Nearly Optimal Algorithms for Contextual Dueling Bandits from Adversarial Feedback

Qiwei Di [1] **Jiafan He** [1] **Quanquan Gu** [1]

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 $\widetilde{O}(d\sqrt{T}/\kappa + dC/\kappa)$ regret bound, where T is the number of rounds, d is the dimension of the context, κ is the lower bound of the derivative of the link function, 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. Our work is the first to achieve nearly minimax optimal regret for dueling bandits in the presence of adversarial preference feedback. Additionally, for the sigmoid link function, we develop a novel algorithm that takes into account the effect of local derivatives into maximum likelihood estimation (MLE) analysis through a refined method for estimating the link function's derivative. This method helps us to eliminate the κ dependence in the leading term with respect to T, which reduces the exponential dependence on the parameter radius B to a polynomial dependence. We conduct experiments to evaluate our proposed algorithm RCDB 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 does

[1]Department of Computer Science, University of California, Los Angeles, CA 90095, USA. Correspondence to: Quanquan Gu <qgu@cs.ucla.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

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. For known adversarial level, our result consists of two terms: a C-independent term $\widetilde{O}(d\sqrt{T}/\kappa)$, and a C-dependent term $\widetilde{O}(dC/\kappa)$, where κ is the lower bound of the derivative of the link function. Additionally, we establish a lower bound for dueling bandits with adversarial feedback, which demonstrates that our algorithm achieves optimal regret both in settings with and without adversarial feedback.

- When the adversarial level is unknown, we conduct our algorithm with an optimistic estimator of the number of adversarial feedback and prove the optimality of our result in case of a strong adversary. To the best of our knowledge, our work is the first to achieve nearly minimax optimal regret for dueling bandits in the presence of adversarial preference feedback, regardless of whether the amount of adversarial feedback is known.

- For the sigmoid link function, we develop a novel algorithm called Robust Contextual Dueling Bandit for Sigmoid link function (RCDB-S). Rather than using the uniform lower bound κ, we introduce local derivatives in maximum likelihood estimation (MLE) analysis through a refined estimation method of the link function's derivative. Our theoretical analysis establishes that RCDB-S achieves a regret bound of $\widetilde{O}\big(dB^{1.5}\sqrt{T}+dBC/\kappa\big)$, where we eliminate the κ dependence in the leading term with respect to T. This represents a significant improvement over prior works, reducing the exponential dependence on the parameter radius B to a polynomial dependence.

- We conduct experiments to validate the effectiveness of our algorithm RCDB (See Appendix E). To further 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.

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 & Szepesvári, 2020; Agrawal & 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 & Slivkins, 2012; Seldin & Slivkins, 2014; Auer & Chiang, 2016; Seldin & Lugosi, 2017; Zimmert & 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

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

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 & 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 & 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 & 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 $\|\boldsymbol{\phi}(x, a)\|_2 \leq 1$ for any $(x, a) \in \mathcal{X} \times \mathcal{A}$. We define the reward function $r_{\boldsymbol{\theta}}$ parameterized by $\boldsymbol{\theta} \in \Theta$, with $r_{\boldsymbol{\theta}}(x, a) = \langle \boldsymbol{\theta}, \boldsymbol{\phi}(x, a)\rangle$. Moreover, there exists $\boldsymbol{\theta}^*$ satisfying $r_{\boldsymbol{\theta}^*} = r^*$. For all with $\boldsymbol{\theta} \in \Theta$, $\|\boldsymbol{\theta}\|_2 \leq B$.

Similar linear 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 that there exists a constant $\kappa > 0$ such that

$$\dot{\sigma}\big(\langle \boldsymbol{\phi}(x, a) - \boldsymbol{\phi}(x, b), \boldsymbol{\theta}\rangle\big) \geq \kappa,$$

for all $x \in \mathcal{X}, a, b \in \mathcal{A}, \boldsymbol{\theta} \in \Theta$.

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. 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. 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.,

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

Remark 3.3. There are two commonly used corruption models for bandits. One is the total budget model (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 $\sum_{t=1}^{T} |r_t(a_t) - \bar{r}_t| \leq C$. Another considers the number of corrupted rounds (Zhang et al., 2021). In our setting, we consider the label-flipping attack. Thus, the magnitude of adversarial feedback is always 1 and these two types of corruption models are equivalent. Moreover, 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. We use the term "adversarial feedback" to differentiate our work from prior studies on corrupted or adversarial reward settings.

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

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

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:

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

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.,

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

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. Intuitively, even without the weighting, we can show that the solution of (4.2), i.e., $\boldsymbol{\theta}_t$, will approach $\boldsymbol{\theta}^*$, using the arguments similar to Lemma 5.1, what we will present next. This inspires us to consider the approximation of Taylor's expansion:

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

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

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

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

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

Therefore, applying the pulling-out-known-factor property of the conditional expectation, the $\mathcal{F}_t$-measurable part will cancel out when calculating the conditional variance. Then, we can approximate the variance of the estimated preference probability by

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

where the last inequality holds due to $\mathbb{E}[o_t|\mathcal{F}_t] \le 1$. Using $\kappa \le \sigma'(\boldsymbol{\phi}_t^\top\boldsymbol{\theta}^*) \le 1$, $\boldsymbol{\Lambda}_t \succeq \boldsymbol{\Sigma}_{t+1} \succeq \boldsymbol{\Sigma}_t$, where $\boldsymbol{\Sigma}_t$ is defined in Line 1 of Algorithm 1, we can see that $\mathrm{Var}\big[\sigma(\boldsymbol{\phi}_t^\top\boldsymbol{\theta}_t)|\mathcal{F}_t\big] \le \|\boldsymbol{\phi}_t\|^2_{\boldsymbol{\Sigma}_t^{-1}}$. 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.

Remark 4.1. Here, we use approximations to illustrate the motivation of our uncertainty-based weight. Rigorous proof for the algorithm's performance is presented in Section A.1, which relies solely on our specific choice of weights and does not use the approximation.

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

Algorithm 1 Robust Contextual Dueling Bandit (RCDB)

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

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

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**

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.,

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

As the arm-selection rules involving the selection of two arms has already been studied and is not the central contribution of our algorithm, we refer the readers to Appendix C of Di et al. (2023) for a detailed discussion.

Computational Complexity. We assume there is a computation oracle to solve the optimization problems of the action selection over $\mathcal{A}$. A similar oracle is implicitly assumed in almost all existing works for solving standard linear bandit problems with infinite arms (e.g., (Abbasi-Yadkori et al., 2011; He et al., 2022)). In the special case where the decision set is finite, we can iterate across all actions, resulting in $O(k^2 d^2)$ complexity for each iteration, where k is the number of actions, and d is the feature dimension.

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

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

The proof of this lemma is postponed to Section B.1.

Remark 5.2. If we set $\alpha = (\sqrt{d} + \sqrt{\lambda} B)/C$, then the confidence 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 confidence radius to be

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

$\lambda = 1/B^2$, $\alpha = \sqrt{d}/(C\sqrt{\kappa})$, then with probability at least $1 - \delta$, the regret of Algorithm 1 in the first T rounds can be upper bounded by

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

The proof of Theorem 5.3 is postponed to Section A.1. The following theorem provides a lower bound for the dueling bandit problem and demonstrates that our regret guarantee is optimal for general link functions.

Theorem 5.4. For any algorithm **Alg** and any $d, C, T \ge \max\{(4d^2)/25, dC\}, \kappa > 0$, there exists an instance of dueling bandit with link function σ, satisfying

$$\dot{\sigma}\big(\langle \boldsymbol{\phi}(x, a) - \boldsymbol{\phi}(x, b), \boldsymbol{\theta} \rangle\big) \ge \kappa, \forall x \in \mathcal{X}, a, b \in \mathcal{A}, \boldsymbol{\theta} \in \Theta,$$

such that the regret of **Alg** over T rounds is $\Omega\big((d\sqrt{T} + dC)/\kappa\big)$.

For the proof of Theorem 5.4, please refer to Appendix A.2. Theorem 5.4 indicates that the regret guarantee $\widetilde{O}((d\sqrt{T} + dC)/\kappa)$ in Theorem 5.3 for dueling bandits with adversarial feedback is optimal not only in d, T and C but also in κ. Notably, the constructed link function has a uniform first-order derivative, with value κ even around 0. As discussed in the next section, for the sigmoid function $\sigma(x) = 1/(1 + e^{-x})$, whose first-order derivative is constant around 0, the dependency on κ can be further improved.

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

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

with the pre-defined adversarial tolerance threshold $\bar{C}$ and $\lambda = 1/B^2$, $\alpha = \sqrt{d}/(\bar{C}\sqrt{\kappa})$, 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 5.6. 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, which is very common in dealing with strong adversaries (He et al., 2022; Ye et al., 2023). 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 Improved Results for Sigmoid Link Function

While Algorithm 1 can handle all link functions and achieves optimal worst-case regret guarantees, it can be improved for some specific link functions. In this section, we demonstrate this improvement for the sigmoid link functions.

We begin by calculating the coefficient κ for the sigmoid link function and analyzing the corresponding regret guarantee in Theorem 5.3. For the sigmoid function $\sigma(x) = 1/(1+e^{-x})$, its derivative is given by $\dot{\sigma}(x) = e^{-x}/(1+e^{-x})^2$. Under Assumption 3.1, we have

$$\big|\langle \boldsymbol{\phi}(x,a) - \boldsymbol{\phi}(x,b), \boldsymbol{\theta}\rangle\big| \le 2B, \forall x \in \mathcal{X}, a, b \in \mathcal{A}, \boldsymbol{\theta} \in \Theta.$$

In the worst case, we have

$$\min_{x,a,b,\boldsymbol{\theta}} \dot{\sigma}\big(\langle \boldsymbol{\phi}(x,a) - \boldsymbol{\phi}(x,b), \boldsymbol{\theta}\rangle\big) = \frac{1}{2 + e^{-2B} + e^{2B}},$$

which implies $\kappa \le 1/(2 + e^{-2B} + e^{2B})$. Therefore, the previous regret bound exhibits a polynomial dependence on $1/\kappa$, which translates to an exponential dependence on B. This exponential dependence on B is frequently observed in the literature of dueling bandits and RLHF (Zhu et al., 2023; Xiong et al., 2023; Li et al., 2024).

To improve the dependency on κ, we present a new algorithm that exploits the structure of the sigmoid link function. In Section 6.1, we outline the key ideas behind our algorithm design. Subsequently, in Section 6.2, we provide a theoretical analysis to establish the regret guarantee of our algorithm.

6.1 Key Ideas of the Algorithm

In this section, we present the key innovation of our algorithm that improves upon Algorithm 1. Let $\boldsymbol{\phi}_i = \boldsymbol{\phi}(x_i, a_i) - \boldsymbol{\phi}(x_i, b_i)$ for notational simplicity. Following the standard (weighted) MLE analysis (Li et al., 2017), we introduce an auxiliary function:

$$G_t(\boldsymbol{\theta}) = \lambda\boldsymbol{\theta} + \sum_{i=1}^{t-1} w_i\Big[\sigma\big(\boldsymbol{\phi}_i^\top\boldsymbol{\theta}\big) - \sigma\big(\boldsymbol{\phi}_i^\top\boldsymbol{\theta}^*\big)\Big]\boldsymbol{\phi}_i.$$

Using the mean-value theorem, $G_t(\boldsymbol{\theta}_t) - G_t(\boldsymbol{\theta}^*)$ can be represented as:

$$\begin{aligned} &G_t(\boldsymbol{\theta}_t) - G_t(\boldsymbol{\theta}^*) \\ &= \Big[\lambda\mathbf{I} + \sum_{i=1}^{t-1} w_i\dot{\sigma}\big(\boldsymbol{\phi}_i^\top\bar{\boldsymbol{\theta}}\big)\boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top\Big](\boldsymbol{\theta}_t - \boldsymbol{\theta}^*), \end{aligned} \tag{6.1}$$

where $\bar{\boldsymbol{\theta}} = m\boldsymbol{\theta}_t + (1-m)\boldsymbol{\theta}^*$ for some $m \in [0,1]$. In Algorithm 1, we use the unified lower bound κ to replace the local derivative $\dot{\sigma}(\boldsymbol{\phi}_i^\top\bar{\boldsymbol{\theta}})$. Thus, we define $\boldsymbol{\Sigma}_t = \lambda\mathbf{I} + \sum_{i=1}^{t-1} w_i\kappa\boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top$, and the following inequality holds: $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}_t} \le \|G_t(\boldsymbol{\theta}_t) - G_t(\boldsymbol{\theta}^*)\|_{\boldsymbol{\Sigma}_t}$.

Relaxing the local derivative $\dot{\sigma}(\boldsymbol{\phi}_i^\top\bar{\boldsymbol{\theta}})$ to the unified lower bound κ works well for smooth link functions where the derivative has little variation (e.g., the instance in Theorem 5.4). However, this approach becomes less effective for the sigmoid link function, whose derivative $\dot{\sigma}(x) = e^{-x}/(1+e^{-x})^2$ exhibits exponential decay with respect to x. In this case, the local derivative can significantly differ from κ, particularly when both actions a_t and b_t are near-optimal. In such situations, $\boldsymbol{\phi}_i^\top\bar{\boldsymbol{\theta}} \approx 0$ and consequently $\dot{\sigma}(\boldsymbol{\phi}_i^\top\bar{\boldsymbol{\theta}}) \approx 1/4$, a constant value. To address this limitation, our new algorithm introduces a refined method to estimate the local derivative $\dot{\sigma}(\boldsymbol{\phi}_i^\top\bar{\boldsymbol{\theta}})$ and directly uses it as weight in the matrix.

Following Abeille et al. (2021), we begin by expressing $\dot{\sigma}(\boldsymbol{\phi}_i^\top\bar{\boldsymbol{\theta}})$ in integral form:

$$\dot{\sigma}(\boldsymbol{\phi}_i^\top\bar{\boldsymbol{\theta}}) = \left[\int_0^1 \dot{\sigma}\Big(\boldsymbol{\phi}_i^\top\big(\boldsymbol{\theta}_t + v(\boldsymbol{\theta}^* - \boldsymbol{\theta}_t)\big)\Big)\mathrm{d}v\right].$$

By applying Lemma F.5, we have:

$$\dot{\sigma}(\boldsymbol{\phi}_i^\top\bar{\boldsymbol{\theta}}) \ge \frac{\dot{\sigma}(\boldsymbol{\phi}_i^\top\boldsymbol{\theta}^*)}{1 + |\boldsymbol{\phi}_i^\top(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*)|} \ge \frac{\dot{\sigma}(\boldsymbol{\phi}_i^\top\boldsymbol{\theta}^*)}{1+4B}.$$

Based on this inequality, we can estimate the local derivative by constructing a lower bound of $\dot{\sigma}(\boldsymbol{\phi}_i^\top \boldsymbol{\theta}^*)$. To this end, we construct a confidence set for $\boldsymbol{\theta}^*$:

$$\mathcal{E}_1 = \big\{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}_t} \le \beta_t, \forall t\big\}.$$

Supposing $\mathcal{E}_1$ holds, then for any i, the following inequality holds:

$$|\boldsymbol{\phi}_i^\top \boldsymbol{\theta}^*| \le |\boldsymbol{\phi}_i^\top \boldsymbol{\theta}_i| + \beta_i \|\boldsymbol{\phi}_i\|_{\boldsymbol{\Sigma}_i^{-1}} \triangleq \widehat{\Delta}_i.$$

By defining $v_i = \max\{\kappa, \dot{\sigma}(\widehat{\Delta}_i)\}$, we have $\dot{\sigma}(\boldsymbol{\phi}_i^\top \boldsymbol{\theta}^*) \ge v_i$ and consequently $\dot{\sigma}(\boldsymbol{\phi}_i^\top \bar{\boldsymbol{\theta}}) \ge v_i/(1+4B)$. Based on this result, we use v_i as a optimistic estimator of local derivative $\dot{\sigma}(\boldsymbol{\phi}_i^\top \bar{\boldsymbol{\theta}})$ to define a refined covariance matrix $\boldsymbol{\Lambda}_t = \lambda \mathbf{I} + \sum_{i=1}^{t-1} w_i v_i \boldsymbol{\phi}_i \boldsymbol{\phi}_i^\top$ (Line 6). For action selection, we apply the pairwise selection rule from Section 4 using $\boldsymbol{\Lambda}_t$:

$$a_t, b_t = \operatorname*{argmax}_{a,b} \Big\{ \big(\boldsymbol{\phi}(x_t, a) + \boldsymbol{\phi}(x_t, b)\big)^\top \boldsymbol{\theta}_t + \widetilde{\beta}_t \big\|\boldsymbol{\phi}(x_t, a) - \boldsymbol{\phi}(x_t, b)\big\|_{\boldsymbol{\Lambda}_t^{-1}} \Big\}.$$

In Algorithm 2, we highlight the differences from Algorithm 1 with red color for clarity.

6.2 Theoretical Results

In this section, we provide the upper bound of regret for our algorithm with the sigmoid link function.

Theorem 6.1. For the sigmoid link function $\sigma(x) = 1/(1+e^{-x})$, under Assumption 3.1, let $0 < \delta < 1$ and the total number of adversarial feedback be C. If we set the confidence radius to be

$$\beta_t = \sqrt{\lambda} B + \frac{1}{\sqrt{\kappa}} \sqrt{d \log(2(1+2t/\lambda)/\delta)} + \alpha C,$$
$$\widetilde{\beta}_t = (1+4B)\left[\sqrt{\lambda} B + \frac{2}{\sqrt{\lambda}} d \log\left(\frac{d\lambda + 2t}{d\lambda\delta}\right) + \alpha C\right],$$

and $\lambda = d/B$, $\alpha = (\sqrt{d} + \sqrt{\lambda} B)/C$, then with probability at least $1-\delta$, the regret of Algorithm 2 in the first T rounds can be upper bounded by

$$\text{Regret}(T) = \widetilde{O}\big(dB^{1.5}\sqrt{T} + dBC/\kappa + d^2 B^2 (1/\kappa^2 + B/\kappa)\big).$$

Remark 6.2. Compared to the result in Theorem 5.3, our new regret bound has a leading term of $\widetilde{O}(dB^{1.5}\sqrt{T})$ with respect to T, which eliminates the polynomial dependency on $1/\kappa$ by utilizing the local derivative rather than the uniform lower bound κ. This implies a reduction from an exponential dependence to a polynomial dependence on B. To the best of our knowledge, this is the first work to achieve such a reduction in regret for contextual dueling bandits. However, there are two scenarios where κ dependency remains: during the warm-up period when selected actions are far from optimal, the local derivative may approach κ, resulting in a κ-dependent constant factor; And in the corruption term, since the strong adversary, who can observe the actions selected by the agent, can strategically flip labels anytime, particularly when the local derivative is close to κ. This leads to a polynomial dependency on κ in the corruption term.

Algorithm 2 Robust Contextual Dueling Bandit for Sigmoid link function (RCDB-S)

1: **Require:** $\alpha > 0$, regularization parameter λ, derivative lower bound κ, confidence radius β_t, $\widetilde{\beta}_t$.
2: Set $\widehat{\Delta}_0 = 0$
3: **for** $t = 1, \dots, T$ **do**
4: Compute $\boldsymbol{\Sigma}_t = \lambda \mathbf{I} + \sum_{i=1}^{t-1} w_i \kappa \big(\boldsymbol{\phi}(x_i, a_i) - \boldsymbol{\phi}(x_i, b_i)\big)\big(\boldsymbol{\phi}(x_i, a_i) - \boldsymbol{\phi}(x_i, b_i)\big)^\top$.
5: // Construct a new weighted covariance matrix
6: Compute $\boldsymbol{\Lambda}_t = \lambda \mathbf{I} + \sum_{i=1}^{t-1} w_i v_i \big(\boldsymbol{\phi}(x_i, a_i) - \boldsymbol{\phi}(x_i, b_i)\big)\big(\boldsymbol{\phi}(x_i, a_i) - \boldsymbol{\phi}(x_i, b_i)\big)^\top$.
7: Calculate the MLE $\boldsymbol{\theta}_t$ by solving the following equation:

$$\lambda \boldsymbol{\theta} + \sum_{i=1}^{t-1} w_i \Big[\sigma\Big(\big(\boldsymbol{\phi}(x_i, a_i) - \boldsymbol{\phi}(x_i, b_i)\big)^\top \boldsymbol{\theta}\Big) - o_i\Big] \big(\boldsymbol{\phi}(x_i, a_i) - \boldsymbol{\phi}(x_i, b_i)\big) = \mathbf{0}. \tag{6.2}$$

8: Observe the context vector x_t.
9: // Use $\boldsymbol{\Lambda}_t$ to do exploration
10: Choose $a_t, b_t = \operatorname{argmax}_{a,b} \Big\{ \big(\boldsymbol{\phi}(x_t, a) + \boldsymbol{\phi}(x_t, b)\big)^\top \boldsymbol{\theta}_t + \widetilde{\beta}_t \big\|\boldsymbol{\phi}(x_t, a) - \boldsymbol{\phi}(x_t, b)\big\|_{\boldsymbol{\Lambda}_t^{-1}} \Big\}$.
11: 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 $o_t = 1 - l_t$.
12: Set weight w_t as (4.3).
13: // Calculate the weight v_t
14: $\widehat{\Delta}_t = \big|\big[\boldsymbol{\phi}(x_t, a_t) - \boldsymbol{\phi}(x_t, b_t)\big]^\top \boldsymbol{\theta}_t\big| + \beta_t \big\|\boldsymbol{\phi}(x_t, a_t) - \boldsymbol{\phi}(x_t, b_t)\big\|_{\boldsymbol{\Sigma}_t^{-1}}$
15: Set $v_t = \max\{\kappa, \dot{\sigma}(\widehat{\Delta}_t)\}$
16: **end for**

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 the sigmoid link function, we develop a novel algorithm called Robust Contextual Dueling Bandit for Sigmoid link function (RCDB-S). Through a refined estimation method of the link function's derivative, RCDB-S achieves a regret bound of $\widetilde{O}\big(dB^{1.5}\sqrt{T} + dBC/\kappa\big)$, where we eliminate the

κ dependence in the leading term with respect to T.

Limitations and Future Works. 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; Verma et al., 2024). Recently, Verma et al. (2024) studied the problem of approximating reward models using neural networks, addressing nonlinear rewards for dueling bandits. It is an interesting future direction to design robust algorithms for nonlinear reward functions, such as with neural networks. Another promising direction is to design a more computationally efficient algorithm that avoids computing the MLE in each round (Jun et al., 2017; Zhang & Sugiyama, 2024; Sawarni et al., 2024).

Impact Statement

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. Moreover, our study on dueling bandits with a logistic link function suggests that the leading term of the regret can remain unaffected by the derivative lower bound κ. This novel result provides valuable insights and may positively impact the study of machine learning theory and its applications.

Acknowledgements

We thank the anonymous reviewers for their helpful comments. JH is supported by the UCLA Dissertation Year Fellowship. QD and QG are supported in part by the National Science Foundation CPS-2312094. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pp. 2312–2320, 2011.

Abeille, M., Faury, L., and Calauzènes, C. Instance-wise minimax-optimal algorithms for logistic bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 3691–3699. PMLR, 2021.

Agarwal, A., Agarwal, S., and Patil, P. Stochastic dueling bandits with adversarial corruption. In *Algorithmic Learning Theory*, pp. 217–248. PMLR, 2021.

Agrawal, S. and Goyal, N. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pp. 39–1. JMLR Workshop and Conference Proceedings, 2012.

Ailon, N., Karnin, Z., and Joachims, T. Reducing dueling bandits to cardinal bandits. In *International Conference on Machine Learning*, pp. 856–864. PMLR, 2014.

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

Auer, P. and Chiang, C.-K. An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. In *Conference on Learning Theory*, pp. 116–120. PMLR, 2016.

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

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002b.

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

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

Bogunovic, I., Losalka, A., Krause, A., and Scarlett, J. Stochastic linear bandits robust to adversarial attacks. In *International Conference on Artificial Intelligence and Statistics*, pp. 991–999. PMLR, 2021.

Bubeck, S. and Slivkins, A. The best of both worlds: Stochastic and adversarial bandits. In *Conference on Learning Theory*, pp. 42–1. JMLR Workshop and Conference Proceedings, 2012.

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

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

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

Ding, Q., Hsieh, C.-J., and Sharpnack, J. Robust stochastic linear contextual bandits under adversarial attacks. In *International Conference on Artificial Intelligence and Statistics*, pp. 7111–7123. PMLR, 2022.

Dudík, M., Hofmann, K., Schapire, R. E., Slivkins, A., and Zoghi, M. Contextual dueling bandits. In *Conference on Learning Theory*, pp. 563–587. PMLR, 2015.

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

Faury, L., Abeille, M., Calauzènes, C., and Fercoq, O. Improved optimistic algorithms for logistic bandits. In *International Conference on Machine Learning*, pp. 3052–3060. PMLR, 2020.

Faury, L., Abeille, M., Jun, K.-S., and Calauzènes, C. Jointly efficient and optimal algorithms for logistic bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 546–580. PMLR, 2022.

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

Gajane, P., Urvoy, T., and Clérot, F. A relative exponential weighing algorithm for adversarial utility-based dueling bandits. In *International Conference on Machine Learning*, pp. 218–227. PMLR, 2015.

Gupta, A., Koren, T., and Talwar, K. Better algorithms for stochastic bandits with adversarial corruptions. In *Conference on Learning Theory*, pp. 1562–1578. PMLR, 2019.

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

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

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

Jun, K.-S., Bhargava, A., Nowak, R., and Willett, R. Scalable generalized linear bandits: Online computation and hashing. *Advances in Neural Information Processing Systems*, 30, 2017.

Jun, K.-S., Jain, L., Mason, B., and Nassif, H. Improved confidence bounds for the linear logistic model and applications to bandits. In *International Conference on Machine Learning*, pp. 5148–5157. PMLR, 2021.

Kalyanakrishnan, S., Tewari, A., Auer, P., and Stone, P. Pac subset selection in stochastic multi-armed bandits. In *ICML*, volume 12, pp. 655–662, 2012.

Komiyama, J., Honda, J., Kashima, H., and Nakagawa, H. Regret lower bound and optimal algorithm in dueling bandit problem. In *Conference on learning theory*, pp. 1141–1154. PMLR, 2015.

Komiyama, J., Honda, J., and Nakagawa, H. Copeland dueling bandit problem: Regret lower bound, optimal algorithm, and computationally efficient algorithm. In *International Conference on Machine Learning*, pp. 1235–1244. PMLR, 2016.

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

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

Lai, T. L., Robbins, H., et al. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6 (1):4–22, 1985.

Lattimore, T. and Szepesvári, C. *Bandit Algorithms*. Cambridge University Press, 2020. doi: 10.1017/9781108571401.

Lee, C.-W., Luo, H., Wei, C.-Y., Zhang, M., and Zhang, X. Achieving near instance-optimality and minimax-optimality in stochastic and adversarial linear bandits simultaneously. In *International Conference on Machine Learning*, pp. 6142–6151. PMLR, 2021.

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

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

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

Lykouris, T., Mirrokni, V., and Paes Leme, R. Stochastic bandits robust to adversarial corruptions. In *Proceedings*

of the 50th Annual ACM SIGACT Symposium on Theory of Computing, pp. 114–122, 2018.

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

Ramamohan, S. Y., Rajkumar, A., and Agarwal, S. Dueling bandits: Beyond condorcet winners to general tournament solutions. *Advances in Neural Information Processing Systems*, 29, 2016.

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

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

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

Saha, A., Koren, T., and Mansour, Y. Adversarial dueling bandits. In *International Conference on Machine Learning*, pp. 9235–9244. PMLR, 2021.

Sawarni, A., Das, N., Barman, S., and Sinha, G. Generalized linear bandits with limited adaptivity. *arXiv preprint arXiv:2404.06831*, 2024.

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

Seldin, Y. and Lugosi, G. An improved parametrization and analysis of the exp3++ algorithm for stochastic and adversarial bandits. In *Conference on Learning Theory*, pp. 1743–1759. PMLR, 2017.

Seldin, Y. and Slivkins, A. One practical algorithm for both stochastic and adversarial bandits. In *International Conference on Machine Learning*, pp. 1287–1295. PMLR, 2014.

Verma, A., Dai, Z., Lin, X., Jaillet, P., and Low, B. K. H. Neural dueling bandits. *arXiv preprint arXiv:2407.17112*, 2024.

Wei, C.-Y., Dann, C., and Zimmert, J. A model selection approach for corruption robust reinforcement learning. In *International Conference on Algorithmic Learning Theory*, pp. 1043–1096. PMLR, 2022.

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

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

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

Ye, C., Xiong, W., Gu, Q., and Zhang, T. Corruption-robust algorithms with uncertainty weighting for nonlinear contextual bandits and markov decision processes. In *International Conference on Machine Learning*, pp. 39834–39863. PMLR, 2023.

Yu, T., Quillen, D., He, Z., Julian, R., Hausman, K., Finn, C., and Levine, S. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on robot learning*, pp. 1094–1100. PMLR, 2020.

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

Zhang, X., Chen, Y., Zhu, X., and Sun, W. Robust policy gradient against strong data corruption. In *International Conference on Machine Learning*, pp. 12391–12401. PMLR, 2021.

Zhang, Y.-J. and Sugiyama, M. Online (multinomial) logistic bandit: Improved regret and constant computation cost. *Advances in Neural Information Processing Systems*, 36, 2024.

Zhao, H., Zhou, D., and Gu, Q. Linear contextual bandits with adversarial corruptions. *arXiv preprint arXiv:2110.12615*, 2021.

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

Zhu, H., Yu, J., Gupta, A., Shah, D., Hartikainen, K., Singh, A., Kumar, V., and Levine, S. The ingredients of real-world robotic reinforcement learning. *arXiv preprint arXiv:2004.12570*, 2020.

Zimmert, J. and Seldin, Y. An optimal algorithm for stochastic and adversarial bandits. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 467–475. PMLR, 2019.

Zoghi, M., Whiteson, S., Munos, R., and Rijke, M. Relative upper confidence bound for the k-armed dueling bandit problem. In *International conference on machine learning*, pp. 10–18. PMLR, 2014.

Zoghi, M., Karnin, Z. S., Whiteson, S., and De Rijke, M. Copeland dueling bandits. *Advances in neural information processing systems*, 28, 2015.

A Proof of Theorems in Section 5

A.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

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

where $\beta = \sqrt{\lambda}B + \alpha C + \sqrt{d\log((1+2T/\lambda)/\delta)/\kappa}$, $\lambda = 1/B^2$.
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 A.1. Suppose the event $\mathcal{E}$ holds. If we set $\beta = \sqrt{\lambda}B + \alpha C + \sqrt{d\log((1+2T/\lambda)/\delta)/\kappa}$, 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 A.1, we can provide the proof of Theorem 5.3.

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

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

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

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

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

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

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 F.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

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

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 F.3. Combining (A.1) and (A.2), we have

$$\text{Regret}(T) \le \frac{4\beta}{\sqrt{\kappa}}\sqrt{dT\log(1 + 2T/\lambda)} + \frac{4d\beta\log(1 + 2T/\lambda)}{\alpha\kappa}.$$

By setting $\beta = \sqrt{\lambda}B + \alpha C + \sqrt{d\log((1 + 2T/\lambda)/\delta)/\kappa}$, $\lambda = 1/B^2$, $\alpha = \sqrt{d}/(C\sqrt{\kappa})$, we complete the proof of Theorem 5.3. □

A.2 Proof of Theorem 5.4

Proof of Theorem 5.4. We consider the following link function

$$\sigma(x) = \begin{cases} 0 & \text{if } x < -\frac{1}{2} \\ \frac{1}{2} + x, & \text{if } x \in [-\frac{1}{2}, \frac{1}{2}] \\ 1, & \text{if } x > \frac{1}{2}. \end{cases} \tag{A.3}$$

Firstly, we prove a minimax lower bound. We follow the proof techniques in Li et al. (2024). The difference lies in that they consider the sigmoid link function, while we consider another. Consider the parameter set $\Theta \in \{-\Delta, \Delta\}^d$, where $\Delta > 0$ is a constant to be determined. Let the action set be $\mathcal{A} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 \le 1\}$. We assume the feature map $\phi(x, \mathbf{a}) = \mathbf{a}$. For any algorithm, denote the trajectories of actions by $\{\mathbf{a}_t, \mathbf{b}_t\}_{t=1}^T$. We require that for any action $\mathbf{x} \in \mathcal{A}$, $\boldsymbol{\theta} \in \Theta$, we have $\mathbf{x}^\top\boldsymbol{\theta} \le 1/8$. For this reason, we assume $\sqrt{d}\Delta \le 1/8$.
We fix $i \in [d]$. Define

$$\tau_i = T \wedge \min\left\{\tau : \sum_{t=1}^{\tau}\left[(\mathbf{a}_t^i)^2 + (\mathbf{b}_t^i)^2\right] \ge \frac{2T}{d}\right\}, \tag{A.4}$$

where $a \wedge b = \min\{a, b\}$ and $\mathbf{a}_t^i$ means the i-th coordinate of $\mathbf{a}_t$. Intuitively, τ_i acts as the first time in which the amount of information gathered for coordinate i until time τ (quantified as the sum of squares of the i-th coordinates of the chosen arms) exceeds the average amount of information expected, which is $2T/d$. For any $\boldsymbol{\theta} \in \Theta$, we denote $\mathbb{P}_{\boldsymbol{\theta}}$ and $\mathbb{E}_{\boldsymbol{\theta}}$ as the distribution and expectation over the trajectories. We define the following function

$$U_{\boldsymbol{\theta},i}(x) = \mathbb{E}_{\boldsymbol{\theta}}\left[\sum_{t=1}^{\tau_i}\left(\frac{1}{\sqrt{d}} - \mathbf{a}_t^i x\right)^2 + \sum_{t=1}^{\tau_i}\left(\frac{1}{\sqrt{d}} - \mathbf{b}_t^i x\right)^2\right],$$

with $x \in \{-1, +1\}$, which can be understood as the lower bounding term that pops up when lower bounding the average dueling regret. We greatly appreciate this insightful explanation above from the autonomous reviewer. Moreover, we define $\boldsymbol{\theta}'$ to be a vector whose i-th coordinate is different from $\boldsymbol{\theta}$, while the other coordinates are the same. Denote the i-th coordinate of $\boldsymbol{\theta}$ by θ_i. We consider the following term:

$$U_{\boldsymbol{\theta},i}(\text{sign}(\theta_i)) + U_{\boldsymbol{\theta}',i}(\text{sign}(\theta_i')) = \underbrace{U_{\boldsymbol{\theta},i}(\text{sign}(\theta_i)) - U_{\boldsymbol{\theta}',i}(\text{sign}(\theta_i))}_{I_1} + \underbrace{U_{\boldsymbol{\theta}',i}(\text{sign}(\theta_i)) + U_{\boldsymbol{\theta}',i}(\text{sign}(\theta_i'))}_{I_2}. \tag{A.5}$$

First, we have

$$\begin{aligned}\sum_{t=1}^{\tau_i}\Big(\frac{1}{\sqrt{d}}-\mathbf{a}_t^i\,\mathrm{sign}(\theta_i)\Big)^2+\sum_{t=1}^{\tau_i}\Big(\frac{1}{\sqrt{d}}-\mathbf{b}_t^i\,\mathrm{sign}(\theta_i)\Big)^2&\le\sum_{t=1}^{\tau_i}\Big[\frac{2}{d}+2(\mathbf{a}_t^i)^2\Big]+\sum_{t=1}^{\tau_i}\Big[\frac{2}{d}+2(\mathbf{b}_t^i)^2\Big]\\&=\frac{4\tau_i}{d}+2\sum_{t=1}^{\tau_i}\Big[(\mathbf{a}_t^i)^2+(\mathbf{b}_t^i)^2\Big]\\&\le\frac{4T}{d}+\frac{4T}{d}+4\\&=\frac{8T}{d}+4,\end{aligned}$$

where the first inequality holds due to $(a-b)^2\le 2a^2+2b^2$. The second inequality holds due to (A.4). Therefore, we have

$$\begin{aligned}I_1=\mathbb{E}_{\boldsymbol{\theta}}\Big[\sum_{t=1}^{\tau_i}\Big(\frac{1}{\sqrt{d}}-\mathbf{a}_t^i\,\mathrm{sign}(\theta_i)\Big)^2+\sum_{t=1}^{\tau_i}\Big(\frac{1}{\sqrt{d}}-\mathbf{b}_t^i\,\mathrm{sign}(\theta_i)\Big)^2\Big]\\-\mathbb{E}_{\boldsymbol{\theta}'}\Big[\sum_{t=1}^{\tau_i}\Big(\frac{1}{\sqrt{d}}-\mathbf{a}_t^i\,\mathrm{sign}(\theta_i)\Big)^2+\sum_{t=1}^{\tau_i}\Big(\frac{1}{\sqrt{d}}-\mathbf{b}_t^i\,\mathrm{sign}(\theta_i)\Big)^2\Big]\\\ge-\Big(\frac{8T}{d}+4\Big)\sqrt{\mathrm{KL}(\mathbb{P}_{\boldsymbol{\theta}}||\mathbb{P}_{\boldsymbol{\theta}'})/2},\end{aligned}\tag{A.6}$$

where the first inequality holds due to the Pinsker's inequality. Next, we compute the KL-divergence.

$$\mathrm{KL}(\mathbb{P}_{\boldsymbol{\theta}}||\mathbb{P}_{\boldsymbol{\theta}'})=\sum_{t=1}^{\tau_i}\mathrm{KL}(\mathbb{P}_{\boldsymbol{\theta}}^t||\mathbb{P}_{\boldsymbol{\theta}'}^t),\tag{A.7}$$

which holds due to the chain rule of KL-divergence. At each step t, the distribution is Bernoulli. Using the link function (A.3), we have $\mathbb{P}_{\boldsymbol{\theta}}^t=\mathrm{Ber}(p_{\boldsymbol{\theta}}^t)$, $\mathbb{P}_{\boldsymbol{\theta}'}^t=\mathrm{Ber}(p_{\boldsymbol{\theta}'}^t)$, where $p_{\boldsymbol{\theta}}^t=1/2+\langle\mathbf{a}_t-\mathbf{b}_t,\boldsymbol{\theta}\rangle$. Direct calculations show that

$$\begin{aligned}\mathrm{KL}(\mathbb{P}_{\boldsymbol{\theta}}^t||\mathbb{P}_{\boldsymbol{\theta}'}^t)&=p_{\boldsymbol{\theta}}^t\log\Big(\frac{p_{\boldsymbol{\theta}}^t}{p_{\boldsymbol{\theta}'}^t}\Big)+(1-p_{\boldsymbol{\theta}}^t)\log\Big(\frac{1-p_{\boldsymbol{\theta}}^t}{1-p_{\boldsymbol{\theta}'}^t}\Big)\\&\le p_{\boldsymbol{\theta}}^t\Big[\frac{p_{\boldsymbol{\theta}}^t}{p_{\boldsymbol{\theta}'}^t}-1\Big]+(1-p_{\boldsymbol{\theta}}^t)\Big[\frac{1-p_{\boldsymbol{\theta}}^t}{1-p_{\boldsymbol{\theta}'}^t}-1\Big]\\&=\frac{(p_{\boldsymbol{\theta}}^t-p_{\boldsymbol{\theta}'}^t)^2}{p_{\boldsymbol{\theta}'}^t(1-p_{\boldsymbol{\theta}'}^t)},\end{aligned}$$

where the first inequality holds due to $\log x\le x-1$. Using $\sqrt{d}\Delta\le 1/8$, we have

$$\mathrm{KL}(\mathbb{P}_{\boldsymbol{\theta}}^t||\mathbb{P}_{\boldsymbol{\theta}'}^t)\le\frac{16}{3}(\langle\mathbf{a}_t-\mathbf{b}_t,\boldsymbol{\theta}-\boldsymbol{\theta}'\rangle)^2.\tag{A.8}$$

Substituting (A.7) and (A.8) into (A.6), we have

$$\begin{aligned}I_1&\ge-3\Big(\frac{8T}{d}+4\Big)\sqrt{\sum_{t=1}^{\tau_i}\big(\langle\mathbf{a}_t-\mathbf{b}_t,\boldsymbol{\theta}-\boldsymbol{\theta}'\rangle\big)^2/2}\\&\ge-3\Big(\frac{8T}{d}+4\Big)\Delta\sqrt{\sum_{t=1}^{\tau_i}|\mathbf{a}_t^i-\mathbf{b}_t^i|^2/2}\\&\ge-3\Big(\frac{8T}{d}+4\Big)\Delta\sqrt{\sum_{t=1}^{\tau_i}\big[(\mathbf{a}_t^i)^2+(\mathbf{b}_t^i)^2\big]}.\end{aligned}$$

where the second inequality holds due to $\boldsymbol{\theta}$ only differs from $\boldsymbol{\theta}'$ at the i-th coordinate. The third inequality holds due to $(a-b)^2 \leq 2a^2 + 2b^2$. Using (A.4), we have

$$I_1 \geq -3\Big(\frac{8T}{d} + 4\Big)\Delta\sqrt{\frac{2T}{d} + 2} \geq -40\Delta(T/d)^{1.5}. \tag{A.9}$$

For I_2, we have

$$\begin{aligned}
I_2 &= U_{\boldsymbol{\theta}',i}(\text{sign}(\theta_i)) + U_{\boldsymbol{\theta}',i}(\text{sign}(\theta_i')) \\
&= \mathbb{E}_{\boldsymbol{\theta}'}\bigg[\sum_{t=1}^{\tau_i}\Big(\frac{1}{\sqrt{d}} - \mathbf{a}_t^i\Big)^2 + \sum_{t=1}^{\tau_i}\Big(\frac{1}{\sqrt{d}} - \mathbf{b}_t^i\Big)^2\bigg] + \mathbb{E}_{\boldsymbol{\theta}'}\bigg[\sum_{t=1}^{\tau_i}\Big(\frac{1}{\sqrt{d}} + \mathbf{a}_t^i\Big)^2 + \sum_{t=1}^{\tau_i}\Big(\frac{1}{\sqrt{d}} + \mathbf{b}_t^i\Big)^2\bigg] \\
&= 2\mathbb{E}_{\boldsymbol{\theta}'}\bigg[\frac{2\tau_i}{d} + \sum_{t=1}^{\tau_i}\Big((\mathbf{a}_t^i)^2 + (\mathbf{b}_t^i)^2\Big)\bigg] \\
&\geq \frac{4T}{d}.
\end{aligned} \tag{A.10}$$

Substituting (A.9) and (A.10) into (A.5), we have

$$U_{\boldsymbol{\theta},i}(\text{sign}(\theta_i)) + U_{\boldsymbol{\theta}',i}(\text{sign}(\theta_i')) \geq \frac{4T}{d} - 40\Delta(T/d)^{1.5}.$$

Using the randomization hammer, we have

$$\frac{1}{|\Theta|}\sum_{\boldsymbol{\theta}\in\Theta}\sum_{i=1}^{d} U_{\boldsymbol{\theta},i}(\text{sign}(\theta_i)) \geq 2T - 20\Delta T^{1.5}d^{-0.5}.$$

Therefore, there exists $\boldsymbol{\theta} \in \Theta$, such that

$$\sum_{i=1}^{d} U_{\boldsymbol{\theta},i}(\text{sign}(\theta_i)) \geq 2T - 20\Delta T^{1.5}d^{-0.5}. \tag{A.11}$$

Using this $\boldsymbol{\theta}$, we can decompose the regret into

$$\text{Regret}(T) = \Delta\mathbb{E}_\theta\bigg[\sum_{t=1}^{T}\sum_{i=1}^{d}\Big(\frac{1}{\sqrt{d}} - \mathbf{a}_t^i\,\text{sign}(\theta_i)\Big)\bigg] + \Delta\mathbb{E}_\theta\bigg[\sum_{t=1}^{T}\sum_{i=1}^{d}\Big(\frac{1}{\sqrt{d}} - \mathbf{b}_t^i\,\text{sign}(\theta_i)\Big)\bigg].$$

Moreover, we have

$$\begin{aligned}
\sum_{i=1}^{d}\Big(\frac{1}{\sqrt{d}} - \mathbf{a}_t^i\,\text{sign}(\theta_i)\Big) &= \sum_{i=1}^{d}\Big(\frac{1}{2\sqrt{d}} - \mathbf{a}_t^i\,\text{sign}(\theta_i)\Big) + \frac{\sqrt{d}}{2} \\
&\geq \sum_{i=1}^{d}\Big(\frac{1}{2\sqrt{d}} - \mathbf{a}_t^i\,\text{sign}(\theta_i)\Big) + \frac{\sqrt{d}}{2}\sum_{i=1}^{d}(\mathbf{a}_t^i)^2 \\
&= \frac{\sqrt{d}}{2}\sum_{i=1}^{d}\bigg(\frac{1}{\sqrt{d}} - \mathbf{a}_t^i\,\text{sign}(\theta_i)\bigg)^2,
\end{aligned}$$

where the first inequality holds due to $\|\mathbf{a}_t\|_2 \leq 1$. Similarly, we have

$$\sum_{i=1}^{d}\Big(\frac{1}{\sqrt{d}} - \mathbf{b}_t^i\,\text{sign}(\theta_i)\Big) \geq \frac{\sqrt{d}}{2}\sum_{i=1}^{d}\bigg(\frac{1}{\sqrt{d}} - \mathbf{b}_t^i\,\text{sign}(\theta_i)\bigg)^2.$$

As a result, we have

$$\begin{aligned}
\text{Regret}(T) &\geq \frac{\sqrt{d}\Delta}{2}\sum_{i=1}^{d}\mathbb{E}_{\boldsymbol{\theta}}\left[\sum_{t=1}^{T}\left(\frac{1}{\sqrt{d}}-\mathbf{a}_t^i\,\text{sign}(\theta_i)\right)^2\right]+\frac{\sqrt{d}\Delta}{2}\sum_{i=1}^{d}\mathbb{E}_{\boldsymbol{\theta}}\left[\sum_{t=1}^{T}\left(\frac{1}{\sqrt{d}}-\mathbf{b}_t^i\,\text{sign}(\theta_i)\right)^2\right]\\
&\geq \frac{\sqrt{d}\Delta}{2}\sum_{i=1}^{d}\mathbb{E}_{\boldsymbol{\theta}}\left[\sum_{t=1}^{\tau_i}\left(\frac{1}{\sqrt{d}}-\mathbf{a}_t^i\,\text{sign}(\theta_i)\right)^2\right]+\frac{\sqrt{d}\Delta}{2}\sum_{i=1}^{d}\mathbb{E}_{\boldsymbol{\theta}}\left[\sum_{t=1}^{\tau_i}\left(\frac{1}{\sqrt{d}}-\mathbf{b}_t^i\,\text{sign}(\theta_i)\right)^2\right]\\
&= \frac{\sqrt{d}\Delta}{2}\sum_{i=1}^{d}U_{\boldsymbol{\theta},i}(\text{sign}(\theta_i))\\
&\geq \frac{\sqrt{d}\Delta}{2}\left[2T-20\Delta T^{1.5}d^{-0.5}\right],
\end{aligned}$$

where the last inequality holds due to (A.11). Set $\Delta=\frac{1}{20}\sqrt{\frac{d}{T}}$. Then we have proved a lower bound $\text{Regret}(T)\geq d\sqrt{T}/40$. Moreover, when $T\geq(4d^2)/25$, we have $\sqrt{d}\Delta\leq 1/8$. In conclusion, for the link function σ defined in (A.3) and any algorithm, there exists a dueling bandit instance such that the regret is $\Omega(d\sqrt{T})$.

Secondly, we consider the lower bound of the corruption term. Our proof adapts the argument in Bogunovic et al. (2021) to dueling bandits. For any dimension d, we construct d instances, each with $\boldsymbol{\theta}_i=\mathbf{e}_i/4$, 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 1/4. 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.

Combining the results above, for the link function σ (defined in (A.3)), the lower bound for dueling bandits is $\Omega(d\sqrt{T}+dC)$. Finally, for any κ, we define $\sigma_\kappa(x)=\sigma(\kappa x)$. Then $\min\dot{\sigma}_\kappa(\cdot)=\kappa$. For any algorithm **Alg**, consider the hard instance $I=(\boldsymbol{\theta}^*,\mathcal{A},\sigma)$ with the link function σ, and $\text{Regret}_{\textbf{Alg}}(T)=\Omega(d\sqrt{T}+dC)$. Define another instance $I'=(\boldsymbol{\theta}^*,\mathcal{A}/\kappa,\sigma_\kappa)$. The interaction with the environment with I and I' are exactly the same. However, the regret with I' is $\text{Regret}_{\textbf{Alg}}(T)/\kappa$. This indicates a lower bound of $\Omega(d\sqrt{T}+dC)/\kappa$, which completes the proof of Theorem 5.4. □

A.3 Proof of Theorem 5.5

Proof of Theorem 5.5. 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 $\text{Regret}(T)\leq 2T$.

- In the scenario where $C\leq\bar{C}$, 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

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

□

B Proof of Lemmas 5.1 and A.1

B.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

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

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

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

Then we have

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

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,

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

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

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

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

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

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

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

where the second inequality holds due to the triangle inequality and $G_t(\boldsymbol{\theta}^*) = \lambda\boldsymbol{\theta}^*$. 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 (B.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,

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

Summing up the two equalities, 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,

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

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

$$\begin{aligned}\det(\boldsymbol{\Sigma}_t) &\leq \bigg(\frac{\mathrm{Tr}(\boldsymbol{\Sigma}_t)}{d}\bigg)^d\\ &= \bigg(\frac{d\lambda + \sum_{i=1}^{t-1} w_i\kappa\|(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i))\|_2^2}{d}\bigg)^d\\ &\leq \bigg(\frac{d\lambda + 2T}{d}\bigg)^d,\end{aligned}$$

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 $\sqrt{w_i\kappa}\|\boldsymbol{\phi}(x_i,a_i)-\boldsymbol{\phi}(x_i,b_i)\|_2 \leq 2$. Easy to see that $\det(\boldsymbol{\Sigma}_0)=\lambda^d$. The term I_1 can be bounded by

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

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

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

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

B.2 Proof of Lemma A.1

Proof of Lemma A.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

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

Adding the above two inequalities, we have

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

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

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

□

C Proof of Theorem 6.1

To start with, we define some high-probability events. Recall that

$$\begin{aligned}\boldsymbol{\Sigma}_t &= \lambda\mathbf{I} + \textstyle\sum_{i=1}^{t-1} w_i \kappa\big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big)\big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big)^\top,\\ \boldsymbol{\Lambda}_t &= \lambda\mathbf{I} + \textstyle\sum_{i=1}^{t-1} w_i v_i \big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big)\big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big)^\top.\end{aligned}$$

Then we define the following events:

$$\begin{aligned}\mathcal{E}_1 &= \{\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_t\|_{\boldsymbol{\Sigma}_t} \le \beta_t, \forall t \in [T]\},\\ \mathcal{E}_2 &= \{\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_t\|_{\boldsymbol{\Lambda}_t} \le \widetilde{\beta}_t, \forall t \in [T]\}.\end{aligned}$$

The following lemma indicates that with some well-chosen confidence radius, both $\mathcal{E}_1$ and $\mathcal{E}_2$ will happen with high probability.

Lemma C.1. When selecting

$$\begin{aligned}\beta_t &= \sqrt{\lambda}B + \frac{1}{\sqrt{\kappa}}\sqrt{d\log(2(1+2t/\lambda)/\delta)} + \alpha C,\\ \widetilde{\beta}_t &= (1+4B)\bigg[\sqrt{\lambda}B + \frac{\sqrt{\lambda}}{2} + \frac{2}{\sqrt{\lambda}} d\log\bigg(\frac{d\lambda+2t}{d\lambda}\bigg) + \frac{2}{\sqrt{\lambda}} d\log(1/\delta) + \alpha C\bigg],\end{aligned}$$

then we have

$$\mathbb{P}[\mathcal{E}_1 \cap \mathcal{E}_2] \ge 1-\delta.$$

Proof of Theorem 6.1. Consider the case when $\mathcal{E}_1 \cap \mathcal{E}_2$ holds. Recall that $\widehat{\Delta}_t$ is defined by

$$\widehat{\Delta}_t = \big|\big(\boldsymbol{\phi}(x_t,a_t) - \boldsymbol{\phi}(x_t,b_t)\big)^\top \boldsymbol{\theta}_t\big| + \beta_t \|\boldsymbol{\phi}(x_t,a_t) - \boldsymbol{\phi}(x_t,b_t)\|_{\boldsymbol{\Sigma}_t^{-1}}$$

then we have

$$\begin{aligned}|\widehat{\Delta}_t| &\ge \big|\big(\boldsymbol{\phi}(x_t,a_t) - \boldsymbol{\phi}(x_t,b_t)\big)^\top \boldsymbol{\theta}^*\big| + \beta \|\boldsymbol{\phi}(x_t,a_t) - \boldsymbol{\phi}(x_t,b_t)\|_{\boldsymbol{\Sigma}_t^{-1}} - \big|\big(\boldsymbol{\phi}(x_t,a_t) - \boldsymbol{\phi}(x_t,b_t)\big)^\top (\boldsymbol{\theta}^* - \boldsymbol{\theta}_t)\big|\\ &\ge \big|\big(\boldsymbol{\phi}(x_t,a_t) - \boldsymbol{\phi}(x_t,b_t)\big)^\top \boldsymbol{\theta}^*\big|,\end{aligned}$$

where the first inequality holds due to the triangle inequality. The second inequality holds due to $\mathcal{E}_1$ and the Cauchy-Schwarz inequality. Let $v_t = \max\{\kappa, \dot{\sigma}(\widehat{\Delta}_t)\}$. Using the fact that $\dot{\sigma}(\cdot)$ is an even function decreasing when $x > 0$, we have $v_t \le \dot{\sigma}\big(\big(\boldsymbol{\phi}(x_t,a_t) - \boldsymbol{\phi}(x_t,b_t)\big)^\top \boldsymbol{\theta}^*\big)$. 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

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

where the first inequality holds due to the Cauchy-Schwarz inequality. The second inequality holds due to the high probability confidence event $\mathcal{E}_2$. Using our action selection rule, we have

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

Adding the above two inequalities, we have

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

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

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

As a result, the total regret can be upper bounded by

$$\text{Regret}(T)\le 2\widetilde{\beta}_T\sum_{t=1}^{T}\|\boldsymbol{\phi}(x_t,a_t)-\boldsymbol{\phi}(x_t,b_t)\|_{\boldsymbol{\Lambda}_t^{-1}},$$

where we use that β_t is increasing in t. Since our weight w_t has two possible values, we can decompose the regret into two terms:

$$\begin{aligned}\text{Regret}(T)\le&\underbrace{2\widetilde{\beta}_T\sum_{t:w_t=1}\|\boldsymbol{\phi}(x_t,a_t)-\boldsymbol{\phi}(x_t,b_t)\|_{\boldsymbol{\Lambda}_t^{-1}}}_{I_1}\\&+\underbrace{2\widetilde{\beta}_T\sum_{t:w_t<1}\|\boldsymbol{\phi}(x_t,a_t)-\boldsymbol{\phi}(x_t,b_t)\|_{\boldsymbol{\Lambda}_t^{-1}}}_{I_2}.\end{aligned}$$

For the term I_1, we consider a partial summation in rounds when $w_t=1$. Let $\widetilde{\boldsymbol{\Lambda}}_t=\lambda\mathbf{I}+\sum_{i\le t-1,w_i=1}v_i\big(\boldsymbol{\phi}(x_i,a_i)-\boldsymbol{\phi}(x_i,b_i)\big)\big(\boldsymbol{\phi}(x_i,a_i)-\boldsymbol{\phi}(x_i,b_i)\big)^\top$. Then we have

$$\begin{aligned}I_1&=2\widetilde{\beta}_T\sum_{t:w_t=1}\frac{1}{\sqrt{v_t}}\big\|\sqrt{v_t}\big(\boldsymbol{\phi}(x_t,a_t)-\boldsymbol{\phi}(x_t,b_t)\big)\big\|_{\boldsymbol{\Lambda}_t^{-1}}\\&\le 2\widetilde{\beta}_T\sqrt{\sum_{t:w_t=1}\frac{1}{v_t}}\cdot\sqrt{\sum_{t:w_t=1}\big\|\sqrt{v_t}\big(\boldsymbol{\phi}(x_t,a_t)-\boldsymbol{\phi}(x_t,b_t)\big)\big\|^2_{\boldsymbol{\Lambda}_t^{-1}}}\\&\le 2\widetilde{\beta}_T\sqrt{\sum_{t:w_t=1}\frac{1}{v_t}}\cdot\sqrt{\sum_{t:w_t=1}\big\|\sqrt{v_t}\big(\boldsymbol{\phi}(x_t,a_t)-\boldsymbol{\phi}(x_t,b_t)\big)\big\|^2_{\widetilde{\boldsymbol{\Lambda}}_t^{-1}}}\\&\le 4\widetilde{\beta}_T\sqrt{d\log(1+2T/\lambda)}\cdot\sqrt{\sum_{t:w_t=1}\frac{1}{v_t}},\end{aligned}\tag{C.1}$$

where the first inequality holds due to the Cauchy-Schwarz inequality. The second inequality holds due to $\boldsymbol{\Lambda}_t\succeq\widetilde{\boldsymbol{\Lambda}}_t$. The last inequality holds due to Lemma F.3 and $\big\|\sqrt{v_t}\big(\boldsymbol{\phi}(x_t,a_t)-\boldsymbol{\phi}(x_t,b_t)\big)\big\|_2\le 2$. Next, we will bound $\sum_{t:w_t=1}1/v_t$. Let $\mathcal{L}_1=\{t:w_t=1,\kappa=\max\{\kappa,\dot{\sigma}(\widehat{\Delta}_t)\}\}$, $\mathcal{L}_2=\{t:w_t=1,\dot{\sigma}(\widehat{\Delta}_t)=\max\{\kappa,\dot{\sigma}(\widehat{\Delta}_t)\}\}$. Then, we have $\{t:w_t=1\}=\mathcal{L}_1\cup\mathcal{L}_2$. Therefore, we have the following decomposition

$$\sum_{w_t=1}\frac{1}{v_t}=\underbrace{\sum_{t\in\mathcal{L}_1}\frac{1}{v_t}}_{J_1}+\underbrace{\sum_{t\in\mathcal{L}_2}\frac{1}{v_t}}_{J_2}.$$

For J_1, we have

$$\begin{aligned}J_1&=\sum_{t\in\mathcal{L}_1}\frac{1}{v_t}\\&=\frac{1}{\kappa}|\mathcal{L}_1|,\end{aligned}\tag{C.2}$$

This equality holds because for $t \in \mathcal{L}_1$, $v_t = \kappa$. Using the high-probability event $\mathcal{E}_1$, the following inequality holds:

$$\begin{aligned}\widehat{\Delta}_t &= \big|\big[\boldsymbol{\phi}(x_t,a_t)-\boldsymbol{\phi}(x_t,b_t)\big]^\top\boldsymbol{\theta}_t\big| + \beta_t\big\|\boldsymbol{\phi}(x_t,a_t)-\boldsymbol{\phi}(x_t,b_t)\big\|_{\boldsymbol{\Sigma}_t^{-1}}\\ &\le \big|\big[\boldsymbol{\phi}(x_t,a_t)-\boldsymbol{\phi}(x_t,b_t)\big]^\top\boldsymbol{\theta}^*\big| + 2\beta_t\big\|\boldsymbol{\phi}(x_t,a_t)-\boldsymbol{\phi}(x_t,b_t)\big\|_{\boldsymbol{\Sigma}_t^{-1}},\end{aligned}$$

On the other hand, for $t \in \mathcal{L}_1$, we have

$$\begin{aligned}\kappa &\ge \dot\sigma(\widehat{\Delta}_t)\\ &\ge \dot\sigma\bigg(\Big|\big[\boldsymbol{\phi}(x_t,a_t)-\boldsymbol{\phi}(x_t,b_t)\big]^\top\boldsymbol{\theta}^*\Big| + 2\beta_t\big\|\boldsymbol{\phi}(x_t,a_t)-\boldsymbol{\phi}(x_t,b_t)\big\|_{\boldsymbol{\Sigma}_t^{-1}}\bigg),\end{aligned}$$

where the first inequality holds due to the definition of $\mathcal{L}_1$. The second inequality holds due to the function $\dot\sigma(\cdot)$ is decreasing when $x > 0$. Using $\dot\sigma(|x|) = e^{-|x|}/(1+e^{-|x|})^2 \ge e^{-|x|}/4$, we have $|x| \ge \log(1/4\dot\sigma(|x|))$. Therefore, the following inequality holds

$$\begin{aligned}\Big|\big[\boldsymbol{\phi}(x_t,a_t)-\boldsymbol{\phi}(x_t,b_t)\big]^\top\boldsymbol{\theta}^*\Big| + 2\beta_t\big\|\boldsymbol{\phi}(x_t,a_t)-\boldsymbol{\phi}(x_t,b_t)\big\|_{\boldsymbol{\Sigma}_t^{-1}} &\ge \log\Big(\frac{1}{4\dot\sigma(\widehat{\Delta}_t)}\Big)\\ &\ge \log(1/4\kappa).\end{aligned}$$

Let $\widetilde{\boldsymbol{\Sigma}}_t = \lambda\mathbf{I} + \sum_{i\le t-1, w_i=1}\kappa\big(\boldsymbol{\phi}(x_i,a_i)-\boldsymbol{\phi}(x_i,b_i)\big)\big(\boldsymbol{\phi}(x_i,a_i)-\boldsymbol{\phi}(x_i,b_i)\big)^\top$. Summing over $t \in \mathcal{L}_1$, we obtain

$$\begin{aligned}\log(1/4\kappa)|\mathcal{L}_1| &\le \sum_{t\in\mathcal{L}_1}\Big|\big[\boldsymbol{\phi}(x_t,a_t)-\boldsymbol{\phi}(x_t,b_t)\big]^\top\boldsymbol{\theta}^*\Big| + 2\beta_T\sum_{t\in\mathcal{L}_1}\big\|\boldsymbol{\phi}(x_t,a_t)-\boldsymbol{\phi}(x_t,b_t)\big\|_{\boldsymbol{\Sigma}_t^{-1}}\\ &\le \sum_{t:w_t=1}\Big|\big[\boldsymbol{\phi}(x_t,a_t)-\boldsymbol{\phi}(x_t,b_t)\big]^\top\boldsymbol{\theta}^*\Big| + 2\beta_T\sum_{t:w_t=1}\big\|\boldsymbol{\phi}(x_t,a_t)-\boldsymbol{\phi}(x_t,b_t)\big\|_{\boldsymbol{\Sigma}_t^{-1}}\\ &\le \sum_{t:w_t=1}\Big|\big[\boldsymbol{\phi}(x_t,a_t)-\boldsymbol{\phi}(x_t,b_t)\big]^\top\boldsymbol{\theta}^*\Big| + 2\beta_T\sqrt{T\sum_{t:w_t=1}\big\|\boldsymbol{\phi}(x_t,a_t)-\boldsymbol{\phi}(x_t,b_t)\big\|^2_{\widetilde{\boldsymbol{\Sigma}}_t^{-1}}}\\ &= \sum_{t:w_t=1}\Big|\big[\boldsymbol{\phi}(x_t,a_t)-\boldsymbol{\phi}(x_t,b_t)\big]^\top\boldsymbol{\theta}^*\Big| + \frac{2\beta_T}{\sqrt{\kappa}}\sqrt{T\sum_{t:w_t=1}\big\|\sqrt{\kappa}\big(\boldsymbol{\phi}(x_t,a_t)-\boldsymbol{\phi}(x_t,b_t)\big)\big\|^2_{\widetilde{\boldsymbol{\Sigma}}_t^{-1}}}\\ &\le \sum_{t=1}^{T}\Big|\big[\boldsymbol{\phi}(x_t,a_t)-\boldsymbol{\phi}(x_t,b_t)\big]^\top\boldsymbol{\theta}^*\Big| + \frac{2\beta_T}{\sqrt{\kappa}}\sqrt{dT\log(1+2T/\lambda)},\end{aligned}$$

where the first inequality holds due to β_t is increasing in t. The second inequality holds because partial summation is less than total summation. The third inequality holds due to the Cauchy-Schwarz inequality and $\boldsymbol{\Sigma}_t \succeq \widetilde{\boldsymbol{\Sigma}}_t$. The last inequality holds due to Lemma F.3. Therefore, we have

$$|\mathcal{L}_1| \le \frac{1}{\log(1/4\kappa)}\bigg[\sum_{t=1}^{T}\Big|\big[\boldsymbol{\phi}(x_t,a_t)-\boldsymbol{\phi}(x_t,b_t)\big]^\top\boldsymbol{\theta}^*\Big| + \frac{2\beta_T}{\sqrt{\kappa}}\sqrt{dT\log(1+2T/\lambda)}\bigg]. \tag{C.3}$$

Substituting (C.3) into (C.2), we have

$$J_1 \le \frac{1}{\kappa\log(1/4\kappa)}\bigg[\sum_{t=1}^{T}\Big|\big[\boldsymbol{\phi}(x_t,a_t)-\boldsymbol{\phi}(x_t,b_t)\big]^\top\boldsymbol{\theta}^*\Big| + \frac{2\beta_T}{\sqrt{\kappa}}\sqrt{dT\log(1+2T/\lambda)}\bigg]. \tag{C.4}$$

For J_2, we have

$$\begin{aligned}J_2 &= \sum_{t\in\mathcal{L}_2}\frac{1}{v_t}\\ &= \sum_{t\in\mathcal{L}_2}\frac{1}{\dot\sigma(|\widehat{\Delta}_t|)}\\ &\le 4\sum_{t\in\mathcal{L}_2}e^{|\widehat{\Delta}_t|},\end{aligned} \tag{C.5}$$

where the last inequality holds due to $\dot{\sigma}(|x|) \geq e^{-|x|}/4$. For $t \in \mathcal{L}_2$, we have

$$\kappa \leq \dot{\sigma}(\widehat{\Delta}_t) \leq e^{-|\widehat{\Delta}_t|},$$

where the last inequality holds due to $\dot{\sigma}(|x|) \leq e^{-|x|}$. Then we have

$$|\widehat{\Delta}_t| \leq \log(1/\kappa).$$

Since the function $f(x) = e^x$ is convex, for any $A > 0$, we have for all $x \in [0, A]$

$$e^x \leq 1 + \frac{e^A - 1}{A} x.$$

Setting $x = \widehat{\Delta}_t$ and $A = \log(1/\kappa)$, we have

$$e^{|\widehat{\Delta}_t|} \leq 1 + \frac{1/\kappa - 1}{\log(1/\kappa)} |\widehat{\Delta}_t|. \tag{C.6}$$

Substituting (C.6) into (C.5), we have

$$J_2 \leq 4 \sum_{t \in \mathcal{L}_2} \left[1 + \frac{1/\kappa - 1}{\log(1/\kappa)} |\widehat{\Delta}_t| \right].$$

Moreover, we know that

$$\begin{aligned}
\widehat{\Delta}_t &= \left| \big(\boldsymbol{\phi}(x_t, a_t) - \boldsymbol{\phi}(x_t, b_t)\big)^\top \boldsymbol{\theta}_t \right| + \beta_t \|\boldsymbol{\phi}(x_t, a_t) - \boldsymbol{\phi}(x_t, b_t)\|_{\boldsymbol{\Sigma}_t^{-1}} \\
&\leq \left| \big(\boldsymbol{\phi}(x_t, a_t) - \boldsymbol{\phi}(x_t, b_t)\big)^\top \boldsymbol{\theta}^* \right| + 2\beta_t \|\boldsymbol{\phi}(x_t, a_t) - \boldsymbol{\phi}(x_t, b_t)\|_{\boldsymbol{\Sigma}_t^{-1}}.
\end{aligned} \tag{C.7}$$

Therefore, we have

$$\begin{aligned}
J_2 &\leq 4 \left[T + \frac{2\beta_T}{\kappa \log(1/\kappa)} \sum_{t=1}^T \|\boldsymbol{\phi}(x_t, a_t) - \boldsymbol{\phi}(x_t, b_t)\|_{\boldsymbol{\Sigma}_t^{-1}} + \frac{1}{\kappa \log(1/\kappa)} \sum_{t=1}^T \left| \big(\boldsymbol{\phi}(x_t, a_t) - \boldsymbol{\phi}(x_t, b_t)\big)^\top \boldsymbol{\theta}^* \right| \right] \\
&\leq 4 \left[T + \frac{2\beta_T}{\kappa^{1.5} \log(1/\kappa)} \sum_{t=1}^T \left\| \sqrt{\kappa} \big(\boldsymbol{\phi}(x_t, a_t) - \boldsymbol{\phi}(x_t, b_t)\big) \right\|_{\boldsymbol{\Sigma}_t^{-1}} + \frac{1}{\kappa \log(1/\kappa)} \sum_{t=1}^T \left| \big(\boldsymbol{\phi}(x_t, a_t) - \boldsymbol{\phi}(x_t, b_t)\big)^\top \boldsymbol{\theta}^* \right| \right] \\
&\leq 4T + \left[\frac{8\beta_T}{\kappa^{1.5} \log(1/\kappa)} \sqrt{dT \log(1 + 2T/\lambda)} + \frac{4}{\kappa \log(1/\kappa)} \sum_{t=1}^T \left| \big(\boldsymbol{\phi}(x_t, a_t) - \boldsymbol{\phi}(x_t, b_t)\big)^\top \boldsymbol{\theta}^* \right| \right],
\end{aligned} \tag{C.8}$$

where the first inequality holds due to (C.7). The last inequality holds due to Lemma F.3 and the Cauchy-Schwarz inequality. Combining (C.4) and (C.8), we have

$$\begin{aligned}
\sum_{t: w_t = 1} \frac{1}{v_t} &\leq \frac{1}{\kappa \log(1/4\kappa)} \left[\sum_{t=1}^T \left| \big[\boldsymbol{\phi}(x_t, a_t) - \boldsymbol{\phi}(x_t, b_t)\big]^\top \boldsymbol{\theta}^* \right| + \frac{2\beta_T}{\sqrt{\kappa}} \sqrt{dT \log(1 + 2T/\lambda)} \right] \\
&\quad + 4T + \left[\frac{8\beta_T}{\kappa^{1.5} \log(1/\kappa)} \sqrt{dT \log(1 + 2T/\lambda)} + \frac{1}{\kappa \log(1/\kappa)} \sum_{t=1}^T \left| \big(\boldsymbol{\phi}(x_t, a_t) - \boldsymbol{\phi}(x_t, b_t)\big)^\top \boldsymbol{\theta}^* \right| \right] \\
&\leq 4T + \frac{10\beta_T}{\kappa^{1.5} \log(1/\kappa)} \sqrt{dT \log(1 + 2T/\lambda)} + \frac{2}{\kappa \log(1/\kappa)} \sum_{t=1}^T \left| \big(\boldsymbol{\phi}(x_t, a_t) - \boldsymbol{\phi}(x_t, b_t)\big)^\top \boldsymbol{\theta}^* \right|
\end{aligned} \tag{C.9}$$

For the term I_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

$$\begin{aligned}
I_2 &= 2\widetilde{\beta}_T \sum_{w_t < 1} \Big(\|\boldsymbol{\phi}(x_t, a_t) - \boldsymbol{\phi}(x_t, b_t)\|_{\boldsymbol{\Lambda}_t^{-1}} w_t \|\boldsymbol{\phi}(x_t, a_t) - \boldsymbol{\phi}(x_t, b_t)\|_{\boldsymbol{\Sigma}_t^{-1}} / \alpha \Big) \\
&\leq \frac{2\widetilde{\beta}_T}{\alpha\kappa} \sum_{t=1}^T \left\| \sqrt{w_t \kappa} (\boldsymbol{\phi}(x_t, a_t) - \boldsymbol{\phi}(x_t, b_t)) \right\|_{\boldsymbol{\Sigma}_t^{-1}}^2 \\
&\leq \frac{4d\widetilde{\beta}_T \log(1 + 2T/\lambda)}{\alpha\kappa},
\end{aligned} \tag{C.10}$$

where the first equality holds due to the choice of w_t. The first inequality holds because $\boldsymbol{\Lambda}_t \succeq \boldsymbol{\Sigma}_t$. The last inequality holds due to Lemma F.3.

Moreover, we notice that the following inequality holds:

$$\begin{aligned}\sum_{t=1}^{T} \big|\big(\boldsymbol{\phi}(x_t, a_t) - \boldsymbol{\phi}(x_t, b_t)\big)^\top \boldsymbol{\theta}^*\big| &= \sum_{t=1}^{T} \Big|\big(\boldsymbol{\phi}(x_t, a_t^*) - \boldsymbol{\phi}(x_t, a_t)\big)^\top \boldsymbol{\theta}^* - \big(\boldsymbol{\phi}(x_t, a_t^*) - \boldsymbol{\phi}(x_t, b_t)\big)^\top \boldsymbol{\theta}^*\Big| \\ &\le \sum_{t=1}^{T} \big(\boldsymbol{\phi}(x_t, a_t^*) - \boldsymbol{\phi}(x_t, a_t)\big)^\top \boldsymbol{\theta}^* + \big(\boldsymbol{\phi}(x_t, a_t^*) - \boldsymbol{\phi}(x_t, b_t)\big)^\top \boldsymbol{\theta}^* \\ &= \text{Regret}(T). \end{aligned} \tag{C.11}$$

Substituting (C.9) and (C.11) into (C.1), we have

$$I_1 \le 4\widetilde{\beta}_T \sqrt{d \log(1 + 2T/\lambda)} \cdot \sqrt{4T + \frac{10\beta_T}{\kappa^{1.5} \log(1/\kappa)} \sqrt{dT \log(1 + 2T/\lambda)} + \frac{2}{\kappa \log(1/\kappa)} \text{Regret}(T)} \tag{C.12}$$

Combining (C.10) and (C.12), we have

$$\begin{aligned}\text{Regret}(T) &\le 4\widetilde{\beta}_T \sqrt{d \log(1 + 2T/\lambda)} \sqrt{4T + \bigg[\frac{10\beta_T}{\kappa^{1.5} \log(1/\kappa)} \sqrt{dT \log(1 + 2T/\lambda)} + \frac{2}{\kappa \log(1/\kappa)} \text{Regret}(T)\bigg]} \\ &\quad + \frac{4d\widetilde{\beta}_T \log(1 + 2T/\lambda)}{\alpha\kappa} \\ &\le 4\widetilde{\beta}_T \sqrt{d \log(1 + 2T/\lambda)} \bigg[3\sqrt{T} + \frac{10\beta_T}{\kappa^{1.5} \log(1/\kappa)} \sqrt{d \log(1 + 2T/\lambda)} + \sqrt{\frac{2}{\kappa \log(1/\kappa)}} \sqrt{\text{Regret}(T)}\bigg] \\ &\quad + \frac{4d\widetilde{\beta}_T \log(1 + 2T/\lambda)}{\alpha\kappa},\end{aligned}$$

where the last inequality holds due to Young's inequality and $\sqrt{a+b+c} \le \sqrt{a} + \sqrt{b} + \sqrt{c}$. Using $x \le a\sqrt{x} + b \Rightarrow x \le a^2 + 2b$, we have

$$\begin{aligned}\text{Regret}(T) &\le 24\widetilde{\beta}_T \sqrt{dT \log(1 + 2T/\lambda)} + \frac{80\big(\widetilde{\beta}_T^2/\kappa + \widetilde{\beta}_T \beta_T/\kappa^{1.5}\big) d \log(1 + 2T/\lambda)}{\log(1/\kappa)} \\ &\quad + \frac{8d\widetilde{\beta}_T \log(1 + 2T/\lambda)}{\alpha\kappa}.\end{aligned} \tag{C.13}$$

Recall that

$$\begin{aligned}\beta_t &= \sqrt{\lambda} B + \frac{1}{\sqrt{\kappa}} \sqrt{d \log(2(1 + 2t/\lambda)/\delta)} + \alpha C, \\ \widetilde{\beta}_t &= (1 + 4B)\bigg[\sqrt{\lambda} B + \frac{\sqrt{\lambda}}{2} + \frac{2}{\sqrt{\lambda}} d \log\bigg(\frac{d\lambda + 2t}{d\lambda}\bigg) + \frac{2}{\sqrt{\lambda}} d \log(1/\delta) + \alpha C\bigg],\end{aligned}$$

Choose $\lambda = d/B$, $\alpha = (\sqrt{d} + \sqrt{\lambda} B)/C$. Then we have

$$\begin{aligned}\beta_t &\le \frac{2}{\sqrt{\kappa}}\Big[\sqrt{dB} + \sqrt{d \log((1 + 2tB/d)/\delta)}\Big] \\ \widetilde{\beta}_t &\le 4(1 + 4B)\sqrt{dB} \log\big((1 + 2tB/d^2)/\delta\big).\end{aligned} \tag{C.14}$$

Substituting (C.14) into (C.13), we have

$$\begin{aligned}\text{Regret}(T) &\le 150 \cdot dB^{1.5} \sqrt{T} \log^{1.5}\big((1 + 2TB/d)/\delta\big) \\ &\quad + \frac{10000(B/\kappa + 1/\kappa^2)}{\log(1/\kappa)} d^2 B^2 \log^3\big((1 + 2TB/d)/\delta\big) \\ &\quad + 32dB \log^2\big((1 + 2TB/d)/\delta\big) C/\kappa.\end{aligned}$$

This completes the proof of Theorem 6.1. □

D Proof of the Lemmas in Section C

We first prove Lemma C.1. Before we start, we will mention that Jun et al. (2021) considered a slightly different but relevant confidence bound. Specifically, they studied the logistic bandit problem with pure exploration and proposed an algorithm with a sample complexity guarantee. In contrast, our work focuses on the regret of dueling bandits, with the approach involving reward estimation complemented by a bonus term. Furthermore, Jun et al. (2021) derived a concentration inequality under a fixed design assumption, given by $|\langle x, \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\rangle| \leq \beta \|x\|_{H_t(\boldsymbol{\theta}^*)^{-1}}$. However, this assumption is too restrictive for our setting with adaptive action selection. In contrast, our concentration inequality (Lemma C.1) is applicable to adaptive designs that holds for any arm, which works for the adaptive arm-selection inherent to bandit algorithms.

D.1 Proof of Lemma C.1

We define some auxiliary quantities

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

In Algorithm 2, $\boldsymbol{\theta}_t$ is chosen to be the solution to the following equation,

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

Then we have

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

First, we will bound $\|Z_t\|_{\boldsymbol{\Sigma}_t^{-1}}$. Following the technique used in Section B.1, we decompose the summation in (B.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,

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

Summing up the two equalities, 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, we have

$$\|Z_t\|_{\boldsymbol{\Sigma}_t^{-1}} \le \frac{1}{\sqrt{\kappa}} \underbrace{\Big\| \sum_{i=1}^{t-1} w_i \sqrt{\kappa} \epsilon_i \big(\boldsymbol{\phi}(x_i, a_i) - \boldsymbol{\phi}(x_i, b_i)\big) \Big\|_{\boldsymbol{\Sigma}_t^{-1}}}_{I_1} + \underbrace{\Big\| \sum_{i<t:c_i=1} w_i \big(\boldsymbol{\phi}(x_i, a_i) - \boldsymbol{\phi}(x_i, b_i)\big) \Big\|_{\boldsymbol{\Sigma}_t^{-1}}}_{I_2}. \tag{D.1}$$

For the term I_1, with probability at least $1 - \delta/2$, for all $t \in [T]$, it can be bounded by

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

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

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

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

$$I_1 \le \sqrt{d \log(2(1 + 2t/\lambda)/\delta)}. \tag{D.2}$$

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

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

where the second inequality holds due to $\boldsymbol{\Sigma}_t \succeq \boldsymbol{\Sigma}_i$. The third inequality holds due to $w_i \le \alpha / \big\|(\boldsymbol{\phi}(x_i, a_i) - \boldsymbol{\phi}(x_i, b_i))\big\|_{\boldsymbol{\Sigma}_i^{-1}}$. The last inequality holds due to the definition of C. Substituting (D.2) and (D.3) into (D.1), we have

$$\|Z_t\|_{\boldsymbol{\Sigma}_t^{-1}} \le \frac{1}{\sqrt{\kappa}} \sqrt{d \log(2(1 + 2t/\lambda)/\delta)} + \alpha C.$$

Therefore, using the triangle inequality, we have

$$\begin{aligned}
\|G_t(\boldsymbol{\theta}^*) - G_t(\boldsymbol{\theta}_t)\|_{\boldsymbol{\Sigma}_t^{-1}} &\le \lambda \|\boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}_t^{-1}} + \frac{1}{\sqrt{\kappa}} \sqrt{d \log(2(1 + 2t/\lambda)/\delta)} + \alpha C \\
&\le \sqrt{\lambda} B + \frac{1}{\sqrt{\kappa}} \sqrt{d \log(2(1 + 2t/\lambda)/\delta)} + \alpha C.
\end{aligned} \tag{D.4}$$

Using the Newton-Leibniz formula, we have

$$G_t(\boldsymbol{\theta}^*) - G_t(\boldsymbol{\theta}_t) = \bigg[\int_0^1 \nabla G_t\big(\boldsymbol{\theta}_t + v(\boldsymbol{\theta}^* - \boldsymbol{\theta}_t)\big)\mathrm{d}v\bigg](\boldsymbol{\theta}^* - \boldsymbol{\theta}_t).$$

Recall the definition

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

then we have

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

We define $\mathbf{V}_t$ as follows:

$$\begin{aligned} \mathbf{V}_t &= \int_0^1 \nabla G_t\big(\boldsymbol{\theta}_t + v(\boldsymbol{\theta}^* - \boldsymbol{\theta}_t)\big)\mathrm{d}v \\ &= \lambda\mathbf{I} + \sum_{i=1}^{t-1} w_t \bigg[\int_0^1 \dot{\sigma}\Big(\big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big)^\top\big(\boldsymbol{\theta}_t + v(\boldsymbol{\theta}^* - \boldsymbol{\theta}_t)\big)\Big)\mathrm{d}v\bigg] \\ &\quad \big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big)\big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big)^\top. \end{aligned} \tag{D.5}$$

Therefore, we have $\mathbf{V}_t \succeq \boldsymbol{\Sigma}_t$. We know that $G_t(\boldsymbol{\theta}^*) - G_t(\boldsymbol{\theta}_t) = \mathbf{V}_t(\boldsymbol{\theta}^* - \boldsymbol{\theta}_t)$. Then the following inequality holds:

$$\begin{aligned} \big\|G_t(\boldsymbol{\theta}_t) - G_t(\boldsymbol{\theta}^*)\big\|_{\boldsymbol{\Sigma}_t^{-1}}^2 &= (\boldsymbol{\theta}_t - \boldsymbol{\theta}^*)^\top\mathbf{V}_t\boldsymbol{\Sigma}_t^{-1}\mathbf{V}_t(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*) \\ &\geq (\boldsymbol{\theta}_t - \boldsymbol{\theta}^*)^\top\boldsymbol{\Sigma}_t(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*). \end{aligned}$$

As a result, we have

$$\begin{aligned} \big\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\big\|_{\boldsymbol{\Sigma}_t} &\leq \big\|G_t(\boldsymbol{\theta}_t) - G_t(\boldsymbol{\theta}^*)\big\|_{\boldsymbol{\Sigma}_t^{-1}} \\ &\leq \sqrt{\lambda}B + \frac{1}{\sqrt{\kappa}}\sqrt{d\log(2(1+2t/\lambda)/\delta)} + \alpha C \\ &= \beta_t, \end{aligned}$$

where the first inequality holds due to (D.4). Consequently, we have $\mathbb{P}[\mathcal{E}_1] \geq 1 - \delta/2$.
Let

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

Next, we consider the following term:

$$\|Z_t\|_{\mathbf{H}_t^{-1}} \leq \underbrace{\bigg\|\sum_{i=1}^{t-1} w_i\epsilon_i\big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big)\bigg\|_{\mathbf{H}_t^{-1}}}_{J_1} + \underbrace{\bigg\|\sum_{i<t:c_i=1} w_i\big(\boldsymbol{\phi}(x_i,a_i) - \boldsymbol{\phi}(x_i,b_i)\big)\bigg\|_{\mathbf{H}_t^{-1}}}_{J_2}.$$

For J_1, let $\mathcal{F}_t = \sigma(x_1,a_1,b_1,o_1,x_2,a_2,b_2,o_2,\ldots,x_t,a_t,b_t)$. We know that ϵ_t It's $\mathcal{F}_{t+1}$-measurable.

$$\begin{aligned} &\mathbb{E}[\epsilon_t|\mathcal{F}_t] = 0 \\ &\mathbb{E}[\epsilon_t^2|\mathcal{F}_t] = \dot{\sigma}\Big(\big(\boldsymbol{\phi}(x_t,a_t) - \boldsymbol{\phi}(x_t,b_t)\big)^\top\boldsymbol{\theta}^*\Big) \end{aligned}$$

Using Lemma F.4, with probability at least $1-\delta/2$, for all $t \in [T]$, we have

$$J_1 = \Big\| \sum_{i=1}^{t-1} w_i \epsilon_i \big(\boldsymbol{\phi}(x_i, a_i) - \boldsymbol{\phi}(x_i, b_i)\big) \Big\|_{\mathbf{H}_t^{-1}} \le \frac{\sqrt{\lambda}}{2} + \frac{2}{\sqrt{\lambda}} \log\left(\frac{2\det(\mathbf{H}_t)^{1/2}\lambda^{-d/2}}{\delta}\right) + \frac{2}{\sqrt{\lambda}} d\log(2).$$

Using $w_i \le 1$, $\dot{\sigma}(\cdot) \le 1$, we have $\sqrt{w_i}\|\boldsymbol{\phi}(x_i, a_i) - \boldsymbol{\phi}(x_i, b_i)\|_2 \le 2$. Moreover, we have

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

where the first inequality holds because for every matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$, $\det \mathbf{A} \le (\mathrm{Tr}(\mathbf{A})/d)^d$. The second inequality holds due to $\sqrt{w_i}\|\boldsymbol{\phi}(x_i, a_i) - \boldsymbol{\phi}(x_i, b_i)\|_2 \le 2$. Therefore, we have

$$J_1 \le \frac{\sqrt{\lambda}}{2} + \frac{2}{\sqrt{\lambda}} d\log\left(\frac{d\lambda + 2T}{d\lambda}\right) + \frac{2}{\sqrt{\lambda}} d\log(1/\delta). \tag{D.6}$$

For J_2, we have

$$\begin{aligned}
J_2 &= \sum_{i<t:c_i=1} w_i \big\|(\boldsymbol{\phi}(x_i, a_i) - \boldsymbol{\phi}(x_i, b_i))\big\|_{\mathbf{H}_t^{-1}} \\
&\le \sum_{i<t:c_i=1} w_i \big\|(\boldsymbol{\phi}(x_i, a_i) - \boldsymbol{\phi}(x_i, b_i))\big\|_{\mathbf{H}_i^{-1}} \\
&\le \sum_{i<t:c_i=1} \alpha \\
&\le \alpha C,
\end{aligned} \tag{D.7}$$

where the first inequality holds due to $\mathbf{H}_t \succeq \mathbf{H}_i$. The second inequality holds due to $w_i \le \alpha/\|(\boldsymbol{\phi}(x_i, a_i) - \boldsymbol{\phi}(x_i, b_i))\big\|_{\boldsymbol{\Sigma}_i^{-1}}$ and $\mathbf{H}_i \succeq \boldsymbol{\Sigma}_i$. The last inequality holds due to the definition of C. Combining (D.6) and (D.7), we have

$$\begin{aligned}
\|G_t(\boldsymbol{\theta}^*) - G_t(\boldsymbol{\theta}_t)\|_{\mathbf{H}_t^{-1}} &\le \lambda\|\boldsymbol{\theta}^*\|_{\mathbf{H}_t^{-1}} + \frac{\sqrt{\lambda}}{2} + \frac{2}{\sqrt{\lambda}} d\log\left(\frac{d\lambda + 2t}{d\lambda}\right) + \frac{2}{\sqrt{\lambda}} d\log(1/\delta) + \alpha C \\
&\le \sqrt{\lambda}B + \frac{\sqrt{\lambda}}{2} + \frac{2}{\sqrt{\lambda}} d\log\left(\frac{d\lambda + 2t}{d\lambda}\right) + \frac{2}{\sqrt{\lambda}} d\log(1/\delta) + \alpha C,
\end{aligned} \tag{D.8}$$

where the first inequality holds due to the triangle inequality. Recall $\mathbf{V}_t$ defined in (D.5). Using Lemma F.5, we have

$$\begin{aligned}
\mathbf{V}_t &\ge \lambda\mathbf{I} + \sum_{i=1}^{t-1} w_i \frac{\dot{\sigma}\big(\big(\boldsymbol{\phi}(x_i, a_i) - \boldsymbol{\phi}(x_i, b_i)\big)^\top \boldsymbol{\theta}^*\big)}{1 + \big|\big(\boldsymbol{\phi}(x_i, a_i) - \boldsymbol{\phi}(x_i, b_i)\big)^\top (\boldsymbol{\theta}^* - \boldsymbol{\theta}_t)\big|} \big(\boldsymbol{\phi}(x_i, a_i) - \boldsymbol{\phi}(x_i, b_i)\big)\big(\boldsymbol{\phi}(x_i, a_i) - \boldsymbol{\phi}(x_i, b_i)\big)^\top \\
&\ge \frac{1}{1+4B}\mathbf{H}_t,
\end{aligned} \tag{D.9}$$

where the last inequality holds due to $\big|\big(\boldsymbol{\phi}(x_i, a_i) - \boldsymbol{\phi}(x_i, b_i)\big)^\top (\boldsymbol{\theta}^* - \boldsymbol{\theta}_t)\big| \le 4B$. This leads to

$$\begin{aligned}
\|G_t(\boldsymbol{\theta}^*) - G_t(\boldsymbol{\theta}_t)\|_{\mathbf{H}_t^{-1}}^2 &= (\boldsymbol{\theta}^* - \boldsymbol{\theta}_t)^\top \mathbf{V}_t \mathbf{H}_t^{-1} \mathbf{V}_t (\boldsymbol{\theta}^* - \boldsymbol{\theta}_t) \\
&\ge \frac{1}{1+4B} (\boldsymbol{\theta}^* - \boldsymbol{\theta}_t)^\top \mathbf{V}_t (\boldsymbol{\theta}^* - \boldsymbol{\theta}_t). \\
&\ge \frac{1}{(1+4B)^2} (\boldsymbol{\theta}^* - \boldsymbol{\theta}_t)^\top \mathbf{H}_t (\boldsymbol{\theta}^* - \boldsymbol{\theta}_t),
\end{aligned} \tag{D.10}$$

where the first equality holds due to $G_t(\boldsymbol{\theta}^*) - G_t(\boldsymbol{\theta}_t) = \mathbf{V}_t(\boldsymbol{\theta}^* - \boldsymbol{\theta}_t)$. The first and second inequalities hold due to (D.9). Conditioned on the event $\mathcal{E}_1$, recall that $\widehat{\Delta}_t$ is defined by

$$\widehat{\Delta}_t = \big|\big(\boldsymbol{\phi}(x_t, a_t) - \boldsymbol{\phi}(x_t, b_t)\big)^\top \boldsymbol{\theta}_t\big| + \beta_t \|\boldsymbol{\phi}(x_t, a_t) - \boldsymbol{\phi}(x_t, b_t)\|_{\boldsymbol{\Sigma}_t^{-1}}.$$

Therefore, we have

$$\begin{aligned}
|\widehat{\Delta}_t| &\geq \big|\big(\boldsymbol{\phi}(x_t, a_t) - \boldsymbol{\phi}(x_t, b_t)\big)^\top \boldsymbol{\theta}^*\big| + \beta \|\boldsymbol{\phi}(x_t, a_t) - \boldsymbol{\phi}(x_t, b_t)\|_{\boldsymbol{\Sigma}_t^{-1}} - \big|\big(\boldsymbol{\phi}(x_t, a_t) - \boldsymbol{\phi}(x_t, b_t)\big)^\top (\boldsymbol{\theta}^* - \boldsymbol{\theta}_t)\big| \\
&\geq \big|\big(\boldsymbol{\phi}(x_t, a_t) - \boldsymbol{\phi}(x_t, b_t)\big)^\top \boldsymbol{\theta}^*\big|,
\end{aligned}$$

where the first inequality holds due to the triangle inequality. The second inequality holds due to $\mathcal{E}_1$ and the Cauchy-Schwarz inequality. Let $v_t = \max\{\kappa, \dot{\sigma}(\widehat{\Delta}_t)\}$. Using the fact that $\dot{\sigma}(\cdot)$ is an even function decreasing when $x > 0$, we have $v_t \leq \dot{\sigma}\big(\big(\boldsymbol{\phi}(x_t, a_t) - \boldsymbol{\phi}(x_t, b_t)\big)^\top \boldsymbol{\theta}^*\big)$. Therefore, we have $\boldsymbol{\Lambda}_t \preceq \mathbf{H}_t$. As a result, (D.10) indicates that

$$\begin{aligned}
\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_t\|_{\boldsymbol{\Lambda}_t}^2 &\leq \|\boldsymbol{\theta}^* - \boldsymbol{\theta}_t\|_{\mathbf{H}_t}^2 \\
&\leq (1 + 4B)^2 \|G_t(\boldsymbol{\theta}^*) - G_t(\boldsymbol{\theta}_t)\|_{\mathbf{H}_t^{-1}}^2.
\end{aligned}$$

Using (D.8), we have

$$\begin{aligned}
\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_t\|_{\boldsymbol{\Lambda}_t} &\leq (1 + 4B) \|G_t(\boldsymbol{\theta}^*) - G_t(\boldsymbol{\theta}_t)\|_{\mathbf{H}_t^{-1}} \\
&\leq (1 + 4B) \bigg[\sqrt{\lambda} B + \frac{\sqrt{\lambda}}{2} + \frac{2}{\sqrt{\lambda}} d \log\bigg(\frac{d\lambda + 2t}{d\lambda}\bigg) + \frac{2}{\sqrt{\lambda}} d \log(1/\delta) + \alpha C\bigg] \\
&= \widetilde{\beta}_t.
\end{aligned}$$

Taking a union bound, we complete the proof of Lemma C.1.

E Experiments

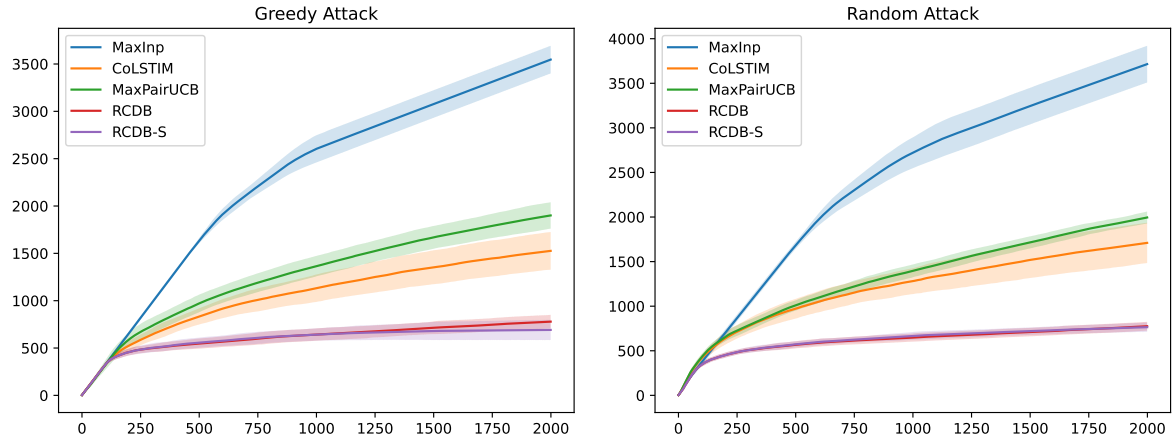
In this section, we conduct simulation experiments to verify our theoretical results.

E.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 $|\mathcal{A}| = 2^d = 32$.

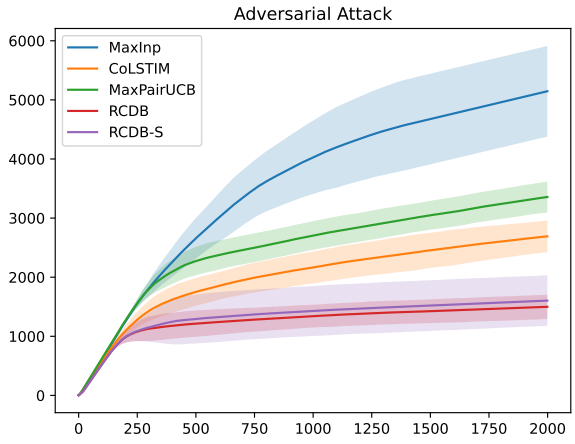
Adversarial Attack Methods. We study the performance of our algorithm using different adversarial attack methods. We categorize the first two methods as "weak" primarily because the adversary in these scenarios does not utilize information about the agent's actions. In contrast, we classify the latter 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.

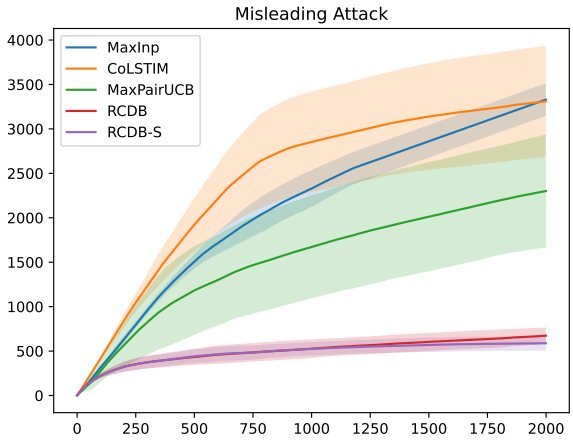


(a) Greedy Attack

(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 $C = \lceil \sqrt{T} \rceil$.

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 cumulative regret averaged across 10 random runs.

E.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.

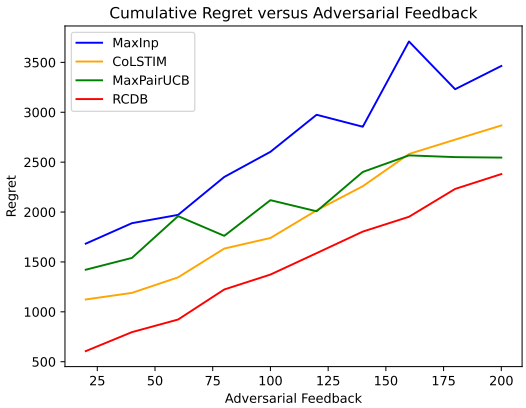


Figure 2. The relationship between cumulative regret and the number of adversarial feedback C. For this specific experiment, we employ the "greedy attack" method to generate the adversarial feedback. C is selected from the set $[20, 40, 60, 80, 100, 120, 140, 160, 180, 200]$ (10 adversarial levels).

E.3 Robustness to Different Numbers of Adversarial Feedback

In this section, we test the performance of algorithms with increasing times of adversarial feedback. As shown in Figure 2, our algorithm has a linear dependency on the number of adversarial feedback C, which is consistent with the theoretical results we have proved in Theorem 5.3. In comparison to other algorithms, RCDB demonstrates superior robustness against adversarial feedback, as evidenced by its notably smaller regret.

F Auxiliary Lemmas

Lemma F.1 (Azuma–Hoeffding inequality, Cesa-Bianchi & Lugosi 2006)**.** Let $\{\eta_k\}_{k=1}^K$ be a martingale difference sequence with respect to a filtration $\{\mathcal{F}_t\}$ satisfying $|\eta_t| \leq R$ 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

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

Lemma F.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.

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

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

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

Lemma F.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 $\|\mathbf{x}_t\|_2 \leq L$ holds for all $t \in [T]$, we have

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

Lemma F.4 (Theorem 1 in Faury et al. 2020)**.** Let $\{\mathcal{F}_t\}_{t=1}^\infty$ be a filtration. Let $\{\mathbf{x}_t\}_{t=1}^\infty$ be a stochastic process, such that $\mathbf{x}_t$ is $\mathcal{F}_t$-measurable. Suppose $\|\mathbf{x}_t\|_2 \leq 1$. Furthermore, let $\{\epsilon_t\}_{t=1}^\infty$ be a stochastic process which is $\mathcal{F}_{t+1}$-measurable. Assume $|\epsilon_t| \leq 1$. $\mathbb{E}[\epsilon_t | \mathcal{F}_t] = 0$. $\mathbb{E}[\epsilon_t^2 | \mathcal{F}_t] = \sigma_t^2$. Let $\lambda > 0$ and for any $t \geq 1$ define:

$$\mathbf{H}_t := \sum_{s=1}^t \sigma_s^2 \mathbf{x}_s \mathbf{x}_s^\top + \lambda \mathbf{I}.$$

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

$$\Big\| \sum_{i=1}^t \mathbf{x}_i \epsilon_i \Big\|_{\mathbf{H}_t^{-1}} \leq \frac{\sqrt{\lambda}}{2} + \frac{2}{\sqrt{\lambda}} \log\left(\frac{\det(\mathbf{H}_t)^{1/2} \lambda^{-d/2}}{\delta} \right) + \frac{2}{\sqrt{\lambda}} d \log(2).$$

Lemma F.5 (Lemma 7 in Abeille et al. 2021)**.** Let f be a strictly increasing function such that $|\ddot{f}| \leq \dot{f}$, and let I be any bounded interval of $\mathbb{R}$. Then, for all $z_1, z_2 \in I$, the following inequality holds:

$$\int_0^1 \dot{f}\big(z_1 + v(z_2 - z_1)\big) \mathrm{d}v \geq \frac{\dot{f}(z)}{1 + |z_1 - z_2|} \text{ for } z \in \{z_1, z_2\}.$$