Versatile Dueling Bandits: Best-of-both World Analyses for Online Learning from Relative Preferences

Aadirupa Saha [1] **Pierre Gaillard** [2]

Abstract

We study the problem of K-armed dueling bandit for both stochastic and adversarial environments, where the goal of the learner is to aggregate information through relative preferences of pair of decision points queried in an online sequential manner. We first propose a novel reduction from any (general) dueling bandits to multi-armed bandits which allows us to improve many existing results in dueling bandits. In particular, *we give the first best-of-both world result for the dueling bandits regret minimization problem*—a unified framework that is guaranteed to perform optimally for both stochastic and adversarial preferences simultaneously. Moreover, our algorithm is also the first to achieve an optimal $O(\sum_{i=1}^{K} \frac{\log T}{\Delta_i})$ regret bound against the Condorcet-winner benchmark, which scales optimally both in terms of the arm-size K and the instance-specific suboptimality gaps $\{\Delta_i\}_{i=1}^{K}$. This resolves the long standing problem of designing an instancewise gap-dependent order optimal regret algorithm for dueling bandits (with matching lower bounds up to small constant factors). We further justify the robustness of our proposed algorithm by proving its optimal regret rate under adversarially corrupted preferences—this outperforms the existing state-of-the-art corrupted dueling results by a large margin. In summary, we believe our reduction idea will find a broader scope in solving a diverse class of dueling bandits setting, which are otherwise studied separately from multi-armed bandits with often more complex solutions and worse guarantees. The efficacy of our proposed algorithms are empirically corroborated against state-of-the-art dueling bandit methods.

[1]Microsoft Research, NYC, US [2]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France.. Correspondence to: Aadirupa Saha <aadirupa.saha@gmail.com>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

1. Introduction

Studies have shown that it is often easier, faster and less expensive to collect feedback on a relative scale rather than asking ratings on an absolute scale. E.g., to understand the liking for a given pair of items, say (A,B), it is easier for the users to answer preference-based queries like: "Do you prefer Item A over B?", rather than their absolute counterparts: "How much do you score items A and B in a scale of [0-10]?". From a system designer's point of view, exploiting such user preference information could greatly aid in improving systems performances, especially when data can be collected on a relative scale and online fashion; such as recommendation systems, crowd-sourcing platforms, training bots, multi-player games, search-engine optimization, online retail, just to name a few. In many real world problems, especially where human preferences are elicited in an online fashion, e.g., design of surveys and expert reviews, assortment selection, search engine optimization, recommender systems, ranking in multiplayer games, etc, or even more general reinforcement learning problems where rewards shaping is often a challenging problem (e.g. if multi-objective rewards etc.), and instead, a preference feedback is much easier to elicit.

Due to the widespread applicability and ease of data collection with relative feedback, learning from preferences has gained much popularity in the machine learning community and widely studied as the problem of *Dueling-Bandits* (DB) over last decade (Yue et al., 2012; Ailon et al., 2014; Zoghi et al., 2014a;b; 2015a), which is an online learning framework that generalizes the standard multiarmed bandit (MAB) (Auer et al., 2002a) setting for identifying a set of 'good' arms from a fixed decision-space (set of items) by querying preference feedback of actively chosen item-pairs.

Dueling Bandits Problem (DB) More formally, in classical dueling bandits with K arms, the learning proceeds in rounds, where at each time step $t \in \{1, 2, \ldots, T\}$, the learner selects a pair of arms $(k_{+1,t}, k_{-1,t})$ and receives the winner of the duel in terms of a binary preference feedback $o_t(k_{+1,t}, k_{-1,t}) \sim \text{Ber}(P_t(k_{+1,t}, k_{-1,t}))$, sampled according to an underlying preference matrix $\mathbf{P}_t \in [0,1]^{K\times K}$ (chosen adversarially in the most general setup). The objective of the learner is to minimize the regret with respect to a (or set of) 'best' arm(s) in hindsight.

Related Works Over the years, the problem of Dueling Bandits has been studied with various objectives and generalizations. This includes analyzing the learning rate under various preference structures, such as total ordering, transitivity, stochastic triangle inequality (Falahatgar et al., 2017; Yue & Joachims, 2011), utility based preference structure (Ailon et al., 2014; Szorenyi et al., 2015; Saha & Gopalan, 2018a; Chen et al., 2018), or under any general pairwise preference matrices (Dudík et al., 2015; Jamieson et al., 2015; Komiyama et al., 2016). Consequently, depending on the underlying preference structure, the learner's performance has been evaluated w.r.t. different benchmarks including among other promising generalizations. The problem of stochastic dueling bandits has been studied for both PAC (Falahatgar et al., 2018; Szorenyi et al., 2015; Sui et al., 2018) as well as regret minimization setting (Zoghi et al., 2014a; Yue & Joachims, 2009; Chen & Frazier, 2017; Zoghi et al., 2015a) under several notions of benchmarks including best arm identification (Saha & Gopalan, 2019b; Busa-Fekete et al., 2014; Falahatgar et al., 2017), top-set detection (Busa-Fekete et al., 2013; Saha & Gopalan, 2019a), ranking (Ren et al., 2018; Saha & Gopalan, 2018b), amongst many. Some recent works have also looked into the problem for adversarial sequence of preference matrices (Saha & Gupta, 2022; Gajane et al., 2015; Saha et al., 2021), robustness to corruptions (Agarwal et al., 2021), or extending dueling bandits to potentially infinite arm sets (Saha, 2021; Kumagai, 2017) and contextual scenarios (Dudík et al., 2015; Saha & Krishnamurthy, 2021). Another interesting line of work along dueling bandits is to study the implications for its subsetwise generalization (Ren et al., 2018; Sui et al., 2018; Brost et al., 2016; Chen et al., 2018), also studied as battling bandits (Saha & Gopalan, 2018a; 2019a; Bengs et al., 2021).

Despite widespread surge of interest along this line of research and multiple attempts there are some fundamental long standing open questions in dueling bandits which are (surprisingly!) yet unresolved.

Unresolved Question #1 One of the longest and most widely studied objective in stochastic dueling bandit is to measure regret w.r.t. the Condorcet winner (CW) arm: Given a preference matrix $\mathbf{P}$, an arm $k^{(\texttt{cw})} \in [K]$ is termed as the CW of $\mathbf{P}$ if $P(k^{(\texttt{cw})}, k) > 0.5 \, \forall k \in [K] \setminus \{k^{(\texttt{cw})}\}$ (Zoghi et al., 2014a). Assuming P contains a CW, there have been several attempts to design an optimal regret dueling bandit algorithm against the CW arm $k^{(\texttt{cw})}$ (see Eq. (2) for details) (Zoghi et al., 2014a; Wu & Liu, 2016; Komiyama et al., 2015; Bengs et al., 2021). Without loss of generality, assuming $k^{(\texttt{cw})} = 1$ and by denoting $\Delta_i = P(1, i) - 0.5$ to be the suboptimality gap of item i w.r.t. the CW, it is well known that the dueling bandit regret lower bound (w.r.t. the CW arm) is $\Omega\big(\sum_{i=2}^{K} \frac{\log T}{\Delta_i}\big)$ (Yue et al., 2012; Komiyama et al., 2015). However, despite several attempts, it is still unknown how to design an order optimal dueling bandit algorithm for the CW regret. Existing upper-bounds suffer all suboptimal $\Delta_{\min}^{-2}$ and/or K^2 dependencies.

Notably, under more restricted structures, e.g. total ordering (Yue & Joachims, 2011), or utility based preferences (Szorenyi et al., 2015; Saha & Gopalan, 2020), or even special preference structures where the suboptimality gaps of all items ($\Delta_i, \;\; i \in [K] \setminus \{1\}$) are equal, the problem is easier to solve and tight regret guarantees are available with matching upper and lower bound analysis. However, for the case of any general preference matrix with CW, none of the existing attempts were able to close this regret analysis gap successfully (Zoghi et al., 2014a; Wu & Liu, 2016; Chen & Frazier, 2017; Komiyama et al., 2015; Saha & Gaillard, 2021). Subsequently, the natural questions to ask are:

(1). Is the lower bound tight? (2). How to close the gap between the upper and lower bound for CW regret?

Unresolved Question #2 Till date, all the proposed algorithms of dueling bandits need to know underlying preference structure ahead of time in order to yield optimal regret bounds. In fact, different algorithms have been proposed based on the nature/structures of the underlying preference matrices, e.g. (Yue & Joachims, 2011) for preferences with total orderings in terms of (relaxed) stochastic transitivity and stochastic triange inequality; (Ailon et al., 2014; Gajane et al., 2015) for linear-utility based preferences, (Szorenyi et al., 2015) for BTL models, (Zoghi et al., 2014a; Komiyama et al., 2015; Wu & Liu, 2016) for stochastic preferences in presence of CW, (Saha et al., 2021; Gajane et al., 2015) for adversarial sequence of preferences, etc. However, it might not always be realistic to assume complete knowledge of the properties underlying preference matrices. Thus the daunting question to ask in this regard is

Is it possible to design an order optimal 'best-of-both-world' algorithm for dueling bandits?

That is, an algorithm that can adapt itself to the underlying structures of the preference environments and give optimal regret for both stochastic and adversarial settings? There has been a series of work on this line for the MAB framework (e.g., Bubeck & Slivkins, 2012; Auer & Chiang, 2016; Zimmert & Seldin, 2021), but unfortunately there has not been any existing 'best-of-both-world' attempt for general dueling bandits.

Unresolved Question #3 In any real world situation, the true feedback are often corrupted with some form of system noise. Undoubtedly, the binary $0/1$ bit dueling preferences are extremely prone to such noises when the learner might get to observe a flipped bit (adversarially corrupted) instead of the true dueling feedback.

Can we design an efficient dueling bandit algorithm which is robust to adversarial corruptions and provably optimal?

Our Contributions In this paper, we answer all of the above three questions affirmatively. The list of our specific contributions can be summarized as follows:

1. **A novel insight on the reduction from DB to MAB.** Ailon et al. (2014) proposed a reduction from utility-based dueling bandits (which is a special class of DB) to standard MAB. We show that the latter can easily be extended to any general dueling bandit problems (including CW), with significant consequences on the state of the art results in dueling bandits theory, as discussed in the points below. We believe that the reduction will find wider application in solving a diverse class of dueling bandit settings, using analyses of their MAB counterparts, which are otherwise studied separately from MAB with often more complex solutions and worse guarantees (Sec. 4).
2. **First Best-of-Both World regret for DB.** Applying the above reduction to a Best-of-Both world algorithm from MAB, we provide an algorithm that simultaneously guarantees a pseudo-regret bound $O(\sqrt{KT})$ in the adversarial setting and $O(K\log(T)/\Delta_{\min})$ in the stochastic one. This leads to the first best-of-both world results for dueling bandits (Sec. 5).
3. **Optimal stochastic gap-dependent Regret.** Our algorithm also provides the first optimal Condorcet regret, which suffers neither from a suboptimal dependence on $\Delta_{\min}^{-2}$ nor from a quadratic dependence on the number of arms (Sec. 6.1).
4. **Robustness to adversarial corruptions.** Our algorithm is robust to adversarial corruptions and significantly improves existing results in DB (Agarwal et al., 2021) (Sec. 6.2).
5. **Another easy algorithm for stochastic DB.** We also propose a new elimination based algorithm with $O(\sum_{i=2}^{K} \frac{K\log T}{\Delta_i})$ Condorcet regret which is regret optimal upto a factor of K (Sec. 3).
6. **Experimental evaluations.** Finally we corroborate our theoretical results with extensive empirical evaluations (Sec. 7).

2. Problem Formulation

Notations. Decision space (or item/arm set) $[K] := \{1, 2, \ldots, K\}$. For any matrix $\mathbf{M} \in \mathbb{R}^{K\times K}$, we define $m_{ij} := M(i,j)$, $\forall i,j \in [K]$. $\mathbf{1}(\cdot)$ denotes the indicator random variable which takes value 1 if the predicate is true and 0 otherwise and $\lesssim$ a rough inequality which holds up to universal constants. For any two items $x, y \in [K]$, we use the symbol $x \succ y$ to denote x is preferred over $\mathbf{y}$. By convention, we set $\frac{0}{0} := 0.5$. For any closed and convex set S, we denote by $\mathcal{I}_S(x) := 0$, if $x \in S$, and $\mathcal{I}_S(x) := \infty$ otherwise; for any $x \in \mathbb{R}_+$ (notation used in Alg. 3).

Consider any function $f : \mathbb{R}^K \mapsto \mathbb{R}$. $\nabla f(\mathbf{x})$ denotes the gradient of function f at point $\mathbf{x} \in \mathbb{R}^K$. The convex conjugate (a.k.a. Fenchel conjugate) of f is defined by:

$$f^*(\mathbf{y}) := \sup_{\mathbf{x}\in\mathbb{R}^k} \big(\langle \mathbf{x}, \mathbf{y}\rangle - f(\mathbf{x})\big).$$

Thus we have for any any closed and convex set S, $(f + \mathcal{I}_S)^*(\mathbf{y}) = \max_{\mathbf{x}\in S} \big(\langle \mathbf{x}, \mathbf{y}\rangle - f(\mathbf{x})\big)$. Further, as claimed in (Zimmert & Seldin, 2019), differentiable and convex f with invertible gradient $(\nabla f)^{-1}$, it holds that

$$\nabla (f + \mathcal{I}_S)^*(\mathbf{y}) = \arg\max_{\mathbf{x}\in S} \big(\langle \mathbf{x}, \mathbf{y}\rangle - f(\mathbf{x})\big).$$

Setup. We assume a decision space of K arms denoted by $\mathcal{A} := [K]$. At each round t, the task of the learner is to select a pair of actions $(k_{+1,t}, k_{-1,t}) \in [K]\times[K]$, upon which a preference feedback $o_t \sim \text{Ber}(P_t(k_{+1,t}, k_{-1,t}))$ is revealed to the learner according to the underlying preference matrix $\mathbf{P}_t \in [0,1]^{K\times K}$ (chosen adversarially), such that the probability of $k_{+1,t}$ being preferred over $k_{-1,t}$ is given by $Pr(o_t = 1) := Pr(k_{+1,t} \succ k_{-1,t}) = P_t(k_{+1,t}, k_{-1,t})$, and hence $Pr(o_t = 0) := Pr(k_{-1,t} \succ k_{+1,t}) = 1 - P_t(k_{+1,t}, k_{-1,t})$.

Objective. Assuming the learner selects the duel $(k_{+1,t}, k_{-1,t})$ at round t, one can measure its performance w.r.t. a single fixed arm $k^* \in [K]$[1] in hindsight by calculating the regret w.r.t. $k^* \in [K]$

$$R_T(k^*) := \sum_{t=1}^{T} \frac{1}{2}\left(P_t(k^*, k_{+1,t}) + P_t(k^*, k_{-1,t}) - 1\right). \tag{1}$$

For the *stochastic setting* where P_ts are fixed across all time steps $t \in [T]$, we denote $P_t = P\ \forall t \in [T]$. Further assuming there exists a Condorcet winner for P, i.e. fixed arm $k^{(\texttt{cw})} \in [K]$ such that $P(k^{(\texttt{cw})}, k) > 0.5\ \forall k \in [K] \setminus \{k^{(\texttt{cw})}\}$, the above notion of regret boils down to the regret with respect to the Condorcet winner for the choice of $k^* = k^{(\texttt{cw})}$, as widely studied in many dueling bandit literature (Zoghi et al., 2014a; Wu & Liu, 2016; Komiyama et al., 2015; Bengs et al., 2021), defined as:

$$R_T^{(\texttt{cw})} := \sum_{t=1}^{T} \frac{1}{2}\big(\Delta(k^{(\texttt{cw})}, k_{+1,t}) + \Delta(k^{(\texttt{cw})}, k_{-1,t})\big), \tag{2}$$

where $\Delta(i,j) := P(i,j) - 1/2$ being the suboptimality gap of item i and j in terms of their relative preferences. Without

[1] Note that this is equivalent to maximizing the expected regret w.r.t. any fixed distribution $\boldsymbol{\pi}^* \in \Delta_K$, i.e. when $k^* \sim \boldsymbol{\pi}^*$. This is because the regret objective is linear in the entries of $\boldsymbol{\pi}^*$, so the maximizer $\boldsymbol{\pi}^*$ is always one hot.

loss of generality we will assume $k^{(cw)} = 1$ throughout the rest of this paper (whenever relevant).

3. Warm-Up: Near-Optimal Algorithm

In this section, we propose an efficient UCB based algorithm for stochastic dueling bandit, which has a nearly optimal gap-dependent Condorcet regret of $R_T^{(\texttt{cw})} = O\big(\sum_{i=2}^{K} i\Delta(k^{(\texttt{cw})}, i)^{-1}\log T\big)$ (assuming $k^{(cw)} = 1$ is the Condorcet Winner). Note that existing dueling bandit algorithms, that satisfy a non-asymptotic Condorcet regret bound, suffer an additional constant of order $\Delta_{\min}^{-2}$ (Bengs et al., 2021), which implies a worst-case regret of order $O(T^{2/3})$ when $\Delta_{\min} \to 0$. The simple elimination algorithm below solves this drawback and depends only on $\Delta_{\min}^{-1}$ but at the cost of a suboptimal quadratic dependence in the number of arms K. Despite our efforts, we could not avoid this suboptimal factor by following the classical stochastic dual bandit analysis. In the following sections, we will show how to easily reach the optimal dependence in both $\Delta_{\min}$ and K connecting the dueling bandit (DB) problem to standard MAB.

Main Ideas: Algorithm `RR-DB` The high-level idea of Algorithm 1 is to sequentially compare arms in a round-robin fashion and eliminate arms when they are significantly suboptimal compared to any other arm. Typically, after t rounds, a suboptimal arm k has been compared at least t/K times with the Condorcet winner. Denoting by Δ_k its suboptimality gap, the arm is eliminated after at most t_k rounds, where $(t_k/K)^{-1/2} \approx \Delta_k$. At that time the arm has been played t_k/K times, yielding a regret of order $(t_k/K) \times \Delta_k \approx (K/\Delta_k^2) \times \Delta_k = K/\Delta_k$. Summing over the arms yields a final regret of order $O(K^2/\Delta_{\min})$.

Algorithm 1 `RR-DB` (Near Optimal DB)

input: Arm set: $[K]$, Confidence parameter $\delta \in (0, 1)$
init: Active arms: $\mathcal{A}_1 := [K]$, $n_{ij}(t) \leftarrow 0$, $\forall i, j \in [K]$
for $t = 1, 2, \ldots, T$ **do**
  Play $(k_{+1,t}, k_{-1,t}) \in \text{argmin}_{i,j\in\mathcal{A}_t}\{n_{ij}(t-1)\}$
  Observe $o_t(k_{+1,t}, k_{-1,t}) = 1 - o_t(k_{+1,t}, k_{-1,t})$
  for $i, j \in \mathcal{A}_t$ **do**
    Define $\mathbb{1}_t(i,j) := \mathbb{1}\big\{\{i,j\} = \{k_{-1,t}, k_{+1,t}\}\big\}$ and
      $n_{ij}(t) := \sum_{s=1}^{t} \mathbb{1}_t(i,j)$
      $\widehat{p}_{ij}(t) := \frac{1}{n_{ij}(t)}\sum_{s=1}^{t} o_t(i,j)\mathbb{1}_t(i,j)$
      $u_{ij}(t) := \widehat{p}_{ij}(t) + \sqrt{\frac{\log(Kt/\delta)}{n_{ij}(t)}}$
    where we assume $x/0 = +\infty$.
  end for
  $\mathcal{A}_{t+1} := \mathcal{A}_t \backslash \big\{i \in \mathcal{A}_t : \exists j \in \mathcal{A}_t \text{ s.t. } \; u_{ij}(t) < \frac{1}{2}\big\}$.
end for

Without loss of generality assume the Condorcet winner $k^{(\texttt{cw})} = 1$, and denote $\Delta_i = \Delta(1, i)$, $\forall i \in [K] \setminus \{1\}$.

Theorem 1 (Regret analysis of `RR-DB` (Alg. 1) for Stochastic Preferences against Condorcet Winner)**.** *Assume any instance of stochastic dueling bandit problem with Condorcet Winner $k^{(cw)} = 1$. Let $\delta \in (0, 1/2)$, for any $T \geq 1$, the regret of Algorithm 1 is upper-bounded with probability at least $1 - \delta$ as*

$$R_T^{(cw)} \leq \frac{K^2}{2} + 4\sum_{i=2}^{K}\frac{(i-1)\log(KT/\delta)}{\Delta_i}.$$

Further, when $T \geq K^2$, in the worst case (over the problem instance, i.e. $\Delta_2, \ldots, \Delta_K$), the regret of Algorithm 1 can be upper bounded as:

$$R_T^{(cw)} \leq 2K\sqrt{T\log(KT/\delta)}.$$

Remark 1. *In particular, our regret analysis shows that, except a logarithmic factor, the regret bound of `RR-DB` (Alg. 1) is off only by a multiplicative factor of K, as follows from the known $\Omega(\sum_{k=1}^{K}\frac{\log T}{\Delta_k})$ Condorcet winner regret lower bound (Yue et al., 2012; Komiyama et al., 2015).*

The proof is postponed to Appendix A.

4. Key Idea: Reducing DB to MAB

We now present a simple reduction from a multi-armed bandit algorithm to a dueling bandit one. The latter was already proposed by Saha & Gupta (2022) to show worst-case guarantees and by Ailon et al. (2014) for utility based dueling bandits only. We recall it here since it is central to our analysis and we believe that it is of important interest for the dueling bandit community that usually uses significantly different algorithms and analysis than the ones from standard multi-armed bandits.

The main idea is to apply a multi-armed bandit algorithm independently to two players $i \in \{-1, +1\}$ respectively with losses defined for any $k \in [K]$ and $t \in [T]$, by

$$\ell_{i,t}(k) := o_t(k_{-i,t}, k)\,,$$

where $o_t(k, k') = 1 - o_t(k', k)$ for $1 < k' \leq K$ follows a Bernoulli with parameter $P_t(k, k')$ (and we assume $o_t(k, k) = 1/2$).

We show below that any MAB regret upper-bound satisfied by $\mathcal{M}_i$ can be transformed into a DB regret upper-bound.

Theorem 2 (DB Regret analysis of Alg. 2 in terms of the regret of the underlying MAB instances $(\mathcal{M}_{-1}, \mathcal{M}_{+1})$)**.** *Define for $i \in \{-1, 1\}$ and $k \in [K]$ by*

$$R_{i,T}(k) := \sum_{t=1}^{T}\ell_{i,t}(k_{i,t}) - \ell_{i,t}(k)$$

Algorithm 2 Reduction from DB to MAB

input: Arm set: $[K]$, two instances $\mathcal{M}_i$ of an algorithm for MAB, $i \in \{-1, +1\}$.
for $t = 1, 2, \dots, T$ **do**
  for $i \in \{+1, -1\}$ **do**
    receive $p_{i,t}$ as suggested by $\mathcal{M}_i$
    sample $k_{i,t}$ from the distribution $p_{i,t}$
  end for
  Play duel $(k_{+1,t}, k_{-1,t})$.
  Observe preference feedback $o_t(k_{+1,t}, k_{-1,t})$ and set $o_t(k_{-1,t}, k_{+1,t}) = 1 - o_t(k_{+1,t}, k_{-1,t})$.
  Feed $\mathcal{M}_i$ with loss $\ell_{i,t}(k) := o_t(k_{-i,t}, k_{i,t})$ for $i \in \{-1, +1\}$.
end for

the regret achieved by algorithm $\mathcal{M}_i$. Then, the expected regret (1) *of Algorithm* 2 *for dueling bandits can be decomposed as*

$$\mathbf{E}\big[R_T(k)\big] = \frac{1}{2}\mathbf{E}\big[R_{-1,T}(k) + R_{+1,T}(k)\big] \, .$$

Proof. The proof follows from

$$\begin{aligned}\mathbf{E}\big[\ell_{-1,t}(k) + \ell_{+1,t}(k)\big] &= \mathbf{E}\big[o_t(k_{+1,t}, k) + o_t(k_{-1,t}, k)\big] \\ &= \mathbf{E}\big[P_t(k_{+1,t}, k) + P_t(k_{-1,t}, k)\big] \\ &= 2 - \mathbf{E}\big[P_t(k, k_{+1,t}) + P_t(k, k_{-1,t})\big]\end{aligned}$$

and

$$\begin{aligned}&\mathbf{E}\big[\ell_{-1,t}(k_{+1,t}) + \ell_{+1,t}(k_{-1,t})\big] \\ &\quad = \mathbf{E}\big[o_t(k_{+1,t}, k_{-1,t}) + o_t(k_{+1,t}, k_{-1,t})\big] = 1 \, .\end{aligned}$$

We conclude by summing over $t = 1, \dots, T$ both equations and by substituting them into the definition of the regret $R_T(k)$ in (1). □

Note that such a reduction can also be used to bound $R_T(k)$ directly rather than its expectation. Saha & Gupta (2022) indeed use this reduction to show a $O(\sqrt{KT})$ high-probability regret upper-bound for adversarial dueling bandit. They also obtain non-stationary regret bounds.

The main message of this paper is that this reduction can be used to transpose many results from standard multi-armed bandit to general dueling bandits. For instance, applying a subroutine $\mathcal{M}_i$ which is robust to delays (Thune et al., 2019; Zimmert & Seldin, 2020, e.g.,), one directly obtains a dueling bandit with the same robustness guarantees.

As we said, this reduction is not new. However, to date, it has only been considered in two specific contexts: the adversarial setting with worst-case regret bounds of order $O(\sqrt{KT})$ and the utility-based setting. Since the losses $\ell_{i,t} = o_t(k_{-i,t}, k)$ are not i.i.d. but depend on an adaptive adversary which chooses $k_{-i,t}$, one cannot use stochastic bandit algorithms. And the dueling bandit community usually needs to resort to more sophisticated algorithms and arguments to obtain logarithmic regret bounds for Condorcet stochastic dueling bandits. The only stochastic dueling bandit for which such a reduction was considered (Ailon et al., 2014; Zimmert & Seldin, 2021, e.g.,) was the utility based-dueling bandits, which is overly restrictive in practice. That is, when the preference matrix if of the form $P_t(k, k') : (1 + u_t(k) + u_t(k'))/2$ for some sequence of utility vectors $(u_t)_{t \geq 1}$. Utility based dueling bandit are known to be easily be reduced to two independent multi-armed bandit problems (Ailon et al., 2014).

Our main contribution is to show that this reduction can in fact be easily extended to the much weaker Condorcet winner hypothesis. To do this, as we show in the next sections, we simply apply the reduction with a best-of-both-worlds multi-armed bandit algorithm. As we will see, this recovers and improves the best existing upper bounds on the Condorcet regret for dueling bandits.

5. Best-of-Both Dueling: Optimal Algorithm for Stochastic and Adversarial DB

This section contains our main result, which extends the best-of-both-worlds result from Zimmert & Seldin (2021) to dueling bandits.Moreover, it allows us to improve the best existing upper bound on the regret for stochastic and corrupted dueling bandits. The main idea is to apply the "reduction idea" of Algorithm 2 with the multi-armed bandit algorithm (Online Mirror Descent with Tsallis regularizer) of Zimmert & Seldin (2021).

Online Mirror Descent (or famously abbreviated as OMD) is essentially a follow the regularized leader (FTRL) type algorithm, for any general choice of regularizer (Srebro et al., 2011; Lattimore & Szepesvári, 2020). The famous exponential weight algorithm in online learning (Cesa-Bianchi & Lugosi, 2006; Arora et al., 2012) (or EXP3 algorithm in MAB (Auer et al., 2002b)) is a special instance of OMD algorithm for (-ve) entropy based regularizer.

It is important to note that, our reduction analysis of Algorithm 3 (see proof of Theorem 3) indicates that we could have used any best-of-both algorithm of MAB as the underlying blackbox algorithm, instead of OMD with Tsallis regularizer (Zimmert & Seldin, 2021). For example, the EXP3 based approach of Bubeck & Slivkins (2012) could have been a potential MAB blackbox algorithm as well, but they do not give optimal regret guarantees due to the use of 'entropic-regularizers' which is not ideal for obtaining the optimal instance (gap)-dependent stochastic regret for MAB. Hence Zimmert & Seldin (2021) motivated the use of OMD based Tsallis-Inf algorithm for the best-of-both MAB algorithm and our analysis of Theorem 3 shows that the same works for the best-of-both DB analysis as well.

Of course, as for the classical adversarial multi-armed bandits, the losses $\ell_{i,t}(k) = o_t(k_{-i,t}, k)$ cannot be observed for all $k \in [K]$. Therefore, they are estimated in the algorithm with the importance weight estimators

$$\widehat{\ell}_{i,t}(k) = \begin{cases} \ell_{i,t}(k)/p_{i,t}(k) & \text{if } k = k_{i,t} \\ 0 & \text{otherwise} \end{cases} . \quad (3)$$

The resulted algorithm is described in Algorithm 3.

Algorithm 3 `Versatile-DB` **(Best-of-Both DB)**

input: Arm set: $[K]$, Regularizers: $(\Psi_t)_{t=1,2,\ldots}$
init: $\widehat{L}_{i,0} \leftarrow \mathbf{0}_K$ for $i \in \{+1, -1\}$
for $t = 1, 2, \ldots$ **do**
  choose $p_{i,t} = \nabla(\Psi_t + \mathcal{I}_\Delta)^*(-\widehat{L}_{i,t-1})$,
  (here Δ denotes the K-simplex. Refer to Sec. 2 for other notations)
  For $i \in \{+1, -1\}$, sample $k_{i,t}$ from the distribution $(p_{i,t}(1), \ldots, p_{i,t}(K))$
  Observe preference feedback $o_t(k_{+1,t}, k_{-1,t})$
  Compute $\widehat{\ell}_{i,t}(k)$ for $i \in \{+1, -1\}$ and $k \in [K]$ using (3)
  update $\widehat{L}_{i,t} = \widehat{L}_{i,t-1} + \widehat{\ell}_{i,t}$
end for

Theorem 3 (Regret analysis of `Versatile-DB` (Alg. 3) for Adversarial and Stochastic Preferences)**.** *For any sequence of preference matrices $\mathbf{P}_t$, the pseudo-regret of Algorithm 3 with $\Psi_t(w) = \sqrt{t}\sum_{k=1}^{K}(\sqrt{w_k} - w_k/2)/8$ satisfies for any $T \geq 1$*

$$\overline{R}_T := \max_{k \in [K]} \mathbf{E}\big[R_T(k)\big] \leq 4\sqrt{KT} + 1.$$

Furthermore, if there exists a gap vector $\Delta \in [0,1]^K$ with a unique zero coordinate $k^* \in [K]$ and $C \geq 0$ such that

$$\overline{R}_T \geq \frac{1}{2}\mathbf{E}\Bigg[\sum_{t=1}^{T}\sum_{k \neq k^*}\big(p_{+1,t}(k) + p_{-1,t}(k)\big)\Delta_k\Bigg] - C\,, \quad (4)$$

the pseudo regret also satisfies

$$\overline{R}_T \leq \sum_{k \neq k^*}\frac{4\log T + 12}{\Delta_k} + 4\log T + \frac{1}{\Delta_{\min}} + \frac{3}{2}\sqrt{K} + 8 + C.$$

where $\Delta_{\min} = \min_{k \neq k^*}\Delta_k$.

The proof is postponed to Appendix B. Note that the theorem largely follows from (and is itself highly similar to) the best-of-both worlds regret-bound of Zimmert & Seldin (2021, Theorem 1) for MAB. The proof is just a clever combination of their MAB analysis with our black box reduction (Theorem 2). But we believe that the simplicity of our approach is its strength that can benefit the community of dueling bandits. As we will see, several state-of-the-art results of dueling bandits can be simultaneously improved as a direct consequence of this theorem, including the order optimal regret bounds for stochastic DB against condorcet winner or DB with adversarial corruption, as discussed in Sec. 6.1 and 6.2 respectively.

Note that for simplicity, we restricted ourselves to importance weighted estimators (3). By using more sophisticated variance reduced estimators, as in Zimmert & Seldin (2021), the multiplicative constants can be reduced. Furthermore, similar to Zimmert & Seldin (2021), the result holds only for the pseudo-regret and not for the true regret. Auer & Chiang (2016) have indeed proven that no optimal adversarial and stochastic high probability regret bounds can be obtained simultaneously for standard stochastic bandits. The learner must pay suboptimal logarithmic factors. The result can be extended to dueling bandits.

It is worth to emphasize that this is the first best-of-both worlds regret bound for general dueling bandits (the stochastic bound follows from the choice $C = 0$, see Sec 6.1). Zimmert & Seldin (2021, Cor. 10) obtain a similar result for the very same algorithm but for utility based dueling bandits only.

Remark 2. *Note that a single sub-routine of OMD to optimize the weights is actually enough to get the same regret guarantee. To do so, one samples both $k_{-1,t}$ and $k_{+1,t}$ independently from the same distribution $\mathbf{p}_t = \nabla(\psi_t + \mathcal{I}_\Delta)^*(-\widehat{L}_{t-1})$. Here, $\widehat{L}_t = \sum_{s=1}^{t}\widehat{\ell}_s \in \mathbb{R}_+^K$ and the importance weight estimator are defined for all $k \in [K]$ by*

$$\widehat{\ell}_t(k) = \frac{1}{2}\big(\widehat{\ell}_{-1,t}(k) + \widehat{\ell}_{+1,t}(k)\big)\,.$$

Noting that $\mathbf{E}[\widehat{\ell}_t(k)] = \mathbf{E}\big[P_t(k_{-1,t}, k) + P_t(k_{+1,t}, k)\big]/2$ and $\mathbf{E}\big[\sum_{k=1}^{K}p_t(k)\widehat{\ell}_t(k)\big] = \mathbf{E}\big[P_t(k_{-1,t}, k_{+1,t}) + P_t(k_{+1,t}, k_{-1,t})\big]/2 = 1/2$, the proof follows similarly to the other one. Though the regret upper-bound is exactly the same, we believe that this version might lead to better performance because the two players share information.

6. Improvements Over Existing Dueling Bandit Results

Also our approach and analysis is rather simple, it allows to outperform the best existing regret-upper bounds for stochastic dueling bandits with or without corruption. We believe that the dueling bandit community will benefit from this reduction and that it may be applied to a wider scope such as to deal with non-stationarity, delays, or non-standard feedbacks (graphs) for which many results already exist in standard multi-armed bandits.

6.1. Stochastic dueling bandits with Condorcet winner

In stochastic dueling bandits, the preference matrices $\mathbf{P}_t$ are fixed over time $\mathbf{P}_t = \mathbf{P}$ for all $t \geq 1$. Under the Condorcet winner assumption there exists $k^{(\mathrm{cw})} \in [K]$ such that $P(k^{(\mathrm{cw})}, k) > \frac{1}{2}$ for all $k \neq k^{(\mathrm{cw})}$. Then, the suboptimality gaps of all actions $k \in [K]$ are defined as $\Delta_k := P(k^{(\mathrm{cw})}, k) - \frac{1}{2}$. Remarking that in this case the self-bounding assumption (4) is satisfied with $C = 0$, since

$$\begin{aligned}\overline{R}_T &= \frac{1}{2}\sum_{t=1}^{T} \mathbf{E}\Big[P_t(k^{(\mathrm{cw})}, k_{+1,t}) + P_t(k^{(\mathrm{cw})}, k_{-1,t}) - 1\Big] \\ &= \frac{1}{2}\sum_{t=1}^{T} \mathbf{E}\Big[\Delta_{k_{+1,t}} + \Delta_{k_{-1,t}}\Big] \\ &= \frac{1}{2}\sum_{t=1}^{T} \mathbf{E}\Big[\sum_{k\neq k^{(\mathrm{cw})}} (p_{-1,t} + p_{+1,t})\Delta_k\Big],\end{aligned}$$

we get the following corollary from Theorem 3.

Corollary 1 (Regret analysis of `Versatile-DB` (Alg. 3) for Stochastic Preferences against Condorcet Winner)**.** *For stochastic dueling bandits with Condorcet winner, the pseudo-regret of Algorithm 3 with well-chosen parameters satisfies*

$$\overline{R}_T \leq \sum_{k\neq k^{(\mathrm{cw})}} \frac{4\log T + 12}{\Delta_k} + 4\log T + \frac{1}{\Delta_{\min}} + \frac{3}{2}\sqrt{K} + 8\,.$$

Note that the above bound is the first pseudo-regret upper-bound for stochastic dueling bandit that does not suffer from a $\Delta_{\min}^{-2}$ dependence under the Condorcet winner assumption, without a quadratic dependence on the number of arms, K, which is a concern when it comes to dealing with large-scale problems. For instance, RUCB (Zoghi et al., 2014a) satisfies a regret bound of order $O(K\log(T)\Delta_{\min}^{-2} + K^2)$, MergeRUCB (Zoghi et al., 2015b) has linear dependence on K but suffers $O(K\log(T)\Delta_{\min}^{-2})$. Finally, RMED from (Komiyama et al., 2015) is asymptotically optimal when $T \to \infty$ but also suffers from large constant terms (K^2 and $\Delta_{\min}^{-2}$) and is only valid for $K \to \infty$. We refer the reader to (Bengs et al., 2021) for existing results on stochastic dueling bandits.

6.2. Corrupted regime

Here, we consider the stochastic dueling bandit problem in the presence of adversarial corruptions. The robustness to adversarial corruption has known recent progress in the MAB setting (Gupta et al., 2019; Lykouris et al., 2018; Zimmert & Seldin, 2021) and has recently been extended to the DB framework (Agarwal et al., 2021). The preference matrices are fixed $\mathbf{P}_t = \mathbf{P}$ for all $t \geq 1$ and we assume the existence of a Condorcet winner $k^{(\mathrm{cw})}$. Furthermore, an adversary may corrupt the outcomes of some duels by replacing the results of the duels $o_t(k, k')$ with corrupted ones $\tilde{o}_t(k, k')$. At the end of each round, the player only observes $\tilde{o}_t(k_{+1,t}, k_{-1,t})$. The objective of the player is to minimize the pseudo-regret $\bar{R}_T$ under a bounded total amount of corruption

$$C := \sum_{t=1}^{T}\sum_{k\neq k^{(\mathrm{cw})}} \big|o_t(k^{(\mathrm{cw})}, k) - \tilde{o}_t(k^{(\mathrm{cw})}, k)\big|.$$

We show here that similarly to what happens for standard Multi-armed bandits in Zimmert & Seldin (2021), this corrupted setting is a special case of the self-bounding assumption (4). Indeed, defining $\tilde{\mathbf{P}}_t$ the corrupted preference matrices by $\tilde{P}_t(k, k') = \mathbf{E}\big[\tilde{o}_t(k, k')\big]$ and the corrupted pseudo-regret

$$\widetilde{R}_T(k) = \frac{1}{2}\mathbf{E}\Bigg[\sum_{t=1}^{T} \tilde{P}_t(k, k_{+1,t}) + P_t(k, k_{-1,t}) - 1\Bigg],$$

we have

$$\begin{aligned}&\widetilde{R}_T(k^{(\mathrm{cw})}) \\ &= \frac{1}{2}\mathbf{E}\Bigg[\sum_{t=1}^{T} \tilde{o}_t(k^{(\mathrm{cw})}, k_{+1,t}) + \tilde{o}_t(k^{(\mathrm{cw})}, k_{-1,t}) - 1\Bigg] \\ &\geq \frac{1}{2}\mathbf{E}\Bigg[\sum_{t=1}^{T} o_t(k^{(\mathrm{cw})}, k_{+1,t}) + o_t(k^{(\mathrm{cw})}, k_{-1,t}) - 1\Bigg] - C \\ &= \overline{R}_T - C \\ &= \frac{1}{2}\sum_{t=1}^{T}\mathbf{E}\Big[\sum_{k\neq k^{(\mathrm{cw})}} (p_{-1,t} + p_{+1,t})\Delta_k\Big] - C\,.\end{aligned}$$

Therefore, the corrupted regime satisfies the self-bounding assumption (4). Applying Theorem 3 on the corrupted regime and using that we also have $\overline{R}_T(k^{(\mathrm{cw})}) \leq \widetilde{R}_T(k^{(\mathrm{cw})}) + C$, we get the following corollary.

Corollary 2 ([Regret analysis of `Versatile-DB` (Alg. 3) for Stochastic Preferences against Condorcet Winner with a maximum adversarial Corruption C)**.** *For stochastic dueling bandits with Condorcet winner and adversarial corruptions (bounded by C), the pseudo-regret $\overline{R}_T$ Algorithm 3 is upper-bounded as*

$$\begin{aligned}\overline{R}_T \leq \sum_{k\neq k^{(\mathrm{cw})}} &\frac{4\log T + 12}{\Delta_k} + 4\log T \\ &+ \frac{1}{\Delta_{\min}} + \frac{3}{2}\sqrt{K} + 8 + 2C.\end{aligned}$$

Although Corollary 2 easily follows from Theorem 3 which itself easily follows from Theorem 1 of Zimmert & Seldin (2021), the latter result significantly improves upon the recent results on dueling bandit with corruptions obtained by Agarwal et al. (2021). Indeed, the latter provide for the same setting and a significantly more sophisticated procedure a high-probability regret bound of order

$$O\left(\frac{K^2 C}{\Delta_{\min}} + \sum_{k \neq k^{(\mathrm{cw})}} \frac{K^2}{\Delta_k^2} \log\left(\frac{K}{\Delta_k}\right) + \sum_{k \neq k^{(\mathrm{cw})}} \frac{\log T}{\Delta_k}\right).$$

As often in the dueling bandit literature, it suffers from both a quadratic dependence on the number of actions and $\Delta_{\min}^{-1}$. Furthermore, our regret bound is sublinear in T as soon as the corruption level is $o(T)$ while Agarwal et al. (2021) can only afford $o(\Delta_{\min} T/K^2)$.

Moreover, Zimmert & Seldin (2021) also provide an upper-bound for stochastic bandits with adversarial corruption. The latter is of order $O\Big(\sum_{k \neq k^{(\mathrm{cw})}} \frac{\log T}{\Delta_k} + \sqrt{C \sum_{k \neq k^{(\mathrm{cw})}} \frac{\log T}{\Delta_k}}\Big)$, which seems to outperform our bound when C is large. The difference is since they upper-bound the corrupted regret $\widetilde{R}_T(k^{(\mathrm{cw})})$ and not $\bar{R}_T$.

7. Experiments

Algorithms. We compared the performances of the following algorithms: 1. VDB: Our proposed `Versatile-DB` (Alg. 3) algorithm. 2. RUCB: The algorithm proposed in Zoghi et al. (2014a) for K-armed stochastic dueling bandits for CW regret. (We set the algorithm parameter $\alpha = 0.6$). 3. RMED: Another algorithm for CW regret as proposed in Komiyama et al. (2015). (We set the algorithm parameter $f(K) = 0.3K^{1.01}$ as suggested in their experimental evaluation). 4. DTS: The double thompson sampling algorithm of Wu & Liu (2016). (Here again we set the similar algorithm parameter $\alpha = 0.5$ as used in the experiments of Wu & Liu (2016)). 5. REX3: As introduced in Gajane et al. (2015). Note that their suggested optimal tuning parameters, i.e. the uniform exploration rate γ as well as the learning rate η requires the knowledge of problem dependent parameters τ (see Thm. 1 of Gajane et al. (2015)) which are unknown to the learner. We used T in place of τ henceforth.

Performance Measures. We report the average cumulative regret (Eqn. (1)) of the algorithms averaged over 20 runs.

7.1. Stochastic Preferences

We compared their regret performance across the following stochastic environments:

Constructing Preference Matrices (P). We use four different utility parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ based preference models where the underlying preference model is defined as $P(i,j) := \frac{\theta_i}{\theta_i + \theta_j}\ \forall i,j \in [K]$. The model is famously studied as BTL model or more generally Plackett-Luce model (Saha & Gopalan, 2019b; Chen et al., 2018; Negahban et al., 2017). Note this ensures P to have *total-ordering* (Yue & Joachims, 2011; Falahatgar et al., 2017).

In particular we consider the following choices of $\boldsymbol{\theta}$: 1. *Trivial* 2. *Easy* 3. *Medium*, and 4. *Hard* with their respective $\boldsymbol{\theta}$ parameters are given by: 1. *Trivial:* $\boldsymbol{\theta}(1) = 1$, $\boldsymbol{\theta}(2 : K) = 0.5$. 2. *Easy:* $\boldsymbol{\theta}(1 : \lfloor K/2 \rfloor) = 1$, $\boldsymbol{\theta}(\lfloor K/2 \rfloor + 1 : K) = 0.5$. 3. *Medium:* $\boldsymbol{\theta}(1 : \lfloor K/3 \rfloor) = 1$, $\boldsymbol{\theta}(\lfloor K/3 \rfloor + 1 : \lfloor 2K/3 \rfloor) = 0.7$, $\boldsymbol{\theta}(\lfloor 2K/3 \rfloor + 1 : K) = 0.4$. 4. *Hard:* $\boldsymbol{\theta}(i) = 1 - (i-1)/K,\ \forall i \in [K]$. Note for each instance the best-arm is arm-1 and the optimal ordering on the arms is $(1 > 2 > \ldots K)$. For the purpose of our experiments we set $K = 10$. We also evaluated the algorithms on two general 10×10 preference matrices *Car* and *Hurdy* as also used in Niranjan & Rajkumar (2017); Saha & Gopalan (2018a).

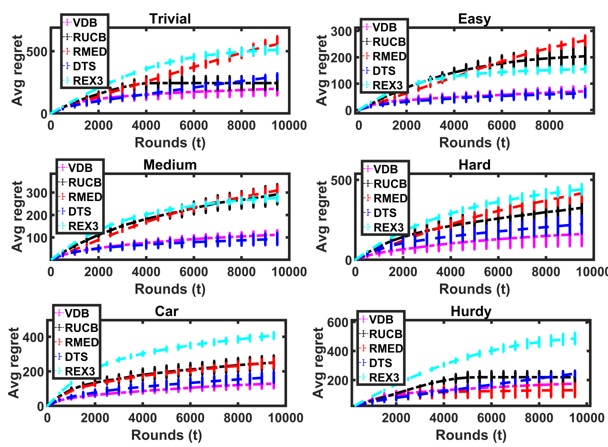

Figure 1. Averaged cumulative regret over time

Regret vs Time. Fig. 1 shows the relative performances of different algorithms with time. As follows from the plots, in general *VDB* (`Versatile-DB`) outperforms the rest in all instances, with *DTS* being closely competitive in some cases. In terms of the problem hardness, as their names suggest too, the *Trivial* and *Easy* instances are easiest to learn as the best-vs-worst item preferences are well separated in these cases and the diversity of the item preferences across different groups are least. Consequently the algorithms yield slightly more regret on *instance-Medium* due to higher preference diversity, and the hardest instance being *Hard* where the algorithms require maximum time to converge, though *VDB* reaching the convergence fastest still.

7.2. Corrupted Preferences

We also evaluated the performances of algorithms in presence of corruption (Sec. 6.2). In particular, Fig. 2 and 3 respectively shows the relative performances of the algorithms

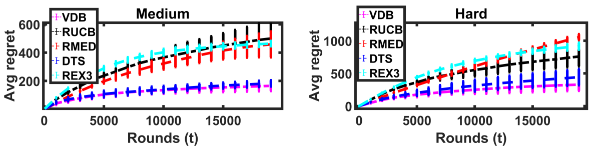

Figure 2. Averaged cumulative regret (20% corrupted feedback)

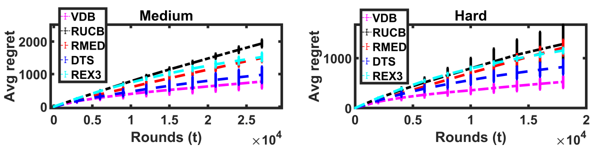

Figure 3. Averaged cumulative regret (40% corrupted feedback)

with 20% and 40% corrupted feedback (at each round, we flip the winner feedback with that certain probability) respectively on *Medium* and *Hard* Plackett-Luce instances. As expected, the performances of all the algorithms decay significantly with increasing degree of feedback-corruption, however as before, *VDB* consistently performed best over all the baselines and tend to converge the fastest among all.

7.3. Adversarial Preferences

In Fig. 4, we also studied the case of adversarial preferences setup (as discussed in Thm. 3 or Sec. 5). The following two plots show the superior performance of VDB on two sequences of adversarial preferences for $K = 10$. (1). *Switching Best-Arm* [Switch(500)]: Here we tweaked P_ts such that the best-arm changes after every 500 rounds, and *Random-Walk* [RW(0.01)]: Here we generated the sequences of preferences $P_t(i, j)$ for all pairs of arms (i, j) as random walks with increment parameter ± 0.01. Any values that fall outside [0, 1] are truncated back to [0, 1]. Both these settings were considered in (Neu & Valko, 2014; Saha et al., 2020; 2021) for generating adversarial sequences (of rewards or preferences).

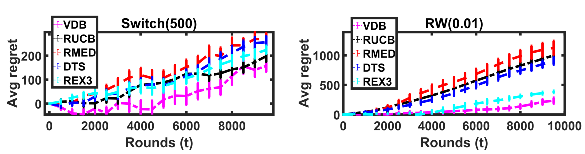

Figure 4. Avg. Regret for Adversarial preferences ($K = 10$)

8. Discussions

We studied the problem of `Versatile Dueling Bandits`, which gives the first *'best-of-both' world* result for the problem of Dueling Bandits. The crux of our analyses relies on a *novel idea of decomposing the dueling bandit regret into multiarmed bandit (MAB) regret* by interpreting the dueling preference feedback as a certain realization of adversarial reward sequence. An important byproduct of our best-of-both dueling analysis is, this gives the first order optimal gap-dependent regret bound for K-armed stochastic dueling bandits, closing the decade-long open problem of tightness of 'Condorcet dueling bandit regret'. Further we also analyze the robustness of our algorithm under corrupted preference feedback setting, which provably improves over the state-of-the art corrupted dueling bandits algorithms.

Future Works. Proving the first *best-of-both world* result for dueling bandits using our novel reduction idea is just a first step towards exploring the possibility of understanding how far this idea can be extended to apply existing multiarmed bandits results to dueling bandits frameworks, instead of putting individual and isolated efforts in developing dueling bandit algorithms, taking inspirations from existing MAB generalizations. Some such extensions could be to analyze dynamic dueling bandit regret under non-stationary preferences (Wei & Luo, 2021; Chen et al., 2019; Besbes et al., 2015), item non-availability (Neu & Valko, 2014; Kanade et al., 2009), delayed feedback (Vernade et al., 2018; Pike-Burke et al., 2018; Thune et al., 2019), budget constraints (Immorlica et al., 2019; Zhou & Tomlin, 2018; Ding et al., 2013), or even the more general reinforcement learning (RL) scenarios (Auer et al., 2009; Talebi & Maillard, 2018; Ng et al., 2006), for which there are already well established theory of works with MAB framework. Also under what settings of Dueling Bandits, its corresponding MAB counterpart based reductions are bound to fail?

Finally it is worth mentioning that, an ambitious (and broad) objective along these line of thoughts is to understand the connection between different learning scenarios to dueling bandits, e.g. feedback graphs (Alon et al., 2015; 2017), partial monitoring problems (Lattimore & Szepesvári, 2019; Mannor et al., 2014; Lin et al., 2014), markov games (Xie et al., 2020; Bai et al., 2020; Bai & Jin, 2020), etc. The obvious motivation being to understand how far we can re-engineer the existing results from related learning literature to solve the online preference bandits problems.

Acknowledgement

We thank the anonymous reviewers for their insightful suggestions to improve the paper.

References

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

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

Alon, N., Cesa-Bianchi, N., Dekel, O., and Koren, T. Online learning with feedback graphs: Beyond bandits. In

JMLR Workshop and Conference Proceedings, volume 40. Microtome Publishing, 2015.

Alon, N., Cesa-Bianchi, N., Gentile, C., Mannor, S., Mansour, Y., and Shamir, O. Nonstochastic multi-armed bandits with graph-structured feedback. *SIAM Journal on Computing*, 46(6):1785–1826, 2017.

Arora, S., Hazan, E., and Kale, S. The multiplicative weights update method: a meta-algorithm and applications. *Theory of computing*, 8(1):121–164, 2012.

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

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002a.

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

Auer, P., Jaksch, T., and Ortner, R. Near-optimal regret bounds for reinforcement learning. In *Advances in neural information processing systems*, pp. 89–96, 2009.

Bai, Y. and Jin, C. Provable self-play algorithms for competitive reinforcement learning. In *International Conference on Machine Learning*, pp. 551–560. PMLR, 2020.

Bai, Y., Jin, C., and Yu, T. Near-optimal reinforcement learning with self-play. In *Advances in Neural Information Processing Systems*, 2020.

Bengs, V., Busa-Fekete, R., El Mesaoudi-Paul, A., and Hüllermeier, E. Preference-based online learning with dueling bandits: A survey. *J. Mach. Learn. Res.*, 22:7–1, 2021.

Besbes, O., Gur, Y., and Zeevi, A. Non-stationary stochastic optimization. *Operations research*, 63(5):1227–1244, 2015.

Brost, B., Seldin, Y., Cox, I. J., and Lioma, C. Multi-dueling bandits and their application to online ranker evaluation. *CoRR*, abs/1608.06253, 2016.

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

Busa-Fekete, R., Szorenyi, B., Cheng, W., Weng, P., and Hüllermeier, E. Top-k selection based on adaptive sampling of noisy preferences. In *International Conference on Machine Learning*, pp. 1094–1102, 2013.

Busa-Fekete, R., Szörényi, B., and Hüllermeier, E. PAC rank elicitation through adaptive sampling of stochastic pairwise preferences. In *AAAI*, pp. 1701–1707, 2014.

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

Chen, B. and Frazier, P. I. Dueling bandits with weak regret. *arXiv preprint arXiv:1706.04304*, 2017.

Chen, X., Li, Y., and Mao, J. A nearly instance optimal algorithm for top-k ranking under the multinomial logit model. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 2504–2522. SIAM, 2018.

Chen, Y., Lee, C.-W., Luo, H., and Wei, C.-Y. A new algorithm for non-stationary contextual bandits: Efficient, optimal, and parameter-free. *In Proceedings of the 32nd Conference on Learning Theory*, 99:1–30, 2019.

Ding, W., Qin, T., Zhang, X.-D., and Liu, T.-Y. Multi-armed bandit with budget constraint and variable costs. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*, 2013.

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

Falahatgar, M., Hao, Y., Orlitsky, A., Pichapati, V., and Ravindrakumar, V. Maxing and ranking with few assumptions. In *Advances in Neural Information Processing Systems*, pp. 7063–7073, 2017.

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

Gajane, P., Urvoy, T., and Clérot, F. A relative exponential weighing algorithm for adversarial utility-based dueling bandits. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 218–227, 2015.

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

Immorlica, N., Sankararaman, K. A., Schapire, R., and Slivkins, A. Adversarial bandits with knapsacks. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 202–219. IEEE, 2019.

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

Kanade, V., McMahan, H. B., and Bryan, B. Sleeping experts and bandits with stochastic action availability and adversarial rewards. 2009.

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

Komiyama, J., Honda, J., and Nakagawa, H. Copeland dueling bandit problem: Regret lower bound, optimal algorithm, and computationally efficient algorithm. *arXiv preprint arXiv:1605.01677*, 2016.

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

Lattimore, T. and Szepesvári, C. An information-theoretic approach to minimax regret in partial monitoring. In *Conference on Learning Theory*, pp. 2111–2139. PMLR, 2019.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Lin, T., Abrahao, B., Kleinberg, R., Lui, J., and Chen, W. Combinatorial partial monitoring game with linear feedback and its applications. In *International Conference on Machine Learning*, pp. 901–909. PMLR, 2014.

Lykouris, T., Mirrokni, V., and Paes Leme, R. Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 114–122, 2018.

Mannor, S., Perchet, V., and Stoltz, G. Set-valued approachability and online learning with partial monitoring. *The Journal of Machine Learning Research*, 15(1):3247–3295, 2014.

Negahban, S., Oh, S., and Shah, D. Rank centrality: Ranking from pairwise comparisons. *Operations Research*, 65 (1):266–287, 2017.

Neu, G. and Valko, M. Online combinatorial optimization with stochastic decision sets and adversarial losses. In *Advances in Neural Information Processing Systems*, pp. 2780–2788, 2014.

Ng, A. Y., Coates, A., Diel, M., Ganapathi, V., Schulte, J., Tse, B., Berger, E., and Liang, E. Autonomous inverted helicopter flight via reinforcement learning. In *Experimental robotics IX*, pp. 363–372. Springer, 2006.

Niranjan, U. and Rajkumar, A. Inductive pairwise ranking: going beyond the n log (n) barrier. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Pike-Burke, C., Agrawal, S., Szepesvari, C., and Grunewalder, S. Bandits with delayed, aggregated anonymous feedback. In *International Conference on Machine Learning*, pp. 4105–4113, 2018.

Ren, W., Liu, J., and Shroff, N. B. PAC ranking from pairwise and listwise queries: Lower bounds and upper bounds. *arXiv preprint arXiv:1806.02970*, 2018.

Saha, A. Optimal algorithms for stochastic contextual dueling bandits. In *Advances in Neural Information Processing Systems*, 2021.

Saha, A. and Gaillard, P. Dueling bandits with adversarial sleeping. *Advances in Neural Information Processing Systems*, 34, 2021.

Saha, A. and Gopalan, A. Battle of bandits. In *Uncertainty in Artificial Intelligence*, 2018a.

Saha, A. and Gopalan, A. Active ranking with subset-wise preferences. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018b.

Saha, A. and Gopalan, A. Combinatorial bandits with relative feedback. In *Advances in Neural Information Processing Systems*, 2019a.

Saha, A. and Gopalan, A. PAC Battling Bandits in the Plackett-Luce Model. In *Algorithmic Learning Theory*, pp. 700–737, 2019b.

Saha, A. and Gopalan, A. From pac to instance-optimal sample complexity in the plackett-luce model. In *International Conference on Machine Learning*, pp. 8367–8376. PMLR, 2020.

Saha, A. and Gupta, S. Optimal and efficient dynamic regret algorithms for non-stationary dueling bandits. In *International Conference on Machine Learning*. PMLR, 2022.

Saha, A. and Krishnamurthy, A. Efficient and optimal algorithms for contextual dueling bandits under realizability. *arXiv preprint arXiv:2111.12306*, 2021.

Saha, A., Gaillard, P., and Valko, M. Improved sleeping bandits with stochastic action sets and adversarial rewards. In *International Conference on Machine Learning*, pp. 8357–8366. PMLR, 2020.

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

Srebro, N., Sridharan, K., and Tewari, A. On the universality of online mirror descent. *Advances in neural information processing systems*, 24, 2011.

Sui, Y., Zoghi, M., Hofmann, K., and Yue, Y. Advancements in dueling bandits. In *IJCAI*, pp. 5502–5510, 2018.

Szorenyi, B., Busa-Fekete, R., Paul, A., and Hullermeier, E. Online rank elicitation for plackett-luce: A dueling bandits approach. In *Advances in Neural Information Processing Systems*, pp. 604–612, 2015.

Talebi, M. S. and Maillard, O.-A. Variance-aware regret bounds for undiscounted reinforcement learning in mdps. *arXiv preprint arXiv:1803.01626*, 2018.

Thune, T. S., Cesa-Bianchi, N., and Seldin, Y. Nonstochastic multiarmed bandits with unrestricted delays. In *Advances in Neural Information Processing Systems*, pp. 6541–6550, 2019.

Vernade, C., Carpentier, A., Lattimore, T., Zappella, G., Ermis, B., and Brueckner, M. Linear bandits with stochastic delayed feedback. *arXiv preprint arXiv:1807.02089*, 2018.

Wei, C.-Y. and Luo, H. Non-stationary reinforcement learning without prior knowledge: An optimal black-box approach. *In Proceedings of the 32nd International Conference on Learning Theory*, 2021.

Wu, H. and Liu, X. Double Thompson sampling for dueling bandits. In *Advances in Neural Information Processing Systems*, pp. 649–657, 2016.

Xie, Q., Chen, Y., Wang, Z., and Yang, Z. Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium. In *Conference on Learning Theory*, 2020.

Yue, Y. and Joachims, T. Interactively optimizing information retrieval systems as a dueling bandits problem. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 1201–1208. ACM, 2009.

Yue, Y. and Joachims, T. Beat the mean bandit. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 241–248, 2011.

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

Zhou, D. P. and Tomlin, C. J. Budget-constrained multi-armed bandits with multiple plays. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

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

Zimmert, J. and Seldin, Y. An optimal algorithm for adversarial bandits with arbitrary delays. In *International Conference on Artificial Intelligence and Statistics*, pp. 3285–3294. PMLR, 2020.

Zimmert, J. and Seldin, Y. Tsallis-inf: An optimal algorithm for stochastic and adversarial bandits. *J. Mach. Learn. Res.*, 22:28–1, 2021.

Zoghi, M., Whiteson, S., Munos, R., Rijke, M. d., et al. Relative upper confidence bound for the k-armed dueling bandit problem. In *JMLR Workshop and Conference Proceedings*, number 32, pp. 10–18. JMLR, 2014a.

Zoghi, M., Whiteson, S. A., De Rijke, M., and Munos, R. Relative confidence sampling for efficient on-line ranker evaluation. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pp. 73–82. ACM, 2014b.

Zoghi, M., Karnin, Z. S., Whiteson, S., and De Rijke, M. Copeland dueling bandits. In *Advances in Neural Information Processing Systems*, pp. 307–315, 2015a.

Zoghi, M., Whiteson, S., and de Rijke, M. Mergerucb: A method for large-scale online ranker evaluation. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pp. 17–26. ACM, 2015b.

Supplementary: Versatile Dueling Bandits: Best-of-both World Analyses for Online Learning from Relative Preferences

A. Regret analysis of Algorithm 1

Theorem 1 (Regret analysis of RR-DB (Alg. 1) for Stochastic Preferences against Condorcet Winner). *Assume any instance of stochastic dueling bandit problem with Condorcet Winner $k^{(cw)} = 1$. Let $\delta \in (0, 1/2)$, for any $T \geq 1$, the regret of Algorithm 1 is upper-bounded with probability at least $1 - \delta$ as*

$$R_T^{(cw)} \leq \frac{K^2}{2} + 4\sum_{i=2}^{K} \frac{(i-1)\log(KT/\delta)}{\Delta_i}.$$

Further, when $T \geq K^2$, in the worst case (over the problem instance, i.e. $\Delta_2, \ldots, \Delta_K$), the regret of Algorithm 1 can be upper bounded as:

$$R_T^{(cw)} \leq 2K\sqrt{T\log(KT/\delta)}.$$

Proof of Theorem 1. Let us denote by

$$u_{ij}(t) := \widehat{p}_{ij}(t) + c_{ij}(t)$$

for any pair (i, j) and time t, where

$$c_{ij}(t) := \sqrt{\frac{\log(Kt/\delta)}{n_{ij}(t)}}\,,$$

and assume $\Delta_2 \leq \Delta_3 \leq \cdots \leq \Delta_K$ without loss of generality. We will also assume the confidence bounds of Lem. 4 holds good for all $t \in [T]$ and all pairs (i, j), which is shown to hold good with probability at least $1 - \delta$. In particular, this implies that the best arm cannot be eliminated, i.e., $1 \in \mathcal{A}_t$ for all $t \geq 1$.

We start by noting that if the worst arm K (since $\Delta_K = \max_{i=2}^K \Delta_i$, arm-K gets maximally beaten by the CW) is played at time t, it means $u_{K1}(t) \geq \frac{1}{2}$. However we also have,

$$\begin{aligned} u_{K1}(t) &= \widehat{p}_{K1}(t) + c_{K1}(t) \\ &\leq p_{K1} + 2c_{K1}(t) = 1/2 - \Delta_K + 2c_{K1}(t), \end{aligned}$$

where the inequality holds due to Lem. 4 and the last equality holds by noting $p_{K1} = 1 - p_{1K} = 1 - (1/2 + \Delta_K) = 1/2 - \Delta_K$.

So $u_{K1}(t) > 1/2$ can only hold good if $c_{K1}(t) > \Delta_K/2$ which implies,

$$n_{K1}(t) \leq \frac{4\log(Kt/\delta)}{\Delta_K^2}. \tag{5}$$

But by the our algorithm design since all the pairs are drawn in a round robin fashion, at any round $t \in [T]$, for any two distinct pairs (i, j) and (i', j') that are in $\mathcal{A}_t$ note that

$$|n_{ij}(t) - n_{i'j'}(t)| \leq 1\,. \tag{6}$$

Thus the total regret incurred by Alg. 1 at rounds where $K \in \{k_{+1,t}, k_{-1,t}\}$, can be upper bounded as:

$$\begin{aligned} &\sum_{t=1}^{T}\sum_{k<K} \mathbf{1}(\{k_{+1,t}, k_{-1,t}\} = \{k, K\})\frac{\Delta_K + \Delta_k}{2} \\ &\qquad \leq \sum_{k=1}^{K-1} n_{kK}(T)\Delta_K \leq (K-1)(1 + n_{K1}(T))\Delta_K \end{aligned}$$

$$\leq (K-1)\Big(1 + \frac{4\log(Kt/\delta)}{\Delta_K^2}\Big)\Delta_K$$
$$= (K-1)\Big(\Delta_k + \frac{4\log(Kt/\delta)}{\Delta_K}\Big).$$

Similarly, note for any $i \in \{2, 3, \dots K-1\}$, we can upper bound the regret of rounds where i was played in the duel as:

$$\sum_{t=1}^{T}\sum_{k<i}\mathbf{1}(\{k_{+1,t}, k_{-1,t}\} = \{k, i\})\frac{\Delta_i + \Delta_k}{2}$$
$$\leq \sum_{k=1}^{i-1} n_{ki}(T)\Delta_i \leq (i-1)(1 + n_{1i}(T))\Delta_i$$
$$\leq (i-1)\Big(\Delta_i + \frac{4\log(Kt/\delta)}{\Delta_i}\Big).$$

Thus we can bound the total regret of Algorithm 1 as:

$$\sum_{t=1}^{T}\sum_{i=2}^{K}\sum_{k=1}^{i-1}\mathbf{1}(\{k_{+1,t}, k_{-1,t}\} = \{k, i\})\frac{\Delta_i + \Delta_k}{2}$$
$$\leq \sum_{i=2}^{K}(i-1)\Big(\Delta_i + \frac{4\log(Kt/\delta)}{\Delta_i}\Big)$$
$$\overset{(\Delta_i \leq 1/2)}{\leq} \frac{K^2}{4} + 4\sum_{i=2}^{K}(i-1)\frac{\log(Kt/\delta)}{\Delta_i}$$

which concludes the first half of the proof. Further, to show the second part of the claim (analyzing worst-case gap-independent regret bound of Algorithm 1), note that Eqn. (5) equivalently implies for any $i \in [K] \setminus \{1\}$ and for any $t \in [T]$:

$$\Delta_i \leq \sqrt{\frac{4\log(Kt/\delta)}{n_{i1}(t)}}.$$

Hence we can alternatively upper bound the regret as:

$$R_T = \sum_{i=2}^{K}\sum_{k=1}^{i-1} n_{ik}(T)\frac{\Delta_i + \Delta_k}{2} \leq \sum_{i=2}^{K}\sum_{k=1}^{i-1} n_{ik}(T)\Delta_i$$
$$\leq \sum_{i=2}^{K}\sum_{k=1}^{i-1} n_{ik}(T)\sqrt{\frac{4\log(KT/\delta)}{n_{i1}(T)}}$$
$$\overset{(a)}{\leq} 2\sum_{i=2}^{K}\sum_{k=1}^{i-1}\sqrt{n_{ik}(T)\log(KT/\delta)}$$
$$\overset{(b)}{\leq} 2\sqrt{K^2\sum_{i=2}^{K}\sum_{k=1}^{i-1} n_{ik}(T)\log(KT/\delta)}$$
$$\leq 2K\sqrt{T\log(KT/\delta)},$$

where (a) follows from the observation of Eqn. (6) which implies $n_{i1}(T) \geq n_{ik}(T)$ when $T \geq K^2$ and (b) from Jensen's inequality and $\sum_{i=2}^{K}(i-1) \leq K^2/2$. □

Lemma 4. *For any $\delta \in (0, 1/2)$. Then, with probability at least $1 - \delta$, for any pair $i, j \in [K]$ and any time $t \in [T]$*

$$\widehat{p}_{ij}(t) - c_{ij}(t) \le p_{ij} \le \widehat{p}_{ij}(t) + c_{ij}(t), \qquad \forall t \in [T],$$

where $c_{ij}(t) := \sqrt{\frac{\log(Kt/\delta)}{n_{ij}(t)}}$.

Proof. Let us denote by $u_{ij}(t) := \widehat{p}_{ij}(t) + c_{ij}(t)$ and $\ell_{ij}(t) := \widehat{p}_{ij}(t) - c_{ij}(t)$. Note the inequality holds trivially at round t, for any pair (i, j) for which $n_{ij}(t) = 0$ since in these cases $\ell_{ij}(t) \le 0$ and $u_{ij}(t) \ge 1$.

Now consider any pair (i, j) and round $t \in [T]$ such that $n_{ij}(t) > 0$. Note in this case by Hoeffding's Inequality:

$$Pr\left(|p_{ij} - \widehat{p}_{ij}(t)| > \sqrt{\frac{\ln(Kt/\delta)}{n_{ij}(t)}}\right) \le 2e^{-2n_{ij}(t)\frac{\ln(Kt/\delta)}{n_{ij}(t)}} = \frac{2\delta^2}{K^2t^2} \le \frac{\delta}{K^2t^2}\,.$$

Taking union bound over all $\binom{K}{2}$ pairs and time $t \in [T]$ we get:

$$Pr\left(\exists i, j \in [K], t \in [T] \text{ s.t. } |p_{ij} - \widehat{p}_{ij}(t)| > \sqrt{\frac{\ln(Kt/\delta)}{n_{ij}(t)}}\right)$$
$$\le \sum_{t=1}^{T}\sum_{i=2}^{K}\sum_{j=1}^{i}\frac{\delta}{K^2t^2} \le \sum_{t=1}^{\infty}\frac{\delta}{2t^2} \le \frac{\delta\pi^2}{12} \le \delta,$$

where in the second last inequality we used $\sum_{t=1}^{\infty}\frac{1}{t^2} < \frac{\pi^2}{6}$. This concludes the claim. □

B. Regret Analysis of Alg. 3

Theorem 3 (Regret analysis of `Versatile-DB` (Alg. 3) for Adversarial and Stochastic Preferences)**.** *For any sequence of preference matrices $\mathbf{P}_t$, the pseudo-regret of Algorithm 3 with $\Psi_t(w) = \sqrt{t}\sum_{k=1}^{K}(\sqrt{w_k} - w_k/2)/8$ satisfies for any $T \ge 1$*

$$\overline{R}_T := \max_{k\in[K]} \mathbf{E}\big[R_T(k)\big] \le 4\sqrt{KT} + 1.$$

Furthermore, if there exists a gap vector $\Delta \in [0, 1]^K$ with a unique zero coordinate $k^* \in [K]$ and $C \ge 0$ such that

$$\overline{R}_T \ge \frac{1}{2}\mathbf{E}\Bigg[\sum_{t=1}^{T}\sum_{k\ne k^*}\big(p_{+1,t}(k) + p_{-1,t}(k)\big)\Delta_k\Bigg] - C\,, \tag{4}$$

the pseudo regret also satisfies

$$\overline{R}_T \le \sum_{k\ne k^*}\frac{4\log T + 12}{\Delta_k} + 4\log T + \frac{1}{\Delta_{\min}} + \frac{3}{2}\sqrt{K} + 8 + C.$$

where $\Delta_{\min} = \min_{k\ne k^*}\Delta_k$.

Proof of Theorem 3. The analysis follows from carefully combining our reduction (Theorem 2) with Theorem 1 of (Zimmert & Seldin, 2021) for MAB to both of the players. Indeed, for each player $i \in \{-1, 1\}$, Algorithm 3 chooses $k_{i,t}$ by following the decisions of Tsallis-INF (Zimmert & Seldin, 2021, Alg. 1) with $\alpha = 1/2$, symmetric regularization, learning rate $\eta_t = 4/\sqrt{t}$ and losses $\ell_{i,t}$ estimated in (3) with standard importance sampling (IW).

Adversarial regime A direct application of Theorem 1 of (Zimmert & Seldin, 2021), upper-bounds the pseudo-regret for each player $i \in \{-1, 1\}$ as

$$\max_{k\in[K]} \mathbf{E}\big[R_{i,T}(k)\big] \le 4\sqrt{KT} + 1\,.$$

Combining the about bounds with the reduction from MAB to DB (Theorem 2) yields the adversarial pseudo-regret upper-bound

$$\mathbf{E}\big[R_T(k)\big] = \frac{1}{2}\mathbf{E}\big[R_{-1,T}(k) + R_{+1,T}(k)\big] \le 4\sqrt{KT} + 1.$$

Adversarial regime with a self-bounding constraint Our self-bounding constraint is slightly different from that of (Zimmert & Seldin, 2021), since it involves both players simultaneously. This is necessary so that our gap vector Δ can recover the standard suboptimality gaps used in stochastic dueling bandits. Thus, we cannot directly combine their result with our black-box reduction in this case. However, the proof largely follows their analysis, except that the upper-bounds on the regret of both players must be combined in the middle of their analysis, just before they apply their self-bounding constraint assumption. Thus, we give here only the modification to the proof of their Theorem 1.

Following their proof of Thm. 1 until their pseudo-regret bound at the top of p. 23, we get for each player $i \in \mathcal{I} := \{-1, +1\}$:

$$\mathbf{E}\big[R_{i,T}(k)\big] \leq \sum_{k \neq k^*} \left(\sum_{t=1}^{T} \frac{\sqrt{\mathbf{E}[p_{i,t}(k)]}}{\sqrt{t}} + \sum_{t=T_0+1}^{T} \frac{\mathbf{E}[p_{i,t}(k)]}{4\sqrt{t}} \right) + M\,,$$

where $M := \sqrt{T_0} + \frac{3}{4}\sqrt{K} + 15 + 14K\log(T)$ and $T_0 := \lceil \Delta_{\min}^{-2}/4 \rceil$. Together with Theorem 2 and taking the $\max$ over k, $\overline{R}_T$ is thus upper-bounded by

$$\frac{1}{2} \sum_{i \in \mathcal{I}} \sum_{k \neq k^*} \left(\sum_{t=1}^{T} \frac{\sqrt{\mathbf{E}[p_{i,t}(k)]}}{\sqrt{t}} + \sum_{t=T_0+1}^{T} \frac{\mathbf{E}[p_{i,t}(k)]}{4\sqrt{t}} \right) + M\,.$$

Now, applying the self-bounding property (4) we get for any $\lambda \in [0,1]$

$$\overline{R}_T \leq \overline{R}_T + \lambda \Bigg(\overline{R}_T - \frac{1}{2}\mathbf{E}\Bigg[\sum_{t=1}^{T} \sum_{k \neq k^*} \big(p_{+1,t}(k) + p_{-1,t}(k)\big)\Delta_k \Bigg] + C \Bigg)$$

Thus, combined with the previous bound using $1 + \lambda \leq 2$

$$\begin{aligned}\overline{R}_T &\leq \frac{1}{2} \sum_{i \in \mathcal{I}} \sum_{k \neq k^*} \left(\sum_{t=1}^{T} \left(\frac{2\sqrt{\mathbf{E}[p_{i,t}(k)]}}{\sqrt{t}} - \lambda \Delta_k \mathbf{E}[p_{i,t}] \right) + \sum_{t=T_0+1}^{T} \frac{\mathbf{E}[p_{i,t}(k)]}{2\sqrt{t}} \right) + 2M + \lambda C\,. \\ &\leq \sum_{k \neq k^*} \left(\sum_{t=1}^{T_0} \max_{z \geq 0} \left\{ \frac{2\sqrt{z}}{\sqrt{t}} - \lambda \Delta_k z \right\} \sum_{t=T_0+1}^{T} \max_{z \geq 0} \left\{ \frac{2\sqrt{z} + \frac{1}{2}z}{\sqrt{t}} - \lambda \Delta_i z \right\} \right) + 2M + \lambda C\end{aligned}$$

Now, we are back with the same upper-bound (Zimmert & Seldin, 2021) have in the middle of their page 23. Following their analysis by solving the optimization problems, summing over t, and optimizing λ concludes. □