Stochastic Dueling Bandits with Adversarial Corruption

Arpit Agarwal AARPIT@SEAS.UPENN.EDU
Shivani Agarwal ASHIVANI@SEAS.UPENN.EDU
Prathamesh Patil PPRATH@SEAS.UPENN.EDU
Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104

Editors: Vitaly Feldman, Katrina Ligett and Sivan Sabato

Abstract

The dueling bandits problem has received a lot of attention in recent years due to its applications in recommendation systems and information retrieval. However, due to the prevalence of malicious users in these systems, it is becoming increasingly important to design dueling bandit algorithms that are robust to corruptions introduced by these malicious users. In this paper we study dueling bandits in the presence of an adversary that can corrupt some of the feedback received by the learner. We propose an algorithm for this problem that is agnostic to the amount of corruption introduced by the adversary: its regret degrades gracefully with the amount of corruption, and in case of no corruption, it essentially matches the optimal regret bounds achievable in the purely stochastic dueling bandits setting.

Keywords: Dueling Bandits, Robustness, Adversarial Corruptions, Multi-armed Bandits

1. Introduction

In the dueling bandits problem there are K arms; in each trial, the learner plays a pair of arms and receives relative feedback between these arms drawn (stochastically) according to an underlying pairwise preference model (Yue and Joachims, 2011). This problem has gained a lot of attention in recent years due to its numerous applications in practical settings such as recommendation systems and information retrieval, where one does not have access to absolute feedback on individual arms, but rather can observe relative feedback gathered implicitly from users through clicks, reviews, ratings etc (Yue et al., 2009; Yue and Joachims, 2011; Yue et al., 2012; Urvoy et al., 2013; Ailon et al., 2014; Zoghi et al., 2014, 2015a,b; Dudik et al., 2015; Jamieson et al., 2015; Komiyama et al., 2015a, 2016; Ramamohan et al., 2016; Chen and Frazier, 2017).

In practice, due to the wide-spread prevalence of click-fraud, and manufactured reviews in recommendation systems and information retrieval, it is becoming increasingly important to guard against adversarial attacks on machine learning algorithms that interact with these systems. However, existing algorithms for dueling bandits are not robust to such adversarial attacks and can fail remarkably in the presence of even small amounts of corruption in the feedback (see Appendix C for more details). With this motivation in mind, we consider the design of algorithms for dueling bandits that are robust to adversarial corruption in the pairwise feedback.

Previously, Gajane et al. (2015) have studied a purely non-stochastic/adversarial version of the dueling bandits problem, however, the regret achievable in this purely adversarial setting is $O(\sqrt{T})$, which can be considerably larger than the distribution dependent $O(\log T)$ regret achievable in the usual stochastic setting. In this paper we ask the following question: can we achieve regret better than $O(\sqrt{T})$ in a setting where most of the pairwise feedback is stochastic, with potentially a small amount of adversarial corruption? In other words can we achieve distribution dependent $O(\log T)$

regret bounds that degrade gracefully with the amount of corruption, while being agnostic to the amount of corruption in the received feedback? Our work is also motivated by progress in classical stochastic bandits in the presence of adversarial corruption, where similar results have recently been achieved (Gupta et al., 2019; Lykouris et al., 2018).

1.1. Problem (Informal)

We consider a setting where there are K arms with an unknown pairwise preference matrix $\boldsymbol{\mu} \in [0,1]^{K \times K}$, where μ_{ij} specifies the probability that arm $i \in [K]$ is preferred to arm $j \in [K]$ in a pairwise comparison between arms i and j. Similar to prior work on dueling bandits, we assume the existence of a *Condorcet winner* (or unique best arm), denoted by $i^* \in [K]$, which is an arm that is preferred to every other arm with probability at least $1/2$ (Urvoy et al., 2013; Zoghi et al., 2014). In each trial t, nature draws the outcome of a pairwise comparison between each pair of arms $i, j \in [K] \times [K]$ (independently) according to $\boldsymbol{\mu}$. Nature then reveals these outcomes to an adversary who can reverse the outcome for some pairs of their choosing (we present the formal corruption model in Section 2). Finally, the player pulls a pair of arms $(u^t, v^t) \in [K] \times [K]$ and observes the outcome of the comparison between them, that has potentially been reversed by the adversary. The goal of the player is to minimize the cumulative regret given by $\sum_{t=1}^{T} (\Delta_{u^t} + \Delta_{v^t})$ over a fixed time horizon T, where $\Delta_i := \mu_{i^* i} - \frac{1}{2}$ for any $i \in [K]$.

1.2. Overview of Results

The following theorem gives our main result:

Theorem 1 (Informal) *There exists an algorithm for dueling bandits with adversarial corruptions whose regret, with high probability, is upper bounded by*

$$O\left(\frac{K^2 C_{i^*}}{\Delta_{\min}}\right) + O\left(\sum_{i \neq i^*} \frac{K^2}{\Delta_i^2} \log \frac{K}{\Delta_i}\right) + \widetilde{O}\left(\sum_{i \neq i^*} \frac{\log T}{\Delta_i}\right)$$

where C_{i^*} is the total number of pairs[1] that include the Condorcet winner arm i^* whose outcomes were adversarially reversed across the entire time horizon, and $\Delta_{\min} := \min_{i \neq i^*} \Delta_i$.

The above theorem shows that the regret of our algorithm is upper bounded by the sum of three terms – the first one depends on the amount of corruption introduced by the adversary, and the last two are comparable to the regret bounds achieved in the completely stochastic setting, for example Zoghi et al. (2014); Komiyama et al. (2015b). Hence, if the amount of corruption introduced by the adversary is zero, then one can recover an asymptotically optimal regret bound that is comparable to the one achievable in the usual stochastic dueling bandits setting. Moreover, the regret degrades linearly with the total amount of corruption introduced by the adversary over the entire time horizon (in Section 5 we also show that such a linear dependence on C_{i^*} is necessary). Most notably, our algorithm does not need to know the value of C_{i^*} ahead of time, and is able to adapt according to the adversary.

1. In particular, the regret achievable by our algorithm is completely independent of corruption introduced into pairs (i, j) of suboptimal arms where $i, j \neq i^*$. This weaker dependence on the corruption is not a-priori obvious, and may come as a surprise to some readers.

Our algorithm builds upon the framework of Gupta et al. (2019) developed for classical bandits with adversarial corruption which is based upon two key tenets: (1) no arm is completely eliminated from play during the execution of the algorithm, (2) the choice of arms to play in any trial is randomized. The former ensures that the adversary cannot simply force the best arm to get eliminated early with a small amount of corruption, causing the algorithm to incur constant regret for the rest of the play; and the latter ensures that the adversary cannot just corrupt the arm(s) that is (are) going to be played in the current round, again forcing constant regret with marginal amounts of corruption.

However, the dueling bandits framework has a fundamentally different structure than classical multi-armed bandits, and consequently requires fundamentally new algorithmic and analytical ideas. In particular, in classical bandits, there is always a total order over the arms (based on the mean rewards μ_i); this means that while playing a suboptimal arm i incurs some regret, it also provides helpful information to distinguish that arm (and possibly others) from the best arm i^*. In other words, one always receives useful feedback during exploration which can be used to improve play in future rounds. On the other hand, in dueling bandits with just a Condorcet winner, there is in general no total order over the arms, due to which playing suboptimal pairs of arms (i, j) that do not involve the Condorcet winner i^* incurs regret without providing any meaningful information about the Condorcet winner. Hence, the objective of minimizing regret due to sub-optimal arms becomes intertwined with the objective of isolating the Condorcet winner, and achieving these objectives simultaneously while allowing for adversarial corruption in addition to sampling noise becomes the central challenge in our setting.

Our proposed algorithm achieves both these objectives by simultaneously estimating the Condorcet winner arm as well as the optimal sampling rates (which are in turn tied to their gap parameters) for each arm, and iteratively refining these estimates over time. Precisely, it maintains a candidate Condorcet winner, which we call the *anchor arm*, and in each round, it samples a *random* arm to play against the anchor arm such that the probability of choosing a given arm is inversely proportional to its estimated gap. The gap of any arm i is then estimated to be the average disadvantage of arm i relative to an anchor arm across all rounds. However, as alluded to earlier, the key difficulty is that in addition to the effects of adversarial corruption, the feedback received in rounds when the anchor arm was not the Condorcet winner would also distort the estimated gaps for other arms, leading to increased regret in future rounds due to incorrect sampling rates.

We control the long term damage caused by adversarial corruptions by dividing the execution of the algorithm into intervals of geometrically increasing lengths (epochs), and using feedback received only from the most recent epoch to estimate the gap parameters of all arms. To account for incorrect identification of the Condorcet winner, we show that the feedback received in such rounds can be thought of as just another form of adversarial corruption, and if the number of such bad rounds can be bounded, then its impact on the total regret would also be small. In order to do so, we have an aggressive replacement strategy for the anchor arm, and one of our key results shows that across the execution of our algorithm, the total number of times the best arm would be replaced as the anchor arm by any bad arm, or fail to replace a bad anchor arm is bounded by a constant $O(KC_{i^*}/\Delta_{\min} + (K/\Delta_{\min})^2 \log(K/\Delta_{\min}))$. We couple this with showing that after sufficiently many rounds, the best arm would necessarily have the highest sampling rate, ensuring that in the event that it were to be replaced as the anchor arm by some suboptimal arm, it would quickly be sampled again, and would replace the bad anchor arm. Thus, the total number of rounds where the best arm is not the anchor arm is then bounded by a quantity which depends on the adversarial corruption.

1.3. Related Work

The dueling bandits problem was first proposed by Yue et al. (2009), motivated by practical applications where one receives relative feedback between arms rather than absolute feedback on individual arms. However, this early work assumed two additional conditions on the underlying pairwise preference matrix – strong stochastic transitivity (SST) and stochastic triangle inequality (STI), which are together much more restrictive than the Condorcet winner condition assumed by us (Yue et al., 2009; Yue and Joachims, 2011). Urvoy et al. (2013) were probably the first ones to consider the dueling bandits problem under the Condorcet winner setting, however, they provided a weak bound of $O(K^2 \log(T))$ for their algorithm. Zoghi et al. (2014) and Komiyama et al. (2015b) gave algorithms with an improved bound of $O(K^2 + K \log(T))$, which is comparable to the bound achieved by our algorithm in the case of no corruption. However, none of these existing algorithms are robust to adversarial corruption; in fact, we show in Appendix C that these algorithms can suffer from linear regret in the presence of just $O(\log T/\Delta_{\min})$ corruption. More recent work on dueling bandits has focused on relaxing the Condorcet winner condition further by considering more general winner concepts (Komiyama et al., 2016; Dudik et al., 2015; Jamieson et al., 2015; Ramamohan et al., 2016), however, these works are still restricted to the purely stochastic setting. Gajane et al. (2016) considered a purely adversarial *utility-based* dueling bandits problem and proposed an algorithm with a $O(\sqrt{T})$ regret bound. Contrary to Gajane et al. (2016), we consider a setting where most of the feedback is stochastic with a bounded number of adversarial corruptions.

Robustness to adversarial corruption has also been considered recently in classical stochastic bandits (Lykouris et al., 2018; Gupta et al., 2019). In particular, Lykouris et al. (2018) proposed a 'multi-layered' active arm elimination algorithm that achieved a $O(CK^2 \log T)$ regret bound. This result was subsequently improved to $O(K \log T + KC)$ by Gupta et al. (2019), who gave a novel algorithm that does not completely eliminate seemingly suboptimal arms but rather plays them a small number of times throughout the time horizon, giving them some recourse. However, as pointed out earlier, the dueling bandits framework is fundamentally different from classical bandits, due to which these results do not directly translate over to our setting. There has also been some work on *best-of-both-worlds* classical bandits where the goal is to achieve the minimum of the two bounds– the adversarial $O(\sqrt{T})$ bound and the stochastic $O(\log T)$ bound – depending on whether the realized instance is adversarial or stochastic while being agnostic to the instance (Zimmert et al., 2019; Zimmert and Seldin, 2019; Seldin and Slivkins, 2014; Seldin and Lugosi, 2017; Bubeck and Slivkins, 2012).

On a related note, the issue of robustness to adversarial corruptions in data has also recently gained interest in the rank aggregation community (Agarwal et al., 2020). Although the type of feedback/observations considered in their setting is the same as ours, namely outcomes of pairwise comparisons some of which may be adversarially corrupted, our goals are quite different; the objective considered in this work is minimizing cumulative regret over a fixed time horizon, whereas the work of Agarwal et al. (2020) largely concerns quantifying the sample complexity of estimating the parameters of the true underlying choice model (assumed to be the Bradley-Terry-Luce model (Bradley and Terry, 1952; Luce, 1959)) from offline comparison data.

More generally, the problem of robustness to adversarial perturbations has been of significant interest in statistics, starting with the work of Huber (1965, 1992). There have been significant advancements in this area since the pioneering work of Huber, the most recent and prominent one being the closing of the long standing gap between computationally efficient estimators and the

information theoretic limit for robust parameter estimation of Gaussians (Diakonikolas et al., 2017, 2018, 2019).

1.4. Organization

We give a formal problem definition in Section 2. We present our algorithm Section 3 and its regret analysis in Section 4. We provide a lower bound in Section 5, and finally present our conclusions in Section 6.

2. Problem Setting

We consider a K-armed dueling bandits problem, where we have a set of K arms indexed $1, 2, \ldots, K$. With each pair of arms i, j, there is associated an unknown probability μ_{ij}, which is the probability with which arm i beats arm j in a pairwise comparison between i, j. We assume that there exists a unique arm, which without loss of generality we assume to be arm 1, which beats every other arm with probability strictly more than $1/2$. Subsequently, for any arm $i \neq 1$, we define the gap $\Delta_i = \mu_{1i} - 1/2$. At every time step t, the player chooses to play a pair of arms $\{u^t, v^t\} \in [K] \times [K]$, and subsequently observes the outcome of a pairwise comparison between arms u^t, v^t, which can potentially be flipped by an adversary. The adversary is assumed to be computationally unbounded, as well as adaptive, i.e. its decisions to flip the outcomes of comparisons between pairs of arms at a given time step can be dependent on the players past choices, as well as the random outcomes of comparisons between all pairs of arms throughout the history of play, including the current time step. Since the adversary can be computationally unbounded, we can further assume the adversary to be deterministic without loss of generality. We describe the exact process as follows

1. At the current time step t, nature draws the outcomes of comparisons Z^t between all pairs of arms $i, j \in [K] \times [K]$, where $Z^t_{ij} = 1(i \succ j)$ is the outcome of a pairwise comparison between arms i, j, which is stochastic with probability μ_{ij}.

2. The adversary observes these outcomes Z^t, and potentially reverses some of them, producing corrupted outcomes $\widetilde{Z}^t$.

3. The player chooses to play a pair of arms $\{u^t, v^t\}$, and observes outcome $\widetilde{Z}^t_{u^t v^t}$.

Our objective is to minimize the strong regret incurred across T time steps, which is defined as

$$\mathcal{R}^T = \sum_{t=1}^{T} (\Delta_{u^t} + \Delta_{v^t})$$

Naturally, one would expect that the incurred regret depends on the total amount of corruption $C := \sum_{t=1}^{T} \sum_{i,j} |Z^t_{ij} - \widetilde{Z}^t_{ij}|$ introduced into the system by the adversary. Since the adversary is adaptive, the amount of corruption C is a random variable dependent on both the outcomes of comparisons between arms and the player's random choices. Hence, we do not assume that the player knows the value of C (or any bounds on C) at the start of play. Our objective is to design a robust strategy that is completely agnostic of C, and whose regret slowly increases as a function of C. However, as we shall soon see, our regret bounds in fact depend upon a quantity that is potentially much smaller than C. Specifically, our regret bounds depend on $C_1 := \sum_{t \in T} \sum_{i \neq 1} |Z^t_{i1} - \widetilde{Z}^t_{i1}|$

which is the amount of corruption in only the comparisons involving the Condorcet winner arm. This weaker dependence on the corruption is not a-priori obvious, and may come as a surprise to some readers.

Remark: Our corruption model is similar to that of Lykouris et al. (2018); Gupta et al. (2019) for the problem of classical stochastic bandits with adversarial corruption. The main difference is in the notion of the corruption level C: both Lykouris et al. (2018) and Gupta et al. (2019) define C in terms of the L_∞ norm difference between the corrupted outcome vector and the stochastic outcome vector in each trial, whereas we define it in terms of the L_1 norm difference between the corrupted outcome vector and the stochastic outcome vector.

3. Algorithm

Our proposed algorithm, termed Winner Isolation With Recourse (WIWR), is a randomized, non active-arm-elimination algorithm that simultaneously estimates the Condorcet winner arm as well as the gap parameter for each arm with respect to the Condorcet winner arm. At trial $t \in [T]$, we denote our estimate of the Condorcet winner by v^t and refer to it as the *anchor arm*. At trial t, the algorithm chooses the anchor arm v^t as the right arm, whereas the left arm is chosen *randomly* based on the estimates of the gap parameters.

The algorithm proceeds in epochs of geometrically increasing lengths [6]. At the start of each epoch m, the algorithm determines a sampling probability for each arm that is inversely proportional to the gap parameter estimated at the end of the previous epoch [5]. These sampling probabilities give a distribution over arms, which is then fixed throughout epoch m. At every time step t in epoch m, a left arm u^t is sampled from the aforementioned distribution over all arms [10], and subsequently played against the current anchor arm v^t [11] . If it is observed that across all comparisons between arms u^t and v^t, u^t had won a majority of them [18], the left arm u^t replaces the current anchor arm v^t to become the new anchor arm v^{t+1} [19]. Therefore, in each trial, the algorithm is randomized with respect to one arm of the pair, and deterministic with respect to the other.

Finally, at the end of epoch m, [28] for every arm i, the gap parameter Δ_i^m is estimated according to the average performance q_i^m of i against anchor arms. However, the exact computation of Δ_i^m [22-26] depends on whether there exists an arm j which was the anchor for more than half of the trials in epoch m. If there exists such an arm j [22], then Δ_i^m is estimated to be *lower confidence of the average disadvantage of arm i only when it was played as the left arm against anchor arm j* in epoch m [24,28]; and if there is no such arm in that epoch [25], then Δ_i^m is estimated to be the *lower confidence of the average disadvantage of arm i against anchor arms when arm i was played as the left arm* in epoch m [26,28]. The former case corresponds to a scenario where the algorithm has good confidence that arm j is the Condorcet winner as it is chosen as the anchor for more than half of the epoch, whereas the latter represents a scenario of uncertainty for the algorithm where it has to compute the gaps against an 'average' anchor arm. Furthermore, these gap parameters are only estimated to a precision of 2^{-m} in epoch m, which gives us this geometric nature of the lengths of the epochs. The pseudo-code for the algorithm is given in Algorithm 1.

4. Regret Analysis

The following theorem gives a high probability bound on the total regret incurred by our WIWR algorithm.

Algorithm 1 Winner Isolation With Recourse (WIWR)

Parameters: Confidence $\delta \in (0,1)$, time horizon T.
Initialize $T^0 = 0$, $v^0 \sim \text{Uniform}[K]$, $\Delta_i^0 = 1/2$ for all $i \in [K]$, $P_{ij} = 0$ for all $i, j \in [K] \times [K]$.
$\lambda \leftarrow 2304 \ln\left(\frac{40K^2}{\delta} \log T\right)$.
for Epochs $m = 1, 2, 3 \ldots$ **do**
  $n_i^m \leftarrow \lambda(\Delta_i^{m-1})^{-2}$ for all arms $i \in [K]$.
  $N^m \leftarrow \sum_{i \in [K]} n_i^m$, and $T^m = T^{m-1} + N^m$.
  $W_i^m \leftarrow 0$, $r_i^m \leftarrow 0$ for all arms $i \in [K]$, and $r_{ij}^m \leftarrow 0$ for all pairs $i, j \in [K] \times [K]$, $i \neq j$
  $\{W_i^m$: number of times arm i was the anchor arm,
    r_i^m: number of times arm i beat the anchor arm when i the left arm,
    r_{ij}^m: number of times arm i beat arm j when i was the left arm and j was the anchor arm$\}$
  for $t = T^{m-1} + 1$ to T^m **do**
    $W_{v^t}^m \leftarrow W_{v^t}^m + 1$.
    Sample $u^t \sim \text{Multinomial}(n_1^m, \ldots, n_K^m)/N^m$.
    Play pair (u^t, v^t), and observe outcome $\widetilde{Z}^t_{u^t v^t} = 1(u^t \succ v^t)$
    Update $P_{u^t v^t} \leftarrow P_{u^t v^t} + (2\widetilde{Z}^t_{u^t v^t} - 1)$, and $P_{v^t u^t} \leftarrow P_{v^t u^t} + (1 - 2\widetilde{Z}^t_{u^t v^t})$.
    if $u^t = v^t$ **then**
      $r_{u^t}^m \leftarrow r_{u^t}^m + 1/2$.
    else
      $r_{u^t v^t}^m \leftarrow r_{u^t v^t}^m + \widetilde{Z}^t_{u^t v^t}$, and $r_{u^t}^m \leftarrow r_{u^t}^m + \widetilde{Z}^t_{u^t v^t}$.
    end if
    if $P_{u^t v^t} > 0$ **then**
      Update $v^{t+1} \leftarrow u^t$.
    end if
  end for
  if $\exists\, j \in [K] : W_j^m \geq N^m/2$ **then**
    $w_j^m \leftarrow W_j^m / N^m$
    $q_i^m \leftarrow r_{ij}^m / (w_j^m n_i^m)$ for all $i \in [K] \setminus \{j\}$, and $q_j^m \leftarrow 1/2$.
  else
    $q_i^m \leftarrow r_i^m / n_i^m$ for all $i \in [K]$.
  end if
  For all $i \in [K]$, $\Delta_i^m \leftarrow \max\left\{2^{-m}, \frac{1}{2} - q_i^m - \frac{\Delta_i^{m-1}}{8}\right\}$
  If $\min_{j \in [K]} \Delta_j^m > 2^{-m}$, round $\Delta_i^m = 2^{-m}$ for an arbitrary $i \in \text{argmin}_{j \in [K]} \Delta_j^m$
end for

Theorem 1 *With probability at least $1 - \delta$, the regret of WIWR (Algorithm 1) is bounded by*

$$O\left(\left(\frac{C_1}{\Delta_{\min}} + \sum_{i \neq 1} \frac{1}{\Delta_i^2} \ln \frac{K}{\delta \Delta_i}\right) K^2 \ln \frac{1}{\delta} + \sum_{i \neq 1} \frac{\log T}{\Delta_i} \ln\left(\frac{K}{\delta} \log T\right)\right),$$

where K is the number of arms, T is the time-horizon, $\Delta_i = \mu_{1i} - \frac{1}{2}$, $\Delta_{\min} = \min_{i \neq 1} \Delta_{1i}$, and $C_1 := \sum_{t \in T} \sum_{i \neq 1} |Z_{i1}^t - \widetilde{Z}_{i1}^t|$ is the total amount of adversarial corruption in comparisons involving the Condorcet winner arm 1.

The proof of the above theorem is broadly structured into two main components. The first component is a high probability bound on the deviation in the fundamental random variables that determine the behavior of our algorithm, which include parameters such as the estimated gaps of all arms in any epoch, the actual number of plays of any arm as the left arm in any epoch, and the number of times any suboptimal arm beats the Condorcet winner arm. This is followed by a bound on the regret incurred by our algorithm assuming that the deviations in the aforementioned variables are small.

Notation. We begin by introducing some notation which will be used extensively in the analysis to follow. For any epoch m, arm i, we define $\tilde{n}_i^m$ to be the actual number of times arm i was sampled as the left arm in epoch m. For any time step t, and pair of arms i, j, we define ϵ_{ij}^t to be a random variable that is 1 if arms i, j were played at time t, i.e. either $u^t = i, v^t = j$ or $u^t = j, v^t = i$. Furthermore, for any time step t, and pair of arms i, j, we define c_{ij}^t to be a random variable that is 1 if the adversary flipped the stochastic outcome of a comparison between arms i, j at time t, and subsequently, for any epoch m, we define $C_{ij}^m = \sum_{t=T^{m-1}+1}^{T^m} c_{ij}^t$ to be the total number of times the adversary flipped the stochastic outcome of a comparison between arms i, j in epoch m. We denote $C_i^m = \sum_{t=T^{m-1}+1}^{T^m} \sum_{j \neq i} c_{ij}^t$ to be the total number of times the adversary flipped the stochastic outcome of any comparison involving arm i in epoch m. For any epoch m, arm i, we define $w_i^m = \sum_{t=T^{m-1}+1}^{T^m} 1(i = v^t)/N^m$ to be the fraction of times arm i was the anchor arm in that epoch. Finally, for any epoch m, we define $D^m = \sum_{t=T^{m-1}+1}^{T^m} 1(1 \neq v^t)$ to be the total number of times arm 1 was not the anchor arm in that epoch. Note that since we consider a fully adaptive adversary, all of these quantities are random variables, and all of our subsequent bounds depend on the realizations of these random variables.

First, observe that the epochs are of bounded lengths.

Lemma 2 *The length N^m of any epoch m is such that*

$$2^{2(m-1)} \leq \frac{N^m}{\lambda} \leq K 2^{2(m-1)}$$

As a consequence, the maximum number of epochs is bounded by $\log T$.

Proof Observe that in any epoch m, there exists at least one arm j for which $\Delta_j^m = 2^{-m}$, which is enforced by our rounding scheme at the end of the epoch. Due to this, $N^{m+1} \geq n_j^{m+1} = \lambda 2^{2m}$, proving the lower bound on the length of epoch $m+1$. The upper bound follows by observing that for any epoch m, $\Delta_i^m \geq 2^{-m}$ for all arms $i \in [K]$, due to which $n_i^{m+1} \leq \lambda 2^{2m}$ for all i. This gives us $N^{m+1} = \sum_{i \in [K]} n_i^{m+1} \leq K \lambda 2^{2m}$. This lower bound on the length of any epoch also implies the upper bound on the number of epochs; the final epoch M will have length at least 2^{2M-2}, which implies $M \leq \log 4T$. ■

We now define a set of *desirable* events that will be later shown to occur with high probability.

Definition 3 *For any given $0 < \delta \leq 1$, we define the following events:*

$$\mathcal{E}^0 := \forall m, i : \tilde{n}_i^m \leq 2n_i^m$$

$$\mathcal{E}^1 := \forall m, (i,j) : \left| \frac{r_{ij}^m}{n_i^m w_j^m} - \mu_{ij} \right| \leq \frac{2C_{ij}^m}{N^m} + \frac{\Delta_i^{m-1}}{8} \text{ or } w_j^m < \frac{1}{2}$$

$$\mathcal{E}^2 := \forall m, i : \left| \frac{r_i^m}{n_i^m} - \mu_{i1} \right| \leq \frac{2(C_{i1}^m + D^m)}{N^m} + \frac{\Delta_i^{m-1}}{8}$$

$$\mathcal{E}^3 := \forall m : \left(\frac{1}{2} + \frac{D^m}{N^m} \Delta_{\min} \right) - \frac{r_1^m}{n_1^m} \leq \frac{C_1^m}{N^m} + \frac{\Delta_i^{m-1}}{8}$$

$$\mathcal{E}^4 := \forall i \neq 1, s \leq T : \sum_{t=1}^{s} \epsilon_{i1}^t (2\tilde{z}_{i1}^t - 1) < 0 \text{ or } \sum_{t=1}^{s} \epsilon_{i1}^t \leq \frac{4C_{i1}}{\Delta_i} + \frac{2}{\Delta_i^2} \ln \frac{20K}{\delta \Delta_i^2}$$

Finally, we define event $\mathcal{E} := \{\mathcal{E}^0 \cap \mathcal{E}^1 \cap \mathcal{E}^2 \cap \mathcal{E}^3 \cap \mathcal{E}^4\}$ to be the intersection of all of the aforementioned events.

At a high level, event $\mathcal{E}^0$ guarantees that in every epoch, the number of times any arm was played as the left arm is not too much more than the expected number of times it was supposed to have been played. Events $\mathcal{E}^1, \mathcal{E}^2, \mathcal{E}^3$ are all guarantees on the deviation in the average reward (probability) estimates q_i^m of any arm in any epoch; $\mathcal{E}^1$ provides guarantees in the event that some arm j was the anchor arm for over half of the epoch, whereas $\mathcal{E}^2, \mathcal{E}^3$ provide a weaker, but more general guarantee, which is useful when no arm remained the anchor for a majority of the epoch. Finally, $\mathcal{E}^4$ guarantees that after we have seen sufficiently many comparisons between arm 1 and any other arm i, in the event that arm 1 is the anchor, and arm i is sampled as the left arm, arm i can never replace arm 1 as the anchor arm, and in the event that arm i is the anchor, and arm 1 is sampled as the left arm, arm i will always be replaced as the anchor arm by arm 1. This bound will be useful to prove that the total number of time steps where arm 1 was not the anchor arm is small, which is crucial as all the time steps where arm 1 was not the anchor not just incur large regret, but also affect the gap estimation for all other arms, which further increases the regret.

Lemma 4 *The event $\mathcal{E} := \{\mathcal{E}^0 \cap \mathcal{E}^1 \cap \mathcal{E}^2 \cap \mathcal{E}^3 \cap \mathcal{E}^4\}$ occurs with probability at least $1 - \delta/2$.*

The proof of this claim starts with the following tail bounds which will be used to bound the probabilities of the complement of each of the above stated events. The detailed proofs of the following tail bounds are technically involved, so in the interest of space, we defer them to Appendix B.

Lemma 5 *For any fixed epoch m, arm i, we have for any $\beta \geq e^{-\lambda/2}$*

$$\Pr(\tilde{n}_i^m \geq 2n_i^m) \leq \beta$$

Applying the above bound with $\beta = \delta/(10K \log T)$, followed by a union bound over all K arms and $\log T$ epochs gives us that $\Pr(\neg\, \mathcal{E}^0) \leq \delta/10$.

Lemma 6 *For any fixed epoch m and an ordered pair of unique arms (i,j), we have for any $\beta > 4e^{-\lambda/2304}$*

$$\Pr\left(\left| \frac{r_{ij}^m}{n_i^m w_j^m} - \mu_{ij} \right| \geq \frac{2C_{ij}^m}{N^m} + \frac{\Delta_i^{m-1}}{8} \text{ and } w_j^m \geq \frac{1}{2} \right) \leq \beta$$

Applying the above bound with $\beta = \delta/(10K^2 \log T)$, followed by a union bound over all $K(K-1)$ choices of unique ordered pairs of arms, and $\log T$ epochs gives us that $\Pr(\neg\, \mathcal{E}^1) \leq \delta/10$.

Lemma 7 *For any fixed epoch m, arm $i \neq 1$, we have for any $\beta > 4e^{-\lambda/2304}$*

$$\Pr\left(\left|\frac{r_i^m}{n_i^m} - p_{i1}\right| \geq \frac{C_{i1}^m + D^m}{N^m} + \frac{\Delta_i^{m-1}}{8}\right) \leq \beta$$

Applying the above bound with $\beta = \delta/(10K \log T)$, followed by a union bound over all $(K-1)$ suboptimal arms, and $\log T$ epochs gives us that $\Pr(\neg\, \mathcal{E}^2) \leq \delta/10$.

Lemma 8 *For any fixed epoch m, we have for any $\beta > 4e^{-\lambda/2304}$,*

$$\Pr\left(\left(\frac{1}{2} + \frac{D^m}{N^m}\Delta_{\min}\right) - \frac{r_1^m}{n_1^m} \geq \frac{C_1^m}{N^m} + \frac{\Delta_1^{m-1}}{8}\right) \leq \beta$$

Applying the above bound with $\beta = \delta/(10 \log T)$, followed by a union bound over all $\log T$ epochs gives us that $\Pr(\neg\, \mathcal{E}^3) \leq \delta/10$. One can verify that all substitutions of β are valid, since $\lambda = 2304 \ln((40K^2 \log T)/\delta)$.

Lemma 9 *We have for any fixed arm i, any $\beta > 0$,*

$$\Pr\left(\exists s \leq T : \sum_{t=1}^{s} \epsilon_{i1}^t (2\tilde{z}_{i1}^t - 1) > 0 \text{ and } \sum_{t=1}^{s} \epsilon_{i1}^t > \frac{4C_{i1}}{\Delta_i} + \frac{2}{\Delta_i^2} \ln \frac{2}{\beta \Delta_{i1}^2}\right) \leq \beta$$

Applying the above bound with $\beta = \delta/(10K)$, followed by a union bound over all K arms gives us that $\Pr(\neg\, \mathcal{E}^4) \leq \delta/10$.

A union bound over all of the above events gives us that $\Pr(\neg\, \mathcal{E}) \leq \delta/2$, implying that the event $\mathcal{E}$ occurs with probability at least $1 - \delta/2$, proving Lemma 4. From this point on, we shall condition on event $\mathcal{E}$.

The key idea behind the subsequent proof is that a given amount of corruption can distort the empirical probability estimates computed by our algorithm only by a bounded amount, as indicated by events $\mathcal{E}^1, \mathcal{E}^2,$ and $\mathcal{E}^3$. Therefore, once the epochs become long enough for the algorithm to collect sufficiently many uncorrupted samples relative to the total amount of corruption, its effect on the computed estimates will become negligible. Therefore, while the algorithm could incur arbitrarily large regret in the initial epochs where the effect of the corruption would be most pronounced, after sufficiently many epochs, the subsequent regret incurred will be small as the algorithm would have collected enough uncorrupted samples in each of the following epochs to compensate for the corruption. The epoch after which this property holds depends on the (realized) amount of total corruption involving the Condorcet winner arm (arm 1), and the algorithm is completely agnostic of this fact. Henceforth, we shall refer to this threshold epoch as $m_0 := \lceil (1/2) \log(8C_1/\lambda \Delta_{\min}) \rceil$.

We begin by showing that for all epochs m after the threshold epoch m_0, the gap estimate for arm 1 will always be good.

Lemma 10 *Conditioned on event $\mathcal{E}$, for any epoch $m \geq m_0$, we have that $\Delta_1^m = 2^{-m}$.*

Proof Observe that for any epoch $m \geq m_0$, we have that $N^m \geq \lambda 2^{2m-2} \geq 2C_1/\Delta_{\min}$. We have one of the following three possibilities: (1) Arm 1 is the anchor arm v for over half the epoch, (2) Some arm $j \neq 1$ is the anchor arm v for over half the epoch, or (3) No arm is the anchor arm v for over half the epoch. We shall prove that $\Delta_1^m = 2^{-m}$ in each of these three cases.

Case 1 (Arm 1 is the anchor arm v for over half the epoch). In this case, by definition of our algorithm, we have that $q_1^m = 1/2$ due to which

$$\Delta_1^m = \max\left\{2^{-m}, \frac{1}{2} - q_1^m - \frac{\Delta^{m-1}}{8}\right\} = \max\left\{2^{-m}, \frac{1}{2} - \frac{1}{2} - \frac{\Delta_i^{m-1}}{8}\right\} = 2^{-m}$$

Case 2 (Arm $i \neq 1$ is the anchor arm v for over half the epoch). In this case, by event $\mathcal{E}^1$, we have that $q_1^m = r_{1i}^m/(w_i^m n_1^m) \geq p_{1i} - 2C_{i1}^m/N^m - \Delta_1^{m-1}/8 \geq 1/2 + \Delta_{\min} - 2C_1/N^m - \Delta_1^{m-1}/8$. Thus, we have

$$\begin{aligned}
\Delta_1^m &= \max\left\{2^{-m}, \frac{1}{2} - q_1^m - \frac{\Delta_1^{m-1}}{8}\right\} \\
&\leq \max\left\{2^{-m}, \frac{1}{2} - \left(\frac{1}{2} + \Delta_{\min} - \frac{2C_1}{N^m} - \frac{\Delta_1^{m-1}}{8}\right) - \frac{\Delta_1^{m-1}}{8}\right\} \\
&\leq \max\{2^{-m}, -\Delta_{\min} + \Delta_{\min}\} = 2^{-m}
\end{aligned}$$

Case 3 (No arm is the anchor arm v for over half the epoch). Observe that when this event occurs, we have that $D^m/N^m > 1/2$. In this case, by event $\mathcal{E}^3$, we have that $r_1^m \geq 1/2 + D^m\Delta_{\min}/N^m - C_1^m/N^m - \Delta_1^{m-1}/8 \geq 1/2 + \Delta_{\min}/2 - C_1/N^m - \Delta_1^{m-1}/8$. Thus, we have

$$\begin{aligned}
\Delta_1^m &= \max\left\{2^{-m}, \frac{1}{2} - r_1^m - \frac{\Delta_1^{m-1}}{8}\right\} \\
&\leq \max\left\{2^{-m}, \frac{1}{2} - \left(\frac{1}{2} + \frac{\Delta_{\min}}{2} - \frac{C_1}{N^m} - \frac{\Delta_1^{m-1}}{8}\right) - \frac{\Delta_1^{m-1}}{8}\right\} \\
&\leq \max\left\{2^{-m}, \frac{-\Delta_{\min} + \Delta_{\min}}{2}\right\} = 2^{-m}
\end{aligned}$$

where the second to last inequality in both Cases 2 and 3 follows from the fact that $N^m \geq 2C_1/\Delta_{\min}$. Thus, we have that $\Delta_1^m = 2^{-m}$ for all epochs $m \geq m_0$. ■

Next, we use the above guarantee to both upper and lower bound the gap estimates Δ_i^m for every arm $i \neq 1$ for all epochs $m \geq m_0$.

Lemma 11 *Conditioned on event $\mathcal{E}$, for any epoch $m \geq m_0$, for all arms $i \neq 1$, we have*

$$\frac{3}{4}\Delta_i - \frac{2(C_{i1}^m + D^m)}{N^m} - \frac{C_{i1}^{m-1} + D^{m-1}}{2N^{m-1}} - \frac{2^{-m}}{2} \leq \Delta_i^m \leq \Delta_i + \frac{2(C_{i1}^m + D^m)}{N^m} + 2^{-m}$$

Proof We first establish the upper bound in our claim. Conditioned on event $\mathcal{E}$, we begin by observing that for any epoch m, the estimate q_i^m is always lower bounded by

$$q_i^m \geq \mu_{i1} - \frac{2(C_{i1}^m + D^m)}{N^m} - \frac{\Delta_i^{m-1}}{8}$$

To see this, observe that in epoch m, one of the following three events must occur: (1) arm 1 is the anchor arm for over half the epoch, in which case $q_i^m = r_{i1}^m/(w_1^m n_i^m)$, (2) some arm $j \neq 1$ is the anchor arm for over half the epoch, in which case $q_i^m = r_{ij}^m/(w_j^m n_i^m)$, or (3) no arm is the anchor arm for over half the epoch, in which case $q_i^m = r_i^m/n_i^m$. In case (1), our claimed lower bound is guaranteed by event $\mathcal{E}^2$, whereas in cases (2), (3), we must have $D^m/N^m > 1/2$ by definition, which corresponds to the trivial lower bound of 0. Therefore, we have that

$$\Delta_i^m = \max\left\{2^{-m}, \frac{1}{2} - q_i^m - \frac{\Delta_i^{m-1}}{8}\right\} \leq 2^{-m} + \frac{1}{2} - q_i^m - \frac{\Delta_i^{m-1}}{8} \leq 2^{-m} + \Delta_i + \frac{2(C_i^m + D^m)}{N^m}$$

proving the upper bound claimed in Lemma 11, where the final inequality follows by our aforementioned lower bound on q_i^m, and by definition of μ_{i1}.

We now use this upper bound on all Δ_i^m to establish the lower bound. Observe that for any epoch m, the estimate q_i^m is always upper bounded by

$$q_i^m \leq \mu_{i1} + \frac{2(C_{i1}^m + D^m)}{N^m} + \frac{\Delta_i^{m-1}}{8}$$

The proof of this follows in an identical manner as the proof of our lower bound; in case (1), the upper bound is guaranteed by event $\mathcal{E}^2$, whereas in cases (2),(3), we must have $D^m/N^m > 1/2$ which corresponds to the trivial upper bound of 1. By Lemma 10, for all epochs $m \geq m_0$, we have $\Delta_1^m = 2^{-m}$ due to which the rounding step is not applied, and the estimate

$$\begin{aligned}
\Delta_i^m &= \max\left\{2^{-m}, \frac{1}{2} - q_i^m - \frac{\Delta_i^{m-1}}{8}\right\} \\
&\geq \frac{1}{2} - \left(\frac{1}{2} - \Delta_i + \frac{2(C_{i1}^m + D^m)}{N^m} + \frac{\Delta_i^{m-1}}{8}\right) - \frac{\Delta_i^{m-1}}{8} \\
&\geq \Delta_i - \frac{2(C_{i1}^m + D^m)}{N^m} - \frac{\Delta_i^{m-1}}{4} \\
&\geq \Delta_i - \frac{2(C_{i1}^m + D^m)}{N^m} - \frac{1}{4}\left(2^{-(m-1)} + \Delta_i + \frac{2(C_{i1}^{m-1} + D^{m-1})}{N^{m-1}}\right) \\
&\geq \frac{3}{4}\Delta_i - \frac{2(C_{i1}^m + D^m)}{N^m} - \frac{C_{i1}^{m-1} + D^{m-1}}{2N^{m-1}} - \frac{2^{-m}}{2}
\end{aligned}$$

proving the lower bound claimed in Lemma 11, where the second inequality follows by our aforementioned upper bound on q_i^m and definition of μ_{i1}, and the second to last inequality follows by substituting the upper bound on Δ_i^{m-1}. ■

We use the guarantee provided by Lemma 10 to also show that after epoch m_0, the total number of time steps where arm 1 is not the anchor arm is bounded by a constant with high probability.

Lemma 12 *Conditioned on event $\mathcal{E}$, we have with probability at least $1 - \delta/2$,*

$$\sum_{m > m_0} D^m \leq \left(\frac{4C_1}{\Delta_{\min}} + \sum_{i \neq 1} \frac{2}{\Delta_i^2} \ln \frac{10K}{\delta \Delta_i^2}\right) K \ln \frac{6}{\delta}$$

Proof First, observe that for all epochs $m > m_0$, we have that the sampling probability of arm 1 $p_1^m \geq 1/K$. This directly follows from Lemma 10, and the upper bound on N^m from Lemma 2. Let us call the event where arm 1 either gets replaced as the anchor arm by some other arm $i \neq 1$, or fails to replace an anchor arm $i \neq 1$ when it is sampled as the left arm as a *violation event*. Then a simple corollary of event $\mathcal{E}^4$ gives us that the total number of violation events across all T time steps is upper bounded by

$$\sum_{t=1}^{T} 1\left[(u^t = 1 \text{ or } v^t = 1) \text{ and } v^{t+1} \neq 1\right] \leq \sum_{i \neq 1} \frac{4C_{i1}}{\Delta_i} + \sum_{i \neq 1} \frac{2}{\Delta_i^2} \ln \frac{10K}{\delta \Delta_i^2} \leq \frac{4C_1}{\Delta_{\min}} + \sum_{i \neq 1} \frac{2}{\Delta_i^2} \ln \frac{10K}{\delta \Delta_i^2}$$

We shall refer to this upper bound on the total number of violation events as V. For each $i \in [V]$, we define a random variable x_i that counts the number of time steps until we see the first draw of arm 1 as the left arm following the i^{th} violation event. In the case that there are strictly fewer than V violation events, we set $x_i = 0$ for the remaining unrealized events. Since at every time step, arm 1 is independently sampled with probability $\geq 1/K$, we can treat $\sum_{i=1}^{V} x_i$ as a sequence of independent Bernoulli trials until we observe V successes, where the probability of success in each trial is $\geq 1/K$. Therefore,

$$\begin{aligned}
\Pr\left(\sum_{i=1}^{V} x_i > VK \ln \frac{6}{\delta}\right) &\leq \Pr\left(\sum_{i=1}^{V} x_i > \left(1 + \ln \frac{2}{\delta}\right) VK\right) \\
&\leq \Pr\left(\text{Bernoulli}\left(VK + VK \ln \frac{2}{\delta}, \frac{1}{K}\right) < V\right) \\
&\leq \Pr\left(\text{Bernoulli}\left(VK + VK \ln \frac{2}{\delta}, \frac{1}{K}\right) - (V + V \ln \frac{2}{\delta}) < -V \ln \frac{2}{\delta}\right) \\
&\leq \exp\left(\frac{-2V^2 \ln^2(2/\delta)}{VK + VK \ln(2/\delta)}\right) \\
&\leq \exp\left(\frac{-V \ln(2/\delta)}{K}\right)
\end{aligned}$$

which is at most $\delta/2$ since $V \geq K$, with the penultimate inequality following from a standard Hoeffding's bound. Once arm 1 is sampled as the left arm, it either replaces the anchor arm at that time step, or fails to replace it, leading to another violation event. In either case, we have that the total number of time steps following all violation events where arm 1 is not drawn as the left arm is bounded by $VK \ln 6/\delta$, and since there are at most V violation events, there are at most $V + VK \ln 6/\delta$ time steps in total where arm 1 is not the anchor arm, proving our claim. ■

Equipped with the upper and lower bounds on the gap estimates, and the upper bound on the number of time steps where arm 1 is not the anchor, we are ready to prove the regret bound of Theorem 1.

Proof (of Theorem 1) We begin by conditioning on both the event $\mathcal{E}$, as well as the event where the upper bound in Lemma 12 holds. We first show that the total regret incurred by our algorithm *after* epoch $m_0 + 2 := \lceil (1/2) \log(8C_1/\lambda \Delta_{\min}) \rceil + 2$ is bounded by

$$\sum_{m > m_0 + 2} \mathcal{R}^m = O\left(C_1 + \sum_{m > m_0} KD^m + \sum_{j \neq 1} \frac{\lambda \log T}{\Delta_j}\right)$$

To see this, we have that the regret incurred by our algorithm after epoch $m_0 + 2$

$$\begin{aligned}\sum_{m>m_0+2} \mathcal{R}^m &= \sum_{m>m_0+2} \sum_{t=T^{m-1}+1}^{T^m} (\Delta_{u^t} + \Delta_{v^t}) \\ &\leq \sum_{m>m_0+2} \left(D^m + \sum_{i\neq 1} \tilde{n}_i^m \Delta_i \right) \\ &\leq \sum_{m>m_0+2} D^m + 2 \sum_{m>m_0+2} \sum_{i\neq 1} n_i^m \Delta_i\end{aligned}$$

where the final inequality follows by event $\mathcal{E}^0$. We shall henceforth refer to the term $n_i^m \Delta_i = \mathcal{R}_i^m$, and we shall bound this term for each epoch, and each arm separately. Remember that $n_i^m = \lambda(\Delta_i^{m-1})^{-2}$. For a fixed epoch m, and arm i, we have the following three cases

Case 1: $\Delta_i \leq 4 \cdot 2^{-m}$
In this case, we have $\Delta_i^{m-1} \geq 2^{-(m-1)} \geq \Delta_i/2$ by definition of our algorithm, giving us

$$\mathcal{R}_i^m \leq \frac{4\Delta_i \lambda}{\Delta_i^2} \leq \frac{4\lambda}{\Delta_i}$$

Case 2: $\Delta_i > 4 \cdot 2^{-m}$ and $2(C_{i1}^{m-1} + D^{m-1})/N^{m-1} + (C_{i1}^{m-2} + D^{m-2})/(2N^{m-2}) \leq \Delta_i/4$
By the lower bound on Δ_i^{m-1} in Lemma 11, we have

$$\begin{aligned}\Delta_i^{m-1} &\geq \frac{3}{4}\Delta_i - \frac{2(C_{i1}^{m-1} + D^{m-1})}{N^{m-1}} - \frac{(C_{i1}^{m-2} + D^{m-2})}{2N^{m-2}} - \frac{2^{-(m-1)}}{2} \\ &\geq \frac{3}{4}\Delta_i - \frac{1}{4}\Delta_i - \frac{1}{4}\Delta_i \\ &\geq \frac{\Delta_i}{4}\end{aligned}$$

In this case, we have

$$\mathcal{R}_i^m \leq \frac{16\lambda\Delta_i}{(\Delta_i)^2} \leq \frac{16\lambda}{\Delta_i}$$

Case 3: $\Delta_i > 4 \cdot 2^{-m}$ and $2(C_{i1}^{m-1} + D^{m-1})/N^{m-1} + (C_{i1}^{m-2} + D^{m-2})/(2N^{m-2}) > \Delta_i/8$
In this case, we have

$$\frac{4}{2^m} < \Delta_i < \frac{16(C_{i1}^{m-1} + D^{m-1})}{N^{m-1}} + \frac{4(C_{i1}^{m-2} + D^{m-2})}{N^{m-2}}$$

Again, by definition of our algorithm, we have $\Delta_i^{m-1} \geq 2^{-(m-1)}$, from which we can upper bound the regret

$$\begin{aligned}\mathcal{R}_i^m &\leq \lambda \Delta_i 2^{2(m-1)} \\ &\leq \lambda 2^{2(m-1)} \left(\frac{16(C_{i1}^{m-1} + D^{m-1})}{N^{m-1}} + \frac{4(C_{i1}^{m-2} + D^{m-2})}{N^{m-2}} \right) \\ &\leq 64(C_{i1}^{m-1} + D^{m-1} + C_{i1}^{m-2} + D^{m-2})\end{aligned}$$

where the final inequality follows from the lower bound $N^m \geq \lambda 2^{2(m-1)}$. From Cases 1, 2 and 3, we get that for any epoch $m > m_0 + 2$, arm i,

$$\mathcal{R}_i^m \leq 64(C_{i1}^{m-1} + D^{m-1} + C_{i1}^{m-2} + D^{m-2}) + \frac{16\lambda}{\Delta_i}$$

Putting this together, we get that

$$\begin{aligned}
\sum_{m>m_0+2} \sum_{i\neq 1} \mathcal{R}_i^m &\leq 64 \sum_{m>m_0+2} \sum_{i\neq 1} (C_{i1}^{m-1} + D^{m-1} + C_{i1}^{m-2} + D^{m-2}) + \sum_{m>m_0+2} \sum_{i\neq 1} \frac{16\lambda}{\Delta_i} \\
&\leq 64 \sum_{m>m_0+2} (C_1^{m-1} + C_1^{m-2}) + 64K \sum_{m>m_0+2} (D^{m-1} + D^{m-2}) + \sum_{i\neq 1} \frac{4\lambda \log T}{\Delta_i} \\
&\leq 128C_1 + 128K \sum_{m>m_0} D^m + \sum_{i\neq 1} \frac{4\lambda \log T}{\Delta_i}
\end{aligned}$$

Thus, the total regret of our algorithm post epoch $m_0 + 2$ is

$$\sum_{m>m_0+2} \mathcal{R}^m = O\left(C_1 + \sum_{m>m_0} KD^m + \sum_{i\neq 1} \frac{\lambda \log T}{\Delta_i}\right)$$

All that remains is to bound the total regret incurred by our algorithm in the first $m_0 + 2$ epochs, which is at most the number of time steps in these epochs. Thus, we have

$$\sum_{m=1}^{m_0+2} \mathcal{R}^m \leq \sum_{m=1}^{m_0+2} N^m \leq K\lambda \sum_{m=1}^{m_0+2} 2^{2m} \leq K\lambda \sum_{n=1}^{2m_0+4} 2^n \leq \frac{256KC_1}{\Delta_{\min}}$$

Thus, the total regret incurred by our algorithm is bounded by

$$\mathcal{R} = O\left(\frac{KC_1}{\Delta_{\min}} + \sum_{m>m_0} KD^m + \sum_{i\neq 1} \frac{\lambda \log T}{\Delta_i}\right)$$

Substituting $\sum_{m>m_0} D^m$ with the bound in Lemma 12, we get our claimed bound on the total regret

$$\mathcal{R} = O\left(\frac{K^2 C_1}{\Delta_{\min}} \ln \frac{1}{\delta} + \sum_{i\neq 1} \frac{K^2}{\Delta_i^2} \ln \frac{K}{\delta \Delta_i} \ln \frac{1}{\delta} + \sum_{i\neq 1} \frac{\log T}{\Delta_i} \ln \frac{K \log T}{\delta}\right)$$

which holds when both event $\mathcal{E}$ occur, as well as the bound in Lemma 12 holds, which together holds with probability at least $(1 - \delta/2)^2 \geq 1 - \delta$.

■

5. Lower Bound

We further show that any algorithm for dueling bandits must incur a regret of $\Omega(C)$, even under a weaker non-adaptive adversary. Consider an instance of dueling bandits with two arms $1, 2$ where $\Pr(1 \succ 2) = 1/2 + \Delta$ for any $\Delta > 0$. Consider a randomized adversary that uses the following corruption mechanism: if adversary chooses to corrupt the outcomes at time step $t \in [T]$, then the corrupted feedback is chosen to be

$$\widetilde{Z}_{12}^t = Z_{12}^t(1 - y^t), \text{ where } y^t \sim \text{Bernoulli}\left(\frac{4\Delta}{1 + 2\Delta}\right)$$

Observe that across all time steps that the adversary chooses to corrupt, the distribution of $\widetilde{Z}_{12}$ is identical to the completely stochastic model with $\Pr(1 \succ 2) = 1/2 - \Delta$. For any choice of C, the adversary corrupts the outcomes of the comparisons between arms $1, 2$ for the first $O(C/\Delta)$ time steps, in which case, we have the realized amount of corruption to be $O(C)$ with high probability. It is now straightforward to see that any algorithm must incur $\Omega(C)$ regret in just these first $O(C/\Delta)$ time steps; if there is an algorithm that incurs $o(C)$ regret in the first $O(C/\Delta)$ time steps in the corrupted instance, then it must incur $\Omega(\Delta(C/\Delta)) - o(C) = \Omega(C)$ regret in just these time steps in an uncorrupted instance where the true underlying probability of arm 1 beating arm 2 is $\Pr(1 \succ 2) = 1/2 - \Delta$. This follows from the fact that in these first $O(C/\Delta)$ time steps, the instance with adversarial corruption is statistically identical to the uncorrupted instance with the roles of the two arms reversed.

6. Conclusion

We study the problem of robustness in stochastic dueling bandits, where we consider a powerful adversarial corruption model in which we allow some of the observed pairwise feedback to be corrupted by an adaptive and computationally unbounded adversary. Furthermore, we assume only the existence of Condorcet winner in our model, which is significantly weaker than some common assumptions in this literature such as SST, STI or BTL that impose an ordering over arms. In this setting, we design a novel algorithm that is agnostic to the amount of corruption introduced into the system by the adversary, which, with high probability, incurs regret $O(K^2C/\Delta_{\min} + \sum_{i \neq i^*}(K^2/\Delta_i)\ln(K/\Delta_i)) + \widetilde{O}(\sum_{i \neq i^*} 1/\Delta_i)$ without any prior knowledge of C, the realized amount of adversarial corruption in the feedback across the time horizon. Most notably, this bound is asymptotically optimal in the completely stochastic setting when there is no adversarial corruption, and degrades gracefully with increasing amounts of corruption. To the best of our knowledge, these are the first guarantees of this kind for dueling bandits.

Our work motivates some natural open problems: in contrast to the regret achievable in classical bandits with adversarial corruption, where the dependence on the adversarial corruption is $O(KC)$, our regret bounds of $O(K^2C/\Delta_{\min})$ are somewhat weaker. However, dueling bandits under just the Condorcet winner condition is a fundamentally more challenging setting than classical bandits, which motivates the question of whether this additional dependence on the minimal gap $\Delta_{\min}$ is inevitable, and whether our lower bound can be improved to reflect it. Secondly, is it possible to improve this dependence on the corruption term under more restrictive assumptions such as SST, STI or BTL on the preference matrix? Another interesting direction would be to extend our results to dueling bandits with other notions of "winner arms", such as Copeland winners, top cycles, Banks

sets etc., each of which have their own separate notion of regret. These in some sense are even more general settings of dueling bandits as, unlike Condorcet winners which may not exist, these notions of winners are guaranteed to exist in any preference matrix (Ramamohan et al. (2016)).

Acknowledgments

This material is based upon work supported in part by the US National Science Foundation (NSF) under Grant Nos. 1717290 and 1934876. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Arpit Agarwal, Shivani Agarwal, Sanjeev Khanna, and Prathamesh Patil. Rank aggregation from pairwise comparisons in the presence of adversarial corruptions. In *ICML*, 2020.

Nir Ailon, Zohar Karnin, and Thorsten Joachims. Reducing Dueling Bandits to Cardinal Bandits. In *Proceedings of the 31st International Conference on Machine Learning*, 2014.

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.

Sébastien Bubeck and Aleksandrs Slivkins. The best of both worlds: Stochastic and adversarial bandits. In *COLT 2012 - The 25th Annual Conference on Learning Theory, June 25-27, 2012, Edinburgh, Scotland*, pages 42.1–42.23, 2012.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Bangrui Chen and Peter I. Frazier. Dueling Bandits with Weak Regret. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Being robust (in high dimensions) can be practical. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 999–1008. JMLR. org, 2017.

Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robustly learning a gaussian: Getting optimal error, efficiently. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2683–2702. SIAM, 2018.

Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, 2019.

Miroslav Dudik, Katja Hofmann, Robert E. Schapire, Aleksandrs Slivkins, and Masrour Zoghi. Contextual Dueling Bandits. In *Proceedings of the 28th Conference on Learning Theory*, 2015.

David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

Pratik Gajane, Tanguy Urvoy, and Fabrice Clérot. A relative exponential weighing algorithm for adversarial utility-based dueling bandits. 2015.

Pratik Gajane, Tanguy Urvoy, and Emilie Kaufmann. Corrupt Bandits. In *13th European Workshop on Reinforcement Learning*, 2016.

Anupam Gupta, Tomer Koren, and Kunal Talwar. Better algorithms for stochastic bandits with adversarial corruptions. In *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, pages 1562–1578, 2019.

Peter J Huber. A robust version of the probability ratio test. *The Annals of Mathematical Statistics*, pages 1753–1758, 1965.

Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics*, pages 492–518. Springer, 1992.

Kevin Jamieson, Sumeet Katariya, Atul Deshpande, and Robert Nowak. Sparse Dueling Bandits. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, 2015.

Junpei Komiyama, Junya Honda, Hisashi Kashima, and Hiroshi Nakagawa. Regret Lower Bound and Optimal Algorithm in Dueling Bandit Problem. In *Proceedings of the 28th Conference on Learning Theory*, 2015a.

Junpei Komiyama, Junya Honda, and Hiroshi Nakagawa. Optimal Regret Analysis of Thompson Sampling in Stochastic Multi-armed Bandit Problem with Multiple Plays. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015b.

Junpei Komiyama, Junya Honda, and Hiroshi Nakagawa. Copeland Dueling Bandit Problem: Regret Lower Bound, Optimal Algorithm, and Computationally Efficient Algorithm. In *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

Robert Duncan Luce. *Individual choice behavior: A theoretical analysis*. Wiley, 1959.

Thodoris Lykouris, Vahab S. Mirrokni, and Renato Paes Leme. Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018, Los Angeles, CA, USA, June 25-29, 2018*, pages 114–122, 2018.

Siddartha Ramamohan, Arun Rajkumar, and Shivani Agarwal. Dueling Bandits : Beyond Condorcet Winners to General Tournament Solutions. In *Advances in Neural Information Processing Systems 29*, 2016.

Yevgeny Seldin and Gábor Lugosi. An improved parametrization and analysis of the EXP3++ algorithm for stochastic and adversarial bandits. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017, Amsterdam, The Netherlands, 7-10 July 2017*, pages 1743–1759, 2017.

Yevgeny Seldin and Aleksandrs Slivkins. One practical algorithm for both stochastic and adversarial bandits. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 1287–1295, 2014.

Tanguy Urvoy, Fabrice Clerot, Raphael Feraud, and Sami Naamane. Generic Exploration and K-armed Voting Bandits. In *Proceedings of the 30th International Conference on Machine Learning*, 2013.

Yisong Yue and Thorsten Joachims. Beat the mean bandit. In *Proceedings of the 28th International Conference on Machine Learning*, 2011.

Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The K-armed Dueling Bandits Problem. In *Proceedings of the 22nd Conference on Learning Theory*, 2009.

Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The k-armed dueling bandits problem. *J. Comput. Syst. Sci.*, 78(5):1538–1556, 2012.

Julian Zimmert and Yevgeny Seldin. An optimal algorithm for stochastic and adversarial bandits. In *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, pages 467–475, 2019.

Julian Zimmert, Haipeng Luo, and Chen-Yu Wei. Beating stochastic and adversarial semi-bandits optimally and simultaneously. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 7683–7692, 2019.

Masrour Zoghi, Shimon Whiteson, Remi Munos, and Maarten de Rijke. Relative Upper Confidence Bound for the K-Armed Dueling Bandit Problem. In *Proceedings of the 31st International Conference on Machine Learning*, 2014.

Masrour Zoghi, Zohar Karnin, Shimon Whiteson, and Maarten de Rijke. Copeland Dueling Bandits. In *Advances in Neural Information Processing Systems 28*, 2015a.

Masrour Zoghi, Shimon Whiteson, and Maarten de Rijke. MergeRUCB: A method for large-scale online ranker evaluation. In *Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, 2015b.

Appendix A. Concentration Inequalities

In this section, we record all of the concentration inequalities used in this this paper. These are all well known inequalities; see Cesa-Bianchi and Lugosi (2006) for the former two inequalities, and Freedman (1975) (Theorem 1.6) for the last inequality.

Theorem 13 (Multiplicative Chernoff Bounds) *Let $X_1, \ldots, X_n$ be independent Bernoulli random variables, with X denoting their sum, and $\mathbb{E}(X) = \mu$ denoting their mean. Then for any $\delta > 0$, we have that*

$$\Pr\left(|X - \mu| \geq \delta\mu\right) \leq 2\exp\left(\frac{-\delta^2\mu}{3}\right)$$

Theorem 14 (Bernstein's Martingale Inequality) *Let $X_1, \ldots, X_n$ be a bounded martingale difference sequence with respect to a certain filtration $\{\mathcal{F}_i\}_{0,\leq i<n}$, and with $|X_i| \leq K$. Let*

$$S_i = \sum_{j=1}^{i} X_j$$

be the associated martingale. Denote the sum of the conditional variances by

$$\Sigma_n^2 = \sum_{j=1}^{n} \mathbb{E}[X_j^2 | \mathcal{F}_{j-1}]$$

Then for all constants $t, v > 0$,

$$\Pr(S_n > \sqrt{2vt} + (\sqrt{2}/3)Kt \text{ and } \Sigma_n^2 \leq v) \leq e^{-t}$$

Theorem 15 (Freedman's Inequality) *Let $X_1, \ldots, X_n$ be a bounded martingale difference sequence with respect to a certain filtration $\{\mathcal{F}_i\}_{0,\leq i<n}$, and with $|X_i| \leq K$ almost surely. Let*

$$S_i = \sum_{j=1}^{i} X_j$$

be the associated martingale. Denote the sum of the conditional variances by

$$\Sigma_i^2 = \sum_{j=1}^{i} \mathbb{E}[X_j^2 | \mathcal{F}_{j-1}]$$

Then for all constants $t, v > 0$,

$$\Pr(\exists\, i \leq n : S_i > t \text{ and } \Sigma_i^2 \leq v) \leq \exp\left(\frac{-t^2}{2v + 2Kt/3}\right)$$

Appendix B. Proofs

B.1. Proof of Lemma 5

Lemma 5 *For any fixed epoch m, arm i, we have for any $\beta \geq e^{-\lambda/2}$*

$$\Pr(\tilde{n}_i^m \geq 2n_i^m) \leq \beta$$

Proof For any fixed epoch m, we first begin by conditioning on all the random variables prior to epoch m, due to which the quantities N^m, T^{m-1}, and n_i^m for all arms $i \in [K]$ become constants. For any epoch m, arm i, for every time step $t \in T^m$, let Y_i^t be an indicator for arm i being sampled as the left arm. Then we have that $\tilde{n}_i^m = \sum_{t \in T^m} Y_i^t$. Furthermore, we have that the expected value $\mathbb{E}(\tilde{n}_i^m) = n_i^m \geq \lambda$, and that all Y_i^t are independent random variables. A standard application of Chernoff Bounds gives us that the probability of the said event (conditioned on all random variables prior to epoch m) is at most $e^{-\lambda/2} \leq \beta$. Finally, observe that this bound holds for any realization of these random variables we conditioned on, due to which it also holds unconditionally. ■

B.2. Proof of Lemma 6

Lemma 6 *For any fixed epoch m and an ordered pair of unique arms (i, j), we have for any $\beta > 4e^{-\lambda/2304}$*

$$\Pr\left(\left|\frac{r_{ij}^m}{n_i^m w_j^m} - \mu_{ij}\right| \geq \frac{2C_{ij}^m}{N^m} + \frac{\Delta_i^{m-1}}{8} \text{ and } w_j^m \geq \frac{1}{2}\right) \leq \beta$$

Proof For any fixed epoch m, we first begin by conditioning on all the random variables prior to epoch m, due to which the quantities N^m, T^{m-1}, and n_i^m for all arms $i \in [K]$ become constants. For epoch m, let $E^m = [T^{m-1}+1, \ldots, T^m]$ be the N^m time steps within that epoch. For the fixed arm i, anchor arm j, for every time step $t \in E^m$, let ϵ_j^t be an indicator for the anchor arm $v^t = j$, and Y_i^t be an indicator for the left arm $u^t = i$. Furthermore, let Z_{ij}^t be the stochastic outcome of the comparison $1(i \succ j)$, and let $c_{ij}^t = \widetilde{Z}_{ij}^t - Z_{ij}^t$ be the corruption introduced by the adversary. Thus, we have that the relative reward r_{ij}^m of arm i with respect to anchor arm j in epoch m as

$$r_{ij}^m = \sum_{t \in E^m} Y_i^t \epsilon_j^t (Z_{ij}^t + c_{ij}^t)$$

Let

$$A_{ij}^m = \sum_{t \in E^m} Y_i^t \epsilon_j^t Z_{ij}^t, \quad B_{ij}^m = \sum_{t \in E^m} Y_i^t \epsilon_j^t c_{ij}^t$$

To analyze the deviation in A_{ij}^m, consider the sequence of random variables $X^1, \ldots, X^{N^m}$, where $X^t = \epsilon_j^t (Y_i^t Z_{ij}^t - p_i^m \mu_{ij})$ for any $t \in E^m$. Then $\{X^t\}_{t=1}^{N^m}$ is a martingale with respect to the filtration $\{\mathcal{F}^t\}_{t=1}^{N^m}$ generated by the random variables $\{Y_j^s\}_{j \in [K], s \leq t}$, $\{Z_{ij}^s\}_{i,j \in [K] \times [K], s \leq t+1}$, and $\{\epsilon_j^s\}_{j \in [K], s \leq t+1}$. This is because conditioned on $\mathcal{F}^{t-1}$, Y_i^t and Z_{ij}^t are independent random variables with mean p_i^m, and μ_{ij} respectively and ϵ_j^t becomes a deterministic quantity which is 1 iff the anchor arm $v^t = j$. Thus,

$$\mathbb{E}[X^t | \mathcal{F}^{t-1}] = \epsilon_j^t \mathbb{E}[Y_i^t Z_{ij}^t - p_i^m \mu_{ij}] = 0$$

Next, we can bound the conditional variance

$$\Sigma = \sum_{t \in E^m} \mathbb{E}[(X^t)^2 | \mathcal{F}^{t-1}] = \sum_{t \in E^m} \epsilon_j^t \mathbb{E}[(Y_i^t Z_{ij}^t - p_i^m \mu_{ij})^2 | \mathcal{F}^{t-1}] \leq W_j^m p_i^m = w_j^m n_i^m \leq n_i^m$$

Observe that

$$\sum_{t \in E^m} X^t = A_{ij}^m - W_j^m p_i^m \mu_{ij} = A_{ij}^m - w_j^m n_i^m \mu_{ij}$$

We have that

$$\begin{aligned}
&\Pr\left(\left|\frac{A_{ij}^m}{w_j^m n_i^m} - \mu_{ij}\right| \geq \sqrt{\frac{8 \ln 4/\beta}{n_i^m}} + \frac{\sqrt{8} \ln 4/\beta}{3 n_i^m} \text{ and } w_j^m \geq 1/2\right) \\
&= \Pr\left(\left|\sum_{t \in T^m} X^t\right| \geq w_j^m \sqrt{8 n_i^m \ln 4/\beta} + \frac{\sqrt{8} w_j^m \ln 4/\beta}{3} \text{ and } w_j^m \geq 1/2\right) \\
&\leq \Pr\left(\left|\sum_{t \in T^m} X^t\right| \geq \sqrt{2 n_i^m \ln 4/\beta} + \frac{\sqrt{2} \ln 4/\beta}{3} \text{ and } \Sigma \leq n_i^m\right) \\
&\leq \beta/2
\end{aligned}$$

Where the final inequality follows by Bernstein's inequality for Martingales. Finally, since we have that $n_i^m \geq \lambda \geq 2304 \ln 4/\beta$, we have that

$$\Pr\left(\left|\frac{A_{ij}^m}{w_j^m n_i^m} - \mu_{ij}\right| \geq \sqrt{\frac{9\ln 4/\beta}{n_m^i}} \text{ and } w_j^m \geq 1/2\right) \leq \beta/2$$

To bound the deviation in B_{ij}^m, consider the sequence of random variables $X^1, \ldots, X^{N^m}$, where $X^t = \epsilon_j^t c_{ij}^t (Y_i^t - p_i^m)$ for any $t \in E^m$. Then $\{X^t\}_{t=1}^{N^m}$ is a martingale with respect to the filtration $\{\mathcal{F}^t\}_{t=1}^{N^m}$ generated by the random variables $\{Y_j^s\}_{j\in[K],s\leq t}$,$\{Z_{ij}^s\}_{i,j\in[K]\times[K],s\leq t+1}$, and $\{\epsilon_j^s\}_{j\in[K],s\leq t+1}$. This is because conditioned on $\mathcal{F}^{t-1}$, Y_i^t is an independent random variable with mean p_i^m, and ϵ_j^t, c_{ij}^t becomes deterministic quantities. Thus,

$$\mathbb{E}[X^t|\mathcal{F}^{t-1}] = \epsilon_j^t c_{ij}^t \mathbb{E}[Y_i^t - p_i^m] = 0$$

Next, we can bound the conditional variance

$$\Sigma = \sum_{t\in E^m} \mathbb{E}[(X^t)^2|\mathcal{F}^{t-1}] = \sum_{t\in E^m} \epsilon_j^t \left|c_{ij}^t\right| \mathbb{E}[(Y_i^t - p_i^m)^2|\mathcal{F}^{t-1}] \leq p_i^m \sum_{t\in E^m} \left|c_{ij}^t\right| \epsilon_j^t \leq n_i^m$$

Observe that

$$\sum_{t\in E^m} X^t = B_{ij}^m - p_i^m \sum_{t\in E^m} \epsilon_j^t c_{ij}^t$$

We have that

$$\begin{aligned}
&\Pr\left(\frac{B_{ij}^m}{w_j^m n_i^m} \geq \frac{2C_{ij}^m}{N^m} + \sqrt{\frac{8\ln 4/\beta}{n_i^m}} + \frac{\sqrt{8}\ln 4/\beta}{3n_i^m} \text{ and } w_j^m \geq 1/2\right) \\
&\leq \Pr\left(\frac{B_{ij}^m}{w_j^m n_i^m} - \frac{\sum_{t\in E^m}\epsilon_j^t c_{ij}^t}{w_j^m N^m} \geq \sqrt{\frac{8\ln 4/\beta}{n_i^m}} + \frac{\sqrt{8}\ln 4/\beta}{3n_i^m} \text{ and } w_j^m \geq 1/2\right) \\
&= \Pr\left(\sum_{t\in E^m} X^t \geq w_j^m\sqrt{8n_i^m \ln 4/\beta} + \frac{\sqrt{8}w_j^m \ln 4/\beta}{3} \text{ and } w_j^m \geq 1/2\right) \\
&\leq \Pr\left(\sum_{t\in E^m} X^t \geq \sqrt{2n_i^m \ln 4/\beta} + \frac{\sqrt{2}\ln 4/\beta}{3} \text{ and } \Sigma \leq n_i^m\right) \\
&\leq \beta/4
\end{aligned}$$

Where the final inequality follows by Bernstein's inequality for Martingales. Finally, since we have that $n_i^m \geq \lambda \geq 2304 \ln 4/\beta$, we have that

$$\Pr\left(\frac{B_{ij}^m}{w_j^m n_i^m} \geq \frac{2C_{ij}^m}{N^m} + \sqrt{\frac{9\ln 4/\beta}{n_i^m}} \text{ and } w_j^m \geq 1/2\right) \leq \beta/4$$

We can similarly bound $-B_{ij}^m/w_j^m n_i^m$, giving a final absolute deviation bound on B as

$$\Pr\left(\left|\frac{B_{ij}^m}{w_j^m n_i^m}\right| \geq \frac{2C_{ij}^m}{N^m} + \sqrt{\frac{9\ln 4/\beta}{n_i^m}} \text{ and } w_j^m \geq 1/2\right) \leq \beta/2$$

Combining the deviation bounds on A_{ij}^m, B_{ij}^m through a union bound, we have that

$$\Pr\left(\left|\frac{r_{ij}^m}{n_i^m w_j^m} - \mu_{ij}\right| \geq \frac{2C_{ij}^m}{N^m} + \sqrt{\frac{36\ln 4/\beta}{n_i^m}} \text{ and } w_j^m \geq \frac{1}{2}\right) \leq \beta$$

Following the observation that $n_i^m = \lambda/(\Delta_i^{m-1})^2$, and $\lambda \geq 2304 \ln 4/\beta$, we have that the probability of the said event (conditioned on all random variables prior to epoch m) is at most β. Finally, observe that this bound holds for any realization of these random variables we conditioned on, due to which it also holds unconditionally, hence proving our claim. ■

B.3. Proof of Lemma 7

Lemma 7 *For any fixed epoch m, arm $i \neq 1$, we have for any $\beta > 4e^{-\lambda/2304}$*

$$\Pr\left(\left|\frac{r_i^m}{n_i^m} - \mu_{i1}\right| \geq \frac{C_{i1}^m + D^m}{N^m} + \frac{\Delta_i^{m-1}}{8}\right) \leq \beta$$

Proof For any fixed epoch m, we first begin by conditioning on all the random variables prior to epoch m, due to which the quantities N^m, T^{m-1}, and n_i^m for all arms $i \in [K]$ become constants. For epoch m, let $E^m = [T^{m-1}+1, \ldots, T^m]$ be the N^m time steps within that epoch. For every time step $t \in E^m$, let Y_i^t be an indicator for the left arm $u^t = i$. Furthermore, let Z_{i1}^t be the stochastic outcome of the comparison $1(i \succ 1)$, and let $c_i^t = \widetilde{Z}_{iv^t}^t - Z_{i1}^t$ be the corruption, either introduced by the adversary or caused by arm i playing against an anchor arm $v^t \neq 1$ (we shall simply treat the latter as another source of adversarial error). Then we have

$$r_i^m = \sum_{t \in E^m} Y_i^t (Z_{i1}^t + c_i^t)$$

We define

$$A_i^m := \sum_{t \in E^m} Y_i^t Z_{i1}^t \quad \text{and} \quad B_i^m := \sum_{t \in E^m} Y_i^t c_i^t$$

To bound the deviation in A_i^m, we have that Z_{i1}^t is an independent $\{0,1\}$ random variable with mean μ_{i1}, and Y_i^t is an independent $\{0,1\}$ random variable with mean p_i^m. Thus, we have $\mathbb{E}(A_i^m) = n_i^m \mu_{i1}$. Thus, by a standard multiplicative Chernoff bound, we have,

$$\Pr\left(\left|\frac{A_i^m}{n_i^m} - \mu_{i1}\right| \geq \sqrt{\frac{3\ln 4/\beta}{n_i^m}}\right) = \Pr\left(|A_i^m - \mathbb{E}(A_i^m)| \geq \sqrt{\frac{3\mathbb{E}(A_i^m)\ln 4/\beta}{\mu_{i1}}}\right) \leq \beta/2$$

To bound B_i^m, consider the sequence of random variables $X^1, \ldots, X^{N^m}$, where $X^t := (Y_i^t - p_i^m)c_i^t$ for all t. Then $\{X^t\}_{t=1}^{T^m}$ is a martingale difference sequence w.r.t. the filtration $\{\mathcal{F}^t\}_{t=1}^{N^m}$ generated by r.v. $\{Y_j^s\}_{j\in[K],s\leq t}, \{Z_{ij}^s\}_{i,j\in[K]\times[K],s\leq t+1}$. This is because the corruption c_i^t becomes a deterministic value conditioned on $\mathcal{F}^{t-1}$, due to which

$$\mathbb{E}(X^t|\mathcal{F}^{t-1}) = \mathbb{E}(Y_i^t - p_i^m|\mathcal{F}^{t-1})c_i^t = 0$$

Next, we can bound the conditional variance

$$\Sigma = \sum_{t \in E^m} \mathbb{E}[(X^t)^2 | \mathcal{F}^{t-1}] = \sum_{t \in E^m} |c_i^t| \mathbb{E}[(Y_i^t - p_i^m)^2 | \mathcal{F}^{t-1}] \leq p_i^m \sum_{t \in E^m} |c_i^t| \leq n_i^m$$

Observe that

$$\sum_{t \in E^m} X^t = B_i^m - p_i^m \sum_{t \in E^m} c_i^t$$

We have that

$$\begin{aligned}
&\Pr\left(\frac{B_i^m}{n_i^m} \geq \frac{C_{i1}^m + D^m}{N^m} + \sqrt{\frac{2 \ln 4/\beta}{n_i^m}} + \frac{\sqrt{2} \ln 4/\beta}{3 n_i^m} \right) \\
&\leq \Pr\left(\frac{B_i^m}{n_i^m} - \frac{\sum_{t \in T^m} c_i^t}{N^m} \geq \sqrt{\frac{2 \ln 4/\beta}{n_i^m}} + \frac{\sqrt{2} \ln 4/\beta}{3 n_i^m} \right) \\
&\leq \Pr\left(\sum_{t \in T^m} X^t \geq \sqrt{2 n_i^m \ln 4/\beta} + \frac{\sqrt{2} \ln 4/\beta}{3} \text{ and } \Sigma \leq n_i^m \right) \\
&\leq \beta/4
\end{aligned}$$

where the second inequality follows by observing that $\sum_{t \in E^m} c_i^m \leq C_{i1}^m + D^m$, and the final inequality follows by Bernstein's inequality for Martingales. Finally, since we have that $n_i^m \geq \lambda \geq 2304 \ln 4/\beta$, we have that

$$\Pr\left(\frac{B_i^m}{n_i^m} \geq \frac{C_{i1}^m + D^m}{N^m} + \sqrt{\frac{3 \ln 4/\beta}{n_i^m}} \right) \leq \beta/4$$

We can similarly bound $-B_i^m/n_i^m$, giving a final absolute deviation bound on B_i^m as

$$\Pr\left(\left| \frac{B_i^m}{n_i^m} \right| \geq \frac{C_{i1}^m + D^m}{N^m} + \sqrt{\frac{3 \ln 4/\beta}{n_i^m}} \right) \leq \beta/2$$

Combining the above guarantees through a union bound, we have that

$$\Pr\left(\left| \frac{r_i^m}{n_i^m} - \mu_{i1} \right| \geq \frac{C_{i1}^m + D^m}{N^m} + \sqrt{\frac{12 \ln 4/\beta}{n_i^m}} \right) \leq \beta$$

Following the observation that $n_i^m = \lambda/(\Delta_i^{m-1})^2$, and $\lambda \geq 2304 \ln 4/\beta$, we have that the probability of the said event (conditioned on all random variables prior to epoch m) is at most β. Finally, observe that this bound holds for any realization of these random variables we conditioned on, due to which it also holds unconditionally, hence proving our claim. ■

B.4. Proof of Lemma 8

Lemma 8 *For any fixed epoch m, we have for any $\beta > 4e^{-\lambda/2304}$,*

$$\Pr\left(\left(\frac{1}{2}+\frac{D^m}{N^m}\Delta_{\min}\right)-\frac{r_1^m}{n_1^m}\geq\frac{C_1^m}{N^m}+\frac{\Delta_1^{m-1}}{8}\right)\leq\beta$$

Proof For any fixed epoch m, we first begin by conditioning on all the random variables prior to epoch m, due to which the quantities N^m, T^{m-1}, and n_i^m for all arms $i \in [K]$ become constants. For epoch m, let $E^m = [T^{m-1}+1, \ldots, T^m]$ be the N^m time steps within that epoch. For every time step $t \in E^m$, let Y_1^t be an indicator for the left arm $u^t = 1$. Furthermore, let Z_{1j}^t be the stochastic outcome of the comparison $1(1 \succ j)$, and let $c_{1j}^t = \widetilde{Z}_{1j}^t - Z_{1j}^t$ be the corruption due to the adversary flipping the outcome between the pair $(1, j)$ in round t. Then we have

$$r_1^m = \sum_{t\in E^m} Y_1^t \sum_{j\in[K]} \epsilon_j^t (Z_{1j}^t + c_{1j}^t)$$

We define

$$A_1^m := \sum_{t\in E^m} Y_1^t \sum_{j\in[K]} \epsilon_j^t Z_{1j}^t \quad \text{and} \quad B_1^m := \sum_{t\in E^m} Y_1^t \sum_{j\in[K]} \epsilon_j^t c_{1j}^t$$

To bound the deviation in A_1^m, consider the sequence of random variables $X^1, \ldots, X^{N^m}$, where $X^t = \sum_{j\in[K]} \epsilon_j^t (p_1^m \mu_{1j} - Y_1^t Z_{1j}^t)$ for any $t \in E^m$. Then $\{X^t\}_{t=1}^{T^m}$ is a martingale with respect to the filtration $\{\mathcal{F}^t\}_{t=1}^{T^m}$ generated by the random variables $\{Y_j^s\}_{j\in[K],s\leq t}$, and $\{Z_{ij}^s\}_{i,j\in[K]\times[K],s\leq t+1}$. This is because conditioned on $\mathcal{F}^{t-1}$, Y_1^t is an independent random variable with mean p_1^m, and all the ϵ_j^t become deterministic quantities. Thus,

$$\mathbb{E}[X^t|\mathcal{F}^{t-1}] = \sum_{j\in[K]} \epsilon_j^t \mathbb{E}[p_1^m \mu_{1j} - Y_1^t Z_{1j}^t] = 0$$

To bound the conditional variance, observe that conditioned on $\mathcal{F}^{t-1}$, all the ϵ_j^t become deterministic quantities. Furthermore, for a fixed t exactly one of ϵ_j^t take value 1, and the rest take value 0

$$\Sigma = \sum_{t\in E^m} \mathbb{E}[(X^t)^2|\mathcal{F}^{t-1}] = \sum_{t\in E^m}\sum_{j\in[K]} \epsilon_j^t \mathbb{E}[(p_1^m\mu_{1j} - Y_1^t Z_{1j}^t)^2|\mathcal{F}^{t-1}] \leq p_1^m \sum_{t\in E^m}\sum_{j\in[K]} \epsilon_j^t \leq n_1^m$$

Observe that

$$\sum_{t\in E^m} X^t = p_1^m \sum_{t\in E^m}\sum_{j\in[K]} \epsilon_j^t \mu_{1j} - A_1^m = p_1^m \left(\frac{N^m}{2} + \sum_{t\in E^m}\sum_{j\neq 1} \epsilon_j^t \Delta_j\right) - A_1^m$$

Thus, we have that

$$\begin{aligned}
&\Pr\left(\left(\frac{1}{2}+\frac{1}{N^m}\sum_{t\in E^m}\sum_{j\neq 1}\epsilon_j^t\Delta_{\min}\right)-\frac{A_1^m}{n_1^m}\geq\sqrt{\frac{2\ln 4/\beta}{n_1^m}}+\frac{\sqrt{2}\ln 4/\beta}{3n_1^m}\right)\\
&\leq \Pr\left(\left(\frac{1}{2}+\frac{1}{N^m}\sum_{t\in E^m}\sum_{j\neq 1}\epsilon_j^t\Delta_{j}\right)-\frac{A_1^m}{n_1^m}\geq\sqrt{\frac{2\ln 4/\beta}{n_1^m}}+\frac{\sqrt{2}\ln 4/\beta}{3n_1^m}\right)\\
&\leq \Pr\left(\sum_{t\in E^m}X^t\geq\sqrt{2n_1^m\ln 4/\beta}+\frac{\sqrt{2}\ln 4/\beta}{3}\text{ and }\Sigma\leq n_1^m\right)\\
&\leq \beta/4
\end{aligned}$$

where the first inequality follows by observing that $\Delta_{\min}\leq\Delta_j$ for all j, and the final inequality follows by Bernstein's inequality for Martingales. Following the observation that $\sum_{t\in T^m}\sum_{j\neq 1}\epsilon_j^t = D^m$, and that $n_1^m\geq\lambda\geq 2304\ln 4/\beta$, we have that

$$\Pr\left(\left(\frac{1}{2}-\frac{D^m}{N^m}\Delta_{\min}\right)-\frac{A_1^m}{n_1^m}\geq\sqrt{\frac{3\ln 4/\beta}{n_1^m}}\right)\leq\beta/4$$

To bound the deviation in B_1^m, consider the sequence of random variables $X^1,\ldots,X^{N^m}$, where $X^t=\sum_{j\in K}\epsilon_j^t c_{1j}^t(Y_1^t-p_1^m)$ for any $t\in E^m$. Then $\{X^t\}_{t=1}^{T^m}$ is a martingale with respect to the filtration $\{\mathcal{F}^t\}_{t=1}^{T^m}$ generated by the random variables $\{Y_j^s\}_{j\in[K],s\leq t}$, and $\{Z_{ij}^s\}_{i,j\in[K]\times[K],s\leq t+1}$. This is because conditioned on $\mathcal{F}^{t-1}$, Y_1^t is an independent random variable with mean p_1^m, and c_{1j}^t, ϵ_j^t all become deterministic quantities. Thus,

$$\mathbb{E}[X^t|\mathcal{F}^{t-1}]=\sum_{j\in[K]}\epsilon_j^t c_{1j}^t\mathbb{E}[Y_1^t-p_1^m]=0$$

To bound the conditional variance, observe that conditioned on $\mathcal{F}^{t-1}$, all the c_{1j}^t, ϵ_j^t become deterministic quantities. Furthermore, exactly one of them takes value 1, and the rest take value 0

$$\Sigma=\sum_{t\in E^m}\mathbb{E}[(X^t)^2|\mathcal{F}^{t-1}]=\sum_{t\in E^m}\sum_{j\in[K]}\epsilon_j^t\left|c_{1j}^t\right|\mathbb{E}[(Y_1^tZ_{1j}^t-p_1^m\mu_{1j})^2|\mathcal{F}^{t-1}]\leq p_1^m\sum_{t\in E^m}\sum_{j\in[K]}\epsilon_j^t\left|c_{1j}^t\right|\leq n_1^m$$

Observe that

$$\sum_{t\in E^m}X^t=B_1^m-p_i^m\sum_{t\in E^m}\sum_{j\in[K]}\epsilon_j^t c_{1j}^t$$

We have that

$$\begin{aligned}
&\Pr\left(\frac{B_1^m}{n_1^m} \geq \frac{C_1^m}{N^m} + \sqrt{\frac{2\ln 4/\beta}{n_1^m}} + \frac{\sqrt{2}\ln 4/\beta}{3n_1^m}\right) \\
&\leq \Pr\left(\frac{B_1^m}{n_1^m} - \frac{1}{N^m}\sum_{t\in E^m}\sum_{j\in[K]}\epsilon_j^t c_{1j}^t \geq \sqrt{\frac{2\ln 4/\beta}{n_1^m}} + \frac{\sqrt{2}\ln 4/\beta}{3n_1^m}\right) \\
&= \Pr\left(\sum_{t\in E^m} X^t \geq \sqrt{2n_1^m \ln 4/\beta} + \frac{\sqrt{2}\ln 4/\beta}{3}\right) \\
&\leq \Pr\left(\sum_{t\in E^m} X^t \geq \sqrt{2n_1^m \ln 4/\beta} + \frac{\sqrt{2}\ln 4/\beta}{3} \text{ and } \Sigma \leq n_1^m\right) \\
&\leq \beta/4
\end{aligned}$$

Where the final inequality follows by Bernstein's inequality for Martingales. Finally, since we have that $n_i^m \geq \lambda \geq 2304 \ln 4/\beta$, we have that

$$\Pr\left(\frac{B_1^m}{n_1^m} \geq \frac{C_1^m}{N^m} + \sqrt{\frac{3\ln 4/\beta}{n_1^m}}\right) \leq \beta/4$$

We can similarly bound $-B_1^m/n_1^m$, giving a final absolute deviation bound on B as

$$\Pr\left(\left|\frac{B_1^m}{n_1^m}\right| \geq \frac{C_1^m}{N^m} + \sqrt{\frac{3\ln 4/\beta}{n_1^m}}\right) \leq \beta/2$$

Combining the above two guarantees through a union bound, we have that

$$\Pr\left(\left(\frac{1}{2} + \frac{D^m}{N^m}\Delta_{\min}\right) - \frac{r_1^m}{n_1^m} \geq \frac{C_1^m}{N^m} + \sqrt{\frac{12\ln 4/\beta}{n_1^m}}\right) \leq \beta$$

Following the observation that $n_1^m = \lambda/(\Delta_1^{m-1})^2$, and $\lambda \geq 2304 \ln 4/\beta$, we have that the probability of the said event (conditioned on all random variables prior to epoch m) is at most β. Finally, observe that this bound holds for any realization of these random variables we conditioned on, due to which it also holds unconditionally, hence proving our claim. ■

B.5. Proof of Lemma 9

Lemma 9 *We have for any fixed arm i, and $\beta > 0$,*

$$\Pr\left(\exists s \leq T : \sum_{t=1}^{s}\epsilon_{i1}^t(2\widetilde{Z}_{i1}^t - 1) > 0 \text{ and } \sum_{t=1}^{s}\epsilon_{i1}^t > \frac{4C_{i1}}{\Delta_i} + \frac{2}{\Delta_i^2}\ln\frac{2}{\beta\Delta_{i1}^2}\right) \leq \beta$$

Proof For ease of exposition, for any fixed arm i, every time step t, we define new random variables $\tilde{y}_{i1}^t = 2\widetilde{Z}_{i1}^t - 1$, $y_{i1}^t = 2Z_{i1}^t - 1$. We start by observing that

$$\begin{aligned}
&\Pr\left(\exists s \le T : \sum_{t=1}^{s} \epsilon_{i1}^t \tilde{y}_{i1}^t > 0 \text{ and } \sum_{t=1}^{s} \epsilon_{i1}^t > \frac{4C_{i1}}{\Delta_i} + \frac{2}{\Delta_i^2}\ln\frac{2}{\beta\Delta_i^2}\right) \\
&= \Pr\left(\exists s \le T : \sum_{t=1}^{s} \epsilon_{i1}^t (y_{i1}^t + c_{i1}^t) > 0 \text{ and } \sum_{t=1}^{s} \epsilon_{i1}^t > \frac{4C_{i1}}{\Delta_i} + \frac{2}{\Delta_i^2}\ln\frac{2}{\beta\Delta_i^2}\right) \\
&\le \Pr\left(\exists s \le T : \sum_{t=1}^{s} \epsilon_{i1}^t y_{i1}^t + 2C_{i1} > 0 \text{ and } \sum_{t=1}^{s} \epsilon_{i1}^t > \frac{4C_{i1}}{\Delta_i} + \frac{2}{\Delta_i^2}\ln\frac{2}{\beta\Delta_i^2}\right) \\
&\le \Pr\left(\exists s \le T : \sum_{t=1}^{s} \epsilon_{i1}^t y_{i1}^t + \frac{1}{2}\sum_{t=1}^{s} \epsilon_{i1}^t \Delta_i > 0 \text{ and } \sum_{t=1}^{s} \epsilon_{i1}^t > \frac{4C_{i1}}{\Delta_i} + \frac{2}{\Delta_i^2}\ln\frac{2}{\beta\Delta_i^2}\right) \\
&\le \Pr\left(\exists s \le T : \sum_{t=1}^{s} \epsilon_{i1}^t y_{i1}^t + \frac{1}{2}\sum_{t=1}^{s} \epsilon_{i1}^t \Delta_i > 0 \text{ and } \sum_{t=1}^{s} \epsilon_{i1}^t > \frac{2}{\Delta_i^2}\ln\frac{2}{\beta\Delta_i^2}\right)
\end{aligned}$$

where the penultimate inequality follows from the fact that $\sum_{t=1}^{s} \epsilon_{i1}^t \ge 4C_{i1}/\Delta_i$. Now consider the sequence of random variables $X^1, \dots X^T$, where $X^t = \epsilon_{i1}^t(y_{i1}^t + 2\Delta_i)$. Then $\{X_t\}_{t=1}^T$ is a martingale with respect to the filtration $\{\mathcal{F}^t\}_{t=1}^T$ generated by the variables $\{\epsilon_{i1}^r\}_{r \le t+1}$. This is because conditioned on $\mathcal{F}^{t-1}$, ϵ_{i1}^t becomes a deterministic quantity, giving us

$$\mathbb{E}[X^t | \mathcal{F}^{t-1}] = \epsilon_{i1}^t \mathbb{E}[y_{i1}^t + 2\Delta_i] = 0$$

To bound the partial conditional variance, we have

$$\Sigma_s = \sum_{t=1}^{s} \mathbb{E}[(X^t)^2 | \mathcal{F}^{t-1}] = 4\sum_{t=1}^{s} \epsilon_{i1}^t \left(\frac{1}{4} - \Delta_i^2\right) \le \sum_{t=1}^{s} \epsilon_{i1}^t$$

Now we have that

$$\begin{aligned}
&\Pr\left(\exists s \le T : \sum_{t=1}^{s} \epsilon_{i1}^t y_{i1}^t + \frac{1}{2}\sum_{t=1}^{s} \epsilon_{i1}^t \Delta_i > 0 \text{ and } \sum_{t=1}^{s} \epsilon_{i1}^t > \frac{2}{\Delta_i^2}\ln\frac{2}{\beta\Delta_i^2}\right) \\
&= \Pr\left(\exists s \le T : \sum_{t=1}^{s} X^t > 2\sum_{t=1}^{s} \epsilon_{i1}^t \Delta_i - \frac{1}{2}\sum_{t=1}^{s} \epsilon_{i1}^t \Delta_i \text{ and } \sum_{t=1}^{s} \epsilon_{i1}^t > \frac{2}{\Delta_i^2}\ln\frac{2}{\beta\Delta_i^2}\right)
\end{aligned}$$

Taking a union bound over all values of $\sum_{t=1}^{s} \epsilon_{i1}^t$ gives us

$$\begin{aligned}
&\Pr\left(\exists s \le T : \sum_{t=1}^{s} X^t > 2\sum_{t=1}^{s} \epsilon_{i1}^t \Delta_i - \frac{1}{2}\sum_{t=1}^{s} \epsilon_{i1}^t \Delta_i \text{ and } \sum_{t=1}^{s} \epsilon_{i1}^t > \frac{2}{\Delta_i^2}\ln\frac{2}{\beta\Delta_i^2}\right) \\
&= \sum_{z=\frac{2}{\Delta_i^2}\ln\frac{2}{\beta\Delta_i^2}+1}^{T} \Pr\left(\exists s \le T : \sum_{t=1}^{s} X^t > \frac{3}{2}z\Delta_i \text{ and } \sum_{t=1}^{s} \epsilon_{i1}^t = z\right) \\
&\le \sum_{z=\frac{2}{\Delta_i^2}\ln\frac{2}{\beta\Delta_i^2}+1}^{T} \Pr\left(\exists s \le T : \sum_{t=1}^{s} X^t > \frac{3}{2}z\Delta_i \text{ and } \Sigma_s \le z\right)
\end{aligned}$$

We can bound the each individual term inside the summation using Freedman's inequality, giving us for any z,

$$\Pr\left(\exists s \le T : \sum_{t=1}^{s} X^t > \frac{3}{2} z\Delta_i \text{ and } \Sigma \le z\right) \le \exp\left(-\left(\frac{3}{2}\right)^2 \frac{z^2\Delta_i^2/2}{(3/2)(z\Delta_i) + z}\right) \le \exp\left(-z\Delta_i^2/2\right)$$

which subsequently gives us

$$\begin{aligned}\Pr\left(\exists s \le T : \sum_{t=1}^{s} \epsilon_{i1}^t \tilde{y}_{i1}^t > 0 \text{ and } \sum_{t=1}^{s} \epsilon_{i1}^t > \frac{4C_{i1}}{\Delta_i} + \frac{2}{\Delta_i^2} \ln \frac{2}{\beta\Delta_i^2}\right) &\le \sum_{z=\frac{2}{\Delta_i^2}\ln\frac{2}{\beta\Delta_i^2}+1}^{T} e^{-z\Delta_i^2/2} \\ &\le \exp\left(-\ln\frac{2}{\beta\Delta_i^2}\right)\sum_{x=1}^{T} e^{-x\Delta_i^2/2} \\ &\le \beta\end{aligned}$$

where the final inequality follows from the fact that $\sum_{x=1}^{T} e^{-x\Delta_i^2/2} \le \sum_{x=1}^{\infty} e^{-x\Delta_i^2/2} \le 1/(e^{\Delta_i^2/2} - 1) \le 2/\Delta_i^2$. ■

Appendix C. Limitations of Existing Dueling Bandit Algorithms

In this section, we shall demonstrate the limitations of the four canonical algorithms for stochastic dueling bandits in the presence of adversarial corruptions. In particular, we show an explicit construction of an adversary that can force a regret of $\Omega(T)$ with just $O(\log T)$ corruption for each of the algorithms discussed below. For all four algorithms, we consider an instance of dueling bandits with just two arms $1, 2$, where arm 1 beats arm 2 with probability $\Pr(1 \succ 2) = 1/2 + \Delta$ for any choice of constant $\Delta > 0$. We begin by constructing a common randomized adversary which will be used to foil all of the existing algorithms.

The adversary. Consider a randomized adversary that uses the following corruption mechanism: Suppose the adversary chooses to corrupt the outcomes at time step $t \in [T]$, then given an outcome Z_{12}^t, if $Z_{12}^t = 1$, i.e. arm 1 beat arm 2, then the adversary tosses a coin with bias $4\Delta/(1 + 2\Delta)$, and if this coin turns up heads, then the adversary reverses the outcome to make $\widetilde{Z}_{12}^t = 0$. If it turns up tails, or if $Z_{12}^t = 0$, then the adversary leaves Z_{12}^t untouched. The exact time steps when the adversary chooses to apply this corruption will vary depending on the algorithm under consideration. Observe that across all time steps the adversary chooses to corrupt, the distribution of $\widetilde{Z}_{12}$ is identical to the case where arm 2 beats arm 1 with probability $\Pr(2 \succ 1) = 1/2 + \Delta$.

C.1. The IF algorithm

The IF algorithm of Yue et al. (2009) is a tournament style arm-elimination algorithm for dueling bandits under the STI, SST assumption that uses an explore-then-exploit type framework. It begins by selecting a random "anchor" arm $\hat{b}$ and playing all other arms against it in a round-robin fashion. If it is observed that some arm i loses to the anchor arm $\hat{b}$ with confidence, then arm i is removed from play. If it is observed that some arm i beats the anchor arm $\hat{b}$ with confidence, then the anchor arm, and all arms that have empirically lost to the anchor arm are removed from play, and this arm

i becomes the new anchor arm $\hat{b}$. One of the key lemmas (Lemma 1) of Yue et al. (2009) states that the total number of comparisons between any pair of arms i, j will, with high probability, be at most $O((\log KT)/\Delta_{ij}^2)$ until the inferior arm gets eliminated. Thus, our adversary simply corrupts the comparisons between arms $1, 2$ for the first $O((\log T)/\Delta^2)$ time steps. With high probability, arm 1 will be removed from play within these time steps since it appears inferior to arm 2 by a factor of Δ, following which the algorithm will incur a constant regret for the remaining time horizon without any additional corruption by the adversary. Thus, the total amount of realized corruption is $O((\log T)/\Delta)$ for which the algorithm suffers a total regret of $\Omega(\Delta T)$.

C.2. The BTM algorithm

The BTM algorithm of Yue and Joachims (2011) is an arm-elimination algorithm for dueling bandits under the STI, relaxed SST assumption. At all times, it maintains a set of active arms, and in each round, it selects a "left arm" to be the one that has be played the least (as the left arm) amongst all arms in the active set, and plays it against a uniformly at random arm from the active set. For every arm in the active set, the algorithm records the empirical probability of that arm winning when it was selected as the left arm, and discards the worst performing arm from the set of active arms when the lower confidence of the best performing arm is larger than the upper confidence of the worst performing arm. The key lemma (Lemma 3) of Yue and Joachims (2011) states that with high probability, the number of comparisons each arm in the active set accumulates before the worst arm b_k in the active set gets discarded is bounded by $O((\log KT)/\Delta_{1k}^2)$. Thus, as in the case of the IF algorithm, our adversary simply corrupts the comparisons between arms $1, 2$ for the first $O(\log T/\Delta^2)$ time steps. With high probability, arm 1 will be removed from play within these time steps since it appears inferior to arm 1 by a factor of Δ, following which the algorithm will incur constant regret for the remaining time horizon without need for any additional corruption by the adversary. Thus, the total amount of realized corruption is $O((\log T)/\Delta)$ for which the algorithm suffers a total regret of $\Omega(\Delta T)$.

C.3. The RUCB algorithm

The RUCB algorithm of Zoghi et al. (2014) is an algorithm for dueling bandits under the general Condorcet winner assumption. Unlike BTM, and IF, it is not an arm elimination algorithm, and at a high level, can be thought of as the dueling bandits counterpart of the UCB algorithm for classical bandits. It essentially maintains a matrix $\mathbf{W}$ of wins where W_{ij} is the number of times arm i beat arm j, using which it computes the relative upper confidence bounds u_{ij} on $\Pr(i \succ j)$ for every pair i, j. At any time step, it chooses to play a "left arm" that has the highest upper confidence bound amongst all arms i that have a relative upper confidence bound $u_{ij} > 1/2$ for all $j \neq i$. The right arm is then chosen to be the arm j with the highest relative upper confidence bound u_{ji}, where $u_{ii} = 1/2$. The key proposition (Proposition 2) of Zoghi et al. (2014) states that with high probability, the total number of times any pair of arms i, j will be played is bounded by $O((\log KT)/\min\{\Delta_i^2, \Delta_j^2\})$. In our case, the adversary will corrupt the outcomes only in time steps where the BTM algorithm plays the pair $1, 2$, which can be determined at the start of any time step $t \in [T]$ by conditioning on the past outcomes of comparisons between arms $1, 2$. Thus, in the presence of adversarial corruption, the effective distribution looks identical to one where $\Pr(1 \succ 2) = 1/2 - \Delta$, due to which with high probability, the algorithm will incur a cumulative regret of $\Omega(\Delta T)$. Furthermore, with high

probability, arms $1, 2$ will only be played $O((\log T)/\Delta^2)$ times across the entire time horizon, due to which the realized amount of adversarial corruption is just $O((\log T)/\Delta)$.

C.4. The RMED algorithm

The RMED algorithm of Komiyama et al. (2015a) is an instance optimal algorithm for dueling bandits under the general Condorcet winner assumption. Like the RUCB algorithm, it is not an arm elimination algorithm, and the choice of arms it chooses to play in any time step is deterministic upon conditioning on the past outcomes between all pairs of arms. At a high level, at every time step, it maintains a set of arms that are likely to be the Condorcet winner arm, and cycles through them in a fixed order, playing each of these candidate arms against the arm that is most likely to beat it (the exact choice of the opponent arm depends on the specific variant of the RMED algorithm under consideration i.e. RMED1, RMED2, or RMED2FH, but the key fact that we exploit in each of these variants is that this choice is deterministic for every candidate arm being played upon conditioning on the past outcomes of comparisons between all pairs of arms). The key lemmas (Lemma 5,6) of Komiyama et al. (2015a) imply that with high probability, the total number of times any suboptimal arm i will be played against the Condorcet winner arm is bounded by $O((K^{1+\epsilon} + \log T)/d(\mu_{i1}, 1/2))$ for any constant $\epsilon > 0$, where $d(\mu_{i1}, 1/2)$ is the KL divergence between μ_{i1} and $1/2$. This upper bound can be further simplified to $O((K^{1+\epsilon} + \log T)/\Delta_i^2)$ using Pinsker's inequality. In our example, the adversary will corrupt the outcomes only in time steps where the RMED algorithm plays the pair $1, 2$, which can be determined at the start of any time step $t \in [T]$ by conditioning on past outcomes between arms $1, 2$. Using an identical argument as in the case of the RUCB algorithm, we have that the RMED algorithm (all variants) will incur a cumulative regret of $\Omega(\Delta T)$, and furthermore, the realized amount of adversarial corruption will be just $O((\log T)/\Delta)$.