Learning from *Imperfect* Human Feedback: a Tale from Corruption-Robust Dueling*

Yuwei Cheng[1] yuweicheng@uchicago.edu
Fan Yao[2,3†‡] fy4bc@virginia.edu
Xuefeng Liu[2†] xuefeng@uchicago.edu
Haifeng Xu[2] haifengxu@uchicago.edu
[1]*Department of Statistics, University of Chicago, USA*
[2]*Department of Computer Science, University of Chicago, USA*
[3]*Department of Computer Science, University of Virginia, USA*

Abstract

This paper studies Learning from *Imperfect* Human Feedback (LIHF), motivated by humans' potential irrationality or imperfect perception of true preference. We revisit the classic dueling bandit problem as a model of learning from comparative human feedback, and enrich it by casting the imperfection in human feedback as *agnostic* corruption to user utilities. We start by identifying the fundamental limits of LIHF and prove a regret lower bound of $\Omega(\max\{T^{1/2}, C\})$, even when the total corruption C is known and when the corruption decays gracefully over time (i.e., user feedback becomes increasingly more accurate). We then turn to design robust algorithms applicable in real-world scenarios with arbitrary corruption and unknown C. Our key finding is that gradient-based algorithms enjoy a smooth efficiency-robustness tradeoff under corruption by varying their learning rates. Specifically, under general concave user utility, Dueling Bandit Gradient Descent (DBGD) of Yue and Joachims (2009) can be tuned to achieve regret $O(T^{1-\alpha} + T^{\alpha}C)$ for any given parameter $\alpha \in (0, \frac{1}{4}]$. Additionally, this result enables us to pin down the regret lower bound of the standard DBGD (the $\alpha = 1/4$ case) as $\Omega(T^{3/4})$ for the first time, to the best of our knowledge. For strongly concave user utility we show a better tradeoff: there is an algorithm that achieves $O(T^{\alpha} + T^{\frac{1}{2}(1-\alpha)}C)$ for any given $\alpha \in [\frac{1}{2}, 1)$. Our theoretical insights are corroborated by extensive experiments on real-world recommendation data.

1. Introduction

Many real-world problems, such as personalized recommendation (Yue and Joachims, 2009; Immorlica et al., 2020; Yao et al., 2022a) and fine-tuning generative models (Bai et al., 2022; Casper et al., 2023; Han et al., 2024), require learning from human feedback. Expressing preferences as numerical values or concrete functions is generally challenging for humans. Therefore, an approach that has achieved significant real-world success is to learn humans' preferences by eliciting their comparative feedback between two options (Bai et al., 2022).

†Equal contribution.

‡Work done while visiting University of Chicago.

*This work is supported by the AI2050 program at Schmidt Sciences (Grant G-24-66104), Army Research Office Award W911NF-23-1-0030, ONR Award N00014-23-1-2802 and NSF Award CCF-2303372.

A natural theoretical framework capturing such learning task is the seminal dueling bandits framework introduced by Yue and Joachims (2009). The dueling bandit problem features an sequential online learning problem, during which a pair of actions are selected at each round and only their comparison result is revealed to the learner. The comparison outcome between two actions is modeled using a utility function of actions, alongside a link function that determines the probability of each action winning based on their utility difference. This modeling approach is termed as utility-based dueling bandit and achieves great success in generating summaries closely aligned with human preferences (Stiennon et al., 2020).

However, human feedback is often *imperfect*, a crucial factor that has been largely overlooked in previous studies of dueling bandit. Ample behavioral studies show that humans often refine their preferences through interactions with the system, navigating a complex tradeoff between exploration and exploitation (Wilson et al., 2014; Cohen et al., 2007). Furthermore, their response errors may exhibit systematic patterns rather than merely random noise. As we are all aware of, humans are not always rational (Posner, 1997) neither perfectly know our preferences (Pu et al., 2012).

In this paper, we study learning from *imperfect* human feedback (LIHF) by casting the imperfection in the feedback as certain form of corruption to the utilities. This leads us to study a natural variant of the continuous dueling bandit with corrupted utilities. Notably, our study differs from recent robust algorithms designed for corrupted dueling bandits (Di et al., 2024; Saha and Gaillard, 2022; Komiyama et al., 2015) in three key aspects. Firstly, these previous algorithms all focus on a finite action space whereas, to the best of our knowledge, ours is the first to investigate a robust dueling bandit algorithm under corruption for continuous action space. This focus is motivated by the fact that many real-world applications have continuous or extremely dense action spaces, such as token embeddings or content recommendations (Chen et al., 2021, 2019). Second, our algorithms are designed for *agnostic* corruption without knowing the corruption levels, whereas Di et al. (2024) requires knowledge of the total corruption amount. This focus on agnostic corruption is motivated by realistic considerations, because systems often cannot accurately discern the irrational or imperfect nature of human feedback. Third, our corruption models are very different from the model studied by Di et al. (2024); Saha and Gaillard (2022); Komiyama et al. (2015). The corruption in their model contaminates feedback by flipping the realized outcomes of the duels, whereas the corruption in our model is on utilities which consequently only alter the winning probabilities of the duels. This difference is due to our different motivation of corruption, which is to capture imperfect human behavior while not intentional adversarial attack (see more illustration in Section 3).

Our Contributions. By modeling LIHF as a continuous dueling bandit under agnostic corruption, we start by characterizing its intrinsic hardness. We prove a regret lower bound of $\Omega(d \max\{\sqrt{T}, C\})$, even when the total corruption C is known and when corruption decays gradually over time. Additionally, we prove that this lower bound result is tight by exhibiting an algorithm that indeed achieves a matching upper bound for the aforementioned setting. Next, we turn to design robust algorithms that are applicable to arbitrary and agnostic corruption with unknown C. Our key finding is that gradient-based algorithms offer a smooth efficiency-robustness tradeoff under corruption by carefully tuning their learning rates. Specifically, let d denote the problem dimension. For any given $\alpha \in (0, 1/4]$, Dueling

Bandit Gradient Descent (DBGD) (Yue and Joachims, 2009) with carefully chosen learning rate (that depends on α) can achieve a regret of $O(\sqrt{d}T^{1-\alpha} + \sqrt{d}T^{\alpha}C)$ under arbitrary concave user utility. That is, by decreasing α from $1/4$ to 0, the regret under no corruption (i.e., learning efficiency) decays from $\sqrt{d}T^{3/4}$ to $\sqrt{d}T$ but the algorithm's tolerance to total amount of corruption increases from $O(T^{3/4})$ to $O(T)$ (i.e., more robustness). Perhaps surprisingly, this result turns out to allow us to pin down a regret lower bound for the standard version of DBGD (the $\alpha = 1/4$ case) as $\Omega(T^{3/4})$, marking the first tight lower bound result for the DBGD algorithm to the best of our knowledge. For strongly concave user utility, we show that an improved tradeoff is possible: Noisy Comparison based Stochastic Mirror Descent (NC-SMD) (Kumagai, 2017) can be refined to achieve a regret of $O(dT^{\alpha} + \sqrt{d}T^{\frac{1-\alpha}{2}}C + dC)$ for any given $\alpha \in [1/2, 1)$. We remark that similar robustness-accuracy tradeoff has been studied classification problem with adversarial examples, both empirically (Tsipras et al., 2018) and theoretically (Zhang et al., 2019). However, to the best of our knowledge, this is the first time such tradeoff is formally analyzed for online learning. Finally, we conduct extensive experiments on real-world recommendation data to validate our theoretical insights.

2. Related Work

Hardness of Learning from Imperfect Human Feedback: To design algorithms robust to adversarial corruption, it is crucial to understand the intrinsic hardness of learning from corrupted feedback. Lattimore and Szepesvári (2020) demonstrates that the minimax regret lower bound for stochastic linear bandits is $\Omega(d\sqrt{T})$. Utilizing the regret lower bound for bandit convex optimization (Shamir, 2013), Kumagai (2017) illustrates that dueling bandits with a strongly concave utility function achieve a regret lower bound of $\Omega(d\sqrt{T})$. In scenarios of adversarial corruption, Bogunovic et al. (2020) proves that any algorithm faces regret at least $\Omega(dC)$ for stochastic linear bandits. However, currently there is no literature that has explored the intrinsic hardness of learning from corruption which decays gradually over time (Wilson et al., 2014; Kumagai, 2017; Cohen et al., 2007), despite its prevalence in learning from imperfect human feedback. It still remains unknown whether learning from such corruption is fundamentally easier than learning from arbitrary corruption.

Stochastic Bandits with Unknown Corruption: Applying the principle of optimism in face of uncertainty and constructing estimators involving uncertainty weighting, algorithms devised by He et al. (2022) and Ye et al. (2024a) attain nearly optimal regret bounds for both linear and nonlinear contextual bandits in presence of adversarial corruption. By leveraging the concept of phase elimination, Bogunovic et al. (2022) introduces a robust algorithm capable of handling corruption in rewards drawn from Gaussian processes. However, the designs of these algorithms require explicit knowledge of the total corruption level C as an input. Gupta et al. (2019) employ novel sampling strategy to channel resources to arms with better performance and craft algorithms robust to agnostic corruption, yet restricted to finite action sets. In scenarios involving continuous action set with unknown corruption level, even though algorithms designed by Ding et al. (2022) and Bogunovic et al. (2020) have sublinear regret guarantee, the regret is at least C times worse than the regret without corruption. It remains as an open problem of whether better regret guarantee is achievable—

that is, affording unknown corruption level more than $\Omega(\sqrt{T})$ while maintaining $O(\sqrt{T})$ regret guarantee in scenario without corruption.

Continuous Dueling Bandits with Unknown Corruption: In addition to stochastic bandits, there exists an extensive body of literature studying the dueling bandit problem with unknown corruption, given its prevalence in real-world application (Bengs et al., 2021b; Sui et al., 2018). The literature on dueling bandits with adversarial corruption has been initially explored by Agarwal et al. (2021) and subsequently enhanced by Saha and Gaillard (2022), which achieves optimality in term of C. However, their work is solely applicable to finite action sets, which may not be feasible in many real-world scenarios that have continuous or extremely dense action spaces, such as token embeddings or content recommendations (Chen et al., 2021, 2019). Ailon et al. (2014) introduced a reduction method that translates existing results from conventional Multi-Armed Bandit algorithms to Dueling Bandits. This reduction method is applicable to both finite and continuous action spaces, albeit requiring a linear link function. DBGD (Yue and Joachims, 2009) and NC-SMD (Kumagai, 2017) enjoy flexibility in choosing the link function but only work for continuous dueling bandits. It still remains as an open problem for developing provably efficient algorithm capable of learning from corrupted dueling feedback with unknown corruption in continuous action space. This serves as the driving force behind our theoretical analysis.

3. The Problem of Learning from Imperfect Human Feedback (LIHF)

Conventional Notation. Throughout the paper, for a positive integer T, we use $[T]$ to denote $\{1, 2, \ldots, T\}$. We use $\mathcal{P}_{\mathcal{A}}(a)$ to denote the projection of action a on $\mathcal{A}$. We use standard asymptotic notations including $O(\cdot)$, $\Omega(\cdot)$, $\Theta(\cdot)$. We use $\tilde{O}(\cdot)$, $\tilde{\Omega}(\cdot)$, $\tilde{\Theta}(\cdot)$ to hide logarithmic factors. We use $\|\cdot\|_\infty$ to define infinity norm. Reg_T represents regret with dueling feedback over T rounds.

To capture learning from imperfect (comparative) human feedback, we now introduce a natural variant of the seminal dueling bandit framework (Yue and Joachims, 2009) by integrating the "imperfections" in human feedback as certain corruption to human agents' feedback. An immediate critique one might have is that arbitrarily adversarial corruption seems too pessimistic as a way to model humans' inaccuracies which are often caused by uncertainty or ignorance, rather than targeted attacks. While this may appear true at the first glance, we provide strong theoretical and empirical evidences that corruption caused by uncertainties generally is as hard as arbitrary corruption. For such reasons, we coin the following problem *Learning from Imperfect Human Feedback* (LIHF).

Basics of Dueling Bandits and Assumptions. We consider the standard dueling bandit framework with continuous *action* set $\mathcal{A} \subset \mathbb{R}^d$ (Yue and Joachims, 2009). At each round $t \in [T]$, the learner (henceforth denoted as $\mathcal{L}$) chooses two actions a_t, a'_t to present to some human agent, henceforth denoted as the "user". The user receives *utility* $\mu(a_t), \mu(a'_t)$ from the two actions respectively, and will pick one of these actions following a *link* function $\sigma(\cdot)$. That is, in absence of any corruption, the user picks action a_t with probability $\sigma(\mu(a_t) - \mu(a'_t))$, and action a'_t otherwise. As clear from its definition, this choice probability depends only on the utility difference $\mu(a_t) - \mu(a'_t)$. Notably, user's randomized action does carry information about underlying utility function μ, which is why the action is also called the "human feedback" (Bengs et al., 2021a; Maystre and Grossglauser, 2017; Ailon et al., 2014). Following standard assumptions in this literature (Kumagai, 2017; Yue and Joachims, 2009), we also assume

1. the continuous action space $\mathcal{A}$ contains the origin, is compact, convex and is contained in a d-dimensional ball of radius R;
2. the utility function $\mu : \mathcal{A} \to \mathbb{R}$ is concave, differentiable, L_μ-Lipschitz and admits unique optimal within set $\mathcal{A}$, denoted as $a^* := \arg\max_{a \in \mathcal{A}} \mu(a)$;
3. the link function $\sigma : \mathbb{R} \to [0, 1]$ is rotational-symmetric, L_σ-Lipschitz and its first derivative σ' is L_2-Lipschitz.

For convenience, we further let $L = L_\sigma L_\mu$. Notably, some previous works restrict link function to specifically be the logistic link function (Di et al., 2024; Xu et al., 2024; Saha, 2021). However, our results shall be applicable to any link functions that satisfy the natural assumptions above.

Inaccuracies in Human Feedback as Agnostic Corruption. Extending the standard dueling bandit framework above, we introduce a *corruption* term $c_t(a_t, a'_t)$ to the utility difference in order to model potential inaccuracy in the user's perception of underlying true utilities. For all our positive results, we always work with the situation where this corruption is *agnostic* in the sense that the learner does not know the value of each c_t neither their accumulated sum. Such agnostic corruption is natural because in reality it is very difficult to truly estimate how imperfect a user is, hence we aim to develop algorithms whose execution is agnostic to such parameters. Let $a \succ a'$ denote the event that the user chooses a over a'. Hence in our LIHF problem, the probability of event $a \succ a'$ in presence of corruption $c(a, a')$, denoted as $\hat{\mathbb{P}}(a \succ a')$, can be expressed as follows $\hat{\mathbb{P}}(a \succ a') = \sigma(\mu(a) - \mu(a') + c(a, a'))$. We further let $\hat{\mathcal{F}}(a_t, a'_t)$ denote user's preferential feedback under corruption, coined *corrupted dueling feedback*. Mathematically, $\hat{\mathcal{F}}(a_t, a'_t)$ follows a binomial distribution with mean $\hat{\mathbb{P}}(a_t \succ a'_t)$, expressed by the following equation.

$$\hat{\mathcal{F}}(a_t, a'_t) := \begin{cases} 1 \text{ w.p. } \hat{\mathbb{P}}(a_t \succ a'_t) \\ 0 \text{ w.p. } 1 - \hat{\mathbb{P}}(a_t \succ a'_t) \end{cases}$$

The "hat" notation in $\hat{\mathcal{F}}$ and $\hat{\mathbb{P}}$ is to denote corruption. When there is no corruption, we use $\mathbb{P}(a \succ a') = \sigma(\mu(a) - \mu(a'))$ to denote the probability of event $a \succ a'$ and $\mathcal{F}(a, a')$ to denote corresponding dueling feedback. It is important to recognize the different roles of the link function σ and utility corruption c_t — σ captures the inherent (often inevitable) randomness in human choice/behavior even given true utilities (Fishburn et al., 1979; Becker et al., 1963; Bradley and Terry, 1952), whereas our c_t is introduced to model inaccurate perception of utilities due to not perfectly knowing her preferences and, importantly, can be alleviated or even eliminated once the user has tried sufficient actions.

Imperfect user feedback generally is not as wild as arbitrary adversarial corruption. A particularly natural pattern of human's imperfection is that the inaccuracy in our estimations often diminishes as we try more and more actions. That is, humans learn along the way, though how fast the inaccuracy decreases often depends on our learning rate. This hence motivates us to consider the following notion of *ρ-imperfect user* which significantly constrain the corruption sequence $c_1, c_2, \cdots$, and force it to diminish at the rate ρ.

Definition 1 (ρ-Imperfect User). *The user feedback is said to be ρ-imperfect for some $\rho \in [0,1]$ if there exists a constant $C_\kappa > 0$ such that the corruption c_t satisfies $|c_t| \leq C_\kappa t^{\rho-1}, \forall t \in [T]$.*[*]

For a ρ-imperfect user, the accumulated corruption $\sum_{\tau=1}^{t-1} |c_\tau| \leq C_\kappa t^\rho$ for every t, which is why this is coined ρ-imperfect. The diminishing inaccuracy of user estimation in Definition 1 is a natural model for user's learning behavior and has already been used in various previous works to model user's imperfect feedback (Wang et al., 2023; Yao et al., 2022a; Immorlica et al., 2020) (Details are presented at Appendix A). Note that agnostic corruption in this situation means the learner know $|c_t| \leq C_\kappa t^{\rho-1}$ for some ρ, but she does not know the value of ρ.

On the other hand, the corruption are said to be *arbitrary* up to certain budget C, if $\sum_{t=1}^{T} |c_t| \leq C$ for a given (hence fixed) time horizon T. Our designed (positive) learning algorithmic will have guaranteed regret upper bound under arbitrary corruption, though our lower bound (i.e., negative result) will hold even in the weaker ρ-imperfect user feedback.

We remark that another line of recent research studies a different type of corruption that directly flip the outcomes of the duels; i.e., flipping the realized event $a \succ a'$ to its opposite $a' \succ a$ (Di et al., 2024; Saha and Gaillard, 2022; Komiyama et al., 2015). These corruption models are motivated by the existence of true *adversaries*, e.g., malicious users in recommendation systems (Lykouris et al., 2018) or fraudulent clicks (Deshpande and Montanari, 2013), who have the capability of altering user choices. In contrast, only utilities are corrupted in our model. These two models are not comparable. For instance, the adversary can induce a deterministic event $a \succ a'$ at any given round with one corruption, whereas the outcomes in our model is almost always probabilistic (unless the adversary uses an infinite budget c_t). Such modeling difference inherently comes from the different motivations of our LIHF problem. Specifically, our model is to capture imperfect human feedback due to inaccurate perception of utilities or ignorance of true preferences. For example, when recommender systems learn to find the optimal recommendation for a user, the user may not know her true preference at the beginning and often gradually refines her preferences through experiencing different items (Yao et al., 2022b). Similarly, when generative models like ChatGPT learn to generate personalized content for a user, the user may go through a similar procedure of preference refinement during interactions with the model. These modelling differences also necessitate our fundamentally different techniques. Our approaches are based on gradient-based methods whereas Di et al. (2024) and Saha (2021) employ the principle of optimistic in face of uncertainty and construct the confidence set of preferences through querying the maximum information pair. Their approaches inherently require the knowledge of total corruption level C. In contrast, the analysis of our gradient-based methods controls the impact of gradient bias (due to corruption) on regret. By adjusting its learning rate, our algorithm will gracefully "trade" its learning efficiency for more robustness against agnostic corruption. Moreover, our algorithm is suitable for

[*]All our results naturally applies to $\rho < 0$, which is a significantly easier situation since accumulated corruption is a constant in that case. We hence will not explicitly consider it in this paper.

infinite *action* space, whereas previous works of Di et al. (2024); Saha (2021) are for finite action spaces.

Learning goal: agnostic corruption-robust regret minimization. The goal of the learner is to optimize her sequential actions to minimize the following dueling regret against an arbitrary sequence of corruption $\{c_t\}_{t=1}^T$:

$$\mathrm{Reg}_T = \mathbb{E}\left(\sum_{t=1}^{T} \sigma(\mu(a^*) - \mu(a_t)) + \sigma(\mu(a^*) - \mu(a_t')) - 1\right). \tag{1}$$

This regret measure is also studied in many prior literature (Saha and Gaillard, 2022; Kumagai, 2017; Komiyama et al., 2015; Yue and Joachims, 2009). Importantly, all our algorithms are agnostic to the corruption in the sense that the algorithm does *not* depend on any knowledge of $\{c_t\}_{t=1}^T$.

4. The Intrinsic Limit of Learning from Imperfect Human Feedback

To understand the inherent difficulty of LIHF, in this section we develop regret lower bounds for a setting that is even easier to learn – that is, learning from a ρ-imperfect user (Definition 1) with a *known* ρ. While this non-agnostic may appear simpler, our main result of this section exhibits a tight lower bound of $\Omega(d \max\{\sqrt{T}, T^\rho\})$ for strongly concave user utilities. We also show a similar lower bound hold for linear user utility as well. Our lower bound result differs in two key aspects from various previous lower bound results for online learning under corruptions (Agarwal et al., 2021; Bogunovic et al., 2020). Firstly, previous lower bound proofs require the adversaries' capability to corrupt the reward feedback in an *arbitrary* way, whereas the corruption in our proof will follow Definition 1. It is unclear whether such strictly more restrictive corruption will be more tractable since it does offer the learner much more accurate estimation about the amount of corruption at each round. Secondly, previous construction of hard instances that attains regret lower bound relies on linear reward, whereas our lower bound holds for strongly concave utility functions, since they more realistically model human's risk attitude and consumption utilities (Chiappori and Rochet, 1987). Strongly concave utility functions are generally perceived as easier to optimize and to learn, hence it is unclear whether it is also the case under imperfect user feedback. While both differences above seem to hint that our problem might be easier than previous online learning problem under corruption, we show that the answer is actually "No".

Theorem 1 (A Tight Lower Bound)**.** *There exists an LIHF instance with a ρ-imperfect user (Def. 1) and strongly concave user utility such that any learner has to suffer $Reg_T \geq \Omega(d \max\{\sqrt{T}, T^\rho\})$ even* with the knowledge of ρ. *Moreover, this regret lower bound is tight, up to logarithmic terms. That is, for the same setting, there exists an online learning algorithm with $Reg_T \leq \tilde{O}(d \max\{\sqrt{T}, T^\rho\})$.*

This theorem reveals the fundamental limits of learning from ρ-imperfect users, even when the learner knows ρ. The core new lemma needed for proving the regret lower bound in Theorem 1 is to show that even in standard online learning with noisy bandit feedback about any pulled arm's reward (as opposed to comparative feedback in dueling setups), the same regret lower bound holds. This is summarized in the following Lemma 1. We then convert this lower bound for standard bandit reward feedback to our setting of dueling/comparative feedback. Notably, Ye et al. (2024b) recently showed a matching upper bound for the standard bandit reward feedback setting assuming *arbitrary corruption* with total amount C. Hence Lemma 1 and their results together echoes our key insight in this section — that is, learning from ρ-imperfect user is as hard as learning from arbitrarily corrupted rewards in classic bandit reward feedback setting as well.

For linear user utility, perhaps unsurprisingly, we are able to show the same lower bound as follows (see Prop. 1). En route, we similarly show that the problem of learning directly from bandit linear reward feedback also suffers the same regret lower bound $\Omega(d\max\{\sqrt{T}, T^{\rho}\})$. This happens to resolve an open problem proposed by Yao et al. (2022a) and implies that the regret upper bound shown in (Yao et al., 2022a) for the "learning from learning user" problem is tight in T.

Proposition 1. *There exists an LIHF instance with a ρ-imperfect user and linear user utility such that any learner knowing ρ has to suffer $Reg_T \geq \Omega(d\max\{\sqrt{T}, T^{\rho}\})$.*

We defer the full proof of Theorem 1 and Proposition 1 to Appendix B.1 and B.2. Below we only state a representative core lemma for the strongly concave utility function case and its proof sketch, which is the crux of our lower bound proof. Notably, the proof for the matching upper bound in Theorem 1 is also highly non-trivial. Our starting point is the recent NC-SMD algorithm. However, to achieve the tight regret upper bound, we have to employ a novel and highly involved induction argument to identify the optimal learning rate. Through this process, we established a new connection between the convergence of the L_2-norm of the action $\|a_t - a^*\|_2$ and the regret itself (It has been experimentally verified at Appendix D.1.2).

Lemma 1. *Assume that the action space $\mathcal{A}$ is contained in a d-dimensional unit ball. Consider the utility function μ in the form $\mu_\theta(a) = \theta^\top a - \frac{1}{2}\|a\|_2^2$, where $\theta \in \mathbb{R}^d, \|\theta\|_2 \leq 1$, is a random vector. Under corruption induced by ρ-*imperfect *user (Def. 1), for any fixed number of rounds T and $d \leq \frac{1}{C_\kappa}T^{1-\rho}$, there exists a preference parameter θ such that for any learner $\mathcal{L}$ under reward feedback,* even with the knowledge of *ρ, has to suffer regret $Reg_T := \mathbb{E}\{\sum_{t=1}^{T}\mu_\theta(a^*) - \mu_\theta(a_t)\} \geq \frac{d}{4}C_\kappa T^{\rho}$.*

Proof (Proof Sketch of Lemma 1). We aim to show that there exists a corruption strategy such that the regret incurred by any learner $\mathcal{L}$ with *corrupted reward feedback* is no less than $\Omega(dT^{\rho})$. The formulation of such a corruption strategy hinges on the observation that the regret incurred by $\mathcal{L}$ is bounded below by the sum of the Kullback-Leibler (KL) divergence between the distribution of the *corrupted reward feedback* conditioned on different possible values of θ. Specifically, let θ be uniformly drawn from $\{-\beta, \beta\}^d$. Given an action a_t, conditioned on $\theta_i > 0$, the corrupted reward feedback $\hat{v}_t$ is

$$\hat{v}_t = \mu_\theta(a_t) + c_t(a_t|\theta_i > 0) + \xi_{a_t} = \underbrace{\left(-\frac{1}{2}\|a_t\|^2 + \sum_{j\neq i}\theta_j a_{t,j}\right) + \beta a_{t,i} + c_t(a_t|\theta_i > 0)}_{\mu_1} + \xi_{a_t}. \tag{2}$$

Conditioned on $\theta_i < 0$, the corrupted reward feedback is

$$\hat{v}_t' = \mu_\theta(a_t) + c_t(a_t|\theta_i < 0) + \xi_{a_t} = \underbrace{\left(-\frac{1}{2}\|a_t\|^2 + \sum_{j\neq i}\theta_j a_{t,j}\right) - \beta a_{t,i} + c_t(a_t|\theta_i < 0)}_{\mu_2} + \xi_{a_t}. \tag{3}$$

Consider the following corruption strategy which sets $c_t(a_t|\theta_i > 0) = -\beta a_{t,i}$ and $c_t(a_t) = \beta a_{t,i}$ to minimizes the KL divergence between $\hat{v}_t$ and $\hat{v}_t'$. Together with 1-strongly concavity of μ_θ, we can establish

$$\mathbb{E}\left\{\sum_{t=1}^{T}\mu_\theta(a^*) - \mu_\theta(a_t)\right\} \geq \frac{1}{2}\max\left\{\sum_{t=1}^{T}(\frac{C_\kappa T^{\rho-1}}{\beta} - \beta)^2, \frac{dT\beta^2}{2}\right\}. \tag{4}$$

Since $d \leq \frac{1}{C_\kappa}T^{1-\rho}$, we can set $\beta = \sqrt{C_\kappa}T^{\frac{\rho-1}{2}}$ to satisfy $\|\theta\|_2 \leq 1$ and to optimize the lower bound in Equation (4). Then we obtain $\mathbb{E}\left\{\sum_{t=1}^{T}\mu_\theta(a^*) - \mu_\theta(a_t)\right\} \geq dC_\kappa T^\rho/4$. Notice that the corruption level each round is bounded by $\beta\|a_t\|_\infty$. In the scenario when the radius of the *action* space $\mathcal{A}$ is upper bounded by β, the corruption budget each round is bounded by $C_\kappa t^{\rho-1}$. It implies that the corruption is induced by ρ-*imperfect* user, completing the proof. ■

5. The Efficiency-Robustness Tradeoff in LIHF

Section 4 shows that a lower bound regret of $\Omega(d\max\{\sqrt{T}, T^\rho\})$ is intrinsic for learning from ρ-imperfect user, even with known ρ. In this section, we study what would be possible — but in a significantly more challenging learning setting with *arbitrary* and *agnostic* corruption. The reason for fighting for such generality is due to practical considerations; in reality, it is rarely possible to truly predict users' irrationality or behaviors. Technique-wise, gradient-based methods are the most commonly used for addressing continuous action spaces, and so are our approaches. The key novelty in our results is not about the development of fundamentally new algorithmic techniques, but rather the conduct of more fine-grained analysis about two popular gradient-based methods — Dueling Bandit Gradient Descent (DBGD (Yue and Joachims, 2009)) and Noisy Comparison-based Stochastic Mirror Descent (NC-SMD) (Kumagai, 2017) — and the proof of how these algorithms can be "tailored" to achieve robustness against corruption. On a conceptual level, our key insight is that one can carefully tune the learning rates of these gradient-based methods to gracefully trade off learning efficiency for adversarial robustness — a interesting concept which we coin the *efficiency-robustness tradeoff*. Notably, en route developing our result, we also pin down a tight lower bound of $\Omega(T^{3/4})$ for the well-known DBGD algorithm which surprisingly remained an open question to date despite extensive studies of this algorithm.

Our finding reveals the intrinsic tradeoff between robustness and efficiency for gradient-based algorithms under agnostic corruption (Theorem 2). In general, the tradeoff is controlled by the exploration size in Algorithm 1 (depends on α) (experimentally verified at

Appendix D.1.4). A higher tolerance of agnostic corruption comes at a cost of worse regret in scenario without corruption, shown by Figure 1.

Algorithm 1: Dueling Bandit Gradient Descent (DBGD)

Input: Exploration rate δ, exploitation rate γ, initial action $a_1 = \mathbf{0} \in \mathcal{A}$.

for $t \in [T]$ **do**

- Sample unit vector u_t uniformly and set $a'_t := \mathcal{P}_{\mathcal{A}}(a_t + \delta u_t)$.
- Present action pair (a_t, a'_t) to user and receive corrupted dueling feedback $\hat{\mathcal{F}}(a'_t, a_t)$.
- Compute gradient $\hat{g}_t := -\frac{d}{\delta}\hat{\mathcal{F}}(a'_t, a_t)u_t$.
- Set step size $\eta := \frac{\gamma\delta}{d}$ and update $a_{t+1} := \mathcal{P}_{\mathcal{A}}(a_t - \eta\hat{g}_t)$.

Theorem 2 (Efficiency-Robustness Tradeoff in DBGD). *For any $\alpha \in (0, \frac{1}{4}]$ and any number of round T satisfying $T > \left(\frac{\sqrt{2Rd}L_\mu L_2}{\sqrt{13L}L_\sigma}\right)^4$, Algorithm 1 parameterized with $\gamma = \frac{R}{\sqrt{T}}$ and $\delta = \frac{\sqrt{2Rd}}{\sqrt{13L}T^\alpha}$ achieves $Reg_T \le O(\sqrt{d}T^{1-\alpha} + \sqrt{d}T^\alpha C)$ for any total corruption level C.*

At the core of this proof is a novel development of a regret decomposition lemma (Lemma 2), where we decompose the regret into two parts: regret of decision making under uncertainty and regret due to feedback imperfection. This decomposition has been proven to be useful for controlling regret of interactive decision-making in reinforcement learning Foster et al. (2023). However, to the best of our knowledge, Lemma 2 is the first to exhibit such a decomposition for dueling feedback, which contains much sparser information than the direct reward feedback as in RL or classic online learning in the aforementioned studies. Hence we believe the lemma may be of independent interest for future research on dueling bandits under corruption or erroneous observation. Our proof employs new techniques and is also more involved. Previous proofs of the decomposition lemma for reward feedback crucially hinges on an online estimation oracle Foster et al. (2023), which however cannot be constructed in our problem with dueling feedback with agnostic corruption. We thus have to develop new proof ideas tailored specifically for dueling bandit feedback. We start from quantifying the bias b_t caused by adversarial corruption in gradient estimation, and then show that we can control the cumulative impact of bias by the exploration size δ, i.e. $\sum_{t=1}^{T} \mathbb{E}\left(b_t^\top(a_t - a^*)\right) \le 2RdL_\sigma C/\delta$. Consequently, Lemma 2 shows that the regret of decision and observation error are jointly controlled by the exploration size δ. We choose δ in order of $\Theta(T^{-\alpha}), \alpha \in (0, 1/4]$ to achieve a balance between optimizing the learning objective (1) and controlling the gradient bias arising from the *corrupted dueling feedback*. Following this, we show Reg_T is upper bounded by $O(\sqrt{d}T^{1-\alpha} + \sqrt{d}T^\alpha C)$.

Lemma 2 (Regret Decomposition). *Let $b_t := \frac{d}{\delta}\left(\hat{\mathbb{P}}(a'_t \succ a_t) - \mathbb{P}(a'_t \succ a_t)\right)u_t, u_t \sim \mathbb{S}$, and choose $\lambda = \frac{L_\sigma}{L_\sigma - \delta L_\mu L_\sigma}$, we have*

$$Reg_T \le \underbrace{\lambda\left(\frac{R^2 d}{\gamma\delta} + \frac{T\gamma d}{\delta} + 13\delta LT\right)}_{\text{Regret of Decision}} + \underbrace{2\lambda\sum_{t=1}^{T}\mathbb{E}\left(b_t^\top(a_t - a^*)\right)}_{\text{Observation Error}}.$$

Note that Algorithm 1 works for general concave user utility. However, if the user displays a strongly concave utility, we can tune learning rate of NC-SMD to get better tradeoff. Specifically, we can get better regret upper bound $\tilde{O}(d\sqrt{T})$ in scenario without corruption at the same time affording $O(T^{3/4})$ agnostic corruption (see Prop 2). The proof of Proposition 2 shares similar structure as the proof Theorem 2, however working with a different algorithm. We direct readers to Appendix C.1 and C.4 to see full statement of Theorem 2, details of the Algorithm NC-SMD and proof of Proposition 2.

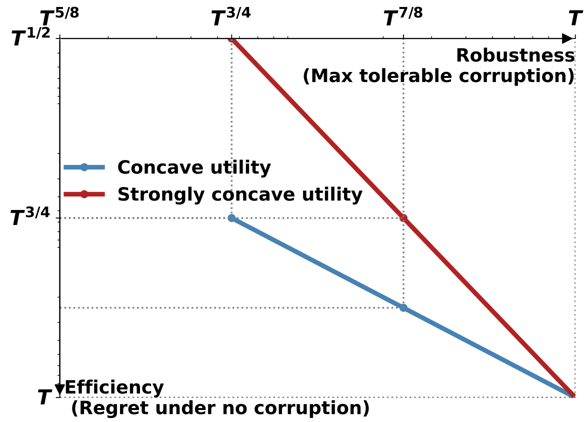

Figure 1: Efficiency-Robustness Tradeoff

Proposition 2 (Efficiency-Robustness Tradeoff in NC-SMD)**.** *For any $\alpha \in [\frac{1}{2}, 1)$, if the utility function is strongly concave, by choosing learning rate $\eta = \frac{\sqrt{\log T}}{2d} T^{-\alpha}$, for the Algorithm NC-SMD, we have $Reg_T \le \tilde{O}(dT^{\alpha} + \sqrt{d}T^{\frac{1}{2}(1-\alpha)}C + dC)$ for any total corruption level C.*

An interesting corollary of Theorem 2 is that we can use it to show the DBGD upper bound proved in (Yue and Joachims, 2009) is tight, which to the best of our knowledge, has not been realized in the previous literature. In the following we provide a sketch of Corollary 1. It has been verified by experiments (see Figure 2, simulation details in Appendix D.1.1). We direct readers to Appendix C.2 for proof details.

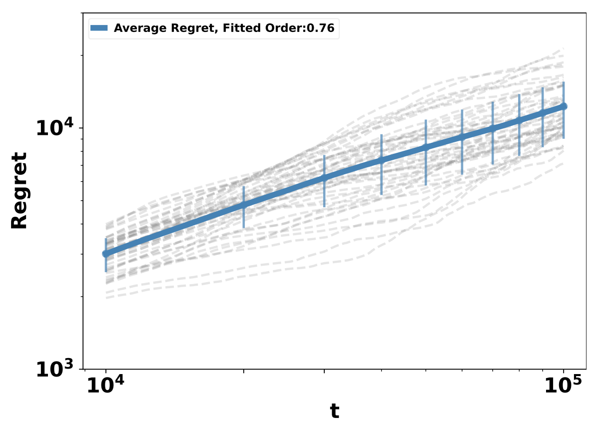

Figure 2: DBGD Lower Bound

Corollary 1 (Minimax Lower Bound for DBGD)**.** *There exists a linear utility function μ and a link function σ such that choosing $\alpha = 1/4$, Reg_T suffered by Algorithm 1 is at least $\Omega\left(T^{3/4}\right)$.*

Proof (Proof Sketch of Corollary 1). Corollary 1 shows that the standard DBGD algorithm with $\alpha = \frac{1}{4}$ can afford $O(T^{3/4})$ agnostic corruption. Moreover, we observe that we can make the sub-optimal arm and the optimal arm indistinguishable by using the total corruption level in the same order of the number of times sub-optimal arm that has been pulled. Therefore, there must exist an problem instance such that the sub-optimal arm has been pulled at least $\Theta(T^{3/4})$ times by DBGD. If not, we can create the parallel world to make DBGD suffer linear regret in one of the environments by using total corruption level less than $\Theta(T^{3/4})$, which contradicts Theorem 2. Intuitively, it says that if an environment is "easy" to attack, the agnostic corruption level which an algorithm could afford can not exceed the worst possible regret suffered by the algorithm in scenario without corruption. ■

6. Experiments

In this section, we study the empirical performance of DBGD and NC-SMD to validate our theoretical analysis by running simulations on both synthetic datasets and real-world recommendation data in comparison with several baseline algorithms. Experimental details can be found at Appendix D.

6.1 Synthetic Data

Experiment Setup. We consider a standard experiment setup, which adopts a strongly concave utility $\mu_\theta(a) := \theta^\top a - \frac{1}{2}\|a\|_2^2$, a logistic link function $\sigma(x) = \frac{1}{1+\exp(-x)}$. We choose $d = 5$, and $T = 10^5$. Our *action* space $\mathcal{A}$ is a d-dimensional ball with radius $R = 10$. The preference parameter θ is randomly sampled from the surface of $\mathcal{A}$. In our problem setting, the optimal action a^* is θ and $\mu_\theta(a^*) = 50$. We simulate *agnostic arbitrary* corruption for $C = T^\rho, \rho \in [0.5, 0.75]$. Specifically, we force the user to submit her least preferred item to the algorithm over the first C rounds (for experiments on corruption induced by ρ-*imperfect* user, see Appendix D.1.3). For each value of ρ, we repeat the experiments for 5 times, each under a different seed. We use Doubler and Sparring proposed by Ailon et al. (2014) as the baseline algorithms for comparison. We set bandit gradient descent (Flaxman et al., 2004) as the black-box bandit algorithm.

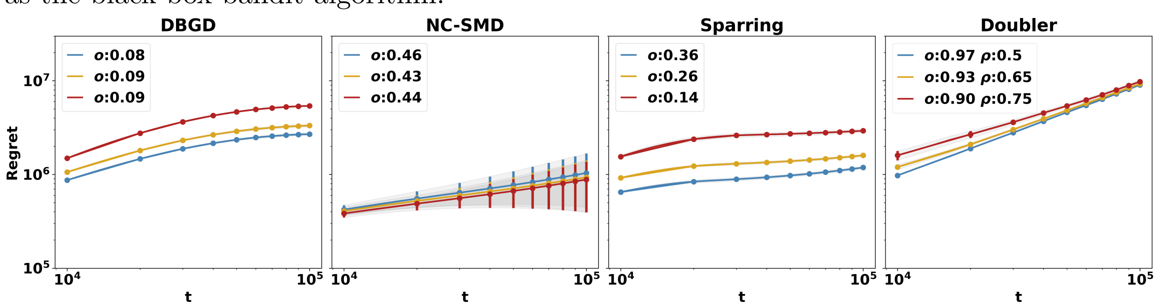

Figure 3: We show the log-log plot for regret occurred by different algorithms under arbitrary agnostic corruption. For each ρ, we present the line plot of average regret over 5 simulations. The shaded region is the average regret $\pm$ one standard deviation. In addition, we fit the order of the average regret, i.e. $\mathbb{E}(\text{Reg}_T) = T^o$, by least squares using last 1% of data.

Results and Discussion. Figure 3 shows that standard DBGD ($\alpha = 1/4$) and NC-SMD ($\alpha = 1/2$) are able to tolerate $O(T^{3/4})$ unknown corruption, aligning with the theoretical result. Experimentally, NC-SMD enjoys the best performance under strongly concave utility. Empirically, Sparring has comparable performance as DBGD, even though it is a heuristic approach with no theoretical guarantee. The worst performance of Doubler might be attributed to the fact that its theoretical guarantee is only applicable to linear link function, which is different from our experiment setup.

6.2 Spotify Recommendation Data

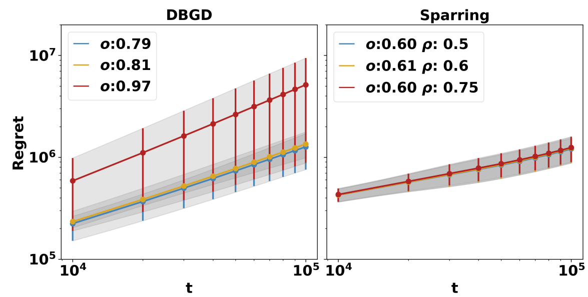

Figure 4: Results on Spotify Recommendation

Evaluation Setup. We also evaluate our approach on Spotify recommendation data (Spotify, 2020). The objective is to recommend songs to incoming users. This dataset includes 17×10^4 songs ($|\mathcal{A}| = 17 \times 10^4$), each is described by 15 distinct features (d = 15). We create 5 different user types by clustering the dataset into 5 groups. We use the average of the song embedding within each group as the user's preference vector. Our aim is to recommend songs to users that have the highest cosine similarity (utility function) with their preference vectors. We rescale the cosine similarity to $[-100, 100]$. We simulate corruption induced by ρ-*imperfect* user to model imperfect human feedback. We only test the performance of DBGD and Sparring because NC-SMD only works under strongly concave utility and Doubler only works under linear link function.

Results and Discussion. Figure 4 show that the performance of DBGD aligns with theoretical prediction on real-world recommendation data, even in the scenario when $\mathcal{A}$ is discrete and nonconvex. Both DBGD and Sparring are robust to agnostic corruption induced by ρ-*imperfect* user, however, empirically it seems that Sparring is invariant to ρ. This might because the song embeddings are highly sparse, which results in the fact that the proposed arms according to the heuristic approach are easy to distinguish, even under agnostic corruption.

Additional Experiments. We also test the performance of Versatile-DB (Saha and Gaillard, 2022), an approach robust to agnostic corruption with theoretical guarantee. However, this approach is not computationally efficient. When $K = 2000$, running 100 iterations takes around 20 CPU hours and it's performance is dominated by DBGD. We direct readers to Appendix D.2.1 to see the details.

7. Conclusion

In this paper, we model learning from *imperfect* human feedback as a corrupted continuous dueling bandit. We highlight the fundamental difficulty of LIHF by developing a regret lower bound. Then we show that by carefully tuning the learning rates of gradient-based methods, we can gracefully trade off learning efficiency for adversarial robustness. Our theoretical predictions also hold on real-world dataset. For future direction, it is interesting to generalize our setting to include contextual information and design algorithms robust to agnostic corruption. In addition, it is also an interesting open problem to design algorithms performs best-of-both-worlds, i.e. performing optimally both with and without agnostic corruption, in continuous action space.

References

Arpit Agarwal, Shivani Agarwal, and Prathamesh Patil. Stochastic dueling bandits with adversarial corruption. In Vitaly Feldman, Katrina Ligett, and Sivan Sabato, editors, *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pages 217–248. PMLR, 16–19 Mar 2021. URL https://proceedings.mlr.press/v132/agarwal21a.html.

Nir Ailon, Thorsten Joachims, and Zohar Karnin. Reducing dueling bandits to cardinal bandits, 2014.

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

Gordon M Becker, Morris H DeGroot, and Jacob Marschak. Stochastic models of choice behavior. *Behavioral science*, 8(1):41–55, 1963.

Viktor Bengs, Róbert Busa-Fekete, Adil El Mesaoudi-Paul, and Eyke Hüllermeier. Preference-based online learning with dueling bandits: A survey. *Journal of Machine Learning Research*, 22(7):1–108, 2021a.

Viktor Bengs, Robert Busa-Fekete, Adil El Mesaoudi-Paul, and Eyke Hüllermeier. Preference-based online learning with dueling bandits: A survey, 2021b.

Ilija Bogunovic, Arpan Losalka, Andreas Krause, and Jonathan Scarlett. Stochastic linear bandits robust to adversarial attacks, 2020.

Ilija Bogunovic, Zihan Li, Andreas Krause, and Jonathan Scarlett. A robust phased elimination algorithm for corruption-tolerant gaussian process bandits, 2022.

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

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*, 2023.

Haokun Chen, Xinyi Dai, Han Cai, Weinan Zhang, Xuejian Wang, Ruiming Tang, Yuzhou Zhang, and Yong Yu. Large-scale interactive recommendation with tree-structured policy gradient. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3312–3320, 2019.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097, 2021.

Pierre-André Chiappori and Jean-Charles Rochet. Revealed preferences and differentiable demand. *Econometrica*, 55(3):687–691, 1987. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1913607.

Jonathan D Cohen, Samuel M McClure, and Angela J Yu. Should i stay or should i go? how the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1481):933–942, 2007.

Yash Deshpande and Andrea Montanari. Linear bandits in high dimension and recommendation systems, 2013.

Qiwei Di, Jiafan He, and Quanquan Gu. Nearly optimal algorithms for contextual dueling bandits from adversarial feedback. *arXiv preprint arXiv:2404.10776*, 2024.

Qin Ding, Cho-Jui Hsieh, and James Sharpnack. Robust stochastic linear contextual bandits under adversarial attacks. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 7111–7123. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/ding22c.html.

Peter C Fishburn, Peter C Fishburn, et al. *Utility theory for decision making*. Krieger NY, 1979.

Abraham D Flaxman, Adam Tauman Kalai, and H Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. *arXiv preprint cs/0408007*, 2004.

Abraham D. Flaxman, Adam Tauman Kalai, and H. Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '05, page 385–394, USA, 2005. Society for Industrial and Applied Mathematics. ISBN 0898715857.

Dylan J. Foster, Sham M. Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making, 2023.

Anupam Gupta, Tomer Koren, and Kunal Talwar. Better algorithms for stochastic bandits with adversarial corruptions. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1562–1578. PMLR, 25–28 Jun 2019. URL https://proceedings.mlr.press/v99/gupta19a.html.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. Parameter-efficient fine-tuning for large models: A comprehensive survey, 2024.

Elad Hazan and Kfir Levy. Bandit convex optimization: Towards tight bounds. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Jiafan He, Dongruo Zhou, Tong Zhang, and Quanquan Gu. Nearly optimal algorithms for linear contextual bandits with adversarial corruptions, 2022.

Nicole Immorlica, Jieming Mao, Aleksandrs Slivkins, and Zhiwei Steven Wu. Incentivizing exploration with selective data disclosure. In *Proceedings of the 21st ACM Conference on Economics and Computation*, pages 647–648, 2020.

Junpei Komiyama, Junya Honda, Hisashi Kashima, and Hiroshi Nakagawa. Regret lower bound and optimal algorithm in dueling bandit problem. In *Conference on learning theory*, pages 1141–1154. PMLR, 2015.

Wataru Kumagai. Regret analysis for continuous dueling bandit. *Advances in Neural Information Processing Systems*, 30, 2017.

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

Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning*, pages 2071–2080. PMLR, 2017.

Thodoris Lykouris, Vahab Mirrokni, and Renato Paes Leme. Stochastic bandits robust to adversarial corruptions, 2018.

Lucas Maystre and Matthias Grossglauser. Just sort it! a simple and effective approach to active preference learning. In *International Conference on Machine Learning*, pages 2344–2353. PMLR, 2017.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Richard A Posner. Rational choice, behavioral economics, and the law. *StAn. l. reV.*, 50: 1551, 1997.

Pearl Pu, Li Chen, and Rong Hu. Evaluating recommender systems from the user's perspective: survey of the state of the art. *User Modeling and User-Adapted Interaction*, 22: 317–355, 2012.

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

Aadirupa Saha and Pierre Gaillard. Versatile dueling bandits: Best-of-both world analyses for learning from relative preferences. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the*

39th International Conference on Machine Learning, volume 162 of *Proceedings of Machine Learning Research*, pages 19011–19026. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/saha22a.html.

Ohad Shamir. On the complexity of bandit and derivative-free stochastic convex optimization, 2013.

Spotify. Music recommendation system using spotify dataset. https://www.kaggle.com/code/vatsalmavani/music-recommendation-system-using-spotify-dataset, 2020.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

Yanan Sui, Masrour Zoghi, Katja Hofmann, and Yisong Yue. Advancements in dueling bandits. In *IJCAI*, pages 5502–5510, 2018.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2018.

Chaoqi Wang, Ziyu Ye, Zhe Feng, Ashwinkumar Badanidiyuru, and Haifeng Xu. Follow-ups also matter: Improving contextual bandits via post-serving contexts, 2023.

WE Williams. A class of integral equations. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 59, pages 589–597. Cambridge University Press, 1963.

Robert C Wilson, Andra Geana, John M White, Elliot A Ludvig, and Jonathan D Cohen. Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of experimental psychology: General*, 143(6):2074, 2014.

Wenjie Xu, Wenbin Wang, Yuning Jiang, Bratislav Svetozarevic, and Colin N Jones. Principled preferential bayesian optimization. *arXiv preprint arXiv:2402.05367*, 2024.

Fan Yao, Chuanhao Li, Denis Nekipelov, Hongning Wang, and Haifeng Xu. Learning from a learning user for optimal recommendations. In *International Conference on Machine Learning*, pages 25382–25406. PMLR, 2022a.

Fan Yao, Chuanhao Li, Denis Nekipelov, Hongning Wang, and Haifeng Xu. Learning the optimal recommendation from explorative users. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9457–9465, 2022b.

Chenlu Ye, Jiafan He, Quanquan Gu, and Tong Zhang. Towards robust model-based reinforcement learning against adversarial corruption, 2024a.

Chenlu Ye, Wei Xiong, Quanquan Gu, and Tong Zhang. Corruption-robust algorithms with uncertainty weighting for nonlinear contextual bandits and markov decision processes, 2024b.

Yisong Yue and Thorsten Joachims. Interactively optimizing information retrieval systems as a dueling bandits problem. In *International Conference on Machine Learning*, 2009. URL https://api.semanticscholar.org/CorpusID:13141692.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019.

Appendix A. Connection between Imperfect User and Generalized Learnability

In this section, we establish the connection between adversarial corruption, corruption induced by generalized learnability, and corruption induced by ρ-*imperfect* user. Broadly speaking, holding total corruption budget C constant, corruption induced by generalized learnability is a specific type of *arbitrary* corruption. Furthermore, if the proposed action pairs exhibit sufficient diversity, corruption induced by ρ-*imperfect* user is equivalent to corruption induced by generalized learnability.

In the following, we introduce the concept of the generalized learnability of the utility function μ in the scenario of pairwise comparison and the associated corruption induced throughout the user's learning process.

Definition A.1 (Corruption induced by Generalized Learnability)**.** *There exists an algorithm $\mathcal{G}$ such that, given any arbitrary t pairs of actions $\{a_s, a'_s\}_{s=1}^t$ and the associated utility differences $\{y_s\}_{s=1}^t$, where $y_s := \mu(a_s) - \mu_(a'_s)$, $\mathcal{G}$ can output an estimated utility function $\mu_t : \mathbb{R}^d \to \mathbb{R}$ of the true utility function $\mu : \mathbb{R}^d \to \mathbb{R}$ such that for any action pair (a, a') the following equation holds*

$$|c_t(a, a')| = |\mu_t(a) - \mu(a) - (\mu_t(a') - \mu(a'))| \le C_0 \left(\|a - a'\|^2_{\bar{V}_t^{-1}} \right)^{1-\rho}.$$

C_0 is some positive constant, $\rho \in [0, 1]$, and $\bar{V}_t := \lambda \mathbb{I}_d + \sum_{s=1}^t (a_s - a'_s)(a_s - a'_s)^\top$, where λ is some positive constant to ensure that $\bar{V}_t$ is full rank.

Def. A.1 is a variant of the generalized learnability assumption proposed by Wang et al. (2023) and Yao et al. (2022a). Essentially, the parameter ρ can be viewed as the user's irrationality level. A higher value of ρ implies lower learning speed of the user's utility function μ, hence a larger magnitude of the total induced corruption. Let's consider *arbitrary* corruption with total corruption budget $C = \Theta(T^\rho)$. Using Lemma A.1, we can show that $c_t(a, a')$ induced by the generalized learnability of μ satisfies $\sum_{t=1}^T |c_t(a_t, a'_t)| \le O(T^\rho)$, implies that it is a specific type of *arbitrary* corruption.

In the scenario when $\lambda_{\min}(\bar{V}_t)$ increases in the order of $\Theta(t)$, $(\|a - a'\|^2_{\bar{V}_t^{-1}})^{1-\rho}$ has the order $\Theta(t^{\rho-1})$, which implies there exists a positive constant C_κ that at round t, given arbitrarily proposed action pairs (a, a'), the magnitude of corruption $|c_t(a, a')| \le C_\kappa t^{\rho-1}$, which means that corruption induced by a ρ-*imperfect* user coincides with corruption induced by the generalized learnability of μ.

We would like to highlight that the condition which $\lambda_{\min}(\bar{V}_t)$ increases in the order of $\Theta(t)$ could be satisfied by gradient-descent based algorithm. Take Algorithm 1 as an example. For simplicity, let's assume a^* is far away from the boundary of $\mathcal{A}$ and projection never happens. It implies that $a'_t = a_t + \delta u_t, \forall t \in [T]$. This assumption is without loss of generality, since most of the proposed action a'_t will belong to the interior of the *action* space $\mathcal{A}$. This is because we start with $a_0 = 0$ and has small exploration size $\delta \sim \Theta(T^{-\alpha})$. Roughly speaking, we can express $\bar{V}_t = \lambda \mathbb{I}_d + \delta^2 \sum_{s=1}^t u_s u_s^\top$, where u_s is uniformly sampled from $\mathbb{S}^d$. Then applying Lemma A.2, we have $\lambda_{\min}(\bar{V}_t) \sim \Theta(t)$ with high probability.

Lemma A.1 (Generalized Elliptical Lemma in Wang et al. (2023))**.** *Suppose $V_0 \in \mathbb{R}^{d\times d}$ is any positive definite matrix, $a_1, \ldots, a_T \in \mathbb{R}^d$ is a sequence of vectors with bounded l_2 norm* *and* $\bar{V}_t := V_0 + \sum_{s=1}^{t} a_s a_s^\top$. *Then for any* $\rho \in [0, 1]$, *the following inequality holds with probability at least* $1 - \delta$

$$\sum_{t=1}^{T} \left(\min \left\{ 1, \|a_t\|^2_{\bar{V}_t^{-1}} \right\} \right)^{1-\rho} \le 2^{1-\rho} T^{\rho} \log^{1-\rho} \left(\frac{\det(\bar{V}_t)}{\det(V_0)} \right).$$

Lemma A.2 (Proposition 1 in Li et al. (2017)). *Define* $V_t := \sum_{s=1}^{t} u_s u_s^\top$, *where* u_s *is drawn iid from some distribution* ν *with support on the unit ball,* $\mathbb{B}^d$. *Furthermore, let* $\Sigma := \mathbb{E}[u_s u_s^\top]$ *be the second moment matrix.* B *and* $\delta > 0$ *are two positive constants. Then, there exist positive, universal constants* C_1 *and* C_2 *such that* $\lambda_{\min}(V_s) \ge B$ *with probability at least* $1 - \delta$, *as long as*

$$t \ge \left(\frac{C_1 \sqrt{d} + C_2 \sqrt{\log(1/\delta)}}{\lambda_{\min}(\Sigma)} \right)^2 + \frac{2B}{\lambda_{\min}(\Sigma)}.$$

Appendix B. Missing Proofs in Theorem 1 and Proposition 1

B.1 Proof for Theorem 1: A Tight Lower Bound

Theorem (Theorem 1 restated). *There exists an LIHF instance with a* ρ*-imperfect user (Def. 1) and strongly concave user utility such that any learner has to suffer* $Reg_T \ge \Omega(d \max\{\sqrt{T}, T^{\rho}\})$ *even* with the knowledge of ρ. *Moreover, this regret lower bound is tight, up to logarithmic terms. That is, for the same setting, there exists an online learning algorithm with* $Reg_T \le \tilde{O}(d \max\{\sqrt{T}, T^{\rho}\})$.

Proof (Proof for Regret Lower Bound in Theorem 1). Before showing the regret lower bound proof for Theorem 1, let's first establish the following Lemma B.1, which builds the connection between regret occurred under bandit reward feedback and dueling feedback. This proof is inspired by Theorem 6 in Yao et al. (2022a). To prove Lemma B.1, in addition to Reg_T, we introduce a new metric to measure the performance of algorithm under dueling feedback, which we coin functional regret Reg_T^{FO}, defined as follows.

$$\text{Reg}_T^{\text{FO}} := \mathbb{E} \left\{ \sum_{t=1}^{T} \mu(a^*) - \mu(a_t) + \mu(a^*) - \mu(a'_t) \right\}. \tag{5}$$

Lemma B.1. *Given a utility function* μ, *if the regret lower bound for algorithm with bandit reward feedback is* $\overline{Reg}$, *then any learner* $\mathcal{L}$ *with dueling feedback has to occur regret* $Reg_T^{FO} \ge 2\overline{Reg}$.

Proof (Proof for Lemma B.1). We prove our claim by contradiction. Let $(a_{0,t}, a_{1,t})$ be the pair of recommendation at round t. Suppose $\text{Reg}_T^{\text{FO}} < 2\overline{\text{Reg}}$. As a result, at least one of the following inequality must hold:

$$\begin{aligned} \mathbb{E} \left\{ \sum_{t=1}^{T} f(a_{0,t}) - f(a^*) \right\} &\le \overline{\text{Reg}}; \\ \mathbb{E} \left\{ \sum_{t=1}^{T} f(a_{1,t}) - f(a^*) \right\} &\le \overline{\text{Reg}}. \end{aligned} \tag{6}$$

Consider a principal who can observe the interaction between a user and the learner $\mathcal{L}$, then Then we can construct two algorithms $\mathcal{L}_0$ and $\mathcal{L}_1$ as follows.

Algorithm B.1: Algorithm $\mathcal{L}_i$

Input: the time horizon T

for $t \leq T$ **do**

- Call algorithm $\mathcal{L}$ to generate two candidates $(a_{0,t}, a_{1,t})$.
- Present $(a_{0,t}, a_{1,t})$ to user and received the feedback.
- Return the feedback $a_{*,t}$ to algorithm $\mathcal{L}$ and update $\mathcal{L}$.

Output: the sequential decision $\{a_{i,t}\}_{t=1}^{T}$

From Equation (6), we know that at least one of $\{\mathcal{L}_0, \mathcal{L}_1\}$ achieves an expected regret lower than $\overline{\text{Reg}}$. However, we know all algorithm with bandit feedback with utility function μ has to occur regret at least $\overline{\text{Reg}}$, contradiction. Therefore, Lemma B.1 must hold, which completes the proof. ■

At the next step, we will prove Theorem 1 by discussing the value of ρ case by case. Consider the scenario when $\rho \leq 0.5$. This is equivalent to say the total corruption level $C \leq \sqrt{T}$. In this scenario, since the utility function is strongly concave, applying Lemma B.2, we know $\overline{\text{Reg}} = 0.02 \min\left\{T, d\sqrt{T}\right\}$. By Lemma B.1, we get $\text{Reg}_T^{\text{FO}} \geq 0.04 \min\left\{T, d\sqrt{T}\right\}$. Consider a linear link function $\sigma(x) = \frac{1}{2} + \frac{1}{2}x$, we have $\text{Reg}_T \geq 0.02 \min\left\{T, d\sqrt{T}\right\} \geq \Omega(d \max\{\sqrt{T}, T^{\rho}\})$, which completes the proof.

Lemma B.2 (Theorem 6 in (Shamir, 2013))**.** *Let the number of rounds T be fixed. The for any learner $\mathcal{L}$, there exists a quadratic function of the form $\mu_\theta(a) = \theta^\top a - \frac{1}{2}\|a\|^2$ which is minimized at θ and $\|\theta\|_2 \leq 0.5$ such that*

$$\mathbb{E}\left(\sum_{t=1}^{T} \mu_\theta(a^*) - \mu_\theta(a_t)\right) \geq 0.02 \min\left\{T, d\sqrt{T}\right\}.$$

Now, let's focus on the scenario when $\rho > 0.5$, which is equivalent to $C > \sqrt{T}$. In essence, we want to extend Lemma B.2 to the scenario in presence of adversarial corruption induced by ρ-*imperfect* user. If we can show the regret lower bound can be generalized to $\Omega(dT^{\rho})$, then applying the same technique which uses Lemma B.1 to connect regret suffered under bandit reward feedback and regret suffered under dueling feedback and choose linear link function to connect Reg_T^{FO} and Reg_T, we will get the desired regret lower bound. The extension of Lemma B.2 is proven at section B.1.1.

B.1.1 Proof for Lemma 1: Regret Lower Bound for Strongly Concave Utility

Lemma (Lemma 1 restated)**.** *Assume that the action space $\mathcal{A}$ is contained in a d-dimensional unit ball. Consider the utility function μ in the form $\mu_\theta(a) = \theta^\top a - \frac{1}{2}\|a\|_2^2$, where $\theta \in \mathbb{R}^d, \|\theta\|_2 \leq 1$ is a random vector. Under corruption induced by ρ-*imperfect *user (Def. 1), for any fixed number of rounds T and $d \leq \frac{1}{C_\kappa} T^{1-\rho}$, there exists a preference parameter θ*

such that for any learner $\mathcal{L}$ under reward feedback, even with the knowledge of ρ, *has to suffer regret* $Reg_T := \mathbb{E}\{\sum_{t=1}^{T} \mu_\theta(a^*) - \mu_\theta(a_t)\} \geq \frac{d}{4} C_\kappa T^\rho$.

Proof (Proof for Lemma 1). To prove Lemma 1, we aim to prove that there exists a corruption strategy such that the regret incurred by any learner $\mathcal{L}$ with *corrupted dueling feedback* is not less than $\Omega(dT^\rho)$. The formulation of such a corruption strategy hinges on the observation that when the utility function μ_θ adopts the quadratic form parameterized by θ, specifically $\mu_\theta(a) = \theta^\top a - \frac{1}{2}\|a\|_2^2$, which is both smooth and strongly concave, then the regret incurred by $\mathcal{L}$ is bounded below by the sum of the Kullback-Leibler (KL) divergence between the distribution of the *corrupted reward* feedback obtained at round t, denoted $\hat{v}_t, t \in [1, T]$, conditioned on different possible values of θ. (see Lemma B.4). Assume that θ is uniformly drawn from $\{-\beta, \beta\}^d$, given an action a_t, conditioned on $\theta_i > 0$, the corrupted reward feedback $\hat{v}_t$ is

$$\hat{v}_t = \mu_\theta(a_t) + c_t(a_t|\theta_i > 0) + \xi_{a_t} = \underbrace{\left(-\frac{1}{2}\|a_t\|^2 + \sum_{j\neq i} \theta_j a_{t,j}\right) + \beta a_{t,i} + c_t(a_t|\theta_i > 0)}_{\mu_1} + \xi_{a_t}.$$

Conditioned on $\theta_i < 0$, the corrupted reward feedback $\hat{v}_t$ is

$$\hat{v}_t = \mu_\theta(a_t) + c_t(a_t|\theta_i < 0) + \xi_{a_t} = \underbrace{\left(-\frac{1}{2}\|a_t\|^2 + \sum_{j\neq i} \theta_j a_{t,j}\right) - \beta a_{t,i} + c_t(a_t|\theta_i < 0)}_{\mu_2} + \xi_{a_t}.$$

We use $c_t(a_t|\theta_i > 0)$ and $c_t(a_t|\theta_i < 0)$ to empathize the fact that the magnitude and sign of corruption can be dependent on θ. ξ_{a_t} follows a standard Gaussian distribution. Therefore, $\hat{v}_t$ in Equation 2 follows $\mathrm{N}(\mu_1, 1)$. $\hat{v}_t$ in Equation 3 follows $\mathrm{N}(\mu_2, 1)$. By using Lemma B.3, we have

$$\mathbb{D}_{t,i} = \mathrm{D}_{\mathrm{KL}}(\mathrm{N}(\mu_1, 1)||\mathrm{N}(\mu_2, 1)) = (\mu_1 - \mu_2)^2.$$

Lemma B.3 (KL Divergence for Normal Distribution)**.** *Let $N(\mu, \sigma^2)$ represent a Gaussian distribution variable with mean μ and variance σ^2. Then*

$$D_{KL}(N(\mu_1, \sigma^2)||N(\mu_2, \sigma^2)) = \frac{(\mu_1 - \mu_2)^2}{2\sigma^2}.$$

To optimize the lower bound in Lemma B.4, the adversary selects $c_t(a_t|\theta_i > 0)$ and $c_t(a_t|\theta_i < 0)$ to minimize the difference between μ_1 and μ_2. Consider the following corruption strategy: when $\theta_i > 0$, set $c_t(a_t|\theta_i > 0) = -\beta a_{t,i}$; when $\theta_i < 0$, set $c_t(a_t|\theta_i < 0) = \beta a_{t,i}$. This strategy ensures that $\mu_1 = \mu_2$.

Next step is to determine the value of β. Notice that the corruption budget is $C_\kappa T^\rho$. To execute the corruption strategy described above, it requires

$$\sum_{t=1}^{T} |c_t(a_t)| \leq \beta \sum_{t=1}^{T} \|a_t\|_\infty \leq C_\kappa T^\rho. \tag{7}$$

Therefore, by exploiting the fact that μ_θ is 1-strongly concave, we establish another lower bound for the regret by

$$\begin{aligned}\mathbb{E}\left\{\sum_{t=1}^{T}\mu_\theta(a^*)-\mu_\theta(a_t)\right\} &\geq \frac{1}{2}\mathbb{E}\left\{\sum_{t=1}^{T}\|a_t-\theta\|_2^2\right\}\\ &\geq \frac{1}{2}\mathbb{E}\left\{\sum_{t=1}^{T}(\|a_t\|_\infty-\beta)^2\right\}\\ &\geq \frac{1}{2}\sum_{t=1}^{T}(C_\kappa T^{\rho-1}/\beta-\beta)^2.\end{aligned}$$

Since $|a_{t,i}-\theta_i|\geq ||a_{t,i}|-\beta|$ for all i, we obtain the second inequality. By utilizing the corruption constraint (7), we derive the last inequality, which is minimized when $\|a_t\|_\infty = \frac{C_\kappa T^{\rho-1}}{\beta}$ for all t. Together with Lemma B.4, we have:

$$\mathbb{E}\left\{\sum_{t=1}^{T}\mu_\theta(a^*)-\mu_\theta(a_t)\right\} \geq \frac{1}{2}\max\left\{\sum_{t=1}^{T}(C_\kappa T^{\rho-1}/\beta-\beta)^2, \frac{dT\beta^2}{2}\right\}. \tag{8}$$

We select β to optimize the lower bound in Equation (8). Since $d\leq \frac{1}{C_\kappa}T^{1-\rho}$, we can set $\beta=\sqrt{C_\kappa}T^{\frac{\rho}{2}-\frac{1}{2}}$ to satisfy $\|\theta\|_2\leq 1$. Then we obtain $\mathbb{E}\left\{\sum_{t=1}^{T}\mu_\theta(a^*)-\mu_\theta(a_t)\right\}\geq dC_\kappa T^\rho/4$. Notice that the corruption level each round is bounded by $\beta\|a_t\|_\infty$. Consider the scenario when the radius of the *action* space $\mathcal{A}$ is upper bounded by β, then the corruption budget each round is bounded by $C_\kappa t^{-1+\rho}$, which satisfies the corruption induced by ρ-*imperfect* user, which completes the proof. ■

Lemma B.4. *Let's consider the utility function* $\mu_\theta(a)=\theta^\top a-\frac{1}{2}\|a\|_2^2$. *Let* $\hat{v}_1,\hat{v}_2,\ldots,\hat{v}_T$ *be a sequence of* corrupted reward feedback *obtained by a learner* $\mathcal{L}$. *Then there exists a* $\theta\in\mathbb{R}^d, \|\theta\|_2\leq 1$, *uniformly drawn from* $\{-\beta,\beta\}^d$, *such that the regret occurred by* $\mathcal{L}$ *is*

$$\mathbb{E}\left\{\sum_{t=1}^{T}\mu_\theta(a^*)-\mu_\theta(a_t)\right\}\geq \frac{dT\beta^2}{4}\left(1-\sqrt{\frac{1}{d}\sum_{i=1}^{d}\sum_{t=1}^{T}\mathbb{D}_{t,i}}\right).$$

$\mathbb{D}_{t,i} := \sup_{\{\theta_j\}_{j\neq i}} D_{KL}\left(\mathbb{P}(\hat{v}_t|\theta_i>0,\{\theta_j\}_{j\neq i},\{\hat{v}_l\}_{l=1}^{t-l})||\mathbb{P}(\hat{v}_t|\theta_i<0,\{\theta_j\}_{j\neq i},\{\hat{v}_l\}_{l=1}^{t-l})\right)$. D_{KL} *is the KL divergence between two distributions.*

Proof (Proof for Lemma B.4). Using the similar argument in Shamir (2013), assume that the leaner $\mathcal{L}$ is deterministic: a_t is a deterministic function of the realized corrupted reward feedback $\hat{v}_1,\hat{v}_2,\ldots,\hat{v}_{t-1}$ at $a_1,a_2,\ldots,a_{t-1}$. This assumption is without loss of generality, since any random learners can be seen as a randomization over deterministic learning algorithms. Thus a lower bound which holds uniformly for any deterministic $\mathcal{L}$ would also hold over a randomization. To lower bound Equation (9), we use Lemma B.5, which relates this to the question of how informative are the query values (as measured by Kullback-Leibler divergence) for determining the sign of θ's coordinates. Intuitively, the more similar the query values are, the smaller is the KL divergence and the harder it is to distinguish the true sign of θ_i, leading to a larger lower bound. In addition, we are facing a powerful adversary who has the complete knowledge of the problem and is able to add corruption on the query value to make they are even more similar, which resulting a even smaller KL divergence, consequently, an even larger lower bound. Let $\bar{a}_T = \frac{1}{T}\sum_{t=1}^{T} a_t$ represent the average action, we have

$$\begin{aligned}
\mathbb{E}\left\{\sum_{t=1}^{T}\mu_\theta(a^*)-\mu_\theta(a_t)\right\} &= T\mathbb{E}\left\{\frac{1}{T}\sum_{t=1}^{T}\mu_\theta(a^*)-\mu_\theta(a_t)\right\} \\
&\geq T\mathbb{E}\left(\mu_\theta(a^*)-\mu_\theta(\bar{a}_T)\right) \\
&\geq T\mathbb{E}\left(\frac{1}{2}\|\bar{a}_T-\theta\|^2\right) \\
&= T\mathbb{E}\left(\frac{1}{2}\sum_{i=1}^{d}(\bar{a}_i-\theta_i)^2\right) \\
&\geq \mathbb{E}\left(\frac{\beta^2 T}{2}\sum_{i=1}^{d}\mathbb{I}_{\bar{a}_i\theta_i<0}\right) \qquad (9)\\
&\geq \frac{dT\beta^2}{4}\left(1-\sqrt{\frac{1}{d}\sum_{i=1}^{d}\sum_{t=1}^{T}\mathbb{D}_{t,i}}\right).
\end{aligned}$$

We get the second inequality by using the fact that μ_θ is 1-strongly concave. We get the last inequality by using Lemma B.5, which completes the proof.

Lemma B.5 (Lemma 4 in (Shamir, 2013))**.** *Let θ be a random vector, none of those coordinates is supported on 0. Let $\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_T$ be a sequence of values obtained by a deterministic learner $\mathcal{L}$ returning a point $\bar{a}_T$ (so that the action a_t is a deterministic function of $\hat{v}_1, \ldots, \hat{v}_{t-1}$ and $\bar{a}_T$ is a deterministic function of $\hat{v}_1, \ldots, \hat{v}_T$). Then we have*

$$\mathbb{E}\left(\sum_{i=1}^{d}\mathbb{I}_{\bar{a}_i\theta_i}\right) \geq \frac{d}{2}\left(1-\sqrt{\frac{1}{d}\sum_{i=1}^{d}\sum_{t=1}^{T}\mathbb{D}_{t,i}}\right),$$

where $\mathbb{D}_{t,i} = \sup_{\{\theta_j\}_{j\neq i}} D_{KL}\left(\mathbb{P}(\hat{v}_t|\theta_i>0,\{\theta_j\}_{j\neq i},\{\hat{v}_l\}_{l=1}^{t-l})||\mathbb{P}(\hat{v}_t|\theta_i<0,\{\theta_j\}_{j\neq i},\{\hat{v}_l\}_{l=1}^{t-l})\right)$, D_{KL} represents the KL divergence between two distributions.

■

We defer reader to Appendix C.3 for the proof of matching regret upper bound and the detailed description of the algorithm which achieves that upper bound. ■

B.2 Proof for Proposition 1: Regret Lower Bound for Linear Utility

Proposition (Proposition 1 restated)**.** *There exists an LIHF instance with a ρ-imperfect user and linear user utility such that any learner knowing ρ has to suffer $Reg_T \geq \Omega(d\max\{\sqrt{T}, T^\rho\})$.*

Proof (Proof for Proposition 1). The proof structure is very similar to the lower bound proof in Theorem 1. We start by discussing the value of ρ. Consider the scenario when $\rho \leq 0.5$. Applying Lemma B.6 and B.1 together with linear link function, we have $\text{Reg}_T \geq \Omega(d \max\{\sqrt{T}, T^\rho\})$, which completes the proof.

Lemma B.6 ((Lattimore and Szepesvári, 2020))**.** *Let $\mathcal{A} = [-1,1]^d$ and $\Theta = [-T^{-\frac{1}{2}}, T^{-\frac{1}{2}}]^d$. Consider the linear reward function $r_t = \theta^\top A_t + \epsilon_t$, ϵ_t is independent Gaussian noise with mean 0 variance 1. Then for any learner $\mathcal{L}$, there exists a vector $\theta \in \Theta$ such that*

$$Reg_T(\mathcal{A}, \theta) \geq \frac{\exp(-2)}{8} d\sqrt{T}.$$

Now, let's focus on the scenario when $\rho > 0.5$ and the essence is to extend Lemma B.6 to the scenario in presence of adversarial corruption induced by ρ-*imperfect* user, which is shown in Lemma B.7

Lemma B.7. *Assume that the* action *space $\mathcal{A} \subset [-1,1]^d$. Given corruption induced by ρ-*imperfect *user (Def. 1), for stochastic linear bandit, there exists a $\theta \in [-C_\kappa T^{\rho-1}, C_\kappa T^{\rho-1}]^d$ such that any learner $\mathcal{L}$ even with the knowledge of ρ has to suffer regret no less than $\frac{d}{8} C_\kappa T^\rho$.*

Proof (Proof for Lemma B.7). The proof extends Lemma B.6 to a scenario in presence of corruption. We want to construct a parameter family and a corruption strategy such that for all algorithm $\mathcal{L}$, it will occur at least $\Omega(dT^\rho)$ regret. Consider the action set $\mathcal{A} \in [-1,1]^d$ and $\Theta = \{-\beta, \beta\}^d$. For any learner $\mathcal{L}$, we can lower bound its regret by

$$\begin{aligned}
\text{Reg}_T(\mathcal{A}, \theta) &= \mathbb{E}_\theta \left[\sum_{t=1}^T \sum_{i=1}^d (\text{sign}(\theta_i) - a_{ti}) \theta_i \right] \\
&\geq \beta \sum_{i=1}^d \mathbb{E}_\theta \left[\sum_{t=1}^T \mathbb{I}\{\text{sign}(a_{ti}) \neq \text{sign}(\theta_i)\} \right] \\
&\geq \frac{T\beta}{2} \sum_{i=1}^d \mathbb{P}_\theta \left(\sum_{t=1}^T \mathbb{I}\{\text{sign}(a_{ti}) \neq \text{sign}(\theta_i)\} \geq \frac{T}{2} \right)
\end{aligned}$$

Let's denote

$$p_{\theta_i} = \mathbb{P}_\theta \left(\sum_{t=1}^T \mathbb{I}\{\text{sign}(a_{ti}) \neq \text{sign}(\theta_i)\} \geq \frac{T}{2} \right)$$

Let $i \in [d]$ and $\theta \in \Theta$ be fixed, and let $\theta'_j = \theta_j$, for $j \neq i$ and $\theta'_i = -\theta_i$. Then using Lemma B.8, we have

$$p_{\theta_i} + p_{\theta'_i} \geq \frac{1}{2} \exp(-\mathbb{D}(\mathbb{P}_\theta, \mathbb{P}_{\theta'})) = \frac{1}{2} \exp\left(-\mathbb{E}\left[\sum_{t=1}^T \text{D}_{\text{KL}}(P_{a_t}, P'_{a_t}) \right] \right).$$

P_{a_t} is the distribution of corrupted reward observed by $\mathcal{L}$ after playing action a_t when reward parameter is θ. Similarly, P'_{a_t} is the distribution of corrupted observed by $\mathcal{L}$ after playing action a_t when the reward parameter is θ'.

Lemma B.8 (Bretagnolle-Huber inequality)**.** *Let $\mathbb{P}$ and $\mathbb{Q}$ be probability measures on the same measurable space $(\Omega, \mathcal{F})$. Let $A \in \mathcal{F}$ be any arbitrary event and A^c is the complement of A. Then we have*

$$\mathbb{P}(A) + \mathbb{Q}(A^c) \geq \frac{1}{2} \exp\left(-D_{KL}(\mathbb{P}, \mathbb{Q})\right). \tag{10}$$

In presence of adversarial corruption, for θ, the corrupted reward is

$$\hat{r}_t = \sum_{j \neq i} a_{tj}\theta_j + a_{ti}\theta_i + c_t(a_t|\theta) + \epsilon_t. \tag{11}$$

For θ', we have

$$\hat{r}_t = \sum_{j \neq i} a_{tj}\theta_j - a_{ti}\theta_i + c_t(a_t|\theta') + \epsilon_t. \tag{12}$$

Consider the corruption strategy such that $c_t(a_t|\theta) = -a_{t,i}\theta_i$ in (11) and $c_t(a_t|\theta') = a_{t,i}\theta_i$ in (12). This is achievable since we assume the adversary has complete knowledge of the problem instance. By doing so, the corrupted reward $\hat{r}_t$ are from the same distribution regardless whether the preference parameter is θ or θ'. This is because θ and θ' only differs in the i-th coordinate with the magnitude β, and this difference could be masked by the corruption, resulting in $\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{D}(P_{A_t}, P'_{A_t})\right] = 0$, which makes it indistinguishable by $\mathcal{L}$. Notice that executing such a corruption strategy requires total corruption budget

$$\sum_{t=1}^{T} |c_t(a_t)| \leq \beta \sum_{t=1}^{T} \|a_t\|_\infty \leq C_\kappa T^\rho. \tag{13}$$

Choose $\beta = C_\kappa T^{\rho-1}$, Equation (13) holds. Therefore, we have

$$p_{\theta_i} + p_{\theta'_i} \geq \frac{1}{2}.$$

Applying an "averaging hammer" over all $\theta \in \Theta$, which satisfies $|\Theta| = 2^d$, we get

$$\sum_{\theta \in \Theta} \frac{1}{|\Theta|} \sum_{i=1}^{d} p_{\theta_i} = \frac{1}{|\Theta|} \sum_{i=1}^{d} \sum_{\theta \in \Theta} p_{\theta_i} \geq \frac{d}{4}.$$

This implies that there exists a $\theta \in \Theta$ such that $\sum_{i=1}^{d} p_{\theta_i} \geq \frac{d}{4}$. Therefore we have

$$\mathrm{Reg}_T(\mathcal{A}, \theta) \geq \frac{dC_\kappa}{8} T^\rho,$$

which completes the proof. We want to highlight that the corruption level each round $|c_t(a_t)|$ is bounded by $\beta\|a_t\|_\infty \leq C_\kappa T^{\rho-1} \leq C_\kappa t^{\rho-1}, \forall t$, satisfying the corruption induced by *ρ-imperfect* user. ∎

∎

Appendix C. Missing Proofs in Theorem 2 and Proposition 2

C.1 Proof for Theorem 2: Efficiency-Robustness Tradeoff in DBGD

In this section, we present the regret upper for dueling bandit gradient descent Yue and Joachims (2009) in presence of *arbitrary* and *agnostic* corruption. Before presenting the proof, we make several remarks.

First, in Yue and Joachims (2009), the utility function μ is assumed to be strictly concave to ensure the uniqueness of the maximizer. However, in scenarios where the maximizer is unique within the *action* set for a concave utility function μ, the requirement for "strictness" can be relaxed.

Second, we highlight the notation differences. We use a to denote actions while Yue and Joachims (2009) uses w. In our proof, we define $P_t(a) := \sigma\left(\mu(a_t) - \mu(a)\right)$ to represent the probability of the event $a_t \succ a$. While in Yue and Joachims (2009), it is denoted as $c_t(w)$. Just to highlight, its corrupted version $\hat{P}_t(a) := \sigma\left(\mu(a_t) - \mu(a) + c_t(a_t, a)\right)$ is newly introduced in our paper. Moreover, $\bar{P}_t(a) := \mathbb{E}_{x\in\mathbb{B}}\left[P_t(\mathcal{P}_\mathcal{A}(a+\delta x))\right]$ is denoted as $\hat{\epsilon}_t(w) + \frac{1}{2}$ in Yue and Joachims (2009). The definition of $\mathcal{F}\left(\mathcal{P}_\mathcal{A}\left(a_t + \delta u_t\right), a_t\right)$ is same as $X_t\left(\mathcal{P}_W(w_t + \delta u_t)\right)$ defined in Yue and Joachims (2009), except different notation.

Theorem (Theorem 2 restated)**.** *For any $\alpha \in (0, \frac{1}{4}]$ and any number of round T satisfying $T > \left(\frac{\sqrt{2Rd}L_\mu L_2}{\sqrt{13L}L_\sigma}\right)^4$, Algorithm 1 parameterized with $\gamma = \frac{R}{\sqrt{T}}$ and $\delta = \frac{\sqrt{2Rd}}{\sqrt{13L}T^\alpha}$ achieves $Reg_T \leq O(\sqrt{d}T^{1-\alpha} + \sqrt{d}T^\alpha C)$ for any total corruption level C.*

Proof (Proof for Theorem 2). The proof of Theorem 2 builds on the proof of Yue and Joachims (2009), where we extend it to a setting in presence of agnostic corruption. The cornerstone of the theoretical analysis relies on developing the regret decomposition lemma (Lemma 2), where we novelly decompose the regret into two parts: regret of decision and observation error. This decomposition is achieved through quantifying the bias, b_t, caused by adversarial corruption, in gradient estimation (Lemma C.1). We further use Lemma C.5 to upper bound the observation error in terms of the exploration size δ. Lemma 2 and Lemma C.5 reveal that the regret of decision and observation error are jointly controlled by the exploration size δ. We choose δ in order of $O(T^{-\alpha})$ to achieve a balance between optimizing the learning objective (1) and controlling the gradient bias arising from the *corrupted dueling feedback*. Consequently, we have Reg_T is upper bounded by $O(\sqrt{d}T^{1-\alpha} + \sqrt{d}T^\alpha C)$.

The development of the proof is based on the observation that we decompose bias in the gradient caused by corruption. Notice that without corruption, DBGD performs expected gradient ascent, where the gradient g_t is defined as follows

$$g_t = -\frac{d}{\delta}\mathcal{F}\left(\mathcal{P}_\mathcal{A}\left(a_t + \delta u_t\right), a_t\right) u_t.$$

In the presence of adversarial corruption, the corrupted gradient is

$$\hat{g}_t = -\frac{d}{\delta}\hat{\mathcal{F}}\left(\mathcal{P}_\mathcal{A}\left(a_t + \delta u_t\right), a_t\right) u_t.$$

Therefore, we quantify the bias in the following lemma.

C.1.1 Proof for Lemma C.1: Gradient Estimation Under Corruption

Lemma C.1. *Let $P_t(a)$ represent the likelihood of action a_t being preferred over a, where $P_t(a) := \sigma(\mu(a_t) - \mu(a))$. Likewise, the corrupted version is defined as $\hat{P}_t(a) = \sigma(\mu(a_t) - \mu(a) + c_t(a_t, a))$. The smoothed version of P_t over $\mathcal{A}$ is $\bar{P}_t(a) := \mathbb{E}_{x\in\mathbb{B}}[P_t(\mathcal{P}_{\mathcal{A}}(a + \delta x))]$. Let $a'_t = \mathcal{P}_{\mathcal{A}}(a_t + \delta u_t)$, where u_t is uniformly sampled from $\mathbb{S}$ and $\hat{g}_t = -\frac{d}{\delta}\hat{\mathcal{F}}(a'_t, a_t)\, u_t$. Let $b_t := \frac{d}{\delta}\left(\hat{P}_t(a'_t) - P_t(a'_t)\right) u_t$. We have $\mathbb{E}(\hat{g}_t|a_t) = \mathbb{E}(\nabla\bar{P}_t(a_t)|a_t) - \mathbb{E}(b_t|a_t)$.*

Proof (Proof for Lemma C.1). Since $a'_t := \mathcal{P}_{\mathcal{A}}(a_t + \delta u_t)$, we have

$$
\begin{aligned}
\mathbb{E}(\hat{g}_t|a_t) &= \mathbb{E}_{u_t}(\mathbb{E}(\hat{g}_t|a_t, u_t)) \\
&= -\frac{d}{\delta}\mathbb{E}_{u_t}\left(\mathbb{E}(\hat{\mathcal{F}}(a'_t, a_t)u_t|a_t, u_t)\right) \\
&= -\frac{d}{\delta}\mathbb{E}\left(\hat{P}_t(a'_t)u_t|a_t\right) \\
&= -\frac{d}{\delta}\mathbb{E}\left(\left[P_t(a'_t) + \hat{P}_t(a'_t) - P_t(a'_t)\right]u_t|a_t\right) \\
&= \nabla\mathbb{E}_{x\in\mathbb{B}}\left(P_t(\mathcal{P}_{\mathcal{A}}(a_t + \delta x))|a_t\right) - \frac{d}{\delta}\mathbb{E}\left[\left(\hat{P}_t(a'_t) - P_t(a'_t)\right)u_t|a_t\right] \qquad (14) \\
&= \nabla\bar{P}_t(a_t) - \frac{d}{\delta}\mathbb{E}\left[\left(\hat{P}_t(a'_t) - P_t(a'_t)\right)u_t|a_t\right] \\
&= \mathbb{E}(g_t|a_t) - \frac{d}{\delta}\mathbb{E}\left[\left(\hat{P}_t(a'_t) - P_t(a'_t)\right)u_t|a_t\right]. \qquad (15)
\end{aligned}
$$

We get the equation (14) by using Lemma C.2. We get equation (15) by using Lemma C.3.

Lemma C.2 (Lemma 2 in Yue and Joachims (2009)). *Fix $\delta > 0$, over random unit vector u, we have*

$$\mathbb{E}[P_t(\mathcal{P}_{\mathcal{A}}(a + \delta u))u] = \frac{\delta}{d}\nabla\bar{P}_t(a).$$

Lemma C.3 (Lemma 1 in Yue and Joachims (2009)). $\mathbb{E}_{\mathcal{F},u}[\mathcal{F}(a'_t, a_t)u] = -\mathbb{E}_u[P_t(a'_t)u]$.

■

C.1.2 Proof for Lemma 2: Regret Decomposition

Lemma (Lemma 2 restated). *Choose $\lambda = \frac{L_\sigma}{L_\sigma - \delta L_\mu L_\sigma}, \gamma = R/\sqrt{T}$, for b_t defined in Lemma C.1, we have*

$$Reg_T^{DB} \le \underbrace{\lambda\left(\frac{2Rd\sqrt{T}}{\delta} + 13\delta LT\right)}_{\textit{Regret of Decision}} + \underbrace{2\lambda\sum_{t=1}^{T}\mathbb{E}\left(b_t^\top(a_t - a^*)\right)}_{\textit{Observation Error}}.$$

Proof (Proof for Lemma 2). Because of Lemma C.4, we have

$$\mathbb{E}\left[\sum_{t=1}^{T}\bar{P}_t(a_t)-\bar{P}_t(a^*)\right]\le\sum_{t=1}^{T}\mathbb{E}\left(\lambda\nabla\bar{P}_t(a_t)(a_t-a^*)+(3+\lambda)\delta L\right)$$
$$=\lambda\sum_{t=1}^{T}\mathbb{E}(\mathbb{E}(g_t|a_t)^\top(a_t-a^*))+(3+\lambda)\delta LT$$
$$=\lambda\sum_{t=1}^{T}\mathbb{E}(\mathbb{E}(\hat{g}_t+b_t|a_t)^\top(a_t-a^*))+(3+\lambda)\delta LT$$
$$=\lambda\sum_{t=1}^{T}\mathbb{E}(\hat{g}_t^\top(a_t-a^*))+\lambda\sum_{t=1}^{T}\mathbb{E}(b_t^\top(a_t-a^*))+(3+\lambda)\delta LT$$

Lemma C.4 (Lemma 4 in Yue and Joachims (2009))**.** *Fix $\delta\in(0,\frac{L_\sigma}{L_\nu L_\sigma})$ and define $\lambda=\frac{L_\sigma}{L_\sigma-\delta L_\nu L_2}$, we have*

$$\mathbb{E}\left[\sum_{t=1}^{T}\bar{P}_t(a_t)-\bar{P}_t(a^*)\right]\le\sum_{t=1}^{T}\mathbb{E}\left(\lambda\nabla\bar{P}_t(a_t)^\top(a_t-a^*)+(3+\lambda)\delta L\right).$$

Since $a_{t+1}=\mathcal{P}_{\mathcal{A}}(a_t-\eta\hat{g}_t)$, and $\eta=\frac{\gamma\delta}{d}$, by applying the telescoping sum and because $a_1=0$, $\|g_t\|_2\le G,\forall t$, we have

$$\lambda\sum_{t=1}^{T}\mathbb{E}\left(\hat{g}_t^\top(a_t-a^*)\right)\le\lambda\left(\frac{R^2}{2\eta}+\frac{T\eta G^2}{2}\right).$$

Noticing that $\|b_t\|_2\le\frac{d}{\delta}\min\left(2,L_\sigma|c_t(a_t,a_t')|\right)$. Moreover, $\|a_t-a^*\|_2\le 2R$. Therefore, we can control the observation error by the following.

C.1.3 Proof for Lemma C.5: Observation Error

Lemma C.5. $\sum_{t=1}^{T}\mathbb{E}\left(b_t^\top(a_t-a^*)\right)\le 2RdL_\sigma C/\delta$.

Proof (Proof for Lemma C.5).

$$\sum_{t=1}^{T}\mathbb{E}(b_t(a_t-a^*))\le 2R\frac{d}{\delta}\sum_{t=1}^{T}\min\left(2,L_\sigma|c_t(a_t,a_t')|\right)$$
$$\le 2R\frac{dL_\sigma}{\delta}\sum_{t=1}^{T}|c_t(a_t,a_t')|$$
$$=2R\frac{dL_\sigma C}{\delta}.$$

■

■

Lemma C.6 (Lemma 3 in Yue and Joachims (2009)). *$Reg_T \le 2\mathbb{E}\left[\sum_{t=1}^T \bar{P}_t(a_t) - \bar{P}_t(a^*)\right] + 5\delta LT$*

Set $G = \frac{d}{\delta}$, and by Lemma C.6, we have

$$\text{Reg}_T \le \lambda\left(\frac{2Rd\sqrt{T}}{\delta} + 13\delta LT\right) + 4R\frac{\lambda d L_\sigma C}{\delta}.$$

Therefore, by setting $\delta = \frac{\sqrt{2Rd}}{\sqrt{13LT^\alpha}}, \gamma = \frac{R}{\sqrt{T}}$, and $T > \left(\frac{\sqrt{2Rd}L_v L_2}{\sqrt{13L}L_\sigma}\right)^4$, we have

$$\text{Reg}_T \le 2\lambda_T\sqrt{26RdLT}T^{1-\alpha} + 2L_\sigma\sqrt{26RdLT}T^\alpha C.$$

$\lambda_T = \frac{L_\sigma\sqrt{13LT}^\alpha}{L_\sigma\sqrt{13LT}^\alpha - L_v L_2\sqrt{2Rd}}$. ■

C.2 Proof for Corollary 1: Minimax Lower Bound for DBGD

Theorem 2 implies that DBGD could afford agnostic corruption level $O(T^{3/4})$. This implies that there exists a hard instance which makes DBGD attain regret in order of $\Omega(T^{3/4})$, formally described as follows.

Corollary (Corollary 1 restated). *There exists a linear utility function μ and a link function σ such that choosing $\alpha = 1/4$, Reg_T suffered by Algorithm 1 is at least $\Omega\left(T^{3/4}\right)$.*

Proof (Proof for Corollary 1). To start a proof by contradiction, we assume that Theorem 2 is false. Specifically, for all problem instance μ, σ, there exist a $\epsilon > 0$ such that $\text{Reg}_T \le O(T^{3/4-\epsilon})$. We assume ϵ to be the least possible. In particular, there exists a pair of instance μ, σ such that $\text{Reg}_T = c_0 T^{3/4-\epsilon}$, c_0 is a positive constant. In other words, it says that there exists a pair of instance μ, σ such that DBGD suffers regret in $\Theta(T^{3/4-\epsilon})$.

Consider the following problem instance. $\mu(a) = \theta^\top a$, specifically μ is linear function. Let $d = 2$ with action set $\mathcal{A}_1 := \{(a_1, a_2) : a_1 \ge 0, a_2 \ge 0, \frac{1}{2}a_1 + a_2 - \frac{1}{4} \le 0\}$ with $\theta = [\frac{1}{2}, \frac{1}{2}]$. The optimal arm is $a_1 = [\frac{1}{2}, 0]$, which is unique. Let the link function $\sigma(x) = \frac{1}{2} + \frac{1}{2}x$. It is easy to verify that it is rotational symmetric and Lipschitz. Denote R_T as the regret occurred by DBGD on this problem instance. We know $R_T \le O(T^{\frac{3}{4}-\epsilon})$. Consequently, it implies that a_2 is at most proposed $8c_0T^{\frac{3}{4}-\epsilon}$ times. This is because $R_T \le c_0T^{\frac{3}{4}-\epsilon}$ and if the proposed action pair (a_t, a'_t) including a_2 at round t, it occurs regret at least $\frac{1}{8}$.

Now consider a different problem instance with the same μ, σ but different action set, which is $\mathcal{A}_2 := \{(a_1, a_2) : a_1 \ge 0, a_2 \ge 0, \frac{3}{2}a_1 + a_2 - \frac{3}{4} \le 0\}$. The optimal arm in this action set $a_2 = [0, \frac{3}{4}]$, which is also unique. Proposing arm $a_1 = [\frac{1}{2}, 0]$ occurs at least $\frac{1}{8}$ regret. Consider an adversary that pulls the utility of a_2 from $\frac{3}{8}$ down to $\frac{1}{8}$ whenever it is pulled at a cost of $c_t(A_t) = \frac{1}{4}$. It is equivalent to say that every time when the proposed arm w falls in the region $\mathcal{W} = \{a_1 \ge 0, a_2 \ge 0, \frac{1}{2}a_1 + a_2 - \frac{1}{4} \ge 0, \frac{3}{2}a_1 + a_2 - \frac{3}{4} \le 0\}$, we assume that the utility is sampled from $\theta^\top\mathcal{P}_{\mathcal{A}_1}(w)$ and $|c_t(w)| = |\theta^\top\mathcal{P}_{\mathcal{A}_1}(w) - \theta^\top w| \le \frac{1}{4}$. Choosing $C = 2c_0T^{\frac{3}{4}-\epsilon}$, the adversary can afford to corrupt $8c_0T^{\frac{3}{4}-\epsilon}$ times.

This means that as long as the adversary is corrupting, the utility observed in the first problem instance is exactly the same as the utility observed in the second problem instance, in which we pull a_2 as most $8c_0T^{\frac{3}{4}-\epsilon}$. However, a_2 is the optimal arm in the second problem instance, which implies that the regret occurred on the second problem instance is at least $T - 8c_0T^{\frac{3}{4}-\epsilon}$, which is linear in T. This contradicts Theorem 2, which says that when agnostic corruption $C = 2c_0T^{\frac{3}{4}-\epsilon}$, the regret upper bound is $O(T^{1-\epsilon})$, which is sublinear. To reconcile the conflicts, we must have Proposition 1 true to make the agnostic corruption $C \geq \Omega(T^{\frac{3}{4}})$ to make the parallel world argument valid. ∎

C.3 Proof for Matching Regret Upper Bound in Theorem 1

In this section, we consider the setting when human has strongly concave utility function. When ρ is known, we optimally adjusts the learning rate of the algorithm, Noisy Comparision-based Stochastic Mirror Descent (NC-SMD), proposed in Kumagai (2017), to achieve a balance between optimizing the learning objective (1) and controlling the gradient bias arising from the *corrupted dueling feedback*. Then we show that it attains the matching regret upper bound of $\tilde{O}(d \max\{\sqrt{T}, T^{\rho}\})$ under corruption induced by ρ-*imperfect* user.

For the ease of delivering proof, in this section, we work with the cost function f, opposite of the utility function μ, We restate the modelling of inaccuracies from human feedback and the learning objective using f as follows.

Assumption C.1. ***(Corrupted Utility-Based Dueling Bandit).*** *For a pair of actions $(a, a') \in \mathcal{A} \times \mathcal{A}$ and the corruption level $c(a, a')$, there exists function $f : \mathcal{A} \to \mathbb{R}$ and $\sigma : \mathbb{R} \to [0, 1]$ such that the probability that $a \succ a'$ happens of receiving* corrupted dueling feedback *can be represented follows*

$$\hat{\mathbb{P}}(a \succ a') = \sigma(f(a') - f(a) + c(a, a)).$$

Additionally, we assume that the cost function f, the link function σ, and the action space $\mathcal{A}$ exhibit the following properties, which align precisely with those assumed in Kumagai (2017).

Assumption C.2. *The cost function $f : \mathcal{A} \to \mathbb{R}$ is twice continuously differentiable, L-Lipschitz, α-strongly convex and β-smooth with respect to the Euclidean norm.*

Definition C.1. ***(Strong Convexity).*** *A function $f : \mathbb{R}^d \to \mathbb{R}$ is α-strongly convex over the set $\mathcal{A} \subset \mathbb{R}^d$ if for all $x, y \in \mathcal{A}$ it holds that*

$$f(y) \geq f(x) + \nabla f(x)^{\top}(y - x) + \frac{\alpha}{2}\|y - x\|_2^2.$$

Definition C.2. ***(Smoothness).*** *A function $f : \mathbb{R}^d \to \mathbb{R}$ is β-smooth over the set $\mathcal{A} \subset \mathbb{R}^d$ if for all $x, y \in \mathcal{A}$ it holds that*

$$f(y) \leq f(x) + \nabla f(x)^{\top}(y - x) + \frac{\beta}{2}\|y - x\|_2^2.$$

Assumption C.3. *Let $B := \sup_{a,a' \in \mathcal{A}} f(a') - f(a)$, f is defined in Assumption C.2. The link function $\sigma : \mathbb{R} \to [0, 1]$ is three times differentiable and rotational-symmetric. Its first derivative is positive and monotonically non-increasing on $[0, B]$.*

Assumption C.4. *We assume the action space $\mathcal{A} \in \mathbb{R}^d$ is convex, compact and has non-empty interior.*

Definition C.3. ***(Self-concordance.)*** *A function $\mathcal{R} : int(\mathcal{A}) \to \mathbb{R}$ is self-concordant if it satisfies*

1. *$\mathcal{R}$ is three times continuously differentiable, convex, and approaches infinity along any sequence of points approaching the boundary of $int(\mathcal{A})$.*
2. *For every $h \in \mathbb{R}^d$ and $x \in int(\mathcal{A})$, $|\nabla^3 \mathcal{R}(x)[h,h,h]| \leq 2\left(h^\top \nabla^2 \mathcal{R}(x) h\right)^{\frac{3}{2}}$ holds, where $|\nabla^3 \mathcal{R}(x)[h,h,h] := \frac{\partial^3 \mathcal{R}}{\partial t_1 \partial t_2 \partial t_3}(x + t_1 h + t_2 h + t_3 h)|_{t_1=t_2=t_3=0}$.*

In addition to these two conditions, if for every $h \in \mathbb{R}^d$ and $x \in int(\mathcal{A})$, $|\nabla \mathcal{R}(x)^\top h \leq \nu^{\frac{1}{2}} (h^\top \nabla^2 \mathcal{R}(x) h)^{\frac{1}{2}}|$ for a positive real number ν, $\mathcal{R}$ is ν-self-concordant.

Through interactions with the user, the learner $\mathcal{L}$ endeavors to minimize the dueling regret, as defined below.

$$\mathrm{Reg}_T^{\mathrm{DB}} = \mathbb{E}\left(\sum_{t=1}^{T} \sigma(f(a_t) - f(a^*)) + \sigma(f(a_t') - f(a^*)) - 1\right). \tag{16}$$

Our minimization objective in (16) aligns with the regret definition in Yue and Joachims (2009). Under Assumptions C.1, C.2, and C.3, the regret defined in (16) is equivalent to the regret defined in Kumagai (2017). Assumption C.2 suggests the existence of a unique minimizer a^* for the cost function f due to its strict convexity. Assumptions C.2 and C.4 imply that the diameter $R := \sup_{a,a' \in \mathcal{A}} |a - a'|$ and B are finite. Assumption C.3 implies the existence of positive constants l_0, L_0, B_2, and L_2 such that the first derivative σ' of the link function is bounded, satisfying $l_0 \leq \sigma' \leq L_0$ on $[-B, B]$. Furthermore, the second derivative σ'' is bounded above by B_2 and is L_2-Lipschitz on $[-B, B]$. Lemma C.7 establishes the relationship between $\mathrm{Reg}_T^{\mathrm{FO}}$ and $\mathrm{Reg}_T^{\mathrm{DB}}$.

Lemma C.7 (Lemma 12 in Kumagai (2017))**.**

$$\frac{Reg_T^{DB}}{L_0} \leq Reg_T^{FO} \leq \frac{Reg_T^{DB}}{l_0}. \tag{17}$$

C.3.1 Noisy Comparison-based Stochastic Mirror Descent

Algorithm C.1: Noisy Comparison-based Stochastic Mirror Descent (NC-SMD)

Input: Learning rate η_ρ, ν-self-concordant function $\mathcal{R}$, time horizon T, tuning parameters λ, μ.

Initialize $a_1 = \arg\min_{a \in A} \mathcal{R}(a)$

for $t \in [T]$ **do**

- Update $\mathcal{R}_t(a) = \mathcal{R}(a) + \frac{\lambda \eta_\rho}{2} \sum_{i=1}^{t} \|a - a_i\|^2 + \mu \|a\|^2$
- Pick a unit vector u_t uniformly at random
- Receive the corrupted feedback $\hat{\mathcal{F}}(a_t', a_t)$, for a_t and $a_t' = a_t + \nabla^2 \mathcal{R}_t(a_t)^{-1/2} u_t$
- Compute the corrupted gradient: $\hat{g}_t = \hat{\mathcal{F}}(a_t', a_t) d \nabla^2 \mathcal{R}_t(a_t)^{1/2} u_t$
- Set $a_{t+1} = \nabla \mathcal{R}_t^{-1}(\nabla \mathcal{R}_t(a_t) - \eta_\rho \hat{g}_t)$

Output: a_{T+1}

Remark C.1. *Note that $a_t' = a_t + \nabla^2\mathcal{R}_t(a_t)^{-\frac{1}{2}}x$ for $x \in \mathbb{B}$ is included in $\mathcal{A}$ due to the properties of the Dikin ellipsoid.*

Now we begin to prove the matching regret upper bound in Theorem 1.

Proposition C.1 (Full Statement of Matching Regret Upper Bound in Theorem 1)**.** *Let the number of rounds T be fixed and $T > \sqrt{2LL_0R}$. Consider a corruption induced by ρ-imperfect user (Def. 1) with known ρ. Under Assumptions C.2, C.3, and C.4, choose $\lambda \le l_0\alpha/2$, $\mu \ge (L_0^3L_2/\lambda)^2$, and set the learning rate $\eta_\rho := \frac{\sqrt{\log T}}{dT^{\max\{1/2,\rho\}}}$ for Algorithm 3. We have we have:*

$$Reg_T^{DB} \le O\left(\max\left\{d\sqrt{T\log T}, dT^\rho \log T\right\}\right).$$

Proof (Proof for Proposition C.1). The formal proof of Proposition C.1 relies on Lemmas C.8, C.9, and C.10. Lemma C.8 establishes a connection between the gradient estimation corrupted by the bias term b_t and the quantity $\bar{P}_t(a)$ (defined in Lemma C.8), which is the main component we aim to analyze in the regret definition (16). Lemma C.9 decomposes the regret into two parts: a regret of decision independent of b_t and an error in observation summarizing the cumulative impact of b_t.

The optimization error can be explicitly upper bounded by T and η_ρ using techniques similar to those employed in Kumagai (2017). The upper bound of the estimation error is further established in Lemma C.10, wherein we utilize the α-strong convexity of f to establish a connection between the convergence of the L_2-norm of the action $\|a_t - a^*\|_2$ to itself.

Lemma C.8 (Gradient Estimation under Corruption)**.** *Let $P_t(a)$ represent the likelihood of action a_t being preferred over a, where $P_t(a) := \sigma(f(a) - f(a_t))$. Likewise, the corrupted version is defined as $\hat{P}_t(a) = \sigma(f(a) - f(a_t) + c_t(a_t, a_t'))$. The smoothed version of P_t over int($\mathcal{A}$) is $\bar{P}_t(a) := \mathbb{E}_{x\in\mathbb{B}}\left[P_t(a + \nabla^2 R_t(a_t)^{-\frac{1}{2}}x)\right]$. $\mathcal{R}_t(a)$ is defined in Algorithm 3. Let $a_t' = a_t + \nabla^2\mathcal{R}_t(a_t)^{-\frac{1}{2}}u_t$, where u_t is uniformly sampled from $\mathbb{S}$. Let $b_t := d\left(P_t(a_t') - \hat{P}_t(a_t')\right)\nabla^2\mathcal{R}_t(a_t)^{\frac{1}{2}}u_t$, we have*

$$\mathbb{E}\left(\hat{g}_t|a_t\right) = \nabla\bar{P}_t(a_t) - \mathbb{E}\left(b_t|a_t\right).$$

Lemma C.9 (Regret Decomposition)**.** *Define $C_0 := (2\nu + 4L_0\beta + 4B_2L^2 + LL_0\beta)/\lambda$, and let $a_T^* = \frac{1}{T}a_1 + (1 - \frac{1}{T})a^*$. By Assumption C.3, executing Algorithm 3, we obtain:*

$$Reg_T^{DB} \le \underbrace{\frac{C_0}{\eta_\rho}\log(T) + 8d^2\eta_\rho T + 2LL_0R}_{\textit{Regret of Decision}} + \underbrace{2\mathbb{E}\left\{\sum_{t=1}^T b_t^\top(a_t - a_T^*)\right\}}_{\textit{Observation Error}}.$$

Lemma C.10 (Estimation Error Upper Bound)**.** *Under Assumption C.2, and C.4, we have*

$$\mathbb{E}\left\{\sum_{t=1}^T b_t^\top(a_t - a_T^*)\right\} \le d\sqrt{\lambda\eta_\rho}\mathbb{E}\left\{\sum_{t=1}^T \sqrt{t}m_t n_t\right\} + C_2 dC + 2d\sqrt{\lambda\eta_\rho T},$$

where we have $\lambda^*_{\mathcal{R}} := \sup_{a\in\mathcal{A}} \lambda_{\max}\left(\nabla^2 \mathcal{R}(a)\right)$*, and* m_t, n_t, C_2 *defined as*

$$m_t := \min\left(2, L_\sigma |c_t(a_t, a_t')|\right), n_t := \min\left(2R, \sqrt{\frac{2}{\alpha}\left(f(a_t) - f(a^*)\right)}\right), C_2 := 2L_\sigma R\sqrt{\lambda^*_R + 2\mu}.$$

Now, we present the proof of Proposition C.1. The key technical innovation lies in realizing that the upper bound of $\mathrm{Reg}_T^{\mathrm{DB}}$ contains itself. This implies that we can initially obtain a rough upper bound and then refine it until it becomes tight. The proof follows a mathematical induction argument. For the ease of notation, we reparameterize $C = C_\kappa T^\rho$. In the following proof, we consider $\rho > 0.5$ to make $T^\rho = \max\{\sqrt{T}, T^\rho\}$. If $\rho \le 0.5$, the following proof can be easily modified to make upper bounds involving T^ρ as $\sqrt{T}$.

In the first step, we employ a rough upper bound, specifically $\mathbb{E}\left(n_t\right) \le 2R$. This yields $\mathrm{Reg}_T^{DB} \le O(dT^{\frac{\rho+1}{2}})$. Since Reg_T^{DB} and $\mathrm{Reg}_T^{\mathrm{FO}}$ share the same order (see Equation (17)), it implies that $\sum_{t=1}^T \mathbb{E}\left(n_t\right)$ must be sublinear as well. Therefore, in the second step, we can use Abel's Summation Equation (Lemma C.16) to refine the upper bound of $\mathbb{E}\left(\sum_{t=1}^T \sqrt{t} m_t n_t\right)$.

In particular, we rigorously prove the following claim:

The Induction Claim: Given the number of iteration T. $T > \sqrt{2LL_0R}$. At k-th step, we have $\mathrm{Reg}_T^{\mathrm{DB}} \le 144dK\sqrt{\log T}T^{\rho+\frac{\frac{3}{2}-\rho}{2^{k+1}}}$, where $K := \max\left\{C_0, \frac{C_2C_\kappa}{\sqrt{\log T}}, L_\sigma^2 C_\kappa^2, 2\sqrt{\lambda}RC_\kappa L_\sigma\right\}$.

The Base Case: When $k = 1$, let $C_3 = C_2dC + 2d\sqrt{\lambda\eta_\rho T}$, we have

$$\mathrm{Reg}_T^{\mathrm{DB}} \le \frac{C_0}{\eta_\rho}\log(T) + 8d^2\eta_\rho T + 2LL_0R + 2C_3 + 2d\sqrt{\lambda\eta_\rho}\mathbb{E}\left(\sum_{t=1}^T \sqrt{t} m_t n_t\right). \tag{18}$$

$$\le \frac{C_0}{\eta_\rho}\log(T) + 8d^2\eta_\rho T + 2LL_0R + 2C_3 + 4dR\sqrt{\lambda\eta_\rho}\mathbb{E}\left\{\sum_{t=1}^T \sqrt{t} m_t\right\} \tag{19}$$

$$\le \frac{C_0}{\eta_\rho}\log(T) + 8d^2\eta_\rho T + 2LL_0R + 2C_3 + 4dR\sqrt{\lambda\eta_\rho T}\left(\sum_{t=1}^T m_t\right) \tag{20}$$

$$\le \frac{C_0}{\eta_\rho}\log(T) + 8d^2\eta_\rho T + 2LL_0R + 2C_3 + 4dL_\sigma R\sqrt{\lambda\eta_\rho}C_\kappa T^\rho. \tag{21}$$

We get Equation (18) according to Lemma C.9 and C.10. We get Equation (19) because of $\mathbb{E}\left(n_t\right) \le 2R$. We get Equation (20) by Abel's Summation Equation (see Lemma C.16). We get get Equation (21) because of corruption budget. Choosing the learning rate η_ρ according to Proposition C.1, we obtain $\mathrm{Reg}_T^{\mathrm{DB}} \le 144\sqrt{\log T}KdT^{\frac{\rho+1}{2}}$. This confirms the validity of the claim when $k = 1$.

The Induction Argument: Let's assume that the claim holds true for a general step k. In the following, we will demonstrate that the claim also holds true for step $k+1$. Because of Equation (17) and the induction claim, we obtain $\mathrm{Reg}_T^{\mathrm{FO}} \le \frac{144}{l_0}dK\sqrt{\log T}T^{\rho+\frac{3/2-\rho}{2^{k+1}}}$. This suggests that

$$\sum_{t=1}^{T} \mathbb{E}\left(f(a_t) - f(a^*)\right) \leq \frac{144}{l_0} dK \sqrt{\log T} T^{\rho + \frac{3/2-\rho}{2^{k+1}}}.$$

By Jensen's inequality, we have

$$\mathbb{E}(n_t) = \min\left\{2, \sqrt{\frac{2}{\alpha}} \mathbb{E}\left(\sqrt{f(a_t) - f(a^*)}\right)\right\} \leq \sqrt{\frac{2}{\alpha}} \sqrt{\mathbb{E}\left(f(a_t) - f(a^*)\right)}.$$

Therefore we have

$$\mathbb{E}\left(\sum_{t=1}^{T} \sqrt{t} m_t n_t\right) \leq \sqrt{\frac{2}{\alpha}} \sum_{t=1}^{T} \sqrt{t} m_t \sqrt{\mathbb{E}\left(f(a_t) - f(a^*)\right)}.$$

By the definition of monotone adversary, we have $|c_t(a_t, a_t')|$ is upper bounded by $c_k t^{-1+\rho}$.

Lemma C.11. *Consider $\sum_{t=1}^{T} n_t^2 \leq C' T^{\frac{1}{2}+\alpha}$, with constants $\alpha \geq 0$, $C' > 0$. In addition, n_t is uniformly upper bounded by a constant K. Specifically, we have $0 \leq n_t \leq K, \forall t$. $m_t \leq c_k t^{\rho-1}$ with constant $c_k \geq 0$. Then we have $\sum_{t=1}^{T} \sqrt{t} m_t n_t \leq 5\sqrt{C'} c_k T^{-\frac{1}{4}+\frac{\alpha}{2}+\rho}$.*

Therefore, using Lemma C.11 (details of proof is listed at Appendix C.3.5), we can upper bound $\sqrt{\lambda} \sum_{t=1}^{T} \sqrt{t} m_t n_t$ by

$$\sqrt{\lambda} \sum_{t=1}^{T} \sqrt{t} m_t n_t \leq 60 L_\sigma \sqrt{\frac{2\lambda}{\alpha l_0}} (\log T)^{0.25} C_k \sqrt{dK} T^{\frac{3}{2}\rho + \frac{3/2-\rho}{2^{k+2}}}$$

Since $\lambda \leq \alpha l_0/2$, choosing $\eta_\rho = \frac{1}{2d} \frac{\sqrt{\log T}}{T^\rho}$, we have

$$\begin{aligned}
\mathrm{Reg}_T^{\mathrm{DB}} &\leq 2d\left(C_0 + \frac{C_2}{\sqrt{\log T}} + 60 C_\kappa \sqrt{K} T^{\frac{3/2-\rho}{2^{k+2}}}\right) \sqrt{\log T} T^\rho + 12Kd\sqrt{\log T} T^\rho \\
&\leq 144 dK \sqrt{\log T} T^{\rho + \frac{3/2-\rho}{2^{k+2}}}.
\end{aligned}$$

The claim holds true for step $k+1$. Therefore, it implies that the claim holds true for all k. Repeating this process infinitely many times, we obtain

$$\mathrm{Reg}_T^{\mathrm{DB}} \leq O\left(d\sqrt{\log T} T^\rho\right),$$

which completes the proof when $\rho \in [0.5, 1]$. ■

C.3.2 Proof for Lemma C.8: Gradient Estimation Under Corruption

Proof (Proof for Lemma C.8). If we do not have corruption (i.e. the probability for observing noisy comparative feedback $\mathcal{F}(a, a') = 1$ is based on true cost difference $f(a') - f(a)$), then the uncorrupted gradient $g_t := \mathcal{F}(a_t', a_t) d \nabla^2 \mathcal{R}_t(a_t)^{1/2} u_t$ should be an unbiased estimate of $\nabla \bar{P}_t(a_t)$, i.e.

$$\mathbb{E}(g_t | a_t) = \nabla \bar{P}_t(a_t).$$

The proof is similar to Lemma 2.1 in Flaxman et al. (2005). Using the Law of total expectation, we have

$$\begin{aligned}\mathbb{E}(g_t|a_t) &= \mathbb{E}_{u_t}(\mathbb{E}(g_t|a_t,u_t))\\ &= \mathbb{E}_{u_t}\left(d\mathbb{E}(P_t(a_t+\nabla^2\mathcal{R}_t(a_t)^{-\frac{1}{2}}u_t)\nabla^2\mathcal{R}_t(a_t)^{\frac{1}{2}}u_t|a_t,u_t)\right)\\ &= d\mathbb{E}(P_t(a_t+\nabla^2\mathcal{R}_t(a_t)^{-\frac{1}{2}}u_t)\nabla^2\mathcal{R}_t(a_t)^{\frac{1}{2}}u_t|a_t)\\ &= \nabla\mathbb{E}_{x\in\mathbb{B}}\left(P_t(a_t+\nabla^2\mathcal{R}_t(a_t)^{-\frac{1}{2}}x|a_t)\right)\\ &= \nabla\bar{P}_t(a_t).\end{aligned}$$

We get the second inequality by using the definition of $\mathcal{F}(a_t', a_t)$. We use the Stroke's Theorem to get the second last equality. The gradient which we get to perform gradient descent $\hat{g}_t$ is corrupted. If we let $a_t' = a_t + \nabla^2\mathcal{R}_t(a_t)^{-\frac{1}{2}}u_t$, we have

$$\begin{aligned}\mathbb{E}(\hat{g}_t|a_t) &= \mathbb{E}_{u_t}(\mathbb{E}(\hat{g}_t|a_t,u_t))\\ &= \mathbb{E}_{u_t}\left(d\mathbb{E}(\hat{P}_t(a_t')\nabla^2\mathcal{R}_t(a_t)^{\frac{1}{2}}u_t|a_t,u_t)\right)\\ &= d\mathbb{E}\left(\hat{P}_t(a_t')\nabla^2\mathcal{R}_t(a_t)^{\frac{1}{2}}u_t|a_t\right)\\ &= d\mathbb{E}\left(\left[P_t(a_t')+\hat{P}_t(a_t')-P_t(a_t')\right]\nabla^2 R_t(a_t)^{\frac{1}{2}}u_t|a_t\right)\\ &= \nabla\mathbb{E}_{x\in\mathbb{B}}\left(P_t(a_t+\nabla^2\mathcal{R}_t(a_t)^{-\frac{1}{2}}x|a_t)\right)+d\mathbb{E}\left[\left(\hat{P}_t(a_t')-P_t(a_t')\right)\nabla^2\mathcal{R}_t(a_t)^{\frac{1}{2}}u_t|a_t\right]\\ &= \nabla\bar{P}_t(a_t)+d\mathbb{E}\left[\left(\hat{P}_t(a_t')-P_t(a_t')\right)\nabla^2\mathcal{R}_t(a_t)^{\frac{1}{2}}u_t|a_t\right]\\ &= \mathbb{E}(g_t|a_t)+d\mathbb{E}\left[\left(\hat{P}_t(a_t')-P_t(a_t')\right)\nabla^2\mathcal{R}_t(a_t)^{\frac{1}{2}}u_t|a_t\right].\end{aligned}$$

We get the second last equality by using the fact that g_t is unbiased, which we proved earlier. If we defined b_t as

$$b_t := d\left(P_t(a_t+\nabla^2\mathcal{R}_t(a_t)^{-\frac{1}{2}}u_t)-\hat{P}_t(a_t+\nabla^2\mathcal{R}_t(a_t)^{-\frac{1}{2}}u_t)\right)\nabla^2\mathcal{R}_t(a_t)^{\frac{1}{2}}u_t,$$

we get

$$\mathbb{E}(g_t|a_t) = \mathbb{E}(\hat{g}_t|a_t)+\mathbb{E}\left[b_t|a_t\right].$$

■

C.3.3 Proof for Lemma C.9: Regret Decomposition

Proof (**Proof for Lemma C.9**). The cornerstone of the analysis below is to separate the impact of b_t from the total regret.

$$\begin{aligned}
&\mathrm{Reg}_T^{\mathrm{DB}} \le 2\mathbb{E}\left[\sum_{t=1}^{T}(P_t(a_t) - P_t(a_T^*)\right] + \frac{LL_0\beta}{\lambda\eta} + 2LL_0R \\
&\le 2\left(\mathbb{E}\left\{\sum_{t=1}^{T}(\bar{P}_t(a_t) - \bar{P}_t(a_T^*))\right\} + \mathbb{E}\left\{\sum_{t=1}^{T}(P_t(a_t) - \bar{P}_t(a_t))\right\} + \mathbb{E}\left\{\sum_{t=1}^{T}(\bar{P}_t(a_T^*) - P_t(a_T^*))\right\}\right) \\
&+ \frac{LL_0\beta}{\lambda\eta} + 2LL_0R \\
&\le 2\mathbb{E}\left\{\sum_{t=1}^{T}(\bar{P}_t(a_t) - \bar{P}_t(a_T^*))\right\} + \frac{4L_0\beta + 4B_2L^2 + LL_0\beta}{\lambda\eta_\rho}\log T + 2LL_0R.
\end{aligned}$$

We get the first inequality by using Lemma C.12. We get the last inequality by because

$$\bar{P}_t(a) - P_t(a) \le \frac{L_0\beta + B_2L^2}{2}\|\nabla^2\mathcal{R}_t(a_t)^{-\frac{1}{2}}u_t\|^2 \le \frac{L_0\beta + B_2L^2}{\lambda\eta_\rho t},$$

and

$$\mathbb{E}\left\{\sum_{t=1}^{T}(P_t(a_t) - P_t(a_T^*))\right\} \le \mathbb{E}\left\{\sum_{t=1}^{T}(\bar{P}_t(a_t) - \bar{P}_t(a_T^*))\right\} + 2\frac{L_0\beta + B_2L^2}{\lambda\eta_\rho}\log T.$$

Lemma C.12 (Lemma 5 in Kumagai (2017))**.**

$$Reg_T^{DB} \le 2\mathbb{E}\left[\sum_{t=1}^{T}(P_t(a_t) - P_t(a_T^*)\right] + \frac{LL_0\beta}{\lambda\eta} + 2LL_0R.$$

Then it remains to bound $\mathbb{E}\{\sum_{t=1}^{T}\bar{P}_t(a_t) - \bar{P}_t(a_T^*)\}$. And we have the following.

$$\begin{aligned}
\mathbb{E}\{\bar{P}_t(a_t) - \bar{P}_t(a_T^*)\} &\le \mathbb{E}\left[\nabla\hat{P}_t^\top(a_t)(a_t - a_T^*) - \frac{l_0\alpha}{4}\|a_t - a_T^*\|^2\right] \\
&= \mathbb{E}\left[g_t^\top(a_t - a_T^*) - \frac{l_0\alpha}{4}\|a_t - a_T^*\|^2\right] \\
&= \mathbb{E}\left[(\hat{g}_t + b_t)^\top(a_t - a_T^*) - \frac{l_0\alpha}{4}\|a_t - a_T^*\|^2\right] \\
&= \mathbb{E}\left[\hat{g}_t^\top(a_t - a_T^*) - \frac{l_0\alpha}{4}\|a_t - a_T^*\|^2\right] + \mathbb{E}\{b_t^\top(a_t - a_T^*)\},
\end{aligned}$$

where the first and second equality is resulted from Lemma C.8. Using the definition of a_{t+1} in Algorithm 3, we have

$$\nabla\mathcal{R}_t(a_{t+1}) - \nabla\mathcal{R}_t(a_t) = \eta_\rho\hat{g}_t.$$

Therefore we have

$$\mathbb{E}\left[\hat{g}_t^\top(a_t - a_T^*) - \frac{l_0\alpha}{4}\|a_t - a_T^*\|^2\right] = \frac{1}{\eta_\rho}\mathbb{E}\left[(\nabla\mathcal{R}_t(a_{t+1}) - \nabla\mathcal{R}_t(a_t))^\top(a_t - a_T^*) - \frac{l_0\alpha\eta_\rho}{4}\|a_t - a_T^*\|^2\right]$$
$$= \frac{1}{\eta_\rho}\mathbb{E}\left[D_{\mathcal{R}_t}(a_T^*, a_t) + D_{\mathcal{R}}(a_t, a_{t+1}) - D_{\mathcal{R}_t}(a_T^*, a_{t+1}) - \frac{l_0\alpha\eta_\rho}{4}\|a_t - a_T^*\|^2\right].$$

$D_{\mathcal{R}}(a, b)$ is the Bregman divergence associated with $\mathcal{R}$, defined by

$$D_{\mathcal{R}}(a, b) = \mathcal{R}(a) - \mathcal{R}(b) - \nabla\mathcal{R}(b)^\top(a - b).$$

Then we can get Equation (22) by using Lemma C.13, C.14, C.15.

$$\mathbb{E}\left\{\sum_{t=1}^{T}(\bar{P}_t(a_t) - \bar{P}_t(a_T^*))\right\} \leq \frac{\nu\log(T)}{\eta_\rho} + 4d^2\eta_\rho T + \mathbb{E}\{\sum_{t=1}^{T} b_t^\top(a_t - a_T^*)\}. \tag{22}$$

Lemma C.13 (Lemma 10 in Kumagai (2017)). *Let $\mathcal{R}_t^*(a) = \sup_{x\in\mathbb{R}^d} x^\top a - \mathcal{R}_t(a)$ denote the Frenchel dual of $\mathcal{R}_t$. Then we have*

$$\sum_{t=1}^{T}\mathbb{E}\left[\hat{g}_t^\top(a_t - a_T^*) - \frac{l_0\alpha}{4}\|a_t - a_T^*\|^2\right] \leq \frac{1}{\eta_\rho}\left(\mathcal{R}(a_T^*) - \mathcal{R}(a_1) + \mathbb{E}\left[\sum_{t=1}^{T} D_{\mathcal{R}_t^*}(\nabla\mathcal{R}_t(a_t) - \eta_\rho\hat{g}_t, \nabla\mathcal{R}_t(a_t))\right]\right).$$

Lemma C.14 (Lemma 11 in Kumagai (2017)). *When $\eta_\rho \leq \frac{1}{2d}$, we have*

$$D_{\mathcal{R}_t^*}(\nabla\mathcal{R}_t(a_t) - \eta_\rho\hat{g}_t, \nabla\mathcal{R}_t(a_t)) \leq 4d^2\eta^2.$$

Lemma C.15 (Lemma 4 in Hazan and Levy (2014)). $\mathcal{R}(a_T^*) - \mathcal{R}(a_1) \leq \nu\log(T)$.

If we let $C_0 = (2\nu + 4L_0\beta + 4B_2L^2 + LL_0\beta)/\lambda$, we have

$$\text{Reg}_T^{\text{DB}} \leq \frac{2\nu\log(T)}{\eta_\rho} + 8d^2\eta_\rho T + 2\mathbb{E}\left\{\sum_{t=1}^{T} b_t^\top(a_t - a_T^*)\right\} + \frac{4L_0\beta + 4B_2L^2 + LL_0\beta}{\lambda\eta_\rho}\log T + 2LL_0R$$
$$= \underbrace{\frac{C_0}{\eta_\rho}\log(T) + 8d^2\eta_\rho T + 2LL_0R}_{\text{Regret of Decision}} + \underbrace{2\mathbb{E}\left\{\sum_{t=1}^{T} b_t^\top(a_t - a_T^*)\right\}}_{\text{Observation Error}},$$

which completes the proof. ■

C.3.4 Proof for Lemma C.10: Observation Error

Proof (**Proof for Lemma C.10**) In the following, we will analyze the expected cumulative sum of the bias b_t on the total regret, which is $\mathbb{E}\left\{\sum_{t=1}^{T} b_t^\top(a_t - a_T^*)\right\}$. By Cauchy-Schwarz inequality, $b_t^\top(a_t - a_T^*) \leq \|b_t\|_2\|(a_t - a_T^*)\|_2$. We can bound the $\|b_t\|_2$ by the following.

$$b_t^\top b_t = d^2\left(P_t(a_t + \nabla^2\mathcal{R}_t(a_t)^{-\frac{1}{2}}u_t) - \hat{P}_t(a_t + \nabla^2\mathcal{R}_t(a_t)^{-\frac{1}{2}}u_t)\right)^2 u_t^\top\nabla^2\mathcal{R}_t(a_t)u_t$$
$$\leq \left(P_t(a_t + \nabla^2\mathcal{R}_t(a_t)^{-\frac{1}{2}}u_t) - \hat{P}_t(a_t + \nabla^2\mathcal{R}_t(a_t)^{-\frac{1}{2}}u_t)\right)^2 d^2\lambda_{\max}(\nabla^2\mathcal{R}_t(a_t)).$$

We get the first equality by Lemma *C*.8. We get the inequality by using the fact $\|u_t\|_2 = 1$. If we let $a_t' = a_t + \nabla^2 \mathcal{R}_t(a_t)^{-\frac{1}{2}} u_t$, then we have

$$\begin{aligned}|P_t(a_t') - \hat{P}_t(a_t')| &= |\sigma(f(a_t) - f(a_t')) - \sigma(f(a_t) - f(a_t') + c_t(a_t, a_t'))| \\ &\le \min\left(2, L_\sigma |c_t(a_t, a_t')|\right).\end{aligned}$$

We get the first equality by using the definition of $P_t(a_t)$ and $\hat{P}_t(a_t)$. We get the first inequality by using the Lipschitz property of σ (see Assumption C.3) Therefore, we have

$$\|b_t\|_2 \le d\sqrt{\lambda_{\max}(\nabla^2 \mathcal{R}_t(a_t))} \min\left\{2, L_\sigma |c_t(a_t, a_t')|\right\}.$$

If we use $\lambda_R^* := \sup_{a \in \mathcal{A}} \lambda_{\max}\left(\nabla^2 \mathcal{R}(a)\right)$, then we have

$$\lambda_{\max}(\nabla^2 \mathcal{R}_t(a_t)) = \lambda_{\max}(\nabla^2 \mathcal{R}(a_t)) + \lambda\eta_\rho t + 2\mu \le \lambda_R^* + 2\mu + \lambda\eta_\rho t.$$

Therefore we have

$$\begin{aligned}
&\mathbb{E}\left\{\sum_{t=1}^T b_t^\top (a_t - a_T^*)\right\} \le \mathbb{E}\left\{\sum_{t=1}^T \|b_t\|_2 \|a_t - a_T^*\|_2\right\} \\
&\le \mathbb{E}\left\{\sum_{t=1}^T (d\sqrt{\lambda_{\max}(\nabla^2 \mathcal{R}_t(a_t))} \min\left(2, L_\sigma |c_t(a_t, a_t')|\right)) \|a_t - a_T^*\|_2\right\} \\
&\le \mathbb{E}\left\{\sum_{t=1}^T (d\sqrt{\lambda_R^* + 2\mu + \lambda\eta_\rho t} \min\left(2, L_\sigma |c_t(a_t, a_t')|\right)) \|a_t - a_T^*\|_2\right\} \\
&\le \mathbb{E}\left\{\sum_{t=1}^T (d\sqrt{\lambda_R^* + 2\mu} \min\left(2, L_\sigma |c_t(a_t, a_t')|\right) \|a_t - a_T^*\|_2\right\} + \mathbb{E}\left\{\sum_{t=1}^T (d\sqrt{\lambda\eta_\rho t} \min\left(2, L_\sigma |c_t(a_t, a_t')|\right) \|a_t - a_T^*\|_2\right\} \\
&\le 2RdL_\sigma \sqrt{\lambda_R^* + 2\mu} C + d\sqrt{\lambda\eta_\rho} \mathbb{E}\left\{\sum_{t=1}^T \sqrt{t} \min\left(2, L_\sigma |c_t(a_t, a_t')|\right) \|a_t - a_T^*\|_2\right\} \qquad (23) \\
&\le 2RdL_\sigma \sqrt{\lambda_R^* + 2\mu} C + 2d\sqrt{\lambda\eta_\rho T} + d\sqrt{\lambda\eta_\rho} \mathbb{E}\left\{\sum_{t=1}^T \sqrt{t} \min\left(2, L_\sigma |c_t(a_t, a_t')|\right) \|a_t - a^*\|_2\right\} \qquad (24) \\
&\le 2RdL_\sigma \sqrt{\lambda_R^* + 2\mu} C + 2d\sqrt{\lambda\eta_\rho T} + d\sqrt{\lambda\eta_\rho} \mathbb{E}\left\{\sum_{t=1}^T \sqrt{t} \min\left(2, L_\sigma |c_t(a_t, a_t')|\right) \min\left(2R, \sqrt{\frac{2}{\alpha}\left(f(a_t) - f(a^*)\right)}\right)\right\}
\end{aligned}$$

We get Equation (23) by using the fact that $\|a_t - a_T^*\|_2 \le 2R$, $\sum_{t=1}^T |c_t(a_t, a_t')| \le C$. We get Equation (24) the definition of a_T^*, and the fact $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$. Using the α-strong convexity of f (see Assumption C.2), we have $\|a_t - a_*\|_2 \le \min\left(2R, \sqrt{\frac{2}{\alpha}\left(f(a_t) - f(a^*)\right)}\right)$ we get the last inequality, which completes the proof. ■

C.3.5 Proof for Lemma C.11

Proof (Proof for Lemma C.11). Using Abel's Summation Equation (Lemma C.16), we have

$$\sum_{t=1}^{T} \sqrt{t} m_t n_t \leq \sqrt{T} \left(\sum_{t=1}^{T} m_t n_t \right)$$

Since we have $\sum_{t=1}^{T} n_t^2 \leq C' T^{\frac{1}{2}+\alpha}$ and $0 \leq n_t \leq K, \forall t$, we have $\sum_{t=1}^{T} n_t \leq \sqrt{C'} T^{\frac{3}{4}+\frac{\alpha}{2}}, \forall t$. Moreover, for $t \geq 1$, $t^{-1+\rho} - (t+1)^{-1+\rho} \leq t^{\rho - \frac{1}{2}}$ holds for all $t \geq 1$. This is because

$$f(t) = t^{-1+\rho} \left((1 + \frac{1}{t})^{-1+\rho} + \frac{1}{t} - 1 \right).$$

is decreasing over $t \geq 1$ and $\lim_{t \to \infty} f(t) = 0$. Consequently, applying Abel's Summation Equation again, we have

$$\begin{aligned}
\sum_{t=1}^{T} m_t n_t &= \left(\sum_{t=1}^{T} n_t \right) m_T + \sum_{t=1}^{T-1} \left(\sum_{i=1}^{t} n_i \right) (m_t - m_{t+1}) \\
&\leq \sqrt{C'} c_k T^{\frac{3}{4}+\frac{\alpha}{2}} T^{-1+\rho} + \sqrt{C'} c_k \sum_{t=1}^{T-1} \left(\sum_{i=1}^{t} n_i \right) \left(t^{-1+\rho} - (t+1)^{-1+\rho} \right) \\
&\leq \sqrt{C'} c_k T^{\frac{3}{4}+\frac{\alpha}{2}} T^{-1+\rho} + \sqrt{C'} c_k \sum_{t=1}^{T-1} \left(\sum_{i=1}^{t} n_i \right) t^{-\frac{1}{2}+\rho} \\
&\leq \sqrt{C'} c_k T^{\frac{3}{4}+\frac{\alpha}{2}} T^{-1+\rho} + \sqrt{C'} c_k \sum_{t=1}^{T-1} t^{\frac{3}{4}+\frac{\alpha}{2}} t^{-\frac{1}{2}+\rho} \\
&\leq 5\sqrt{C'} c_k T^{-\frac{1}{4}+\frac{\alpha}{2}+\rho}.
\end{aligned}$$

Since $0 \leq n_t \leq K$, $\sum_{i=1}^{t} n_i$ is increasing and the increasing rate is upper bound by t. Together with the constraint $\sum_{t=1}^{T} n_t \leq \sqrt{C'} T^{\frac{3}{4}+\frac{\alpha}{2}}$, $\sum_{t=1}^{T-1} \left(\sum_{i=1}^{t} n_i \right) t^{-\frac{1}{2}+\rho}$ is optimized when $\sum_{i=1}^{t} n_i$ increasing in the speed of $\sqrt{C'} t^{\frac{3}{4}+\frac{\alpha}{2}}$. Because of this, the second last inequality holds.

Lemma C.16. ***(Abel's Summation Equation (Williams, 1963)).*** *For any numbers a_k, b_k, we have*

$$\sum_{k=1}^{n} a_k b_k = \left(\sum_{k=1}^{n} b_k \right) a_n + \sum_{k=1}^{n-1} \left(\sum_{i=1}^{k} b_i \right) (a_k - a_{k+1}).$$

■

C.4 Proof for Proposition 2: Efficiency-Robustness Tradeoff in NC-SMD

Theorem (Full Statement of Proposition 2)**.** *Let the number of rounds T be fixed and $T > \sqrt{2LL_0R}$. Under Assumptions C.2, C.3, and C.4, by choosing $\lambda \leq l_0\alpha/2$, $\mu \geq (L_0^3 L_2/\lambda)^2$,* *and setting the learning parameter* $\eta_\rho = \frac{\sqrt{\log T}}{2d} T^{-\alpha}, \alpha \in [0.5, 1]$, *regret occurred by Algorithm 3 satisfies*

$$Reg_T \leq \tilde{O}\left(dT^\alpha + \sqrt{d}T^{\frac{1}{2}(1-\alpha)}C + dC\right)$$

for any corruption with budget C.

Proof (Proof for Proposition 2). From Lemma C.10, we know

$$\begin{aligned}
\mathbb{E}\left\{\sum_{t=1}^T b_t^\top (a_t - a_T^*)\right\} &\leq 2RdL_\sigma\sqrt{\lambda_R^* + 2\mu}C + d\sqrt{\lambda\eta_\rho}\mathbb{E}\left\{\sum_{t=1}^T \sqrt{t}\min\left(2, L_\sigma|c_t(a_t, a_t')|\right)\|a_t - a_T^*\|_2\right\} \\
&\leq 2RdL_\sigma\sqrt{\lambda_R^* + 2\mu}C + 2Rd\sqrt{\lambda\eta_\rho}L_\sigma\sum_{t=1}^T \sqrt{t}|c_t(a_t, a_t')| \\
&\leq 2RdL_\sigma\sqrt{\lambda_R^* + 2\mu}C + 2Rd\sqrt{\lambda\eta_\rho T}L_\sigma C.
\end{aligned}$$

We get the second inequality by using the fact that the diameter of the action space of R. We get the last inequality by using Abel Summation Equation (see Lemma C.16). Therefore, by Lemma C.9, we have

$$\text{Reg}_T \leq \frac{C_0}{\eta_\rho}\log(T) + 8d^2\eta_\rho T + 2LL_0R + 4RdL_\sigma\sqrt{\lambda_R^* + 2\mu}C + 4Rd\sqrt{\lambda\eta_\rho T}L_\sigma C.$$

Choosing $\eta_\rho = \frac{\sqrt{\log T}}{2d}T^{-\alpha}, \alpha \in [0.5, 1]$, we have

$$\begin{aligned}
\text{Reg}_T &\leq 2dC_0\sqrt{\log T}T^\alpha + 4d\sqrt{\log T}T^{1-\alpha} + 4RdL_\sigma\sqrt{\lambda_R^* + 2\mu}C + 4R\sqrt{d}L_\sigma(\log T)^{0.25}T^{\frac{1}{2}(1-\alpha)}C + 2LL_0R \\
&\leq O\left(d\sqrt{\log T}T^\alpha + \sqrt{d}(\log T)^{\frac{1}{4}}T^{\frac{1}{2}(1-\alpha)}C + dC\right),
\end{aligned}$$

which completes the proof. ■

Appendix D. Additional Experiments

In this section, we list all the experiment details. At the beginning, we introduce the regret order fitting method, different types of corruption that we simulate in the experiments, and baseline algorithms for comparison.

Fitted Order of Regret. We introduce the methodology that we use to compute the order of Reg_T in terms of the number of iteration, T. To fit the order of the regret, we first convert it into log scale. Then we input the last 1% of data, run linear regression, and use ordinary least squares to estimate the slope of the line, which is the *fitted order* of Reg_T.

Corruption. In the experiments, we consider two types of corruption: corruption induced by *ρ-imperfect* user and *arbitrary* corruption. For corruption induced by *ρ-imperfect* user, we simulate $c_t(a_t, a_t')$ according to Definition A.1. We set $\lambda = 2$ and $C_0 = 0.1(\max_{a\in\mathcal{A}}\mu(a) - \min_{a\in\mathcal{A}}\mu(a))$. Moreover, we add $c_t(a_t, a_t')$ to the utility difference of a_t and a_t' adversarially. In particular, if $\mu(a_t) > \mu(a_t')$, then $\hat{\mathbb{P}}(a_t \succ a_t') = \sigma(\mu(a_t) - \mu(a_t') - c_t(a_t, a_t'))$, then vice versa. For *arbitrary* corruption, we force the user to submit her least preferred item each round to the algorithm over the first C rounds.

Baseline Algorithms. We consider three baseline algorithms for comparison, Doubler (Ailon et al., 2014), Sparring (Ailon et al., 2014), and Versatile-DB (Saha and Gaillard, 2022).

Doubler is the first approach that transforms a dueling bandit problem into a standard multi-armed bandit (MAB) problem. It operates in epochs of exponentially increasing length: in each epoch, the left arm is sampled from a fixed distribution, and the right arm is selected using a MAB algorithm to minimize regret against the left arm. The feedback received by the MAB algorithm is the number of wins the right arm achieves compared to the left arm. Under linear link assumption, Doubler has been proven to experience regret as the same order as underlying MAB algorithm. For continuous *action* space and general concave utility, we choose Bandit Gradient Descent (BGD, (Flaxman et al., 2004)), with regret $O(T^{3/4})$, as the underlying MAB algorithm.

Sparring initializes two MAB instances and lets them compete against each other. It is a heuristic improvement over Doubler. Although it does not come with a regret upper bound guarantee, it is reported to enjoy better performance compared to Doubler (Ailon et al., 2014). We also choose BGD as the underlying MAB algorithm.

Versatile-DB applies novel reduction approach which converts dueling bandit to MAB. Essentially, it designs simplex over K arms using the idea of follow-the-regularized-leader and updates the simplex using information of whether the arm wins. It has been proved that Versatile-DB is robust to corruption defined as the number of flips of duels and Versatile-DB enjoys regret linear in C (Saha and Gaillard, 2022).

D.1 Experiments on Synthetic Data

D.1.1 DBGD Lower Bound Simulation

Experiment Setup. To validate Corollary 1, we construct the following problem instance. Consider the linear utility function $\mu(a) := \theta^\top a$. Choose $d = 2, \theta = [\frac{1}{2}, \frac{1}{2}], T = 10^5$. Let *action* set $\mathcal{A} := \{(a_1, a_2) : a_1 \geq 0, a_2 \geq 0, \frac{1}{2}a_1 + a_2 - \frac{1}{4} \leq 0\}$. We repeat the experiments for 50 times, each time with a different seed.

Results and Discussion. Each grey-colored dotted line in Figure 2 indicate the cumulative regret for each simulation. The blue line is the average cumulative regret over 50 repeats. The error bar indicates $\pm$ one standard deviation of the regret. The fitted order over the last 1% of the average cumulative regret is 0.76, which is close 0.75 and supports Corollary 1.

D.1.2 Regret Upper Bound in Theorem 1

Experiment Setup. To validate regret upper bound in Theorem 1, we consider a strongly concave utility $\mu_\theta(a) := \theta^\top a - \frac{1}{2}\|a\|_2^2$, a logistic link function $\sigma(x) = \frac{1}{1+\exp(-x)}$. We choose $d = 5$, and $T = 10^5$. Our *action* space $\mathcal{A}$ is a d-dimensional ball with radius $R = 10$.

The preference parameter θ is randomly sampled from the surface of $\mathcal{A}$. In our problem setting, the optimal action a^* is θ and $\mu_\theta(a^*) = 50$. We simulate corruption induced by ρ-*imperfect* user and conduct experiments for $\rho \in [0.5, 1.0]$. For each value of ρ, we repeat the experiments for 5 times, each under a different seed. We set the learning rate η_ρ according to Proposition C.1.

Results and Discussion. Figure D.1 shows that when ρ is known, NC-SMD can tolerate linear ρ-*imperfect* user corruption, which confirms theoretical analysis in Proposition C.1 and implies that lower bound in Theorem 1 is tight.

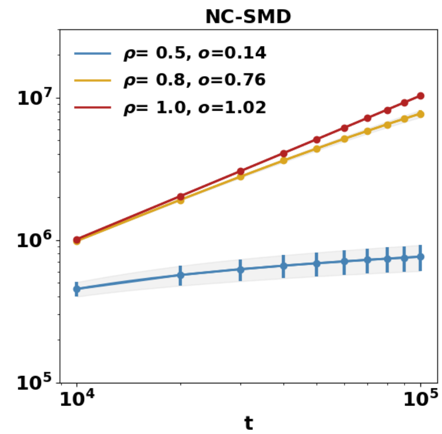

Figure D.1: Performance under Known C

D.1.3 Robustness under Corruption induced by Imperfect User

In Figure D.2, we simulated the performance of DBGD, NC-SMD, Sparring, Doubler in the same experiment setting as in Section 6.1 expect using corruption induced by ρ-Imperfect User. We notice that the experiment results also align with theoretical prediction (Theorem 2 and Proposition 2), which predicts when $\alpha = 0.25$ for DBGD and $\alpha = 0.5$ for NC-SMD, they can tolerate $O(T^{0.75})$ agnostic corruption.

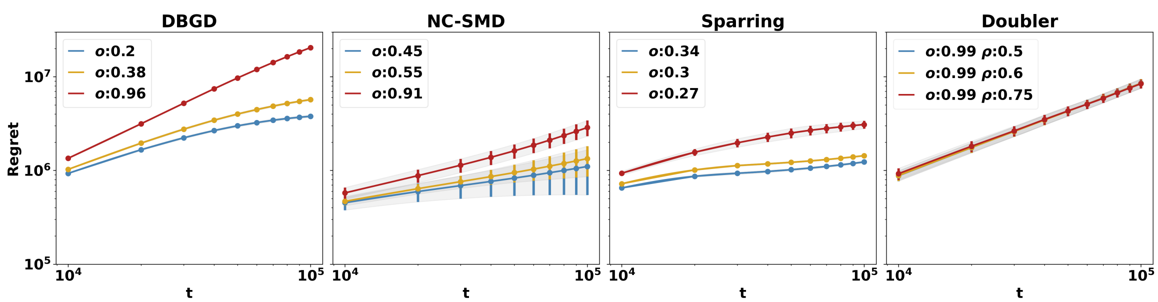

Figure D.2: Robustness under Corruption induced by ρ-Imperfect User

D.1.4 Efficiency-Robustness Tradeoff in Theorem 2

To validate efficiency-robustness trade-off in Theorem 2 and Proposition 2, we adopt the same experiment setup as in Section 6.1 except different α values. In the first row of Figure D.3, we consider $\alpha = 0.05$ for DBGD and $\alpha = 0.9$ for NC-SMD. We consider $\rho \in [0.5, 0.95]$. According to the theoretical prediction, both DBGD and NC-SMD can tolerate at most $O(T^{0.95})$ agnostic corruption, which aligns with experiment results. In the second row of Figure D.3, we consider $\alpha = 0.1$ for DBGD and $\alpha = 0.8$ for NC-SMD. According to the theoretical prediction, both DBGD and NC-SMD can tolerate at most $O(T^{0.9})$ agnostic corruption. From the figure, we can see they have smaller fitted order of regret when

$\rho = 0.5$ while at a cost of tolerating smaller magnitude of agnostic corruption (it has linear regret when $\rho = 0.95$), which reveals intrinsic tradeoff between efficiency and robustness.

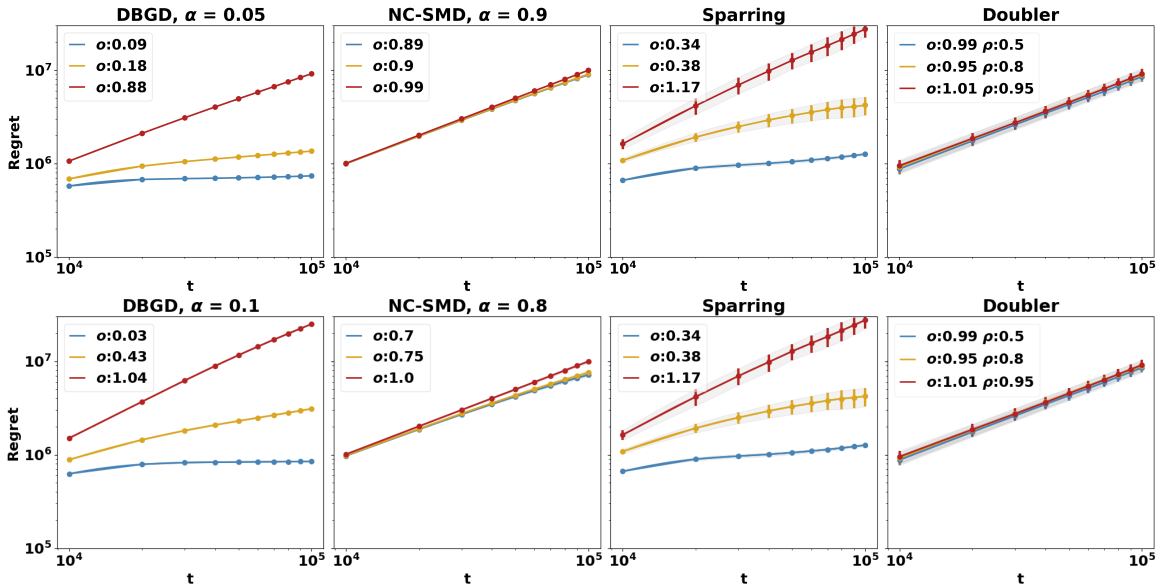

Figure D.3: Efficiency-Robustness Tradeoff

D.2 Experiments on Spotify Recommendation Data

Evaluation Setup. We evaluate our approach on Spotify recommendation data (Spotify, 2020). The objective is to recommend songs to incoming users. This dataset includes 17×10^4 songs ($|\mathcal{A}| = 17 \times 10^4$), each is described by 15 distinct features (d = 15).

User Profile Identification. In order to model different types of user, we first use Standard Scaler to standardize the dataset. Then we use Kmeans to conduct clustering (Pedregosa et al., 2011).

User Utility. We use the average of the song embedding within each group as the preference vector. Our aim is to recommend songs to users that have the highest cosine similarity (utility function) with their preference vectors. We rescale the cosine similarity to $[-100, 100]$.

D.2.1 Performance of Versatile-DB on Spotify Recommendation Data

To test the performance of Versatile-DB, we randomly sample $K = 2000$ songs from the Spotify Recommendation Dataset. We identify 2 different user types by clustering the dataset into 2 groups. We use the average of the song embedding within each group as the preference vector and compute the user utility for each recommendation. We consider corruption induced by ρ-*imperfect* user for $\rho \in [0.5, 0.75]$. We set $\alpha = 0.25$ for DBGD and run both algorithms for $T = 100$ iterations. This is because Versatile-DB is extremely computationally intensive and running $K = 2000, T = 100$ takes around 20 CPU hours. For each user type and for each value of ρ, we repeat the experiment over 5 times, each under a different seed. From Figure D.4, we notice that DBGD outperforms Versatile-DB in all ρ, which highlights DBGD's robustness and applicability in real-world data.

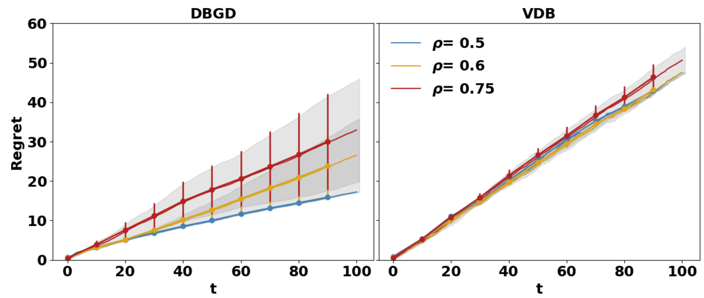

Figure D.4: Performance Comparison between DBGD and Versatile-DB