Nearly Optimal Algorithms for Contextual Dueling Bandits from Adversarial Feedback

Anonymous Author(s)
Affiliation
Address
email

Abstract

Learning from human feedback plays an important role in aligning generative models, such as large language models (LLM). However, the effectiveness of this approach can be influenced by adversaries, who may intentionally provide misleading preferences to manipulate the output in an undesirable or harmful direction. To tackle this challenge, we study a specific model within this problem domain–contextual dueling bandits with adversarial feedback, where the true preference label can be flipped by an adversary. We propose an algorithm namely robust contextual dueling bandits (`RCDB`), which is based on uncertainty-weighted maximum likelihood estimation. Our algorithm achieves an $\widetilde{O}(d\sqrt{T} + dC)$ regret bound, where T is the number of rounds, d is the dimension of the context, and $0 \leq C \leq T$ is the total number of adversarial feedback. We also prove a lower bound to show that our regret bound is nearly optimal, both in scenarios with and without ($C = 0$) adversarial feedback. Additionally, we conduct experiments to evaluate our proposed algorithm against various types of adversarial feedback. Experimental results demonstrate its superiority over the state-of-the-art dueling bandit algorithms in the presence of adversarial feedback.

1 Introduction

Acquiring an appropriate reward proves challenging in numerous real-world applications, often necessitating intricate instrumentation (Zhu et al., 2020) and time-consuming calibration (Yu et al., 2020) to achieve satisfactory levels of sample efficiency. For instance, in training large language models (LLM) using reinforcement learning from human feedback (RLHF), the diverse values and perspectives of humans can lead to uncalibrated and noisy rewards (Ouyang et al., 2022). In contrast, preference-based data, which involves comparing or ranking various actions, is a more straightforward method for capturing human judgments and decisions. In this context, the dueling bandit model (Yue et al., 2012) provides a problem framework that focuses on optimal decision-making through pairwise comparisons, rather than relying on the absolute reward for each action.

However, human feedback may not always be reliable. In real-world applications, human feedback is particularly vulnerable to manipulation through preference label flip. Adversarial feedback can significantly increase the risk of misleading a large language model (LLM) into erroneously prioritizing harmful content, under the false belief that it reflects human preference. Despite the significant influence of adversarial feedback, there is limited existing research on the impact of adversarial feedback specifically within the context of dueling bandits. A notable exception is Agarwal et al. (2021), which studies dueling bandits when an adversary can flip some of the preference labels received by the learner. They proposed an algorithm that is agnostic to the amount of adversarial feedback introduced by the adversary. However, their setting has the following two limitations. First, their study was confined to a finite-armed setting, which renders their results less applicable to modern applications such as RLHF. Second, their adversarial feedback is defined on the whole comparison matrix. In each round, the adversary observes the outcomes of all pairwise comparisons and then decides to corrupt some of the pairs before the agent selects the actions. This assumption

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

does not align well with the real-world scenario, where the adversary often flips the preference label based on the information of the selected actions.

In this paper, to address the above challenge, we aim to develop contextual dueling bandit algorithms that are robust to adversarial feedback. This enables us to effectively tackle problems involving a large number of actions while also taking advantage of contextual information. We specifically consider a scenario where the adversary knows the selected action pair and the true preference of their comparison. In this setting, the adversary's only decision is whether to flip the preference label or not. We highlight our contributions as follows:

- We propose a new algorithm called robust contextual dueling bandits (RCDB), which integrates uncertainty-dependent weights into the Maximum Likelihood Estimator (MLE). Intuitively, our choice of weight is designed to induce a higher degree of skepticism about potentially "untrustworthy" feedback. The agent is encouraged to focus more on feedback that is more likely to be genuine, effectively diminishing the impact of any adversarial feedback.
- We analyze the regret of our algorithm under at most C number of adversarial feedback. Our result consists of two terms: a C-independent term $\widetilde{O}(d\sqrt{T})$, which matches the lower bound established in Bengs et al. (2022) for uncorrupted linear contextual dueling bandits, and a C-dependent term $\widetilde{O}(dC)$. Furthermore, we establish a lower bound for dueling bandits with adversarial feedback, demonstrating the optimality of our adversarial term. Consequently, our algorithm for dueling bandits attains the optimal regret in both scenarios, with and without adversarial feedback.
- We conduct extensive experiments to validate the effectiveness of our algorithm RCDB. To comprehensively assess RCDB's robustness against adversarial feedback, we evaluate its performance under various types of adversarial feedback and compare the results with state-of-the-art dueling bandit algorithms. Experimental results demonstrate the superiority of our algorithm in the presence of adversarial feedback, which corroborate our theoretical analysis.

Table 1: Comparison of algorithms for robust bandits and dueling bandits.

Model	Algorithm	Setting	Regret
Bandits	Multi-layer Active Arm Elimination Race (Lykouris et al., 2018)	K-armed Bandits	$\widetilde{O}(K^{1.5}C\sqrt{T})$
	BARBAR (Gupta et al., 2019)	K-armed Bandits	$\widetilde{O}(\sqrt{KT}+KC)$
	SBE (Li et al., 2019)	Linear Bandits	$\widetilde{O}(d^2C/\Delta+d^5/\Delta^2)$
	Robust Phased Elimination (Bogunovic et al., 2021)	Linear Bandits	$\widetilde{O}(\sqrt{dT}+d^{1.5}C+C^2)$
	Robust weighted OFUL (Zhao et al., 2021)	Linear Contextual Bandits	$\widetilde{O}(dC\sqrt{T})$
	CW-OFUL (He et al., 2022)	Linear Contextual Bandits	$\widetilde{O}(d\sqrt{T}+dC)$
Dueling Bandits	WIWR (Agarwal et al., 2021)	K-armed Dueling Bandits	$\widetilde{O}(K^2C/\Delta_{\min}+\sum_{i\neq i^*}K^2/\Delta_i^2)$
	Versatile-DB (Saha and Gaillard, 2022)	K-armed Dueling Bandits	$\widetilde{O}(C+\sum_{i\neq i^*}1/\Delta_i+\sqrt{K})$
	RCDB **(Our work)**	Contextual Dueling Bandits	$\widetilde{O}(d\sqrt{T}+dC)$

Notation. In this paper, we use plain letters such as x to denote scalars, lowercase bold letters such as $\mathbf{x}$ to denote vectors and uppercase bold letters such as $\mathbf{X}$ to denote matrices. For a vector $\mathbf{x}$, $\|\mathbf{x}\|_2$ denotes its ℓ_2-norm. The weighted ℓ_2-norm associated with a positive-definite matrix $\mathbf{A}$ is defined as $\|\mathbf{x}\|_{\mathbf{A}}=\sqrt{\mathbf{x}^\top\mathbf{A}\mathbf{x}}$. For two symmetric matrices $\mathbf{A}$ and $\mathbf{B}$, we use $\mathbf{A}\succeq\mathbf{B}$ to denote $\mathbf{A}-\mathbf{B}$ is positive semidefinite. We use $\mathbb{1}$ to denote the indicator function and $\mathbf{0}$ to denote the zero vector. For two actions a, b, we use $a\succ b$ to denote a is more preferable to b. For a postive integer N, we use $[N]$ to denote $\{1,2,\ldots,N\}$. We use standard asymptotic notations including $O(\cdot),\Omega(\cdot),\Theta(\cdot)$, and $\widetilde{O}(\cdot),\widetilde{\Omega}(\cdot),\widetilde{\Theta}(\cdot)$ will hide logarithmic factors.

2 Related Work

Bandits with Adversarial Reward. The multi-armed bandit problem, involving an agent making sequential decisions among multiple arms, has been studied with both stochastic rewards (Lai et al., 1985; Lai, 1987; Auer, 2002; Auer et al., 2002a; Kalyanakrishnan et al., 2012; Lattimore and Szepesvári, 2020; Agrawal and Goyal, 2012), and adversarial rewards (Auer et al., 2002b; Bubeck et al., 2012). Moreover, a line of works focuses on designing algorithms that can achieve near-optimal regret bounds for both stochastic bandits and adversarial bandits simultaneously (Bubeck and Slivkins, 2012; Seldin and Slivkins, 2014; Auer and Chiang, 2016; Seldin and Lugosi, 2017; Zimmert and Seldin, 2019; Lee et al., 2021), which is known as "the best of both worlds" guarantee. Distinct from fully stochastic and fully adversarial models, Lykouris et al. (2018) studied a setting, where only a portion of the rewards is subject to corruption. They proposed an algorithm with a regret dependent on the corruption level C, defined as the cumulative sum of the corruption magnitudes in each round. Their result is C times worse than the regret without corruption. Gupta et al. (2019) improved the result by providing a regret guarantee comprising two terms, a corruption-independent term that matches the regret lower bound without corruption, and a corruption-dependent term that is linear in C. In addition, Gupta et al. (2019) proved a lower bound demonstrating the optimality of the linear dependency on C.

Contextual Bandits with Corruption. Li et al. (2019) studied stochastic linear bandits with corruption and presented an instance-dependent regret bound linearly dependent on the corruption level C. Bogunovic et al. (2021) studied the same problem and proposed an algorithm with near-optimal regret in the non-corrupted case. Lee et al. (2021) studied this problem in a different setting, where the adversarial corruptions are generated through the inner product of a corrupted vector and the context vector. For linear contextual bandits, Bogunovic et al. (2021) proved that under an additional context diversity assumption, the regret of a simple greedy algorithm is nearly optimal with an additive corruption term. Zhao et al. (2021) and Ding et al. (2022) extended the OFUL algorithm (Abbasi-Yadkori et al., 2011) and proved a regret with a corruption term polynomially dependent on the total number of rounds T. He et al. (2022) proposed an algorithm for known corruption level C to remove the polynomial dependency on T in the corruption term, which only has a linear dependency on C. They also proved a lower bound showing the optimality of linear dependency on C for linear contextual bandits with a known corruption level. Additionally, He et al. (2022) extended the proposed algorithm to an unknown corruption level and provided a near-optimal performance guarantee that matches the lower bound. For more extensions, Kuroki et al. (2023) studied best-of-both-worlds algorithms for linear contextual bandits. Ye et al. (2023) proposed a corruption robust algorithm for nonlinear contextual bandits.

Dueling Bandits and Logistic Bandits. The dueling bandit model was first proposed in Yue et al. (2012). Compared with bandits, the agent will select two arms and receive the preference feedback between the two arms from the environment. For general preference, there may not exist the "best" arm that always wins in the pairwise comparison. Therefore, various alternative winners are considered, including Condorcet winner (Zoghi et al., 2014; Komiyama et al., 2015), Copeland winner (Zoghi et al., 2015; Wu and Liu, 2016; Komiyama et al., 2016), Borda winner (Jamieson et al., 2015; Falahatgar et al., 2017; Heckel et al., 2018; Saha et al., 2021; Wu et al., 2023) and von Neumann winner (Ramamohan et al., 2016; Dudík et al., 2015; Balsubramani et al., 2016), along with their corresponding performance metrics. To handle potentially large action space or context information, Saha (2021) studied a structured contextual dueling bandit setting. In this setting, each arm possesses an unknown intrinsic reward. The comparison is determined based on a logistic function of the relative rewards. In a similar setting, Bengs et al. (2022) studied contextual linear stochastic transitivity model with contextualized utilities. Di et al. (2023) proposed a layered algorithm with variance aware regret bound. Another line of works does not make the reward assumption. Instead, they assume the preference feedback can be represented by a function class. Saha and Krishnamurthy (2022) designed an algorithm that achieves the optimal regret for K-armed contextual dueling bandit problem. Sekhari et al. (2023) studied contextual dueling bandits in a more general setting and proposed an algorithm the provides guarantees for both regret and the number of queries. Another related area of research is the logistic bandits, where the agent selects one arm in each round and receives a Bernoulli reward. Faury et al. (2020) studied the dependency with respect to the degree of non-linearity of the logistic function κ. They proposed an algorithm with no dependency in κ. Abeille et al. (2021) further improved the dependency on κ and proved a problem dependent lower bound. Faury et al. (2022) proposed a computationally efficient algorithm with regret performance still matching the lower-bound proved in Abeille et al. (2021).

Dueling Bandits with Adversarial Feedback. A line of work has focused on dueling bandits with adversarial feedback or corruption. Gajane et al. (2015) studied a fully adversarial utility-based version of dueling bandits, which was proposed in Ailon et al. (2014). Saha et al. (2021) considered the Borda regret for adversarial dueling bandits without the assumption of utility. In a setting parallel to that in Lykouris et al. (2018); Gupta et al. (2019), Agarwal et al. (2021) studied K-armed dueling bandits in a scenario where an adversary has the capability to corrupt part of the feedback received by the learner. They designed an algorithm whose regret comprises two terms: one that is optimal in uncorrupted scenarios, and another that is linearly dependent on the total times of adversarial feedback C. Later on, Saha and Gaillard (2022) achieved "best-of-both world" result for noncontextual dueling bandits and improved the adversarial term of Agarwal et al. (2021) in the same setting. For contextual dueling bandits, Wu et al. (2023) proposed an EXP3-type algorithm for the adversarial linear setting using Borda regret. For a comparison of the most related works for robust bandits and dueling bandits, please refer to Table 1. In this paper, we study the influence of adversarial feedback within contextual dueling bandits, particularly in a setting where only a minority of the feedback is adversarial. Compared to previous studies, most studies have focused on the multi-armed dueling bandit framework without integrating context information. The notable exception is Wu et al. (2023); however, this study does not provide guarantees regarding the dependency on the number of adversarial feedback instances.

3 Preliminaries

In this work, we study linear contextual dueling bandits with adversarial feedback. In each round $t \in [T]$, the agent observes the context information x_t from a context set $\mathcal{X}$ and the corresponding action set $\mathcal{A}$. Utilizing this context information, the agent selects two actions, a_t and b_t. Subsequently, the environment will generate a binary feedback (i.e., preference label) $l_t = \mathbb{1}(a_t \succ b_t) \in \{0, 1\}$ indicating the preferable action. We assume the existence of a reward function $r^*(x, a)$ dependent on the context information x and action a, and a monotonically increasing link function σ satisfying $\sigma(x) + \sigma(-x) = 1$. The preference probability will be determined by the link function and the difference between the rewards of the selected arms, i.e.,

$$\mathbb{P}(a \succ b|x) = \sigma\big(r^*(x, a) - r^*(x, b)\big). \tag{3.1}$$

We assume that the reward function is linear with respect to some known feature map $\phi(x, a)$. To be more specific, we make the following assumption:

Assumption 3.1. Let $\boldsymbol{\phi} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$ be a known feature map, with $\|\boldsymbol{\phi}(x, a)\|_2 \le 1$ for any $(x, a) \in \mathcal{X} \times \mathcal{A}$. We define the reward function $r_{\boldsymbol{\theta}}$ parameterized by $\boldsymbol{\theta} \in \mathbb{R}^d$, with $r_{\boldsymbol{\theta}}(x, a) = \langle \boldsymbol{\theta}, \boldsymbol{\phi}(x, a)\rangle$. Moreover, there exists $\boldsymbol{\theta}^*$ satisfying $r_{\boldsymbol{\theta}^*} = r^*$, with $\|\boldsymbol{\theta}^*\|_2 \le B$.

Similar assumptions have been made in the literature of dueling bandits (Saha, 2021; Bengs et al., 2022; Xiong et al., 2023). We also make an assumption on the derivative of the link function, which is common in the study of generalized linear models for bandits (Filippi et al., 2010).

Assumption 3.2. The link function σ is differentiable. Furthermore, its first-order derivative satisfies:

$$\dot{\sigma}(\cdot) \ge \kappa$$

for some constant $\kappa > 0$.

In our setting, however, the agent does not directly observe the true binary feedback. Instead, an adversary will see both the choice of the agent and the true feedback. Based on the information, the adversary can decide whether to corrupt the binary feedback or not.[1] We represent the adversary's decision in round t by an adversarial indicator c_t, which takes values from the set $\{0, 1\}$. If the adversary chooses not to corrupt the result, we have $c_t = 0$. Otherwise, we have $c_t = 1$, which means adversarial feedback in this round. As a result, the agent will observe a flipped preference label, i.e., the observation $o_t = 1 - l_t$. We define C as the total level of adversarial feedback, i.e.,

$$\textstyle\sum_{t=1}^T c_t \le C.$$

Remark 3.3. Adversarial corruption has been firstly studied in bandits (Lykouris et al., 2018), where in each round t, the agent selects an action a_t and the environment generates a numerical reward $r_t(a_t)$. The adversary observes the reward and returns a corrupted reward $\bar{r}_t$. The corruption level C is defined by $\sum_{t=1}^T |r_t(a_t) - \bar{r}_t| \le C$. Compared with the continuous perturbation of rewards in bandits, the adversary's label flipping attack method in our model is quite different. The cost of obtaining adversarial feedback is uniformly 1, unlike in bandits where the cost depends on the intensity of the perturbation. Additionally, adversarial feedback in our setting involves comparing two arms, whereas in bandits it pertains to the reward of a single arm. The only previous work that studied label-flipping is (Agarwal et al., 2021), where the adversary cannot observe the action selected by the agent. In contrast, our setting focuses on scenarios where this information is available to adversaries, which is common in many real-life applications.

As the context is changing, the optimal action is different in each round, denoted by $a_t^* = \mathrm{argmax}_{a \in \mathcal{A}}\, r^*(x_t, a)$. The goal of our algorithm is to minimize the cumulative gap between the rewards of both selected actions and the optimal action

$$\mathrm{Regret}(T) = \textstyle\sum_{t=1}^T 2r^*(x_t, a_t^*) - r^*(x_t, a_t) - r^*(x_t, b_t). \tag{3.2}$$

[1]Such adversary is referred to as strong adversary (He et al., 2022), compared with the weak adversary who cannot obtain the information before the decision.

This regret definition is the same as that in Saha (2021) and the average regret defined in Bengs et al. (2022). It is typically stronger than weak regret defined in Bengs et al. (2022), which only considers the reward gap of the better action.

4 Algorithm

In this section, we present our new algorithm RCDB, designed for learning contextual linear dueling bandits. The main algorithm is illustrated in Algorithm 1. At a high level, we incorporate uncertainty-dependent weighting into the Maximum Likelihood Estimator (MLE) to counter adversarial feedback. Specifically, in each round $t \in [T]$, we construct the estimator of parameter $\boldsymbol{\theta}$ by solving the following equation:

$$\lambda\kappa\boldsymbol{\theta} + \textstyle\sum_{i=1}^{t-1} w_i\big(\sigma(\boldsymbol{\phi}_i^\top\boldsymbol{\theta}) - o_i\big)\boldsymbol{\phi}_i = \mathbf{0}, \tag{4.1}$$

where we denote $\boldsymbol{\phi}_i = \boldsymbol{\phi}(x_i, a_i) - \boldsymbol{\phi}(x_i, b_i)$ for simplicity, w_i is the uncertainty weight we are going to choose. To obtain an intuitive understanding of our weight, we consider any action-observation sequence $(x_1, a_1, b_1, o_1, x_2, a_2, b_2, o_2, \ldots, x_t, a_t, b_t, o_t)$ up to round t. For simplicity, we denote $\mathcal{F}_t = \sigma(x_1, a_1, b_1, o_1, x_2, a_2, b_2, o_2, \ldots, x_t, a_t, b_t)$ as the filtration. Suppose the estimated parameter $\boldsymbol{\theta}_t$ is the solution to the unweighted version equation of (4.1), i.e.,

$$\lambda\kappa\boldsymbol{\theta}_t + \textstyle\sum_{i=1}^{t}\big(\sigma(\boldsymbol{\phi}_i^\top\boldsymbol{\theta}_t) - o_i\big)\boldsymbol{\phi}_i = \mathbf{0}. \tag{4.2}$$

When we receive $\boldsymbol{\phi}_t = \boldsymbol{\phi}(x_t, a_t) - \boldsymbol{\phi}(x_t, b_t)$, the probability of receiving $l_t = 1$ can be estimated by $\sigma(\boldsymbol{\phi}_t^\top\boldsymbol{\theta}_t)$. We consider the conditional variance of the estimated probability $\sigma(\boldsymbol{\phi}_t^\top\boldsymbol{\theta}_t)$ in round t, i.e.,$\mathrm{Var}\big[\sigma(\boldsymbol{\phi}_t^\top\boldsymbol{\theta}_t)|\mathcal{F}_t\big]$, involving a posterior estimate of the prediction's variance. First, we have

$$\begin{aligned}
\mathbb{E}\big[\sigma(\boldsymbol{\phi}_t^\top\boldsymbol{\theta}_t)|\mathcal{F}_t\big] &\approx \mathbb{E}\big[\sigma(\boldsymbol{\phi}_t^\top\boldsymbol{\theta}^*) + \sigma'(\boldsymbol{\phi}_t^\top\boldsymbol{\theta}^*)\boldsymbol{\phi}_t^\top(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*)|\mathcal{F}_t\big] \\
&= \mathbb{E}\big[\underbrace{\sigma(\boldsymbol{\phi}_t^\top\boldsymbol{\theta}^*) - \sigma'(\boldsymbol{\phi}_t^\top\boldsymbol{\theta}^*)\boldsymbol{\phi}_t^\top\boldsymbol{\theta}^*}_{\mathcal{F}_t-\text{measurable}}|\mathcal{F}_t\big] + \mathbb{E}\big[\sigma'(\boldsymbol{\phi}_t^\top\boldsymbol{\theta}^*)\boldsymbol{\phi}_t^\top\boldsymbol{\theta}_t|\mathcal{F}_t\big].
\end{aligned}$$

Moreover, using the Taylor's expansion to (4.2), we have

$$\begin{aligned}
\mathbf{0} &= \lambda\kappa\boldsymbol{\theta}_t + \textstyle\sum_{i=1}^{t}\big(\sigma(\boldsymbol{\phi}_i^\top\boldsymbol{\theta}_t) - o_i\big)\boldsymbol{\phi}_i \\
&\approx \Big(\lambda\kappa\mathbf{I} + \textstyle\sum_{i=1}^{t}\sigma'(\boldsymbol{\phi}_i^\top\boldsymbol{\theta}^*)\boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top\Big)\boldsymbol{\theta}_t + \sum_{i=1}^{t}\big(\sigma(\boldsymbol{\phi}_i^\top\boldsymbol{\theta}^*) - o_i\big)\boldsymbol{\phi}_i - \sum_{i=1}^{t}\sigma'(\boldsymbol{\phi}_i^\top\boldsymbol{\theta}^*)\boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top\boldsymbol{\theta}^*.
\end{aligned}$$

Let $\boldsymbol{\Lambda}_t = \lambda\kappa\mathbf{I} + \sum_{i=1}^{t}\sigma'(\boldsymbol{\phi}_i^\top\boldsymbol{\theta}^*)\boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top$, we have

$$\begin{aligned}
\boldsymbol{\theta}_t &\approx \boldsymbol{\Lambda}_t^{-1}\Big[\textstyle\sum_{i=1}^{t}\sigma'(\boldsymbol{\phi}_i^\top\boldsymbol{\theta}^*)\boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top\boldsymbol{\theta}^* - \sum_{i=1}^{t}\big(\sigma(\boldsymbol{\phi}_i^\top\boldsymbol{\theta}^*) - o_i\big)\boldsymbol{\phi}_i\Big] \\
&= \underbrace{\boldsymbol{\Lambda}_t^{-1}\Big[\textstyle\sum_{i=1}^{t}\sigma'(\boldsymbol{\phi}_i^\top\boldsymbol{\theta}^*)\boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top\boldsymbol{\theta}^* - \sum_{i=1}^{t-1}\big(\sigma(\boldsymbol{\phi}_i^\top\boldsymbol{\theta}^*) - o_i\big)\boldsymbol{\phi}_i - \sigma(\boldsymbol{\phi}_t^\top\boldsymbol{\theta}^*)\Big]}_{\mathcal{F}_t-\text{measurable}} + o_t\boldsymbol{\Lambda}_t^{-1}\boldsymbol{\phi}_t
\end{aligned}$$

Therefore, the variance of the estimated preference probability can be approximated by

$$\begin{aligned}
\mathrm{Var}\big[\sigma(\boldsymbol{\phi}_t^\top\boldsymbol{\theta}_t)|\mathcal{F}_t\big] &= \mathbb{E}\big[\big(\sigma(\boldsymbol{\phi}_t^\top\boldsymbol{\theta}_t) - \mathbb{E}\big[\sigma(\boldsymbol{\phi}_t^\top\boldsymbol{\theta}_t)|\mathcal{F}_t\big]\big)^2|\mathcal{F}_t\big] \\
&\approx \mathbb{E}\Big[\Big(\mathbb{E}\big[o_t\sigma'(\boldsymbol{\phi}_t^\top\boldsymbol{\theta}^*)\boldsymbol{\phi}_t^\top\boldsymbol{\Lambda}_t^{-1}\boldsymbol{\phi}_t|\mathcal{F}_t\big]\Big)^2\Big|\mathcal{F}_t\Big] \\
&\le \mathbb{E}\big[o_t[\sigma'(\boldsymbol{\phi}_t^\top\boldsymbol{\theta}^*)]^2\|\boldsymbol{\phi}_t\|_{\boldsymbol{\Lambda}_t^{-1}}^2|\mathcal{F}_t\big] \le [\sigma'(\boldsymbol{\phi}_t^\top\boldsymbol{\theta}^*)]^2\|\boldsymbol{\phi}_t\|_{\boldsymbol{\Lambda}_t^{-1}}^2,
\end{aligned}$$

where the first inequality holds due to the Jensen's inequality and $o_t^2 = o_t$, and the last inequality holds due to $\mathbb{E}[o_t|\mathcal{F}_t] \le 1$. Using $\sigma'(\boldsymbol{\phi}_t^\top\boldsymbol{\theta}^*) \le 1$, $\boldsymbol{\phi}_t^\top\boldsymbol{\theta}^* \le 1$, $\boldsymbol{\Lambda}_t \ge \kappa\boldsymbol{\Sigma}_{t+1} \ge \kappa\boldsymbol{\Sigma}_t$, we can see that $\mathrm{Var}\big[\sigma(\boldsymbol{\phi}_t^\top\boldsymbol{\theta}_t)|\mathcal{F}_t\big] \le \kappa^{-1}\|\boldsymbol{\phi}_t\|_{\boldsymbol{\Sigma}_t^{-1}}^2$. Since higher variance leads to larger uncertainty, which harms the credibility of the data, it is natural to assign a smaller weight to the data with high uncertainty. Thus, we choose the weight to cancel out the uncertainty as follows

$$w_i = \min\{1, \alpha/\|\boldsymbol{\phi}_i\|_{\boldsymbol{\Sigma}_i^{-1}}\}, \tag{4.3}$$

where $\alpha/\|\boldsymbol{\phi}_i\|_{\boldsymbol{\Sigma}_i^{-1}}$ normalizes the variance of the estimated probability. To prevent excessively large weights, we apply truncation to this value. A similar weight has been used in He et al. (2022) for linear contextual bandits under corruption. Different from their setting where the weight is an estimate of the variance of the linear model, our weight is an estimate of a generalized linear model.

Furthermore, by selecting a proper threshold parameter, e.g., $\alpha = \sqrt{d}/C$, the weighted MLE shares the same confidence radius with that of the no-adversary scenario.

After constructing the estimator $\boldsymbol{\theta}_t$ from the weighted MLE, the sum of the estimated reward for each duel (a, b) can be calculated as $\big(\boldsymbol{\phi}(x_t, a) + \boldsymbol{\phi}(x_t, b)\big)^\top \boldsymbol{\theta}_t$. To encourage the exploration of duel (a, b) with high uncertainty during the learning process, we introduce an exploration bonus with the following $\beta\big\|\boldsymbol{\phi}(x_t, a) - \boldsymbol{\phi}(x_t, b)\big\|_{\boldsymbol{\Sigma}_t^{-1}}$, which follows a similar spirit to the bonus term in the context of linear bandit problems (Abbasi-Yadkori et al., 2011). However, the reward term and the bonus term exhibit different combinations of the feature maps $\boldsymbol{\phi}(x_t, a)$ and $\boldsymbol{\phi}(x_t, b)$, which is the key difference between bandits and dueling bandits. The selection of action pairs (a, b) is subsequently determined by maximizing the estimated reward with the exploration bonus term, i.e.,

$$\big(\boldsymbol{\phi}(x_t, a) + \boldsymbol{\phi}(x_t, b)\big)^\top \boldsymbol{\theta}_t + \beta\big\|\boldsymbol{\phi}(x_t, a) - \boldsymbol{\phi}(x_t, b)\big\|_{\boldsymbol{\Sigma}_t^{-1}}.$$

More discussion about the selection rule was discussed in Appendix A of Di et al. (2023).

Algorithm 1 Robust Contextual Dueling Bandit (RCDB)

1: **Require:** $\alpha > 0$, Regularization parameter λ, confidence radius β.
2: **for** $t = 1, \ldots, T$ **do**
3: Compute $\boldsymbol{\Sigma}_t = \lambda \mathbf{I} + \sum_{i=1}^{t-1} w_i \big(\boldsymbol{\phi}(x_i, a_i) - \boldsymbol{\phi}(x_i, b_i)\big)\big(\boldsymbol{\phi}(x_i, a_i) - \boldsymbol{\phi}(x_i, b_i)\big)^\top$.
4: Calculate the MLE $\boldsymbol{\theta}_t$ by solving the following equation:

$$\lambda\kappa\boldsymbol{\theta} + \sum_{i=1}^{t-1} w_i \Big[\sigma\Big(\big(\boldsymbol{\phi}(x_i, a_i) - \boldsymbol{\phi}(x_i, b_i)\big)^\top \boldsymbol{\theta}\Big) - o_i\Big]\big(\boldsymbol{\phi}(x_i, a_i) - \boldsymbol{\phi}(x_i, b_i)\big) = \mathbf{0}. \quad (4.4)$$

5: Observe the context vector x_t.
6: Choose $a_t, b_t = \operatorname{argmax}_{a,b} \Big\{\big(\boldsymbol{\phi}(x_t, a) + \boldsymbol{\phi}(x_t, b)\big)^\top \boldsymbol{\theta}_t + \beta\big\|\boldsymbol{\phi}(x_t, a) - \boldsymbol{\phi}(x_t, b)\big\|_{\boldsymbol{\Sigma}_t^{-1}}\Big\}$.
7: The adversary sees the feedback $l_t = \mathbb{1}(a_t \succ b_t)$ and decides the indicator c_t. Observe $o_t = l_t$ when $c_t = 0$, otherwise observe $o_t = 1 - l_t$.
8: Set weight w_t as (4.3).
9: **end for**

5 Main Results

5.1 Known Number of Adversarial Feedback

At the center of our algorithm design is the uncertainty-weighted MLE. When faced with adversarial feedback, the estimation error of the weighted MLE $\boldsymbol{\theta}_t$ can be characterized by the following lemma.

Lemma 5.1. If we set $\beta = \sqrt{\lambda}B + \big(\alpha C + \sqrt{d \log((1 + 2T/\lambda)/\delta)}\big)/\kappa$, then with probability at least $1 - \delta$, for any $t \in [T]$, we have

$$\big\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\big\|_{\boldsymbol{\Sigma}_t} \le \beta.$$

Remark 5.2. If we set $\alpha = (\sqrt{d} + \sqrt{\lambda}B)/C$, then the bonus radius β has no direct dependency on the number of adversarial feedback C. This observation plays a key role in proving the adversarial term in the regret without polynomial dependence on the total number of rounds T.

With Lemma 5.1, we can present the following regret guarantee of our algorithm RCDB in the dueling bandit framework.

Theorem 5.3. Under Assumption 3.1 and 3.2, let $0 < \delta < 1$, the total number of adversarial feedback be C. If we set the bonus radius to be

$$\beta = \sqrt{\lambda}B + \big(\alpha C + \sqrt{d \log((1 + 2T/\lambda)/\delta)}\big)/\kappa,$$

then with probability at least $1 - \delta$, the regret in the first t rounds can be upper bounded by

$$\begin{aligned}\text{Regret}(T) \le\, & 4\big(\sqrt{\lambda}B + \alpha C/\kappa\big)\sqrt{dT \log(1 + 2T/\lambda)} \\ & + 4d\big(\sqrt{T}/\kappa + \sqrt{\lambda}B/\alpha + 4C/\kappa\big)\log\big((1 + 2T/\lambda)/\delta\big) \\ & + 4d^{1.5}\sqrt{\log^3\big((1 + 2T/\lambda)/\delta\big)}/(\alpha\kappa).\end{aligned}$$

Moreover, if we set $\alpha = (\sqrt{d} + \sqrt{\lambda}B)/C$, $\lambda = 1/B^2$, the regret upper bound can be simplified to

$$\text{Regret}(T) = \widetilde{O}\big(d\sqrt{T}/\kappa + dC/\kappa\big).$$

Remark 5.4. Our regret bound consists of two terms. The first one is a C-independent term $\widetilde{O}(d\sqrt{T})$, which matches the lower bound $\widetilde{\Omega}(d\sqrt{T})$ proved in Bengs et al. (2022). This indicates that our result is optimal in scenarios without adversarial feedback ($C = 0$). Additionally, our result includes an additive term that is linearly dependent on the number of adversarial feedback C. When $C = O(\sqrt{T})$, the order of regret will be the same as the stochastic setting. It indicates the robustness of our algorithm to adversarial feedback. Additionally, the following theorem we present establishes a lower bound for this adversarial term, indicating that our dependency on the number of adversarial feedback C and the context dimension d is also optimal.

Theorem 5.5. For any dimension d, there exists an instance of dueling bandits with $|\mathcal{A}| = d$, such that any algorithm with the knowledge of the number of adversarial feedback C must incur $\Omega(dC)$ regret with probability at least 1/2.

Remark 5.6. The proof of Theorem 5.5 follows Bogunovic et al. (2021). In the constructed instances, only one action has reward 1, while others have 0. Compared with linear bandits, where the feedback is an exact reward, dueling bandits deal with the comparison between a pair of actions. A critical observation from our preference model, as formulated in (3.1), is that two actions with identical rewards result in a pair that is challenging to differentiate. The lower bound can be proved by corrupting every comparison into a random guess until the total times of adversarial feedback have been used up. For detailed proof, please refer to Section B.2. Our proved lower bound $\Omega(dC)$ shows that our result is nearly optimal because of the linear dependency on C, d and only logarithmic dependency on the total number of rounds T.

5.2 Unknown Number of Adversarial Feedback

In our previous analysis, the selection of parameters depends on having prior knowledge of the total number of adversarial feedback C. In this subsection, we extend our previous result to address the challenge posed by an unknown number of adversarial feedback C. Our approach to tackle this uncertainty follows He et al. (2022), we introduce an adversarial tolerance threshold $\bar{C}$ for the adversary count. This threshold can be regarded as an optimistic estimator of the actual number of adversarial feedback C. Under this situation, the subsequent theorem provides an upper bound for regret of Algorithm 1 in the case of an unknown number of adversarial feedback C.

Theorem 5.7. Under Assumptions 3.1 and 3.2, if we set the the confidence radius as

$$\beta = \sqrt{\lambda}B + \big[\alpha\bar{C} + \sqrt{d\log\big((1+2T/\lambda)/\delta\big)}\big]/\kappa,$$

with the pre-defined adversarial tolerance threshold $\bar{C}$ and $\alpha = (\sqrt{d}+\sqrt{\lambda}B)/\bar{C}$, then with probability at least $1-\delta$, the regret of Algorithm 1 can be upper bounded as following:

- If the actual number of adversarial feedback C is smaller than the adversarial tolerance threshold $\bar{C}$, then we have

$$\text{Regret}(T) = \widetilde{O}\big(d\sqrt{T}/\kappa + d\bar{C}/\kappa\big).$$

- If the actual number of adversarial feedback C is larger than the adversarial tolerance threshold $\bar{C}$, then we have $\text{Regret}(T) = O(T)$.

Remark 5.8. The COBE framework (Wei et al., 2022) converts any algorithm with the known adversarial level to an algorithm in the unknown case. However, such a framework only works for weak adversaries and does not work in our strong adversary setting. In fact, He et al. (2022) proved that any algorithm cannot simultaneously achieve near-optimal regret when uncorrupted and maintain sublinear regret with corruption level $C = \Omega(\sqrt{T})$. Therefore, there exists a trade-off between robust adversarial defense and near-optimal algorithmic performance. Our algorithm achieves the same nearly optimal $\widetilde{O}(d\sqrt{T})$ regret as the no-adversary case even when $C = \Theta(\sqrt{T})$, which indicates that our results are optimal in the presence of an unknown number of adversarial feedback.

6 Experiments

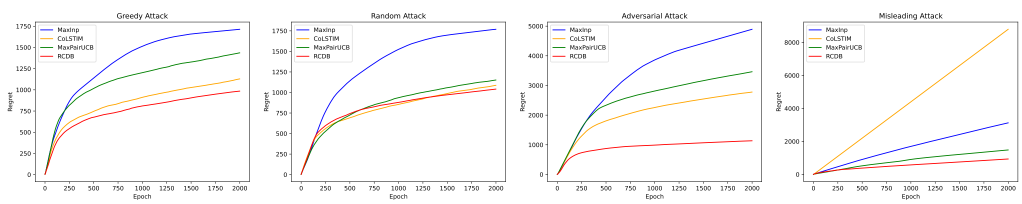
6.1 Experiment Setup

Preference Model. We study the effect of adversarial feedback with the preference model determined by (3.1), where $\sigma(x) = 1/(1+e^{-x})$. We randomly generate the underlying parameter in $[-0.5, 0.5]^d$ and normalize it to be a vector with $\|\boldsymbol{\theta}^*\|_2 = 2$. Then, we set it to be the underlying parameter and construct the reward utilized in the preference model as $r^*(x,a) = \langle\boldsymbol{\theta}^*, \boldsymbol{\phi}(x,a)\rangle$. We set the action set $\mathcal{A} = \big\{-1/\sqrt{d}, 1/\sqrt{d}\big\}^d$. For simplicity, we assume $\boldsymbol{\phi}(x,a) = a$. In our experiment, we set the dimension $d = 5$, with the size of action set $|\mathcal{A}| = 2^d = 32$.

Adversarial Attack Methods. We study the performance of our algorithm using different adversarial attack methods. We categorize the first two methods as "weak" primarily because the adversary in these scenarios does not utilize information about the agent's actions. In contrast, we classify the latter two methods as "strong" attacks. In these cases, the adversary leverages a broader scope of information, including knowledge of the actions selected by the agent and the true preference model. This enables it to devise more targeted adversarial methods.

- "Greedy Attack": The adversary will flip the preference label for the first C rounds. After that, it will not corrupt the result anymore.
- "Random Attack": In each round, the adversary will flip the preference label with the probability of $0 < p < 1$, until the times of adversarial feedback reach C.
- "Adversarial Attack": The adversary can have access to the true preference model. It will only flip the preference label when it aligns with the preference model, i.e., the probability for the preference model to make that decision is larger than 0.5, until the times of adversarial feedback reach C.
- "Misleading Attack": The adversary selects a suboptimal action. It will make sure this arm is always the winner in the comparison until the times of adversarial feedback reach C. In this way, it will mislead the agent to believe this action is the optimal one.

Experiment Setup. For each experiment instance, we simulate the interaction with the environment for $T = 2000$ rounds. In each round, the feedback for the action pair selected by the algorithm is generated according to the defined preference model. Subsequently, the adversary observes both the selected actions and their corresponding feedback and then engages in one of the previously described adversarial attack methods. We report the regret defined in (3.2) averaged across 10 random runs.



(a) Greedy Attack (b) Random Attack (c) Adversarial Attack (d) Misleading Attack

Figure 1: Comparison of `RCDB` (Our Algorithm 1), `MaxInp` (Saha, 2021), `CoLSTIM` (Bengs et al., 2022) and `MaxPairUCB` (Di et al., 2023). We report the cumulative regret with various adversarial attack methods (Greedy, Random, Adversarial, Misleading). For the baselines, the parameters are carefully tuned to achieve better results with different attack methods. The total number of adversarial feedback is $C = \lceil \sqrt{T} \rceil$.

6.2 Performance Comparison

We first introduce the algorithms studied in this section.

- `MaxInP`: Maximum Informative Pair by Saha (2021). It involves maintaining a standard MLE. With the estimated model, it then identifies a set of promising arms possible to beat the rest. The selection of arm pairs is then strategically designed to maximize the uncertainty in the difference between the two arms within this promising set, referred to as "maximum informative".
- `CoLSTIM`: The method by Bengs et al. (2022). It involves maintaining a standard MLE for the estimated model. Based on this model, the first arm is selected as the one with the highest estimated reward, implying it is the most likely to prevail over competitors. The second arm is selected to be the first arm's toughest competitor, with an added uncertainty bonus.
- `MaxPairUCB`: This algorithm was proposed in Di et al. (2023). It uses the regularized MLE to estimate the parameter $\boldsymbol{\theta}^*$. Then it selects the actions based on a symmetric action selection rule, i.e. the actions with the largest estimated reward plus some uncertainty bonus.
- `RCDB`: Algorithm 1 proposed in this paper. The key difference from the other algorithms is the use of uncertainty weight in the calculation of MLE (4.4). The we use the same symmetric action selection rule as `MaxPairUCB`. Our experiment results show that the uncertainty weight is critical in the face of adversarial feedback.

Our results are demonstrated in Figure 1. In Figure 1(a) and Figure 1(b), we observe scenarios where the adversary is "weak" due to the lack of access to information regarding the selected actions and the underlying preference model. Notably, in these situations, our algorithm `RCDB` outperforms all other baseline algorithms, demonstrating its robustness. Among the other algorithms, `CoLSTIM` performs as the strongest competitor.

In Figure 1(c), the adversary employs a 'stronger' adversarial method. Due to the inherent randomness of the model, some labels may naturally be 'incorrect'. An adversary with knowledge of the selected actions and the preference model can strategically neglect these naturally incorrect labels and selectively flip the others. This method proves catastrophic for algorithms to learn the true model, as it results in the agent encountering only incorrect preference labels at the beginning. Our results indicate that this leads to significantly higher regret. However, it's noteworthy that our algorithm `RCDB` demonstrates considerable robustness.

In Figure 1(d), the adversary employs a strategy aimed at misleading algorithms into believing a suboptimal action is the best choice. The algorithm `CoLSTIM` appears to be the most susceptible to being cheated by this method. Despite the deployment of 'strong' adversarial methods, as shown in both Figure 1(c) and Figure 1(d), our algorithm, `RCDB`, consistently demonstrates exceptional robustness against these attacks. A significant advantage of `RCDB` lies in that our parameter is selected solely based on the number of adversarial feedback C, irrespective of the nature of the adversarial methods employed. This contrasts with other algorithms where parameter tuning must be specifically adapted for each distinct adversarial method.

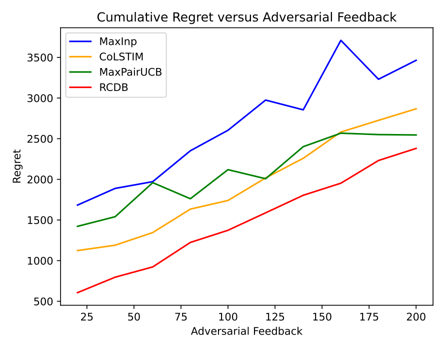


Figure 2: The relationship between cumulative regret and the number of adversarial feedback C. For this specific experiment, we employ the "greedy attack" method to generate the adversarial feedback. C is selected from the set $[20, 40, 60, 80, 100, 120, 140, 160, 180, 200]$ (10 adversarial levels).

6.3 Robustness to Different Numbers of Adversarial Feedback

In this section, we test the performance of algorithms with increasing times of adversarial feedback. Our results show a linear dependency on the number of adversarial feedback C, which is consistent with the theoretical results we have proved in Theorem 5.3 and 5.5. In comparison to other algorithms, `RCDB` demonstrates superior robustness against adversarial feedback, as evidenced by its notably smaller regret.

7 Conclusion

In this paper, we focus on the contextual dueling bandit problem from adversarial feedback. We introduce a novel algorithm, `RCDB`, which utilizes an uncertainty-weighted Maximum Likelihood Estimator (MLE) approach. This algorithm not only achieves optimal theoretical results in scenarios with and without adversarial feedback but also demonstrates superior performance with synthetic data. For future direction, we aim to extend our uncertainty-weighted method to encompass more general settings involving preference-based data. A particularly promising future direction of our research lies in addressing adversarial feedback within the process of aligning large language models using Reinforcement Learning from Human Feedback (RLHF).

Limitations. We assume that the reward is linear with respect to some known feature maps. Although this setting is common in the literature, we observe that some recent works on dueling bandits can deal with nonlinear rewards (Li et al., 2024). Therefore, it's possible to extend our results to a more general setting. Another assumption concerns the lower bound of the derivative of the link function. Notably, in the logistic bandit model, which shares similarities with our setting through Bernoulli variables, some work (Abeille et al., 2021; Faury et al., 2022) can improve the dependency of κ from $1/\kappa$ to $\sqrt{\kappa}$. A similar improvement might be achieved in our setting as well.

References

Abbasi-Yadkori, Y., Pál, D. and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*.

Abeille, M., Faury, L. and Calauzènes, C. (2021). Instance-wise minimax-optimal algorithms for logistic bandits. In *International Conference on Artificial Intelligence and Statistics*. PMLR.

Agarwal, A., Agarwal, S. and Patil, P. (2021). Stochastic dueling bandits with adversarial corruption. In *Algorithmic Learning Theory*. PMLR.

Agrawal, S. and Goyal, N. (2012). Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*. JMLR Workshop and Conference Proceedings.

Ailon, N., Karnin, Z. and Joachims, T. (2014). Reducing dueling bandits to cardinal bandits. In *International Conference on Machine Learning*. PMLR.

Auer, P. (2002). Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research* **3** 397–422.

Auer, P., Cesa-Bianchi, N. and Fischer, P. (2002a). Finite-time analysis of the multiarmed bandit problem. *Machine Learning* **47** 235–256.

Auer, P., Cesa-Bianchi, N., Freund, Y. and Schapire, R. E. (2002b). The nonstochastic multiarmed bandit problem. *SIAM journal on computing* **32** 48–77.

Auer, P. and Chiang, C.-K. (2016). An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. In *Conference on Learning Theory*. PMLR.

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

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

Bogunovic, I., Losalka, A., Krause, A. and Scarlett, J. (2021). Stochastic linear bandits robust to adversarial attacks. In *International Conference on Artificial Intelligence and Statistics*. PMLR.

Bubeck, S., Cesa-Bianchi, N. et al. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning* **5** 1–122.

Bubeck, S. and Slivkins, A. (2012). The best of both worlds: Stochastic and adversarial bandits. In *Conference on Learning Theory*. JMLR Workshop and Conference Proceedings.

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

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

Ding, Q., Hsieh, C.-J. and Sharpnack, J. (2022). Robust stochastic linear contextual bandits under adversarial attacks. In *International Conference on Artificial Intelligence and Statistics*. PMLR.

Dudík, M., Hofmann, K., Schapire, R. E., Slivkins, A. and Zoghi, M. (2015). Contextual dueling bandits. In *Conference on Learning Theory*. PMLR.

Falahatgar, M., Hao, Y., Orlitsky, A., Pichapati, V. and Ravindrakumar, V. (2017). Maxing and ranking with few assumptions. *Advances in Neural Information Processing Systems* **30**.

Faury, L., Abeille, M., Calauzènes, C. and Fercoq, O. (2020). Improved optimistic algorithms for logistic bandits. In *International Conference on Machine Learning*. PMLR.

Faury, L., Abeille, M., Jun, K.-S. and Calauzènes, C. (2022). Jointly efficient and optimal algorithms for logistic bandits. In *International Conference on Artificial Intelligence and Statistics*. PMLR.

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

Gajane, P., Urvoy, T. and Clérot, F. (2015). A relative exponential weighing algorithm for adversarial utility-based dueling bandits. In *International Conference on Machine Learning*. PMLR.

Gupta, A., Koren, T. and Talwar, K. (2019). Better algorithms for stochastic bandits with adversarial corruptions. In *Conference on Learning Theory*. PMLR.

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

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

Jamieson, K., Katariya, S., Deshpande, A. and Nowak, R. (2015). Sparse dueling bandits. In *Artificial Intelligence and Statistics*. PMLR.

Kalyanakrishnan, S., Tewari, A., Auer, P. and Stone, P. (2012). Pac subset selection in stochastic multi-armed bandits. In *ICML*, vol. 12.

Komiyama, J., Honda, J., Kashima, H. and Nakagawa, H. (2015). Regret lower bound and optimal algorithm in dueling bandit problem. In *Conference on learning theory*. PMLR.

Komiyama, J., Honda, J. and Nakagawa, H. (2016). Copeland dueling bandit problem: Regret lower bound, optimal algorithm, and computationally efficient algorithm. In *International Conference on Machine Learning*. PMLR.

Kuroki, Y., Rumi, A., Tsuchiya, T., Vitale, F. and Cesa-Bianchi, N. (2023). Best-of-both-worlds algorithms for linear contextual bandits. *arXiv preprint arXiv:2312.15433* .

Lai, T. L. (1987). Adaptive treatment allocation and the multi-armed bandit problem. *The annals of statistics* 1091–1114.

Lai, T. L., Robbins, H. et al. (1985). Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics* **6** 4–22.

Lattimore, T. and Szepesvári, C. (2020). *Bandit Algorithms*. Cambridge University Press.

Lee, C.-W., Luo, H., Wei, C.-Y., Zhang, M. and Zhang, X. (2021). Achieving near instance-optimality and minimax-optimality in stochastic and adversarial linear bandits simultaneously. In *International Conference on Machine Learning*. PMLR.

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

Li, X., Zhao, H. and Gu, Q. (2024). Feel-good thompson sampling for contextual dueling bandits. *arXiv preprint arXiv:2404.06013* .

Li, Y., Lou, E. Y. and Shan, L. (2019). Stochastic linear optimization with adversarial corruption. *arXiv preprint arXiv:1909.02109* .

Lykouris, T., Mirrokni, V. and Paes Leme, R. (2018). Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A. et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* **35** 27730–27744.

RAMAMOHAN, S. Y., RAJKUMAR, A. and AGARWAL, S. (2016). Dueling bandits: Beyond condorcet winners to general tournament solutions. *Advances in Neural Information Processing Systems* **29**.

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

SAHA, A. and GAILLARD, P. (2022). Versatile dueling bandits: Best-of-both world analyses for learning from relative preferences. In *International Conference on Machine Learning*. PMLR.

SAHA, A., KOREN, T. and MANSOUR, Y. (2021). Adversarial dueling bandits. In *International Conference on Machine Learning*. PMLR.

SAHA, A. and KRISHNAMURTHY, A. (2022). Efficient and optimal algorithms for contextual dueling bandits under realizability. In *International Conference on Algorithmic Learning Theory*. PMLR.

SEKHARI, A., SRIDHARAN, K., SUN, W. and WU, R. (2023). Contextual bandits and imitation learning via preference-based active queries. *arXiv preprint arXiv:2307.12926* .

SELDIN, Y. and LUGOSI, G. (2017). An improved parametrization and analysis of the exp3++ algorithm for stochastic and adversarial bandits. In *Conference on Learning Theory*. PMLR.

SELDIN, Y. and SLIVKINS, A. (2014). One practical algorithm for both stochastic and adversarial bandits. In *International Conference on Machine Learning*. PMLR.

WEI, C.-Y., DANN, C. and ZIMMERT, J. (2022). A model selection approach for corruption robust reinforcement learning. In *International Conference on Algorithmic Learning Theory*. PMLR.

WU, H. and LIU, X. (2016). Double thompson sampling for dueling bandits. *Advances in neural information processing systems* **29**.

WU, Y., JIN, T., LOU, H., FARNOUD, F. and GU, Q. (2023). Borda regret minimization for generalized linear dueling bandits. *arXiv preprint arXiv:2303.08816* .

XIONG, W., DONG, H., YE, C., ZHONG, H., JIANG, N. and ZHANG, T. (2023). Gibbs sampling from human feedback: A provable kl-constrained framework for rlhf. *arXiv preprint arXiv:2312.11456* .

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

YU, T., QUILLEN, D., HE, Z., JULIAN, R., HAUSMAN, K., FINN, C. and LEVINE, S. (2020). Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on robot learning*. PMLR.

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

ZHAO, H., ZHOU, D. and GU, Q. (2021). Linear contextual bandits with adversarial corruptions. *arXiv preprint arXiv:2110.12615* .

ZHU, H., YU, J., GUPTA, A., SHAH, D., HARTIKAINEN, K., SINGH, A., KUMAR, V. and LEVINE, S. (2020). The ingredients of real-world robotic reinforcement learning. *arXiv preprint arXiv:2004.12570* .

ZIMMERT, J. and SELDIN, Y. (2019). An optimal algorithm for stochastic and adversarial bandits. In *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR.

ZOGHI, M., KARNIN, Z. S., WHITESON, S. and DE RIJKE, M. (2015). Copeland dueling bandits. *Advances in neural information processing systems* **28**.

ZOGHI, M., WHITESON, S., MUNOS, R. and RIJKE, M. (2014). Relative upper confidence bound for the k-armed dueling bandit problem. In *International conference on machine learning*. PMLR.

Broader Impact

This paper studies contextual dueling bandits with adversarial feedback. Our primary objective is to propel advancements in bandit theory by introducing a more robust algorithm backed by solid theoretical guarantees. The uncertainty-weighted approach we have developed for dueling bandits holds significant potential to address the issue of adversarial feedback in preference-based data, which could be instrumental in enhancing the robustness of generative models against adversarial attacks, thereby contributing positively to the societal impact and reliability of machine learning applications.

A Roadmap of the Proof

A.1 Uncertainty-weighted MLE with Adversarial Feedback

In this section, we offer an overview of the proof for Lemma 5.1. The general proof idea for the uncertainty-weighted MLE with adversarial feedback lies in decomposing the estimation error into three terms, a stochastic error term, an adversarial term, and an additional regularization term. Following the analysis of standard (weighted) MLE (Li et al., 2017), we introduce an auxiliary function:

$$\begin{aligned} G_t(\boldsymbol{\theta}) = \lambda\kappa\boldsymbol{\theta} + \sum_{i=1}^{t-1} w_i \Big[\sigma\Big(\big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big)^\top \boldsymbol{\theta}\Big) \\ - \sigma\Big(\big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big)^\top \boldsymbol{\theta}^*\Big)\Big]\big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big). \end{aligned}$$

It satisfies two conditions: First, for the true parameter value $\boldsymbol{\theta}^*$, $G_t(\boldsymbol{\theta}^*)$ has a simple expression, i.e.,

$$G_t(\boldsymbol{\theta}^*) = \lambda\kappa\boldsymbol{\theta}^*.$$

Second, according to (4.4), we can get the value of function G_t at the MLE $\boldsymbol{\theta}_t$,

$$G_t(\boldsymbol{\theta}_t) = \sum_{i=1}^{t-1} w_i\gamma_i\big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big), \tag{A.1}$$

where $\gamma_i = o_i - \sigma\big((\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i))^\top \boldsymbol{\theta}^*\big)$. To connect the desired estimation error with the function G_t, we use the mean value theorem. This leads to an upper bound of the estimation error:

$$\begin{aligned} \|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}_t} &\le \frac{1}{\kappa}\big\|G_t(\boldsymbol{\theta}_t) - G_t(\boldsymbol{\theta}^*)\big\|_{\boldsymbol{\Sigma}_t^{-1}} \\ &\le \underbrace{\frac{1}{\kappa}\lambda\|\boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}_t^{-1}}}_{\text{Regularization term}} + \underbrace{\frac{1}{\kappa}\big\|G_t(\boldsymbol{\theta}_t)\big\|_{\boldsymbol{\Sigma}_t^{-1}}}_{I_1}. \end{aligned}$$

For term I_1, we can decompose the summation in (A.1) based on the adversarial feedback c_t, i.e.,

$$G_t(\boldsymbol{\theta}_t) = \sum_{i<t:c_i=0} w_i\gamma_i\big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big) + \underbrace{\sum_{i<t:c_i=1} w_i\gamma_i\big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big)}_{I_2},$$

where I_2 can be further decomposed as

$$I_2 = \sum_{i<t:c_i=1} w_i\epsilon_i\big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big) + \sum_{i<t:c_i=1} w_i(\gamma_i - \epsilon_i)\big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big).$$

where $\epsilon_i = l_i - \sigma\big((\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i))^\top \boldsymbol{\theta}^*\big)$. With our notation of adversarial feedback, when $c_i = 0$, we have $\gamma_i = \epsilon_i$. Therefore, we have $|\gamma_i - \epsilon_i| \le 1$ and

$$I_1 \le \underbrace{\frac{1}{\kappa}\Big\|\sum_{i=1}^{t-1} w_i\epsilon_i\big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big)\Big\|_{\boldsymbol{\Sigma}_t^{-1}}}_{\text{Stochastic term}} + \underbrace{\frac{1}{\kappa}\Big\|\sum_{i<t:c_i=1} w_i\big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big)\Big\|_{\boldsymbol{\Sigma}_t^{-1}}}_{\text{Adversarial term}}.$$

The stochastic term can be upper bounded with the concentration inequality (Lemma D.2). Additionally, by employing our specifically chosen weight, as (4.3), we can control the adversarial term, with $w_i\|\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\|_{\boldsymbol{\Sigma}_t^{-1}} \le \alpha$. Therefore, the adversarial term can be bounded by $\alpha C/\kappa$.

A.2 Regret Upper Bound

With a similar discussion of the symmetric arm selection rule to Di et al. (2023), the regret defined in (3.2) can be bounded by

$$\text{Regret}(T) \le \sum_{t=1}^{T} \min\Big\{4, 2\beta\|\boldsymbol{\phi}(x_t,a_t)-\boldsymbol{\phi}(x_t,b_t)\|_{\boldsymbol{\Sigma}_t^{-1}}\Big\}.$$

Note that in our selection of weight w_t, it has two possible values. We decompose the summation based on the two cases separately. We have

$$\begin{aligned}\text{Regret}(T) \le & \underbrace{\sum_{w_t=1} \min\Big\{4, 2\beta\|\boldsymbol{\phi}(x_t,a_t)-\boldsymbol{\phi}(x_t,b_t)\|_{\boldsymbol{\Sigma}_t^{-1}}\Big\}}_{J_1} \\ & + \underbrace{\sum_{w_t<1} \min\Big\{4, 2\beta\|\boldsymbol{\phi}(x_t,a_t)-\boldsymbol{\phi}(x_t,b_t)\|_{\boldsymbol{\Sigma}_t^{-1}}\Big\}}_{J_2}.\end{aligned}$$

We consider J_1, J_2 separately. For the term J_1, we define $\boldsymbol{\Lambda}_t = \lambda\mathbf{I} + \sum_{i\le t-1, w_i=1}\big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big)\big(\boldsymbol{\phi}(x_i,a_i)-\boldsymbol{\phi}(x_i,b_i)\big)^\top$. Then, we have $\boldsymbol{\Sigma}_t \succeq \boldsymbol{\Lambda}_t$, and therefore

$$\|\boldsymbol{\phi}(x_t,a_t)-\boldsymbol{\phi}(x_t,b_t)\|_{\boldsymbol{\Sigma}_t^{-1}} \le \|\boldsymbol{\phi}(x_t,a_t)-\boldsymbol{\phi}(x_t,b_t)\|_{\boldsymbol{\Lambda}_t^{-1}}.$$

Using Lemma D.3 with $\mathbf{x}_t = \boldsymbol{\phi}(x_t,a_t)-\boldsymbol{\phi}(x_t,b_t)$, we have

$$J_1 \le 4\beta\sqrt{dT\log(1+2T/\lambda)}. \tag{A.2}$$

For term J_2, we note that $w_t < 1$ implies that $w_t = \alpha/\|\boldsymbol{\phi}(x_t,a_t)-\boldsymbol{\phi}(x_t,b_t)\|_{\boldsymbol{\Sigma}_t^{-1}}$. Therefore, we have

$$J_2 \le \sum_{t=1}^{T}\frac{4\beta}{\alpha}\min\Big\{1, \|\sqrt{w_t}(\boldsymbol{\phi}(x_t,a_t)-\boldsymbol{\phi}(x_t,b_t))\|^2_{\boldsymbol{\Sigma}_t^{-1}}\Big\}.$$

Using Lemma D.3 with $\mathbf{x}'_t = \sqrt{w_t}(\boldsymbol{\phi}(x_t,a_t)-\boldsymbol{\phi}(x_t,b_t))$, we have

$$J_2 \le \frac{4d\beta\log(1+2T/\lambda)}{\alpha}. \tag{A.3}$$

We conclude the proof of regret by combining (A.2) and (A.3).

B Proof of Theorems in Section 5

B.1 Proof of Theorem 5.3

In this subsection, we provide the proof of Theorem 5.3. We condition on the high-probability event in Lemma 5.1

$$\mathcal{E} = \Big\{\big\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\big\|_{\boldsymbol{\Sigma}_t} \le \beta, \forall t\in[T]\Big\}.$$

Let $r_t = 2r^*(x_t,a_t^*) - r^*(x_t,a_t) - r^*(x_t,b_t)$ be the regret incurred in round t. The following lemma provides the upper bound of r_t.

Lemma B.1. Let $0<\delta<1$. If we set $\beta = \sqrt{\lambda}B + \big(\alpha C + \sqrt{d\log((1+2T/\lambda)/\delta)}\big)/\kappa$, on event $\mathcal{E}$, the regret of Algorithm 1 incurred in round t can be upper bounded by

$$r_t \le \min\Big\{4, 2\beta\|\boldsymbol{\phi}(x_t,a_t)-\boldsymbol{\phi}(x_t,b_t)\|_{\boldsymbol{\Sigma}_t^{-1}}\Big\}.$$

Moreover, the regret can be upper bounded by

$$\text{Regret}(T) \le \sum_{t=1}^{T} \min\Big\{4, 2\beta\|\boldsymbol{\phi}(x_t,a_t)-\boldsymbol{\phi}(x_t,b_t)\|_{\boldsymbol{\Sigma}_t^{-1}}\Big\}.$$

With Lemma B.1, we can provide the proof of Theorem 5.3.

Proof of Theorem 5.3. Using Lemma B.1, the total regret can be upper bounded by

$$\mathrm{Regret}(T) \leq \sum_{t=1}^{T} \min\Big\{4, 2\beta\|\boldsymbol{\phi}(x_t, a_t) - \boldsymbol{\phi}(x_t, b_t)\|_{\boldsymbol{\Sigma}_t^{-1}}\Big\}.$$

Our weight w_t has two possible values. We decompose the summation based on the two cases separately. We have

$$\begin{aligned}\mathrm{Regret}(T) \leq & \underbrace{\sum_{w_t=1} \min\Big\{4, 2\beta\|\boldsymbol{\phi}(x_t, a_t) - \boldsymbol{\phi}(x_t, b_t)\|_{\boldsymbol{\Sigma}_t^{-1}}\Big\}}_{J_1} \\ & + \underbrace{\sum_{w_t<1} \min\Big\{4, 2\beta\|\boldsymbol{\phi}(x_t, a_t) - \boldsymbol{\phi}(x_t, b_t)\|_{\boldsymbol{\Sigma}_t^{-1}}\Big\}}_{J_2}.\end{aligned}$$

For the term J_1, we consider a partial summation in rounds when $w_t = 1$. Let $\boldsymbol{\Lambda}_t = \lambda\mathbf{I} + \sum_{i\leq k-1, w_i=1}\big(\boldsymbol{\phi}(x_i, a_i) - \boldsymbol{\phi}(x_i, b_i)\big)\big(\boldsymbol{\phi}(x_i, a_i) - \boldsymbol{\phi}(x_i, b_i)\big)^\top$. Then we have

$$\begin{aligned} J_1 &\leq 4\beta \sum_{t:w_t=1} \min\Big\{1, \|\boldsymbol{\phi}(x_t, a_t) - \boldsymbol{\phi}(x_t, b_t)\|_{\boldsymbol{\Sigma}_t^{-1}}\Big\} \\ &\leq 4\beta \sum_{t:w_t=1} \min\Big\{1, \|\boldsymbol{\phi}(x_t, a_t) - \boldsymbol{\phi}(x_t, b_t)\|_{\boldsymbol{\Lambda}_t^{-1}}\Big\} \\ &\leq 4\beta \sqrt{T \sum_{t:w_t=1} \min\big\{1, \|\boldsymbol{\phi}(x_t, a_t) - \boldsymbol{\phi}(x_t, b_t)\|^2_{\boldsymbol{\Lambda}_t^{-1}}\big\}} \\ &\leq 4\beta\sqrt{dT\log(1 + 2T/\lambda)}, \end{aligned} \tag{B.1}$$

where the second inequality holds due to $\boldsymbol{\Sigma}_t \succeq \boldsymbol{\Lambda}_t$. The third inequality holds due to the Cauchy-Schwartz inequality, The last inequality holds due to Lemma D.3.

For the term J_2, the weight in this summation satisfies $w_t < 1$, and therefore $w_t = \alpha/\|\boldsymbol{\phi}(x_t, a_t) - \boldsymbol{\phi}(x_t, b_t)\|_{\boldsymbol{\Sigma}_t^{-1}}$. Then we have

$$\begin{aligned} J_2 &= \sum_{w_t<1} \min\Big\{4, 2\beta\|\boldsymbol{\phi}(x_t, a_t) - \boldsymbol{\phi}(x_t, b_t)\|_{\boldsymbol{\Sigma}_t^{-1}} w_t \|\boldsymbol{\phi}(x_t, a_t) - \boldsymbol{\phi}(x_t, b_t)\|_{\boldsymbol{\Sigma}_t^{-1}}/\alpha\Big\} \\ &\leq \sum_{t=1}^{T} \min\Big\{4, 2\beta/\alpha\|\sqrt{w_t}(\boldsymbol{\phi}(x_t, a_t) - \boldsymbol{\phi}(x_t, b_t))\|^2_{\boldsymbol{\Sigma}_t^{-1}}\Big\} \\ &\leq \sum_{t=1}^{T} \frac{4\beta}{\alpha} \min\Big\{1, \|\sqrt{w_t}(\boldsymbol{\phi}(x_t, a_t) - \boldsymbol{\phi}(x_t, b_t))\|^2_{\boldsymbol{\Sigma}_t^{-1}}\Big\} \\ &\leq \frac{4d\beta\log(1 + 2T/\lambda)}{\alpha}, \end{aligned} \tag{B.2}$$

where the first equality holds due to the choice of w_t. The first inequality holds because each term in the summation is positive. The last inequality holds due to Lemma D.3. Combining (B.1) and (B.2), we complete the proof of Theorem 5.3. □

B.2 Proof of Theorem 5.5

Proof of Theorem 5.5. Our proof adapts the argument in Bogunovic et al. (2021) to dueling bandits. For any dimension d, we construct d instances, each with $\boldsymbol{\theta}_i = \mathbf{e}_i$, where $\mathbf{e}_i$ is the i-th standard basis vector. We set the action set $\mathcal{A} = \{\mathbf{e}_i\}_{i=1}^d$. Therefore, in the i-th instance, the reward for the i-th action will be 1. For the other actions, it will be 0. Therefore, the i-th action will be more preferable to any other action. While for other pairs, the feedback is simply a random guess.

Consider an adversary that knows the exact instance. When the comparison involves the i-th action, it will corrupt the feedback with a random guess. Otherwise, it will not corrupt. In the i-th instance, the adversary stops the adversarial attack only after C times of comparison involving the i-th action. However, after $Cd/4$ rounds, at least $d/2$ actions have not been compared for C times. For the instances corresponding to these actions, the agent learns no information and suffers from $\Omega(dC)$ regret. This completes the proof of Theorem 5.5. □

B.3 Proof of Theorem 5.7

Proof of Theorem 5.7. Here, based on the relationship between C and the threshold $\bar{C}$, we discuss two distinct cases separately.

- In the scenario where $\bar{C} < C$, Algorithm 1 can ensures a trivial regret bound, with the guarantee that $\text{Regret}(T) \le 2T$.
- In the scenario where $C \le \bar{C}$, we know that $\bar{C}$ is remains a valid upper bound on the number of adversarial feedback. Under this situation, Algorithm 1 operates successfully with $\bar{C}$ adversarial feedback. Therefore, according to Theorem 5.3, the regret is upper bounded by

$$\text{Regret}(T) \le \widetilde{O}(d\sqrt{T} + d\bar{C}).$$

$\square$

C Proof of Lemmas 5.1 and B.1

C.1 Proof of Lemma 5.1

Proof of Lemma 5.1. Using a similar reasoning in Li et al. (2017), we define some auxiliary quantities

$$\begin{aligned}
G_t(\boldsymbol{\theta}) &= \lambda\kappa\boldsymbol{\theta} + \sum_{i=1}^{t-1} w_i \Big[\sigma\Big(\big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big)^\top \boldsymbol{\theta}\Big) \\
&\quad - \sigma\Big(\big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big)^\top \boldsymbol{\theta}^*\Big)\Big]\big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big), \\
\epsilon_t &= l_t - \sigma\Big(\big(\boldsymbol{\phi}(x_t,a_t) - \boldsymbol{\phi}(x_t,b_t)\big)^\top \boldsymbol{\theta}^*\Big), \\
\gamma_t &= o_t - \sigma\Big(\big(\boldsymbol{\phi}(x_t,a_t) - \boldsymbol{\phi}(x_t,b_t)\big)^\top \boldsymbol{\theta}^*\Big), \\
Z_t &= \sum_{i=1}^{t-1} w_i\gamma_i\big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big).
\end{aligned}$$

In Algorithm 1, $\boldsymbol{\theta}_t$ is chosen to be the solution to the following equation,

$$\lambda\kappa\boldsymbol{\theta}_t + \sum_{i=1}^{t-1} w_i\Big[\sigma\Big(\big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big)^\top\boldsymbol{\theta}_t\Big) - o_i\Big]\big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big) = \mathbf{0}. \tag{C.1}$$

Then we have

$$\begin{aligned}
G_t(\boldsymbol{\theta}_t) &= \lambda\kappa\boldsymbol{\theta}_t + \sum_{i=1}^{t-1} w_i \Big[\sigma\Big(\big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big)^\top \boldsymbol{\theta}_t\Big) \\
&\quad - \sigma\Big(\big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big)^\top \boldsymbol{\theta}^*\Big)\Big]\big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big) \\
&= \sum_{i=1}^{t-1} w_i\Big[o_i - \sigma\Big(\big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big)^\top \boldsymbol{\theta}^*\Big)\Big]\big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big) \\
&= Z_t.
\end{aligned}$$

The analysis in Li et al. (2017); Di et al. (2023) shows that this equation has a unique solution, with $\boldsymbol{\theta}_t = G_t^{-1}(Z_t)$. Using the mean value theorem, for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^d$, there exists $m \in [0,1]$ and $\bar{\boldsymbol{\theta}} = m\boldsymbol{\theta}_1 + (1-m)\boldsymbol{\theta}_2$, such that the following equation holds,

$$\begin{aligned}
G_t(\boldsymbol{\theta}_1) - G_t(\boldsymbol{\theta}_2) &= \lambda\kappa(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2) + \sum_{i=1}^{t-1} w_i\Big[\sigma\Big(\big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big)^\top\boldsymbol{\theta}_1\Big) \\
&\quad - \sigma\Big(\big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big)^\top\boldsymbol{\theta}_2\Big)\Big]\big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big) \\
&= \Big[\lambda\kappa\mathbf{I} + \sum_{i=1}^{t-1} w_i\dot{\sigma}\Big(\big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big)^\top\bar{\boldsymbol{\theta}}\Big) \\
&\quad \big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big)\big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big)^\top\Big](\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2).
\end{aligned}$$

We define $F(\bar{\boldsymbol{\theta}})$ as

$$F(\bar{\boldsymbol{\theta}}) = \lambda\kappa\mathbf{I} + \sum_{i=1}^{t-1} w_i \dot{\sigma}\Big(\big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big)^\top \bar{\boldsymbol{\theta}}\Big)\big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big)\big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big)^\top\Big].$$

Moreover, we can see that $G_t(\boldsymbol{\theta}^*) = \lambda\kappa\boldsymbol{\theta}^*$. Recall $\boldsymbol{\Sigma}_t = \lambda\mathbf{I} + \sum_{i=1}^{t-1} w_i\big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big)\big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big)^\top$. We have

$$\begin{aligned}
\big\|G_t(\boldsymbol{\theta}_t) - G_t(\boldsymbol{\theta}^*)\big\|_{\boldsymbol{\Sigma}_t^{-1}}^2 &= (\boldsymbol{\theta}_t - \boldsymbol{\theta}^*)^\top F(\bar{\boldsymbol{\theta}})\boldsymbol{\Sigma}_t^{-1} F(\bar{\boldsymbol{\theta}})(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*) \\
&\geq \kappa^2(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*)^\top \boldsymbol{\Sigma}_t(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*) \\
&= \kappa^2\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}_t}^2,
\end{aligned}$$

where the first inequality holds due to $\dot{\mu}(\cdot) \geq \kappa > 0$ and $F(\bar{\boldsymbol{\theta}}) \succeq \kappa\boldsymbol{\Sigma}_t$. Then we have the following estimate of the estimation error:

$$\begin{aligned}
\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}_t} &\leq \frac{1}{\kappa}\big\|G_t(\boldsymbol{\theta}_t) - G_t(\boldsymbol{\theta}^*)\big\|_{\boldsymbol{\Sigma}_t^{-1}} \\
&\leq \lambda\|\boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}_t^{-1}} + \frac{1}{\kappa}\|Z_t\|_{\boldsymbol{\Sigma}_t^{-1}} \\
&\leq \sqrt{\lambda}\|\boldsymbol{\theta}^*\|_2 + \frac{1}{\kappa}\|Z_t\|_{\boldsymbol{\Sigma}_t^{-1}},
\end{aligned}$$

where the second inequality holds due to the triangle inequality and $G_t(\boldsymbol{\theta}^*) = \lambda\kappa\boldsymbol{\theta}^*$. The last inequality holds due to $\boldsymbol{\Sigma}_t \succeq \lambda\mathbf{I}$. Finally, we need to bound the $\|Z_t\|_{\boldsymbol{\Sigma}_t^{-1}}$ term. To study the impact of adversarial feedback, we decompose the summation in (A.1) based on the adversarial feedback c_t, i.e.,

$$Z_t = \sum_{i<t:c_i=0} w_i\gamma_i\big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big) + \sum_{i<t:c_i=1} w_i\gamma_i\big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big),$$

When $c_i = 1$, i.e. with adversarial feedback, $|\gamma_i - \epsilon_i| = 1$. On the contrary, when $c_i = 0$, $\gamma_i = \epsilon_i$. Therefore,

$$\begin{aligned}
\sum_{i<t:c_i=0} w_i\gamma_i\big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big) &= \sum_{i<t:c_i=0} w_i\epsilon_i\big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big), \\
\sum_{i<t:c_i=1} w_i\gamma_i\big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big) &= \sum_{i<t:c_i=1} w_i\epsilon_i\big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big) \\
&\quad + \sum_{i<t:c_i=1} w_i\big(\gamma_i - \epsilon_i)(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big).
\end{aligned}$$

Summing up the two equalties, we have

$$Z_t = \sum_{i=1}^{t-1} w_i\epsilon_i\big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big) + \sum_{i<t:c_i=1} w_i(\gamma_i - \epsilon_i)\big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big).$$

Therefore,

$$\|Z_t\|_{\boldsymbol{\Sigma}_t^{-1}} \leq \underbrace{\bigg\|\sum_{i=1}^{t-1} w_i\epsilon_i\big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big)\bigg\|_{\boldsymbol{\Sigma}_t^{-1}}}_{I_1} + \underbrace{\bigg\|\sum_{i<t:c_i=1} w_i\big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big)\bigg\|_{\boldsymbol{\Sigma}_t^{-1}}}_{I_2}.$$

For the term I_1, with probability at least $1-\delta$, for all $t \in [T]$, it can be bounded by

$$I_1 \leq \sqrt{2\log\Big(\frac{\det(\boldsymbol{\Sigma}_t)^{1/2}\det(\boldsymbol{\Sigma}_0)^{-1/2}}{\delta}\Big)},$$

due to Lemma D.2. Using $w_i \leq 1$, we have $\sqrt{w_i}\|\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\|_2 \leq 2$. Moreover, we have

$$\det(\boldsymbol{\Sigma}_t) \leq \bigg(\frac{\mathrm{Tr}(\boldsymbol{\Sigma}_t)}{d}\bigg)^d$$

$$
\begin{aligned}
&= \left(\frac{d\lambda + \sum_{i=1}^{t-1} w_i \|(\boldsymbol{\phi}(x_i, a_i) - \boldsymbol{\phi}(x_i, b_i))\|_2^2}{d} \right)^d \\
&\leq \left(\frac{d\lambda + 2T}{d} \right)^d,
\end{aligned}
$$

where the first inequality holds because for every matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$, $\det \mathbf{A} \leq (\mathrm{Tr}(\mathbf{A})/d)^d$. The second inequality holds due to $\sqrt{w_i}\|\boldsymbol{\phi}(x_i, a_i) - \boldsymbol{\phi}(x_i, b_i)\|_2 \leq 2$. Easy to see that $\det(\boldsymbol{\Sigma}_0) = \lambda^d$. The term I_1 can be bounded by

$$
I_1 \leq \sqrt{d \log((1 + 2T/\lambda)/\delta)}. \tag{C.2}
$$

For I_2, with our choice of the weight w_i, we have

$$
\begin{aligned}
I_2 &\leq \sum_{i<t: c_i=1} w_i \big\|(\boldsymbol{\phi}(x_i, a_i) - \boldsymbol{\phi}(x_i, b_i))\big\|_{\boldsymbol{\Sigma}_t^{-1}} \\
&\leq \sum_{i<t: c_i=1} w_i \big\|(\boldsymbol{\phi}(x_i, a_i) - \boldsymbol{\phi}(x_i, b_i))\big\|_{\boldsymbol{\Sigma}_i^{-1}} \\
&\leq \sum_{i<t: c_i=1} \alpha \\
&\leq \alpha C,
\end{aligned} \tag{C.3}
$$

where the second inequality holds due to $\boldsymbol{\Sigma}_t \succeq \boldsymbol{\Sigma}_i$. The third inequality holds due to $w_i \leq \alpha / \big\|(\boldsymbol{\phi}(x_i, a_i) - \boldsymbol{\phi}(x_i, b_i))\big\|_{\boldsymbol{\Sigma}_i^{-1}}$. The last inequality holds due to the definition of C. Combining (C.2) and (C.3), we complete the proof of Lemma 5.1. □

C.2 Proof of Lemma B.1

Proof of Lemma B.1. Let the regret incurred in the t-th round by $r_t = 2r^*(x_t, a_t^*) - r^*(x_t, a_t) - r^*(x_t, b_t)$. It can be decomposed as

$$
\begin{aligned}
r_t &= 2r^*(x_t, a_t^*) - r^*(x_t, a_t) - r^*(x_t, b_t) \\
&= \langle \boldsymbol{\phi}(x_t, a_t^*) - \boldsymbol{\phi}(x_t, a_t), \boldsymbol{\theta}^* \rangle + \langle \boldsymbol{\phi}(x_t, a_t^*) - \boldsymbol{\phi}(x_t, b_t), \boldsymbol{\theta}^* \rangle \\
&= \langle \boldsymbol{\phi}(x_t, a_t^*) - \boldsymbol{\phi}(x_t, a_t), \boldsymbol{\theta}^* - \boldsymbol{\theta}_t \rangle + \langle \boldsymbol{\phi}(x_t, a_t^*) - \boldsymbol{\phi}(x_t, b_t), \boldsymbol{\theta}^* - \boldsymbol{\theta}_t \rangle \\
&\quad + \langle 2\boldsymbol{\phi}(x_t, a_t^*) - \boldsymbol{\phi}(x_t, a_t) - \boldsymbol{\phi}(x_t, b_t), \boldsymbol{\theta}_t \rangle \\
&\leq \|\boldsymbol{\phi}(x_t, a_t^*) - \boldsymbol{\phi}(x_t, a_t)\|_{\boldsymbol{\Sigma}_t^{-1}} \|\boldsymbol{\theta}^* - \boldsymbol{\theta}_t\|_{\boldsymbol{\Sigma}_t} + \|\boldsymbol{\phi}(x_t, a_t^*) - \boldsymbol{\phi}(x_t, b_t)\|_{\boldsymbol{\Sigma}_t^{-1}} \|\boldsymbol{\theta}^* - \boldsymbol{\theta}_t\|_{\boldsymbol{\Sigma}_t} \\
&\quad + \langle 2\boldsymbol{\phi}(x_t, a_t^*) - \boldsymbol{\phi}(x_t, a_t) - \boldsymbol{\phi}(x_t, b_t), \boldsymbol{\theta}_t \rangle \\
&\leq \beta \|\boldsymbol{\phi}(x_t, a_t^*) - \boldsymbol{\phi}(x_t, a_t)\|_{\boldsymbol{\Sigma}_t^{-1}} + \beta \|\boldsymbol{\phi}(x_t, a_t^*) - \boldsymbol{\phi}(x_t, b_t)\|_{\boldsymbol{\Sigma}_t^{-1}} \\
&\quad + \langle 2\boldsymbol{\phi}(x_t, a_t^*) - \boldsymbol{\phi}(x_t, a_t) - \boldsymbol{\phi}(x_t, b_t), \boldsymbol{\theta}_t \rangle,
\end{aligned}
$$

where the first inequality holds due to the Cauchy-Schwarz inequality. The second inequality holds due to the high probability confidence event $\mathcal{E}$. Using our action selection rule, we have

$$
\begin{aligned}
&\langle \boldsymbol{\phi}(x_t, a_t^*) - \boldsymbol{\phi}(x_t, a_t), \boldsymbol{\theta}_t \rangle + \beta \|\boldsymbol{\phi}(x_t, a_t^*) - \boldsymbol{\phi}(x_t, a_t)\|_{\boldsymbol{\Sigma}_t^{-1}} \\
&\qquad \leq \langle \boldsymbol{\phi}(x_t, b_t) - \boldsymbol{\phi}(x_t, a_t), \boldsymbol{\theta}_t \rangle + \beta \|\boldsymbol{\phi}(x_t, a_t) - \boldsymbol{\phi}(x_t, b_t)\|_{\boldsymbol{\Sigma}_t^{-1}} \\
&\langle \boldsymbol{\phi}(x_t, a_t^*) - \boldsymbol{\phi}(x_t, b_t), \boldsymbol{\theta}_t \rangle + \beta \|\boldsymbol{\phi}(x_t, a_t^*) - \boldsymbol{\phi}(x_t, b_t)\|_{\boldsymbol{\Sigma}_t^{-1}} \\
&\qquad \leq \langle \boldsymbol{\phi}(x_t, a_t) - \boldsymbol{\phi}(x_t, b_t), \boldsymbol{\theta}_t \rangle + \beta \|\boldsymbol{\phi}(x_t, a_t) - \boldsymbol{\phi}(x_t, b_t)\|_{\boldsymbol{\Sigma}_t^{-1}}.
\end{aligned}
$$

Adding the above two inequalities, we have

$$
\begin{aligned}
&\beta \|\boldsymbol{\phi}(x_t, a_t^*) - \boldsymbol{\phi}(x_t, a_t)\|_{\boldsymbol{\Sigma}_t^{-1}} + \beta \|\boldsymbol{\phi}(x_t, a_t^*) - \boldsymbol{\phi}(x_t, b_t)\|_{\boldsymbol{\Sigma}_t^{-1}} \\
&\qquad \leq \langle \boldsymbol{\phi}(x_t, a_t) + \boldsymbol{\phi}(x_t, b_t) - 2\boldsymbol{\phi}(x_t, a_t^*), \boldsymbol{\theta}_t \rangle + 2\beta \|\boldsymbol{\phi}(x_t, a_t) - \boldsymbol{\phi}(x_t, b_t)\|_{\boldsymbol{\Sigma}_t^{-1}}.
\end{aligned}
$$

Therefore, we prove that the regret in round t can be upper bounded by

$$
r_t \leq 2\beta \|\boldsymbol{\phi}(x_t, a_t) - \boldsymbol{\phi}(x_t, b_t)\|_{\boldsymbol{\Sigma}_t^{-1}}.
$$

With a simple observation, we have $r_t \leq 4$. Therefore, the total regret can be upper bounded by

$$
\mathrm{Regret}(T) \leq \sum_{t=1}^{T} \min \Big\{ 4, 2\beta \|\boldsymbol{\phi}(x_t, a_t) - \boldsymbol{\phi}(x_t, b_t)\|_{\boldsymbol{\Sigma}_t^{-1}} \Big\}.
$$

□

D Auxiliary Lemmas

Lemma D.1 (Azuma–Hoeffding inequality, Cesa-Bianchi and Lugosi 2006)**.** Let $\{\eta_k\}_{k=1}^K$ be a martingale difference sequence with respect to a filtration $\{\mathcal{F}_t\}$ satisfying $|\eta_t| \le R$ for some constant R, η_t is $\mathcal{F}_{t+1}$-measurable, $\mathbb{E}[\eta_t|\mathcal{F}_t] = 0$. Then for any $0 < \delta < 1$, with probability at least $1 - \delta$, we have

$$\sum_{t=1}^{T} \eta_t \le R\sqrt{2T \log 1/\delta}.$$

Lemma D.2 (Lemma 9 Abbasi-Yadkori et al. 2011)**.** Let $\{\epsilon_t\}_{t=1}^T$ be a real-valued stochastic process with corresponding filtration $\{\mathcal{F}_t\}_{t=0}^T$ such that ϵ_t is $\mathcal{F}_t$-measurable and ϵ_t is conditionally R-sub-Gaussian, i.e.

$$\forall \lambda \in \mathbb{R}, \mathbb{E}[e^{\lambda \epsilon_t}|\mathcal{F}_{t-1}] \le \exp\left(\frac{\lambda^2 R^2}{2}\right).$$

Let $\{\mathbf{x}_t\}_{t=1}^T$ be an $\mathbb{R}^d$-valued stochastic process where $\mathbf{x}_t$ is $\mathcal{F}_{t-1}$-measurable and for any $t \in [T]$, we further define $\mathbf{\Sigma}_t = \lambda \mathbf{I} + \sum_{i=1}^t \mathbf{x}_i \mathbf{x}_i^\top$. Then with probability at least $1 - \delta$, for all $t \in [T]$, we have

$$\left\| \sum_{i=1}^{T} \mathbf{x}_i \eta_i \right\|_{\mathbf{\Sigma}_t^{-1}}^2 \le 2R^2 \log\left(\frac{\det(\mathbf{\Sigma}_t)^{1/2} \det(\mathbf{\Sigma}_0)^{-1/2}}{\delta}\right).$$

Lemma D.3 (Lemma 11, Abbasi-Yadkori et al. 2011)**.** For any $\lambda > 0$ and sequence $\{\mathbf{x}_t\}_{t=1}^T \subseteq \mathbb{R}^d$ for $t \in [T]$, define $\mathbf{Z}_t = \lambda \mathbf{I} + \sum_{i=1}^{t-1} \mathbf{x}_i \mathbf{x}_i^\top$. Then, provided that $\|\mathbf{x}_t\|_2 \le L$ holds for all $t \in [T]$, we have

$$\sum_{t=1}^{T} \min\left\{1, \|\mathbf{x}_t\|_{\mathbf{Z}_t^{-1}}^2\right\} \le 2d \log(1 + TL^2/(d\lambda)).$$

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: The primary contribution of this paper is addressing the challenge of adversarial feedback within the dueling bandit model, where feedback is represented as a binary preference label. Our research introduces a new perspective to machine learning. Unlike previous works on corruption-robust bandits, where corruption in each round affects the single-arm exploration and exploitation process. Flipping the preference label potentially impacts the expected reward of both actions chosen in a duel. This interaction can further affect subsequent decisions involving only one of these arms. Compared with previous adversarial dueling bandit work, we study the most direct label-flipping attack, which is aligned with many real-life preference-based learning scenarios. Our uncertainty-weighted maximum likelihood estimation method helps to solve this novel problem, in scenarios with known and unknown adversarial feedback. All the scope has been discussed clearly in our abstract and introduction.

   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 have added a Limitations setting in our main paper. We assume that the reward is linear with respect to some known feature maps. Although this setting is common in the literature, we observe that some recent works on dueling bandits can deal with nonlinear rewards (Li et al., 2024). Therefore, it's possible to extend our results to a more general setting. Another assumption concerns the lower bound of the derivative of the link function. Notably, in the logistic bandit model, which shares similarities with our setting through Bernoulli variables, some work (Abeille et al., 2021; Faury et al., 2022) can improve the dependency of κ from $1/\kappa$ to $\sqrt{\kappa}$. A similar improvement might be achieved in our setting as well.

   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 have clearly stated and proved all the lemmas and theorems used in our theoretical results. To help readers understand the proof without checking all the details, we provide a roadmap of our proof in Appendix A. We also write explanation and clarification for every formula in our paper.

   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 paper is mainly theoretical but we also do numerical experiments to justify the correctness of our results. We provide all the information to reproduce our results in Section 6.

   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: [No]

   Justification: Our experiments involve synthetic data generated from a generalized linear model, which is quite simple and easy to reproduce. That's why we do not provide access to our data and code. All the information required to reproduce the results is provided in Section 6. 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:

   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: [No]

   Justification:

   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 only have synthetic experiments and it can be reproduced on CPUs.

   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:

   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: [Yes]

    Justification: This paper studies contextual dueling bandits with adversarial feedback. Our primary objective is to propel advancements in bandit theory by introducing a more robust algorithm backed by solid theoretical guarantees. The uncertainty-weighted approach we have developed for dueling bandits holds significant potential to address the issue of adversarial feedback in preference-based data, which could be instrumental in enhancing the robustness of generative models against adversarial attacks, thereby contributing positively to the societal impact and reliability of machine learning applications.

    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:

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [NA]

    Justification:

    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:

    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:

    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:

    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.