926*, 2018.

Sekhari, A., Sridharan, K., Sun, W., and Wu, R. Contextual bandits and imitation learning via preference-based active queries. *arXiv preprint arXiv:2307.12926*, 2023.

Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Simchowitz, M. and Jamieson, K. G. Non-asymptotic gap-dependent regret bounds for tabular mdps. In *Advances in Neural Information Processing Systems*, pp. 1153–1162, 2019.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Tucker, A. D., Biddulph, C., Wang, C., and Joachims, T. Bandits with costly reward observations. In *Uncertainty in Artificial Intelligence*, pp. 2147–2156. PMLR, 2023.

Tunstall, L., Beeching, E., Lambert, N., Rajani, N., Rasul, K., Belkada, Y., Huang, S., von Werra, L., Fourrier, C., Habib, N., Sarrazin, N., Sanseviero, O., Rush, A. M., and Wolf, T. Zephyr: Direct distillation of lm alignment, 2023.

Wei, X., Cui, X., Cheng, N., Wang, X., Zhang, X., Huang, S., Xie, P., Xu, J., Chen, Y., Zhang, M., et al. Zero-shot information extraction via chatting with chatgpt. *arXiv preprint arXiv:2302.10205*, 2023.

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

Wu, H. and Liu, X. Double thompson sampling for dueling bandits. *Advances in neural information processing systems*, 29, 2016.

Wu, Y., Jin, T., Lou, H., Xu, P., Farnoud, F., and Gu, Q. Adaptive sampling for heterogeneous rank aggregation from noisy pairwise comparisons. In *International Conference on Artificial Intelligence and Statistics*, pp. 11014–11036. PMLR, 2022.

Wu, Y., Jin, T., Lou, H., Farnoud, F., and Gu, Q. Borda regret minimization for generalized linear dueling bandits. *arXiv preprint arXiv:2303.08816*, 2023.

Xiong, W., Dong, H., Ye, C., Zhong, H., Jiang, N., and Zhang, T. Gibbs sampling from human feedback: A provable kl-constrained framework for rlhf. *arXiv preprint arXiv:2312.11456*, 2023.

Yang, K., Yang, L. F., and Du, S. S. q-learning with logarithmic regret. *arXiv preprint arXiv:2006.09118*, 2020.

Yuan, C., Xie, Q., and Ananiadou, S. Zero-shot temporal relation extraction with chatgpt. *arXiv preprint arXiv:2304.05454*, 2023.

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

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi, Y. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019.

Zhang, T. and Oles, F. The value of unlabeled data for classification problems. In *Proceedings of the Seventeenth International Conference on Machine Learning,(Langley, P., ed.)*, volume 20, pp. 0. Citeseer, 2000.

Zhang, W., He, J., Fan, Z., and Gu, Q. On the interplay between misspecification and sub-optimality gap in linear contextual bandits. *arXiv preprint arXiv:2303.09390*, 2023.

Zhao, H., He, J., Zhou, D., Zhang, T., and Gu, Q. Variance-dependent regret bounds for linear bandits and reinforcement learning: Adaptivity and computational efficiency. *arXiv preprint arXiv:2302.10371*, 2023.

Zhou, D. and Gu, Q. Computationally efficient horizon-free reinforcement learning for linear mixture mdps. *Advances in neural information processing systems*, 35: 36337–36349, 2022.

Zhu, B., Jiao, J., and Jordan, M. I. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. *arXiv preprint arXiv:2301.11270*, 2023.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., and Irving, G. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

A Proof of Theorems in Section

In this section, we provide the proof of Theorem 5.1 and we first introduce several lemmas. The following lemma provides an upper bound of the query complexity and corresponding dataset size $|\mathcal{C}_T|$.

Lemma A.1 (Lemma 4.5, Zhang et al., 2023)**.** Given $0 < \Gamma \leq 1$, if we set $\lambda = B^{-2}$, then for all round $t \in [T]$, we have $|\mathcal{C}_t| \leq |\mathcal{C}_T| \leq 16d\Gamma^{-2}\log(3LB\Gamma^{-1})$.

For a finite dataset $\mathcal{C}_T$, the following lemma provides a upper bound for the estimation error between $\widehat{\boldsymbol{\theta}}_t$ and $\boldsymbol{\theta}^*$.

Lemma A.2. Suppose we have $\|\boldsymbol{\theta}^*\| \leq B$, $\|\boldsymbol{\phi}(x,y)\| \leq L/2$. Then with probability at least $1-\delta$, for each round $t \in [T]$, we have

$$\|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_t\|_{\Sigma_{t-1}} \leq \frac{1}{\kappa_\sigma} \cdot \big(\sqrt{\lambda}B + \sqrt{2d\log(\lambda + |\mathcal{C}_T|L^2/d\lambda\delta)}\big),$$

Based on the Lemmas A.1 and A.2, the next auxiliary lemma proposed a proper choice of the uncertainty threshold Γ and confidence radius β in Algorithm 1.

Lemma A.3. If we set the uncertainty threshold $\Gamma = \kappa_\sigma\Delta/2d\iota_1$ and confidence radius $\beta = \kappa_\sigma^{-1}(1 + 4\sqrt{d\iota_2} + \sqrt{2d\iota_3})$, where $\iota_1 = 42\log(126LB\sqrt{d}\Delta^{-1}\kappa_\sigma^{-1}) + \sqrt{8\log(1/\delta)}$, $\iota_2 = \log(3LB\Gamma^{-1})$ and $\iota_3 = \log((1+16L^2B^1\Gamma^{-2}\iota_2)/\delta)$, then we have $2\beta\Gamma < \Delta$ and

$$\beta \geq \frac{1}{\kappa_\sigma} \cdot \big(\sqrt{\lambda}B + \sqrt{2d\log(\lambda + |\mathcal{C}_T|L^2/d\lambda\delta)}\big).$$

With these parameters, we now define the event $\mathcal{E}_1$ as

$$\mathcal{E}_1 = \{\forall t \in [T], \|\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}\|_{\boldsymbol{\Sigma}_{t-1}^{-1}} \leq \beta\}.$$

According to Lemma A.2 and Lemma A.3, we have $\Pr(\mathcal{E}_1) \geq 1-\delta$. Conditioned on the event $\mathcal{E}_1$, the following lemma suggests that our estimated discrepancy is no less than the actual discrepancy.

Lemma A.4. On the event $\mathcal{E}_1$, for each round $t \in [T]$, context $x \in \mathcal{C}$ and any action $y \in \mathcal{A}$, the estimated discrepancy $\widehat{D}_t(x,y)$ satisfied

$$\widehat{D}_t(x,y) \geq D_t(x,y) = \langle \boldsymbol{\theta}^*, \boldsymbol{\phi}(x,y) - \boldsymbol{\phi}_t^2\rangle.$$

On the other hand, we have

$$\widehat{D}_t(x,y) \leq D_t(x,y) + 2\beta\|\boldsymbol{\phi}(x,y) - \boldsymbol{\phi}_t^2\|_{\boldsymbol{\Sigma}_{t-1}^{-1}}.$$

It is worth to notice that in Algorithm 1 (Line 1), we update the policy π_t with online mirror descent and the following lemma provides the regret guarantee for these process.

Lemma A.5 (Lemma 6.2, He et al., 2022b)**.** For any estimated value function $\widehat{D}_t(x_{y+1}, \cdot)$, if we update the policy $\pi_{t+1}(\cdot|x)$ by the exponential rule:

$$\pi_{t+1}(\cdot|x) \propto \pi_t(\cdot|x) \cdot \exp\big(\eta\widehat{D}_t(x,\cdot)\big),$$

then the expected sub-optimality gap at round T can be upper bounded as following:

$$\mathbb{E}_{x\sim\mathcal{D},y\sim\pi^*(\cdot|x)}[\widehat{D}_t(x,y)] - \mathbb{E}_{x\sim\mathcal{D},y\sim\pi_t(\cdot|x)}[\widehat{D}_t(x,y)] \leq 2\eta + \eta^{-1}\mathbb{E}_{x\sim\mathcal{D}}\Big[\mathrm{KL}\big(\pi^*(\cdot|x)\|\pi_t(\cdot|x)\big) - \mathrm{KL}\big(\pi^*(\cdot|x)\|\pi_{t+1}(\cdot|x)\big)\Big]$$

With the help of these lemmas, we are ready to prove our main theorem.

Proof of Theorem 5.1. Now we start the regret analysis. For simplicity, for each round $t \in [T]$, we use $\boldsymbol{\phi}_t$ to denote $\boldsymbol{\phi}(x_t, y_t)$. Initially, the episodes and their corresponding regret can be decomposed into two groups based on whether episode t is added to the dataset $\mathcal{C}_T$:

$$\mathrm{Regret}(T) = \sum_{t=1}^{T}\langle\boldsymbol{\theta}^*, \boldsymbol{\phi}_t^*\rangle - \langle\boldsymbol{\theta}^*, \boldsymbol{\phi}_t^1\rangle$$

$$
\begin{aligned}
&= \sum_{t=1}^{T} D_t(x_t, y_t^*) - D_t(x_t, y_t^1) \\
&= \underbrace{\sum_{t \notin \mathcal{C}_T} D_t(x_t, y_t^*) - D_t(x_t, y_t^1)}_{I_1} + \underbrace{\sum_{t \notin \mathcal{C}_T} D_t(x_t, y_t^*) - D_t(x_t, y_t^1)}_{I_2}
\end{aligned} \tag{A.1}
$$

where $D_t(x, y) = \langle \boldsymbol{\theta}^*, \boldsymbol{\phi}(x, y) - \boldsymbol{\phi}_t^2 \rangle$ denotes the reward gap between action $y \in \mathcal{A}$ and selected action $\boldsymbol{\phi}_t^2$ at round t. Now, we bound this two term separately. For the term I_1, we have

$$
\begin{aligned}
I_1 &= \sum_{t \in \mathcal{C}_T} D_t(x_t, y_t^*) - D_t(x_t, y_t^1) \\
&\le \underbrace{\sum_{t \in \mathcal{C}_T} \widehat{D}_t(x_t, y_t^1) - D_t(x_t, y_t^1)}_{J_1} + \sum_{t \in \mathcal{C}_T} \widehat{D}_t(x_t, y_t^*) - \widehat{D}_t(x_t, y_t^1) \\
&= J_1 + \underbrace{\sum_{t \in \mathcal{C}_T} \mathbb{E}_{x_t \sim \mathcal{D}, y \sim \pi^*(\cdot|x)}[\widehat{D}_t(x_t, y)] - \mathbb{E}_{x_t \sim \mathcal{D}, y \sim \pi_t(\cdot|x)}[\widehat{D}_t(x_t, y)]}_{J_2} \\
&\quad + \underbrace{\sum_{t \in \mathcal{C}_T} \widehat{D}_t(x_t, y_t^*) - \widehat{D}_t(x_t, y_t^1) - \sum_{t \in \mathcal{C}_T} \mathbb{E}_{x_t \sim \mathcal{D}, y \sim \pi^*(\cdot|x)}[\widehat{D}_t(x_t, y)] - \mathbb{E}_{x_t \sim \mathcal{D}, y \sim \pi_t(\cdot|x)}[\widehat{D}_t(x_t, y_t)]}_{J_3},
\end{aligned} \tag{A.2}
$$

where the inequality holds due to Lemma A.4.
For the term J_1, we have

$$
\begin{aligned}
J_1 &= \sum_{t \in \mathcal{C}_T} \widehat{D}_t(x_t, y_t^1) - D_t(x_t, y_t^1) \\
&\le \sum_{t \in \mathcal{C}_T} \min\{4, 2\beta \|\boldsymbol{\phi}_t^1 - \boldsymbol{\phi}_t^2\|_{\boldsymbol{\Sigma}_{t-1}^{-1}}\} \\
&\le 4\beta \sqrt{|\mathcal{C}_T| \cdot \sum_{t \in \mathcal{C}_T} \min\{1, \|\boldsymbol{\phi}_t^1 - \boldsymbol{\phi}_t^2\|^2_{\boldsymbol{\Sigma}_{t-1}^{-1}}\}} \\
&\le 8\beta \sqrt{|\mathcal{C}_T| d \log\left(\frac{\lambda d + |\mathcal{C}_T| L^2}{\lambda d}\right)},
\end{aligned} \tag{A.3}
$$

where the first inequality holds due to Lemma A.4 with the fact that $-2 \le D_t(x_y, y_t^1) \le 2$, the second inequality holds due to Cauchy–Schwarz inequality and the last inequality holds due to the elliptical potential lemma (Lemma C.5).
The term J_2 reflects the sub-optimality from the online mirror descent process and can be upper bounded by Lemma A.5. For simplicity, we denote $\mathcal{C}_T = \{t_1, ..., t_K\}$ where $K = |\mathcal{C}_T|$. Thus, we have

$$
\begin{aligned}
J_2 &= \sum_{k=1}^{K} \mathbb{E}_{x_{t_k} \sim \mathcal{D}, y \sim \pi^*(\cdot|x)}[\widehat{D}_{t_k}(x_{t_k}, y)] - \mathbb{E}_{x_{t_k} \sim \mathcal{D}, y \sim \pi_{t_k}(\cdot|x)}[\widehat{D}_{t_k}(x_{t_k}, y)] \\
&\le \sum_{k=1}^{K} \big(2\eta + \eta^{-1} \mathbb{E}_{x \sim \mathcal{D}}\big[\mathrm{KL}(\pi^*(\cdot|x) \| \pi_{t_k}(\cdot|x)) - \mathrm{KL}(\pi^*(\cdot|x) \| \pi_{t_k+1}(\cdot|x)\big]\big) \\
&= 2\eta K + \eta^{-1} \mathbb{E}_{x \sim \mathcal{D}}\big[\mathrm{KL}(\pi^*(\cdot|x) \| \pi_1(\cdot|x)) - \mathrm{KL}(\pi^*(\cdot|x) \| \pi_{t_K+1}(\cdot|x)\big] \\
&\le 2\eta K + \eta^{-1} \mathbb{E}_{x \sim \mathcal{D}}\big[\mathrm{KL}(\pi^*(\cdot|x) \| \pi_1(\cdot|x))\big] \\
&\le 2\sqrt{2|\mathcal{C}_T| \log |\mathcal{A}|},
\end{aligned} \tag{A.4}
$$

where the first inequality holds due to Lemma A.5, the second equation holds due to policy π keeps unchanged for $t \in \mathcal{C}_T$, the second inequality holds due to $\mathrm{KL}(\cdot\|\cdot) \ge 0$ and the last inequality holds due to $\eta = \sqrt{\Gamma^2 \log \mathcal{A} / \big(32 d \log(3LB\Gamma^{-1})\big)}$ with the fact that π_1 is uniform policy.

According to Azuma-Hoeffding's inequality (Lemma C.4), with probability at least $1-\delta$, the term J_3 can be upper bounded by

$$J_3 \leq 2\sqrt{2|\mathcal{C}_T|\log(1/\delta)}. \tag{A.5}$$

Substituting (A.3), (A.4) and (A.5) into (A.2), we have

$$\begin{aligned} I_1 = J_1 + J_2 + J_3 &\leq 8\beta\sqrt{|\mathcal{C}_T| d \log\left(\frac{\lambda d + |\mathcal{C}_T|L^2}{\lambda d}\right)} + 2\sqrt{2|\mathcal{C}_T|\log|\mathcal{A}|} + 2\sqrt{2|\mathcal{C}_T|\log(1/\delta)} \\ &\leq \widetilde{O}(\beta d/\Gamma) \\ &= \frac{d^2}{\Delta}. \end{aligned} \tag{A.6}$$

where the last inequality holds due to Lemma A.1.

Now, we only need to focus on the term I_2. For each round $t \notin \mathcal{C}_T$, we have

$$\begin{aligned} D_t(x_t, y_t^*) - D_t(x_t, y_t^1) &= \langle \boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_t, \boldsymbol{\phi}_t^* - \boldsymbol{\phi}_t^2\rangle + \langle \widehat{\boldsymbol{\theta}}_t, \boldsymbol{\phi}_t^* - \boldsymbol{\phi}_t^2\rangle - \langle \boldsymbol{\theta}^*, \boldsymbol{\phi}_t^1 - \boldsymbol{\phi}_t^2\rangle \\ &\leq \beta\|\boldsymbol{\phi}_t^* - \boldsymbol{\phi}_t^2\|_{\boldsymbol{\Sigma}_{t-1}^{-1}} + \langle \widehat{\boldsymbol{\theta}}_t, \boldsymbol{\phi}_t^* - \boldsymbol{\phi}_t^2\rangle - \langle \boldsymbol{\theta}^*, \boldsymbol{\phi}_t^1 - \boldsymbol{\phi}_t^2\rangle \\ &\leq \beta\|\boldsymbol{\phi}_t^1 - \boldsymbol{\phi}_t^2\|_{\boldsymbol{\Sigma}_{t-1}^{-1}} + \langle \widehat{\boldsymbol{\theta}}_t, \boldsymbol{\phi}_t^1 - \boldsymbol{\phi}_t^2\rangle - \langle \boldsymbol{\theta}^*, \boldsymbol{\phi}_t^1 - \boldsymbol{\phi}_t^2\rangle \\ &\leq 2\beta\|\boldsymbol{\phi}_t^1 - \boldsymbol{\phi}_t^2\|_{\boldsymbol{\Sigma}_{t-1}^{-1}}, \end{aligned}$$

where the first inequality holds due to Lemma A.4, the second inequality holds due to the selection rule of action $\boldsymbol{\phi}_t^1$ and the last inequality holds due to Lemma A.4. According to the definition of set $\mathcal{C}_T$ in Algorithm 1, for each round $t \notin \mathcal{C}_T$, we have $\|\boldsymbol{\phi}_t^1 - \boldsymbol{\phi}_t\|_{\boldsymbol{\Sigma}_{t-1}^{-1}} \leq \Gamma$. Therefore, the sub-optimality gap at round t is upper bounded by

$$2\beta\|\boldsymbol{\phi}_t^1 - \boldsymbol{\phi}_t\|_{\boldsymbol{\Sigma}_{t-1}^{-1}} \leq 2\beta\Gamma < \Delta,$$

where the second inequality holds due to Lemma A.3. According to the minimal sub-optimality assumption (Assumption 3.4), this indicates that the regret yielded in round $t \notin \mathcal{C}_T$ is 0. Summing up over $t \notin \mathcal{C}_T$, we have

$$I_2 = \sum_{t \in \mathcal{T}_t} D_t(x_t, y_t^*) - D_t(x_t, y_t^1) = 0. \tag{A.7}$$

Combining the results in (A.6) and (A.7), we complete the proof of Theorem 5.1. □

B Proof of Lemmas in Appendix A

In this section, we provide the proofs of the lemmas in Appendix A.

B.1 Proof of Lemma A.2

Proof of Lemma A.2. This proof follows the proof in Di et al. (2023). Let $t \in \Psi_{T+1,\ell}$, $t \geq 2$, we define some auxiliary quantities:

$$\begin{aligned} G_t(\boldsymbol{\theta}) &= \lambda\boldsymbol{\theta} + \sum_{\tau \in \mathcal{C}_{t-1}} \Big[\sigma\big((\boldsymbol{\phi}_\tau^1 - \boldsymbol{\phi}_\tau^2)^\top\boldsymbol{\theta}\big) - \sigma\big((\boldsymbol{\phi}_\tau^1 - \boldsymbol{\phi}_\tau^2)^\top\boldsymbol{\theta}^*\big)\Big](\boldsymbol{\phi}_\tau^1 - \boldsymbol{\phi}_\tau^2) \\ \epsilon_t &= o_t - \sigma\big((\boldsymbol{\phi}_t^1 - \boldsymbol{\phi}_t^2)^\top\boldsymbol{\theta}^*\big) \\ Z_t &= \sum_{\tau \in \mathcal{C}_{t-1}} \epsilon_\tau(\boldsymbol{\phi}_\tau^1 - \boldsymbol{\phi}_\tau^2). \end{aligned}$$

By the definition that $\widehat{\boldsymbol{\theta}}_t$ is the solution to (4.1), we plugging the equation to the definition of G_t and we have that

$$G_t(\widehat{\boldsymbol{\theta}}_t) = \lambda\widehat{\boldsymbol{\theta}}_t + \sum_{\tau \in \mathcal{C}_{t-1}} \Big[\sigma\big((\boldsymbol{\phi}_\tau^1 - \boldsymbol{\phi}_\tau^2)^\top\widehat{\boldsymbol{\theta}}_t\big) - o_\tau + o_\tau - \sigma\big((\boldsymbol{\phi}_\tau^1 - \boldsymbol{\phi}_\tau^2)^\top\boldsymbol{\theta}^*\big)\Big](\boldsymbol{\phi}_\tau^1 - \boldsymbol{\phi}_\tau^2)$$

$$= \lambda\widehat{\boldsymbol{\theta}}_t + \sum_{\tau\in\mathcal{C}_{t-1}} \Big[\sigma\big((\boldsymbol{\phi}_\tau^1 - \boldsymbol{\phi}_\tau^2)^\top\widehat{\boldsymbol{\theta}}_t\big) - o_\tau\Big] + \sum_{\tau\in\mathcal{C}_{t-1}} \Big[o_\tau - \sigma\big((\boldsymbol{\phi}_\tau^1 - \boldsymbol{\phi}_\tau^2)^\top\boldsymbol{\theta}^*\big)\Big](\boldsymbol{\phi}_\tau^1 - \boldsymbol{\phi}_\tau^2)$$
$$= Z_t.$$

Therefore, we have that

$$G_t(\widehat{\boldsymbol{\theta}}_t) - G_t(\boldsymbol{\theta}^*) = Z_t - \lambda\boldsymbol{\theta}^*.$$

On the other hand, by Taylor's expansion, we have

$$\begin{aligned} G_t(\widehat{\boldsymbol{\theta}}_t) - G_t(\boldsymbol{\theta}^*) &= \lambda(\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*) + \sum_{\tau\in\mathcal{C}_{t-1}} \Big[\sigma\big((\boldsymbol{\phi}_\tau^1 - \boldsymbol{\phi}_\tau^2)^\top\boldsymbol{\theta}\big) - \sigma\big((\boldsymbol{\phi}_\tau^1 - \boldsymbol{\phi}_\tau^2)^\top\boldsymbol{\theta}^*\big)\Big](\boldsymbol{\phi}_\tau^1 - \boldsymbol{\phi}_\tau^2) \\ &= \Big[\lambda\boldsymbol{I} + \sum_{\tau\in\mathcal{C}_{t-1}} \sigma'\big((\boldsymbol{\phi}_\tau^1 - \boldsymbol{\phi}_\tau^2)^\top\widetilde{\boldsymbol{\theta}}_\tau\big)(\boldsymbol{\phi}_\tau^1 - \boldsymbol{\phi}_\tau^2)(\boldsymbol{\phi}_\tau^1 - \boldsymbol{\phi}_\tau^2)^\top\Big](\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*) \\ &= F(\widetilde{\boldsymbol{\theta}}_t)(\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*), \end{aligned}$$

where we know that $F(\widetilde{\boldsymbol{\theta}}_t) \succeq \kappa_\sigma\widehat{\boldsymbol{\Sigma}}_{t,\ell}$. Now we have

$$\begin{aligned} \|\widehat{\boldsymbol{\theta}}_{t,\ell} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}_{t-1}}^1 &= (Z_t - \lambda\boldsymbol{\theta}^*)^\top F(\widetilde{\boldsymbol{\theta}}_t)^{-1}\boldsymbol{\Sigma}_{t-1}F(\widetilde{\boldsymbol{\theta}}_t)^{-1}(Z_t - \lambda\boldsymbol{\theta}^*) \\ &\le \frac{1}{\kappa_\sigma^1}(Z_t - \lambda\boldsymbol{\theta}^*)^\top\boldsymbol{\Sigma}_{t-1}^{-1}(Z_t - \lambda\boldsymbol{\theta}^*) \\ &\le \frac{1}{\kappa_\sigma^1}\big(\|Z_t\|_{\boldsymbol{\Sigma}_{t-1}^{-1}}^1 + \lambda\|\boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}_{t-1}^{-1}}^1\big) \\ &\le \frac{1}{\kappa_\sigma^1}\big(\|Z_t\|_{\boldsymbol{\Sigma}_{t-1}^{-1}}^1 + \lambda B^1\big) \end{aligned}$$

where the first inequality holds due to $F(\widetilde{\boldsymbol{\theta}}_t) \succeq \kappa_\sigma\widehat{\boldsymbol{\Sigma}}_{t,\ell}$ and the second holds due to triangle inequality. Now it remains to bound $\|Z_t\|_{\boldsymbol{\Sigma}_{t-1}^{-1}}^1$.

By Lemma C.2, we have

$$\|Z_t\|_{\boldsymbol{\Sigma}_{t-1}^{-1}} \le \sqrt{2\log\left(\frac{\sqrt{\det(\boldsymbol{\Sigma}_{t-1})}}{\sqrt{\det(\boldsymbol{\Sigma}_0)}\delta}\right)} \le \sqrt{2\log\left(\frac{\det(\boldsymbol{\Sigma}_{t-1})}{\lambda^d\delta}\right)} \le \sqrt{2d\log\left(\frac{\lambda + |\mathcal{C}_t|L^2/d}{\lambda^d\delta}\right)}$$

where the first inequality holds due to Lemma C.2 and the last inequality holds due to Lemma C.3. Now we combine the two term and take the union bound over all $t \in \mathcal{C}_T$, we have

$$\|\widehat{\boldsymbol{\theta}}_{t,\ell} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}_{t-1}} \le \frac{1}{\kappa_\sigma}(\sqrt{\lambda}B + \sqrt{2d\log\left(\frac{\lambda + |\mathcal{C}_t|L^2/d}{\lambda^d\delta}\right)},$$

which concludes our statement. □

B.2 Proof of Lemma A.3

Proof of Lemma A.3. This proof follows the proof in Zhang et al. (2023). First we recall that $\Gamma = \Delta\kappa_\sigma/2d\iota_1$ and $\beta = \kappa_\sigma^{-1}(1 + 4\sqrt{d\iota_2} + \sqrt{2d\iota_3})$, We first prove that the selection of β satisfy the requirement in Lemma A.3. Recalling that $\lambda = B^{-2}$, by basic calculation, we have

$$\begin{aligned} \kappa_\sigma\beta &\ge 1 + \sqrt{2d\log((1 + L^2B^1 16d\Gamma^{-2}\iota_2)/d\delta)} \\ &\ge 1 + \sqrt{2d\log((1 + L^2B^1|\mathcal{C}_T|)/d\delta)} \\ &= \sqrt{\lambda}B + \sqrt{2d\log(\lambda + |\mathcal{C}_T|L^2/d\lambda\delta)}, \end{aligned}$$

where the first inequality hold by neglecting the positive term $4\sqrt{d\iota_2}$ and $d \ge 1$, the second inequality holds due to Lemma A.1 and the last equation holds due by plugging in $\lambda = B^{-2}$. Now we have come to the second statement. First, by basic computation, we have

$$\sqrt{2\iota_3} \le \sqrt{2\log((1 + 16L^2B^1\Gamma^{-2}\iota_2)} + \sqrt{2\log(1/\delta))}.$$

Notice that we have $L \geq 1$, $B \geq 1$, and $\Gamma \leq 1$, which further concludes that $LB\Gamma^{-1} \geq 1$, which induces that

$$2 + 4\sqrt{\iota_2} \leq 6\iota_2, \quad \sqrt{2\log((1 + 16L^2B^1\Gamma^{-2}\iota_2)} \leq 3\iota_2.$$

Therefore, we have:

$$\begin{aligned} 2 + 4\sqrt{\iota_2} + \sqrt{2\iota_3} &\leq 9\iota_2 + 2\sqrt{\log(1/\delta)} \\ &\leq 9\log(6LB\sqrt{d}\Delta^{-1}\kappa_\sigma^{-1}\iota_1) + 2\sqrt{\log(1/\delta)}. \end{aligned}$$

By Lemma C.1, we know the sufficient condition for the following inequality

$$(6LB\sqrt{d}\Delta^{-1}\kappa_\sigma^{-1})\iota_1 \geq 9(6LB\sqrt{d}\Delta^{-1}\kappa_\sigma^{-1})\log(6LB\sqrt{d}\Delta^{-1}\kappa_\sigma^{-1}\iota_1) + 2(6LB\sqrt{d}\Delta^{-1}\kappa_\sigma^{-1})\sqrt{\log(1/\delta)} \tag{B.1}$$

is that

$$\iota_1 \geq 36\log(108LB\sqrt{d}\Delta^{-1}\kappa_{mu}^{-1}) + \sqrt{8\log(1/\delta)},$$

which naturally holds due to our definition of ι_1. Eliminating the $6LB\sqrt{d}\Delta^{-1}\kappa_\sigma^{-1}$ term in (B.1) yields that

$$\iota_1 \geq 2 + 4\sqrt{\iota_2} + \sqrt{2\iota_3},$$

which implies that

$$2\beta\Gamma = \frac{\Delta\kappa_\sigma}{2\sqrt{d\iota_1}}\frac{1}{\kappa_\sigma}(1 + 2\sqrt{d\iota_2} + \sqrt{2\iota_3}) \geq \Delta.$$

□

C Auxiliary Lemmas

Lemma C.1 (Lemma A.2, Shalev-Shwartz & Ben-David, 2014)**.** Let $a \geq 1$ and $b \geq 0$, then $x \geq 4a\log(2a) + 2b$ results in $x \geq a\log x + b$.

Lemma C.2 (Theorem 1, Abbasi-Yadkori et al., 2011)**.** Let $\{\mathcal{F}_t\}_{t=0}^{\infty}$ be a filtration. Let $\{\epsilon_t\}_{t=1}^{\infty}$ be a real-valued stochastic process such that ϵ_t is $\mathcal{F}_t$-measurable and ϵ_t is conditionally R-sub-Gaussian for some $R \geq 0$. Let $\{\boldsymbol{\phi}_t\}_{t=1}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process such that $\boldsymbol{\phi}_t$ is $\mathcal{F}_{t-1}$ measurable and $\|\boldsymbol{\phi}\|_2 \leq L$ for all t. For any $t \geq 0$, define $\mathbf{U}_t = \lambda\mathbf{I} + \sum_{k=1}^{t}\boldsymbol{\phi}_t\boldsymbol{\phi}_t$. Then for any $\delta > 0$, with probability at least $1 - \delta$, for all $t \geq 0$

$$\left\|\sum_{k=1}^{t}\boldsymbol{\phi}_t\epsilon_t\right\|_{\mathbf{U}_t^{-1}}^1 \leq 2R^1\log\left(\frac{\sqrt{\det(\mathbf{U}_t)}}{\sqrt{\det(\mathbf{U}_0)\delta}}\right).$$

Lemma C.3 (Lemma C.7, Zhang et al., 2023)**.** Suppose sequence $\{\mathbf{x}_t\}_{k=1}^{K} \subset \mathbb{R}^d$ and for any $k \in [K]$, $\|\mathbf{x}_t\|_2 \leq L$. For any index subset $\mathcal{C} \subseteq [K]$, define $\mathbf{U} = \lambda\mathbf{I} + \sum_{k\in\mathcal{C}}\mathbf{x}_t\mathbf{x}_t^\top$ for some $\lambda > 0$, then $\det(\mathbf{U}) \leq (\lambda + |\mathcal{C}|L^2/d)^d$.

Lemma C.4 (Azuma–Hoeffding inequality, Cesa-Bianchi & Lugosi 2006)**.** Let $\{x_i\}_{i=1}^{n}$ be a martingale difference sequence with respect to a filtration $\{\mathcal{G}_i\}$ satisfying $|x_i| \leq M$ for some constant M, x_i is $\mathcal{G}_{i+1}$-measurable, $\mathbb{E}[x_i|\mathcal{G}_i] = 0$. Then for any $0 < \delta < 1$, with probability at least $1 - \delta$, we have

$$\sum_{i=1}^{n}x_i \leq M\sqrt{2n\log(1/\delta)}.$$

Lemma C.5 (Lemma 11 in (Abbasi-Yadkori et al., 2011))**.** Let $\{X_t\}_{t=1}^{+\infty}$ be a sequence in $\mathbb{R}^d$, V a $d \times d$ positive definite matrix and define $V_t = V + \sum_{i=1}^{t}X_t^\top X_t$. If $\|X_t\|_2 \leq L$ and $\lambda_{\min}(V) \geq \max(1, L^2)$, then we have

$$\sum_{i=1}^{t}X_i^\top(V_{i-1})^{-1}X_i \leq 2\log\left(\frac{\det V_t}{\det V}\right).$$