Best-of-Both-Worlds Multi-Dueling Bandits: Unified Algorithms for Stochastic and Adversarial Preferences under Condorcet and Borda Objectives

S Akash, Pratik Gajane, Jawar Singh

Keywords: Dueling bandits, best-of-both-worlds, Condorcet winner, Borda winner

Summary

Multi-dueling bandits, where a learner selects $m \geq 2$ arms per round and observes only the winner, arise naturally in many applications including ranking and recommendation systems, yet a fundamental question has remained open: can a single algorithm perform optimally in both stochastic and adversarial environments, without knowing which regime it faces? We answer this affirmatively, providing the first best-of-both-worlds algorithms for multi-dueling bandits under both Condorcet and Borda objectives. For the Condorcet setting, we propose `MetaDueling`, a black-box reduction that converts any dueling bandit algorithm into a multi-dueling bandit algorithm by transforming multi-way winner feedback into an unbiased pairwise signal. Instantiating our reduction with `Versatile-DB` yields the first best-of-both-worlds algorithm for multi-dueling bandits: it achieves $O(\sqrt{KT})$ pseudo-regret against adversarial preferences and the instance-optimal $O\left(\sum_{i \neq a^\star} \frac{\log T}{\Delta_i}\right)$ pseudo-regret under stochastic preferences, both simultaneously and without prior knowledge of the regime. For the Borda setting, we propose `SA-MiDEX`, a stochastic-and-adversarial algorithm that achieves $O\left(K^2 \log KT + K \log^2 T + \sum_{i:\Delta_i^{\mathrm{B}}>0} \frac{K \log KT}{(\Delta_i^{\mathrm{B}})^2}\right)$ regret in stochastic environments and $O\left(K\sqrt{T \log KT} + K^{1/3}T^{2/3}(\log K)^{1/3}\right)$ regret against adversaries, without prior knowledge of the regime. We complement our upper bounds with matching lower bounds for the Condorcet setting. For the Borda setting, our stochastic upper bound exceeds the lower bound by a factor of $O(K/\Delta_i^B)$ per suboptimal arm, while, for $T > K^4$, our adversarial upper bound matches the lower bound up to logarithmic factors and matches the best-known result in the literature.

Contribution(s)

1. A black-box reduction (`MetaDueling`) from multi-dueling bandits to dueling bandits that preserves expected pseudo-regret under the Condorcet winner assumption.
   Context: The reduction builds on ideas from Gajane (2024) and Saha & Gopalan (2018).
2. The first best-of-both-worlds algorithm for multi-dueling bandits in the Condorcet setting, achieving $O(\sqrt{KT})$ adversarial pseudo-regret and $O\left(\sum_{i \neq a^\star} \frac{\log T}{\Delta_i}\right)$ instance-optimal stochastic pseudo-regret, simultaneously and without prior knowledge of the regime.
   Context: Obtained by combining our `MetaDueling` reduction with `Versatile-DB` (Saha & Gaillard, 2022).
3. A stochastic-and-adversarial algorithm (`SA-MiDEX`) for multi-dueling bandits with Borda objectives, achieving $O\left(K^2 \log KT + K \log^2 T + \sum_{i:\Delta_i^{\mathrm{B}}>0} \frac{K \log KT}{(\Delta_i^{\mathrm{B}})^2}\right)$ regret in stochastic settings and $O\left(K\sqrt{T \log KT} + K^{1/3}T^{2/3}(\log K)^{1/3}\right)$ regret in adversarial settings.
   Context: Adapts the stochastic-and-adversarial paradigm to multi-dueling bandits using Successive Elimination for Borda estimation and an EXP3-style fallback (Gajane, 2024).
4. Condorcet lower bounds establishing the optimality of our guarantees in both adversarial and stochastic regimes, and Borda lower bounds against which our adversarial upper bound is near-optimal, matching up to logarithmic factors for $T > K^4$.
   Context: Derived via reduction from known dueling bandit lower bounds, and consistent with the established $\Omega(K^{1/3}T^{2/3})$ baseline for adversarial Borda objectives.

Best-of-Both-Worlds Multi-Dueling Bandits: Unified Algorithms for Stochastic and Adversarial Preferences under Condorcet and Borda Objectives

S Akash[1], **Pratik Gajane**[2], **Jawar Singh**[1]

[1]**Indian Institute of Technology Patna, India**
[2]**Laboratoire d'Informatique Fondamentale d'Orléans, University of Orléans, France**

Abstract

Multi-dueling bandits, where a learner selects $m \geq 2$ arms per round and observes only the winner, arise naturally in many applications including ranking and recommendation systems, yet a fundamental question has remained open: can a single algorithm perform optimally in both stochastic and adversarial environments, without knowing which regime it faces? We answer this affirmatively, providing the first best-of-both-worlds algorithms for multi-dueling bandits under both Condorcet and Borda objectives. For the Condorcet setting, we propose `MetaDueling`, a black-box reduction that converts any dueling bandit algorithm into a multi-dueling bandit algorithm by transforming multi-way winner feedback into an unbiased pairwise signal. Instantiating our reduction with `Versatile-DB` yields the first best-of-both-worlds algorithm for multi-dueling bandits: it achieves $O(\sqrt{KT})$ pseudo-regret against adversarial preferences and the instance-optimal $O\left(\sum_{i \neq a^\star} \frac{\log T}{\Delta_i}\right)$ pseudo-regret under stochastic preferences, both simultaneously and without prior knowledge of the regime. For the Borda setting, we propose `SA-MiDEX`, a stochastic-and-adversarial algorithm that achieves $O\left(K^2 \log KT + K \log^2 T + \sum_{i:\Delta_i^{\mathrm{B}}>0} \frac{K \log KT}{(\Delta_i^{\mathrm{B}})^2}\right)$ regret in stochastic environments and $O\left(K\sqrt{T \log KT} + K^{1/3}T^{2/3}(\log K)^{1/3}\right)$ regret against adversaries, again without prior knowledge of the regime. We complement our upper bounds with matching lower bounds for the Condorcet setting. For the Borda setting, our stochastic upper bound exceeds the lower bound by a factor of $O(K/\Delta_i^B)$ per suboptimal arm, while, for $T > K^4$, our adversarial upper bound matches the lower bound up to logarithmic factors and matches the best-known result in the literature.

1 Introduction

Multi-armed bandits with preference feedback have emerged as a powerful framework for applications where absolute numerical rewards are difficult to obtain, but relative comparisons are natural, such as information retrieval (Yue & Joachims, 2009) and recommendation systems. In the classical *dueling bandits* setting, a learner selects two arms per round and observes which is preferred. Recently, *multi-dueling bandits* (Saha & Gopalan, 2018; Brost et al., 2016) have extended this framework to settings where $m \geq 2$ arms are compared simultaneously, with only the winner revealed (e.g., ranker evaluation based on user clicks).

A fundamental question in online learning is whether algorithms can achieve optimal performance in both stochastic and adversarial environments without prior knowledge of the regime. Such *best-of-both-worlds* guarantees are established for multi-armed (Zimmert & Seldin, 2021) and dueling bandits (Saha & Gaillard, 2022), but remain unexplored for multi-dueling bandits. Existing work addresses these regimes separately: Saha & Gopalan (2018) achieve $O(K \log T/\Delta)$ regret

in stochastic multi-dueling bandits under the *Condorcet* assumption, while Gajane (2024) achieve $O((K \log K)^{1/3} T^{2/3})$ regret in adversarial settings using *Borda* scores.

1.1 Contributions

We provide the first best-of-both-worlds algorithms for multi-dueling bandits, addressing both Condorcet and Borda objectives:

1. **Condorcet Setting: The `MetaDueling` Reduction.** We propose `MetaDueling`, a black-box reduction from multi-dueling to dueling bandits. By constructing multisets containing only two distinct arms, we extract unbiased pairwise comparison information from multi-way winner feedback, as the bias from duplicate arms corresponds to an affine rescaling of the preference matrix that preserves the Condorcet winner and gap ordering (Lemma C.1). Instantiating this with `Versatile-DB` (Saha & Gaillard, 2022) yields the first best-of-both-worlds multi-dueling guarantee (Theorem 4.1): $O(\sqrt{KT})$ adversarial pseudo-regret and instance-optimal $O(\sum_{i \neq a^\star} \log T / \Delta_i)$ stochastic pseudo-regret, independently of the subset size m.

2. **Borda Setting: The `SA-MiDEX` Algorithm.** We propose `SA-MiDEX` for Borda winners. Since the Borda objective requires estimating performance against all opponents, `SA-MiDEX` initially assumes stochastic preferences using successive elimination strategy to obtain unbiased estimates. It monitors for adversarial deviations via martingale concentration; upon detection, it irreversibly switches to adversarial mode. This yields $O\left(K^2 \log KT + K \log^2 T + \sum_{i:\Delta_i^{\mathrm{B}}>0} \frac{K \log KT}{(\Delta_i^{\mathrm{B}})^2}\right)$ stochastic regret and $O\left(K\sqrt{T \log KT} + K^{1/3} T^{2/3} (\log K)^{1/3}\right)$ adversarial regret (Theorems 5.3 and 5.4).

3. **Lower Bounds.** We prove matching lower bounds of $\Omega(\sqrt{KT})$ for adversarial Condorcet regret and $\Omega(\sum_{i \neq a^\star} \log T / \Delta_i)$ for stochastic Condorcet regret (Theorems 4.4 and 4.5), establishing the optimality of our Condorcet guarantees. We also establish a lower bound of $\Omega\big(\sum_{i:\Delta_i^{\mathrm{B}}>0} \frac{\log T}{(\Delta_i^{\mathrm{B}})}\big)$ for stochastic regret under Borda objectives.

1.2 Related work

Multi-dueling bandits (Brost et al., 2016) generalize the standard dueling framework (Yue & Joachims, 2009) to simultaneous m-way comparisons. Under the challenging winner-only feedback model, Saha & Gopalan (2018) established gap-dependent $O(K \log T / \Delta)$ regret bounds for purely stochastic environments with Condorcet winners. More recently, Gajane (2024) studied the adversarial multi-dueling regime, achieving $O((K \log K)^{1/3} T^{2/3})$ regret under the Borda objective. However, neither of these approaches can automatically adapt when the nature of the environment (stochastic versus adversarial) is unknown.

The quest for algorithms that adapt seamlessly to either regime, termed *best-of-both-worlds*, has seen significant breakthroughs via Tsallis entropy regularization (Zimmert & Seldin, 2019; 2021). Saha & Gaillard (2022) successfully extended this machinery to standard dueling bandits ($m = 2$), achieving optimal rates in both regimes simultaneously. Our `MetaDueling` reduction directly leverages their `Versatile-DB` algorithm to elevate these dual guarantees to the multi-dueling setting. For the Borda setting, we draw upon the stochastic-and-adversarial paradigm introduced by Auer & Chiang (2016). Rather than relying on regularization, this approach actively monitors empirical observations for adversarial deviations, permanently switching from a stochastic successive elimination strategy to a robust adversarial one upon detection. A comprehensive discussion of the broader literature, including alternative choice models and preference structures, is deferred to Appendix A.

2 Preliminaries

We consider online learning over a finite set of K arms $[K] := \{1, \dots, K\}$ for T rounds. At each round $t \in [T]$, the environment obliviously selects a preference matrix $P_t \in [0,1]^{K \times K}$ satisfying reciprocity ($P_t(i,j) + P_t(j,i) = 1$) and self-comparison ($P_t(i,i) = 1/2$) for all $i, j \in [K]$. The learner selects a multiset $\mathcal{A}_t \subseteq [K]$ of size $|\mathcal{A}_t| = m$ ($2 \le m \le K$) and observes a winning arm. In the *stochastic setting*, preferences are stationary ($P_t = P$ for all t); in the *adversarial setting*, the sequence $\{P_t\}_{t=1}^T$ is arbitrary.

Choice Model. Following Saha & Gopalan (2018), we assume a pairwise-subset choice model. Given a multiset $\mathcal{A}$ and preference matrix P, the probability that the arm at index $i \in [m]$ wins is determined by averaging pairwise preferences: $W(i \mid \mathcal{A}, P) := \sum_{j \neq i} \frac{2P(\mathcal{A}(i), \mathcal{A}(j))}{m(m-1)}$.

Condorcet Setting. The Condorcet framework assumes a globally dominant arm.

Assumption 2.1 (Condorcet Winner and Gap)**.** *We assume there exists a Condorcet winner $a^\star \in [K]$ such that $P_t(a^\star, i) \ge 1/2$ for all $i \in [K]$ and $t \in [T]$. We define the sub-optimality gap for arm i at round t as $\Delta_t(i) := P_t(a^\star, i) - 1/2$. In the stochastic setting, we write $\Delta(i) := \Delta_t(i)$ and let $\Delta_{\min} := \min_{i \neq a^\star} \Delta(i) > 0$.*

The cumulative Condorcet pseudo-regret measures the average gap of selected arms: $R_T^{\mathrm{C}} := \sum_{t=1}^T \frac{1}{m} \sum_{j=1}^m \Delta_t(\mathcal{A}_t(j))$. Note that standard dueling bandit regret is simply the special case with $m = 2$.

Borda Setting. Alternatively, optimality can be measured by average pairwise performance. The Borda score of arm i at round t is the probability it beats an opponent selected uniformly at random: $b_t(i) := \frac{1}{K-1} \sum_{j \neq i} P_t(i,j)$.

For the relationship between Condorcet setting and Borda setting, see Appendix B.

3 The MetaDueling Reduction

The transition from pairwise to multi-way preferences introduces a significant structural challenge: most multi-dueling algorithms are designed for specific, narrow environments. This lack of modularity often prevents the direct application of well-optimized dueling bandit algorithms to multi-player settings, requiring new, complex strategies for every different scenarios.

We present `MetaDueling`, a black-box reduction that decouples the multi-way feedback mechanism from the core learning strategy. Our approach constructs the multiset $\mathcal{A}_t$ by populating its m available slots with only two distinct arms, i and j. By isolating these two arms, we collapse the multi-player interaction into a controlled pairwise signal, allowing multi-way winner probabilities to be mapped directly back to pairwise preferences without needing to model the internal dynamics of the selection process.

The primary advantage of `MetaDueling` is its inherent modularity. By treating the multi-dueling environment as a black-box interface, our framework can be seamlessly integrated with any standard dueling bandit learner—including stochastic, adversarial, or contextual variants. This plug-and-play architecture ensures that `MetaDueling` inherits the regret guarantees of the base learner while remaining extensible to any preference learning setting where multi-way feedback is available.

3.1 Algorithm Description

The reduction operates as follows: at each round, the base dueling bandit algorithm proposes two arms (x_t, y_t). We construct a multiset containing only these two arms, with counts (n_x, n_y) summing to m. When m is even, we set $n_x = n_y = m/2$. When m is odd, we randomize: with

Algorithm 1 `MetaDueling`: Reduction from Multi-Dueling to Dueling Bandits

Require: Base dueling bandit algorithm $\mathcal{D}$, horizon T, arms K, subset size m

Initialize base algorithm $\mathcal{D}$ (e.g., `Versatile-DB`)
for $t = 1, 2, \ldots, T$ **do**
  // Query base algorithm for arm pair
  Let $(q_t^{(1)}, q_t^{(2)})$ be the distributions maintained by $\mathcal{D}$
  Sample $x_t \sim q_t^{(1)}$ and $y_t \sim q_t^{(2)}$ independently
  // Construct multiset with symmetrization
  Sample $B_t \sim \text{Bernoulli}(1/2)$
  if m is even **then**
    $(n_x, n_y) \leftarrow (m/2, m/2)$
  else if $B_t = 1$ **then**
    $(n_x, n_y) \leftarrow (\lceil m/2 \rceil, \lfloor m/2 \rfloor)$
  else
    $(n_x, n_y) \leftarrow (\lfloor m/2 \rfloor, \lceil m/2 \rceil)$
  end if
  Construct $\mathcal{A}_t = \{\underbrace{x_t, \ldots, x_t}_{n_x}, \underbrace{y_t, \ldots, y_t}_{n_y}\}$
  // Play and observe feedback
  Play $\mathcal{A}_t$ and observe winner index $I_t \in [m]$
  Set $o_t \leftarrow \mathbb{1}[\mathcal{A}_t(I_t) = x_t]$
  // Feed outcome to base algorithm
  Update $\mathcal{D}$ with outcome o_t for pair (x_t, y_t)
end for

probability $1/2$, we set $(n_x, n_y) = (\lceil m/2 \rceil, \lfloor m/2 \rfloor)$, and with probability $1/2$, we swap. This symmetrization ensures unbiased feedback in expectation.

3.2 Rescaled Preferences

The multiset construction introduces bias from ties between identical arms: when multiple copies of x_t compete, some probability mass goes to x_t-vs-x_t comparisons (which x_t wins with probability $1/2$) rather than x_t-vs-y_t comparisons. We show that this bias corresponds exactly to an affine rescaling of the preference matrix.

Definition 3.1 (Rescaling Constants)**.** *Define the rescaling constants:*

$$\beta_m := \begin{cases} \dfrac{m}{2(m-1)} & \text{if } m \text{ is even,} \\ \dfrac{m+1}{2m} & \text{if } m \text{ is odd,} \end{cases} \qquad \alpha_m := \frac{1 - \beta_m}{2}. \tag{1}$$

Definition 3.2 (Rescaled Preference Matrix)**.** *The rescaled preference matrix is:*

$$\widehat{P}_t(i, j) := \alpha_m + \beta_m P_t(i, j). \tag{2}$$

The rescaled preference matrix $\widehat{P}_t$ is a valid preference matrix and preserves the Condorcet winner (see Appendix C for the proof).

The following lemma shows that the observed outcome o_t is an unbiased estimate of the rescaled preference, not the original preference.

Lemma 3.3 (Unbiased Feedback)**.** *Under Algorithm 1, the observed outcome satisfies:*

$$\mathbb{E}[o_t \mid x_t, y_t] = \widehat{P}_t(x_t, y_t) = \alpha_m + \beta_m P_t(x_t, y_t).$$

Proof. See Appendix D.1. □

Remark 3.4 (Implicit Rescaling). *The raw outcome $o_t \in \{0,1\}$ is* not *an unbiased estimate of $P_t(x_t, y_t)$ due to the tie bias. However, o_t* is *an unbiased estimate of $\widehat{P}_t(x_t, y_t) \in [0,1]$. Since the rescaled preferences preserve all structural properties needed by the base algorithm (Lemma C.1), we can feed o_t directly to $\mathcal{D}$ without explicit transformation.*

3.3 Gap Rescaling

The sub-optimality gaps are also rescaled by the factor β_m.

Lemma 3.5 (Gap Rescaling). *For any arm $i \in [K]$, the rescaled gap satisfies:*

$$\widehat{\Delta}_t(i) := \widehat{P}_t(a^\star, i) - \tfrac{1}{2} = \beta_m \, \Delta_t(i).$$

Lemma 3.6 (Bound on Rescaling Factor). *For all $m \geq 2$, the rescaling factor satisfies $\beta_m \geq \frac{1}{2}$, with equality as $m \to \infty$.*

For proofs of Lemma 3.5 and Lemma 3.6, see Appendix D.2.

3.4 Regret Equivalence

The key result of this section establishes that the multi-dueling regret equals the dueling regret in rescaled preferences, up to the factor $1/\beta_m \leq 2$.

Lemma 3.7 (Regret Equivalence). *For any sequence of preference matrices $P_1, \ldots, P_T$ satisfying Assumption 2.1:*

$$\mathbb{E}[R_T^{\mathrm{C}}] = \frac{1}{\beta_m} \mathbb{E}[\widehat{R}_T^{\mathrm{DB}}], \tag{3}$$

where $\widehat{R}_T^{\mathrm{DB}} := \sum_{t=1}^{T} \frac{1}{2}(\widehat{\Delta}_t(x_t) + \widehat{\Delta}_t(y_t))$ is the pseudo-regret of the base algorithm under $\widehat{P}_t$.

For the proof of Lemma 3.7, see Appendix D.3.

Remark 3.8 (Black-Box Nature of the Reduction). *We only require that the base algorithm $\mathcal{D}$ accepts binary outcomes $o_t \in \{0,1\}$ and achieves bounded pseudo-regret on valid preference matrices. No modification to $\mathcal{D}$'s internals is needed. This modularity means that future improvements to dueling bandit algorithms automatically transfer to multi-dueling bandits via our reduction.*

Remark 3.9 (Special Case: $m = 2$). *For $m = 2$, we have $\alpha_2 = 0$ and $\beta_2 = 1$, so $\widehat{P}_t = P_t$. The multiset $\mathcal{A}_t = \{x_t, y_t\}$ is a standard pair, and Algorithm 1 reduces exactly to the base algorithm $\mathcal{D}$.*

4 Main Results: Condorcet Setting

While previous literature has addressed stochastic and adversarial multi-dueling environments in isolation, we demonstrate that feeding our rescaled pairwise signals into a best-of-both-worlds base algorithm seamlessly bridges this divide. This approach eliminates the need to manually adapt to environmental shifts, as Alg. 1 inherits the robust properties of the underlying learner.

By instantiating `MetaDueling` with `Versatile-DB` (Saha & Gaillard, 2022), we obtain the first unified guarantees for multi-dueling bandits: optimal adversarial pseudo-regret and instance-optimal stochastic pseudo-regret, achieved simultaneously and without prior knowledge of the regime. This reduction provides a rigorous theoretical foundation for multi-way feedback that remains valid across settings.

The strength of Theorem 4.1 lies in the rigorous mapping between multi-way winner probabilities and the rescaled pairwise preference matrix $\widehat{P}_t$. By proving that the symmetrized multiset construction in `MetaDueling` yields an unbiased pairwise signal, we ensure that the base learner $\mathcal{D}$ operates on a valid dueling bandit instance. The core of the proof establishes a strict linear regret

equivalence $1/\beta_m$, which allows the $O(\sqrt{KT})$ adversarial and $O(\sum \log T/\Delta_i)$ stochastic guarantees of `Versatile-DB` to be transferred exactly to the multi-dueling setting. This transformation demonstrates that the complexity of the m-way interaction is captured by a universal constant factor ($1/\beta_m \leq 2$), preserving the optimality of the underlying dueling bandit algorithm regardless of the number of arms compared simultaneously.

4.1 Best-of-Both-Worlds Guarantee

Theorem 4.1 (Main Result: Best-of-Both-Worlds for Condorcet Multi-Dueling)**.** *Under Assumption 2.1, `MetaDueling` (Algorithm 1) instantiated with `Versatile-DB` satisfies the following simultaneously for any $T \geq 1$:*

(i) Adversarial pseudo-regret bound. *For any oblivious sequence of preference matrices $P_1, \ldots, P_T$:*

$$\mathbb{E}[R_T^{\mathrm{C}}] \leq \frac{1}{\beta_m}\left(4\sqrt{KT}+1\right) \leq 8\sqrt{KT}+2. \tag{4}$$

(ii) Bound for self-bounding preference matrices. *If the preference matrices satisfy the self-bounding condition*

$$\mathbb{E}[\widehat{R}_T^{\mathrm{DB}}] \geq \frac{1}{2}\,\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\neq a^\star}\left(p_{+1,t}(i)+p_{-1,t}(i)\right)\widehat{\Delta}(i)\right], \tag{5}$$

where $p_{+1,t}$ and $p_{-1,t}$ are the arm distributions of the two players in `Versatile-DB`, then:

$$\mathbb{E}[R_T^{\mathrm{C}}] \leq \frac{1}{\beta_m^2}\sum_{i\neq a^\star}\frac{16\log T+48}{\Delta_i}+\frac{4\log T+12}{\beta_m}+\frac{8\sqrt{K}}{\beta_m}. \tag{6}$$

Both bounds hold simultaneously without prior knowledge of the regime.

We emphasize that condition (5) is not an assumption needed for our framework; it is required by `Versatile-DB` (Saha & Gaillard, 2022, Theorem 3). The condition (5) appears in Theorem 4.1 solely because we instantiated `MetaDueling` with `Versatile-DB`. Our framework is agnostic to the choice of the base learner as mentioned in Remark 3.8. If `MetaDueling` is instantiated with another base learner which does not require condition (5), the latter's guarantees would transfer to the multi-dueling setup seamlessly.

The proof of Theorem 4.1 leverages the insight that the algorithm's observed outcomes can be rescaled to form valid dueling bandit feedback, allowing the application of existing guarantees for the `Versatile-DB` algorithm. In the adversarial case, this directly yields sublinear pseudo-regret bounds, while in the self-bounding setting, the rescaling of gaps enables the translation of known gap-dependent guarantees to the Condorcet setting. This approach simultaneously harnesses prior work on dueling bandits and our novel rescaling technique, producing the first best-of-both-worlds bounds for multi-dueling bandits under Condorcet objectives. For the complete proof of Theorem 4.1, see Appendix D.4.

4.2 Stochastic Condorcet Setting

In the stochastic setting with a fixed Condorcet winner, the self-bounding condition (5) is automatically satisfied, yielding instance-optimal regret.

Corollary 4.2 (Instance-Optimal Stochastic Regret)**.** *In the stochastic setting with a fixed preference matrix $P_t = P$ for all $t \in [T]$ and Condorcet winner $a^\star$, the self-bounding condition (5) holds with equality. Consequently:*

$$\mathbb{E}[R_T^{\mathrm{C}}] \leq \sum_{i\neq a^\star}\frac{64\log T+192}{\Delta_i}+8\log T+24+16\sqrt{K} = O\left(\sum_{i\neq a^\star}\frac{\log T}{\Delta_i}\right). \tag{7}$$

Proof. In the stochastic setting, $\Delta_t(i) = \Delta(i)$ for all $t \in [T]$, and the Condorcet winner $a^\star$ is fixed. The self-bounding condition holds because pulling suboptimal arms directly contributes to regret proportional to their gaps.

The bound follows from Theorem 4.1(ii) with $1/\beta_m \leq 2$ and $1/\beta_m^2 \leq 4$. □

Remark 4.3 (Comparison with Prior Work). *Saha & Gopalan (2018) achieved $O(K \log T/\Delta_{\min})$ regret for stochastic multi-dueling bandits with Condorcet winners. Our bound $O(\sum_{i \neq a^\star} \log T/\Delta_i)$ is instance-optimal: it depends on individual gaps rather than just the minimum gap, providing tighter guarantees when gaps are heterogeneous. Moreover, our algorithm simultaneously achieves $O(\sqrt{KT})$ adversarial regret, which Saha & Gopalan (2018) do not address.*

4.3 Lower Bounds

We establish matching lower bounds that demonstrate the optimality of our guarantees.

Theorem 4.4 (Adversarial Lower Bound). *For any multi-dueling bandit algorithm and any $T \geq K \geq 4$, there exists an adversarial instance satisfying Assumption 2.1 such that:*

$$\mathbb{E}[R_T^{\mathrm{C}}] \geq \Omega(\sqrt{KT}). \tag{8}$$

Proof Sketch. The proof of the adversarial lower bound proceeds via a reduction from dueling bandits. First, any multi-dueling bandit algorithm is converted into a standard dueling bandit algorithm by randomly selecting a pair of arms from the chosen subset and using the observed duel outcome to simulate multi-dueling feedback. This ensures that the expected regret of the constructed dueling algorithm matches that of the original multi-dueling algorithm. Next, the standard minimax lower bound for adversarial multi-armed bandits is applied, and via known reductions, this yields a hard adversarial instance with a best arm in hindsight. Finally, the reduction shows that if a multi-dueling algorithm could achieve lower regret, it would violate the dueling bandit lower bound, establishing the desired $\Omega(\sqrt{KT})$ bound for multi-dueling setting. □

Theorem 4.5 (Stochastic Lower Bound). *For any multi-dueling bandit algorithm, there exists a stochastic instance with Condorcet winner $a^\star$ such that:*

$$\mathbb{E}[R_T^{\mathrm{C}}] \geq \Omega\left(\sum_{i \neq a^\star} \frac{\log T}{\Delta_i}\right). \tag{9}$$

Proof Sketch. The proof follows the same reduction structure as Theorem 4.4, but applies the stochastic dueling bandit lower bound of Komiyama et al. (2015):

$$\liminf_{T \to \infty} \frac{\mathbb{E}[R_T^{\mathrm{DB}}]}{\log T} \geq \sum_{i \neq a^\star} \frac{1}{2\,\mathrm{KL}(P(a^\star, i) \| \frac{1}{2})}.$$

For gaps $\Delta_i = P(a^\star, i) - \frac{1}{2}$, a Taylor expansion gives $\mathrm{KL}(P(a^\star, i) \| \frac{1}{2}) = \Theta(\Delta_i^2)$, yielding the $\Omega(\sum_{i \neq a^\star} \log T/\Delta_i)$ lower bound.

See Appendix D.5 for the complete proof. □

Remark 4.6 (Optimality). *The matching upper and lower bounds establish that our algorithm achieves optimal pseudo-regret in both adversarial and stochastic settings. The adversarial bound $\Theta(\sqrt{KT})$ matches the minimax rate for multi-armed bandits, and the stochastic bound $\Theta(\sum_{i \neq a^\star} \log T/\Delta_i)$ matches the instance-optimal rate for dueling bandits.*

Remark 4.7 (Pseudo-Regret vs. High-Probability Bounds). *Our bounds are in expectation (pseudo-regret). As shown in Auer & Chiang (2016), no algorithm can simultaneously achieve optimal high-probability bounds in both stochastic and adversarial regimes. Our pseudo-regret formulation is therefore the natural target for best-of-both-worlds guarantees.*

5 Borda Setting: The SA-MiDEX Algorithm

We now turn to multi-dueling bandits with Borda objectives. Unlike the Condorcet setting, the Borda objective requires estimating each arm's average performance against *all* opponents, not just the specific arm it was compared against. In this context, we propose SA-MiDEX (stochastic-and-Adversarial Multi-Dueling with EXploration), an algorithm based on the stochastic-and-adversarial paradigm of Auer & Chiang (2016). The algorithm initially assumes a stochastic environment, monitors for adversarial deviations, and switches to a robust adversarial strategy upon detection.

The Borda setting introduces unique challenges: estimating each arm's Borda score requires comparisons against all other arms, no Condorcet winner is guaranteed (so preferences may be cyclic or change over time), and multi-dueling feedback provides limited information for importance-weighted estimation. To address these, the approach combines decaying uniform opponent sampling with probability $p_t = \min(1, \frac{K \log t}{t})$ to maintain unbiased Borda estimates, a two-phase exploration-exploitation scheme, martingale-based deviation detection to identify adversarial shifts, and an EXP3 fallback to ensure robust adversarial regret.

5.1 Algorithm Description

The SA-MiDEX algorithm operates in two modes with an irreversible mode switch, leveraging the MetaDueling reduction (Algorithm 1) as a black box to convert multi-dueling feedback into pairwise comparisons.

Feedback Transformation. The MetaDueling reduction returns a raw outcome $o_t \in \{0, 1\}$ indicating whether arm x_t won. To obtain an unbiased estimate of the true preference probability $P(x_t, y_t)$, we apply the affine transformation:

$$g_t = \frac{o_t - \alpha_m}{\beta_m}, \tag{10}$$

where α_m and β_m are the bias and scaling constants from Lemma 3.3 that depend on the subset size m. This yields $\mathbb{E}[g_t \mid x_t, y_t] = P(x_t, y_t)$, enabling standard bandit estimation techniques despite the non-standard multi-dueling feedback model.

Stochastic Mode. The algorithm begins in stochastic mode, which operates via three mechanisms:

Stage 0: Baseline Exploration ($t \leq T_0$). For the first $T_0 = 4K(K-1)\lceil \log(2K^2T^2/\delta) \rceil$ rounds, the algorithm explores all ordered pairs via a round-robin schedule. This pure exploration phase is necessary to establish a high-confidence empirical baseline $\hat{P}_{T_0}(i, j)$ for the deviation detection mechanism.

Stage 1: Successive Elimination ($t > T_0$). The algorithm maintains an active set of candidate arms $\mathcal{C}$, initialized to $[K]$. In each round t, to guarantee unbiased Borda estimates, the algorithm selects $x_t \in \mathcal{C}$ (e.g., via robin over $\mathcal{C}$) and samples the opponent $y_t \sim \text{Uniform}([K] \setminus \{x_t\})$. Suboptimal arms are permanently eliminated from $\mathcal{C}$ once sufficient evidence accumulates. Specifically, arm $j \in \mathcal{C}$ is eliminated if there exists another arm $i \in \mathcal{C}$ such that $\hat{b}_t(i) - \text{conf}_t(i) > \hat{b}_t(j) + \text{conf}_t(j)$, where $\text{conf}_t(i) = \sqrt{\frac{20K \log(2K^2T/\delta)}{N_t(i)}}$. Once $|\mathcal{C}| = 1$, letting $\mathcal{C} = \{i^\star\}$, the algorithm exploits by playing $x_t = y_t = i^\star$.

Stage 2: Background Monitoring ($t > T_0$). To ensure the martingale deviation detection functions correctly in case of adversarial shifts, the algorithm forces random exploration with probability $p_t = \min(1, \frac{K \log t}{t})$ by sampling $x_t, y_t \sim \text{Uniform}([K])$. These uniform samples feed directly into the deviation detection statistics.

Algorithm 2 SA-MiDEX: Stochastic-and-Adversarial Multi-Dueling Bandits

Require: Arms K, horizon T, subset size m, confidence $\delta = 1/T^2$
Initialize: mode $\leftarrow$ STOCH, $\mathcal{C} \leftarrow [K]$, $T_0 \leftarrow 4K(K-1)\lceil \log(2K^2T^2/\delta) \rceil$
Init Stats: $N, S^{\text{obs}}, \tilde{S} \leftarrow 0$, $\hat{P} \leftarrow 1/2$
for $t = 1, \ldots, T$ **do**
  if mode $=$ STOCH **and** $t \leq T_0$ **then** ▷ Stage 0: Baseline Exploration
    $(x_t, y_t) \leftarrow$ ROUNDROBIN(t)
  else if mode $=$ STOCH **then** ▷ Stages 1 & 2: Elimination & Monitoring
    Sample $u_t \sim \text{Uniform}(0, 1)$
    if $u_t < \min(1, \frac{K \log t}{t})$ **then**
      $(x_t, y_t) \sim \text{Uniform}([K] \times [K])$
    else if $|\mathcal{C}| > 1$ **then**
      Pick $x_t \in \mathcal{C}$ (round-robin), $y_t \sim \text{Uniform}([K] \setminus \{x_t\})$
    else
      $(x_t, y_t) \leftarrow (i^\star, i^\star)$ where $\mathcal{C} = \{i^\star\}$
    end if
  else
    $(x_t, y_t) \leftarrow$ EXP3SELECT ▷ Sample from q in Eq. (12)
  end if
  $o_t \leftarrow \texttt{MetaDueling}(x_t, y_t, m)$; $g_t \leftarrow (o_t - \alpha_m)/\beta_m$; Update $N, S^{\text{obs}}, \hat{b}$
  if $t = T_0$ **then** $\hat{P}_{T_0}(i, j) \leftarrow S^{\text{obs}}(i, j)/N(i, j)$ for all (i, j) ▷ Lock baseline
  if mode $=$ STOCH **and** $t > T_0$ **then**
    Eliminate j from $\mathcal{C}$ if $\exists\, i \in \mathcal{C}$ s.t. $\hat{b}_t(i) - \text{conf}_t(i) > \hat{b}_t(j) + \text{conf}_t(j)$
    if $\exists\,(i, j) : D_t(i, j) > \tau(\sqrt{N_t(i, j) - N_{T_0}(i, j)} + 1)$ **then** mode $\leftarrow$ ADV; $\tilde{S} \leftarrow 0$
  else if mode $=$ ADV **then**
    $\tilde{S}(x_t) \leftarrow \frac{\mathbb{1}(i = x_t)}{K q_t(i)} \sum_{j \in [K]} \frac{\mathbb{1}(j = y_t) g_t}{q_t(j)}$
  end if
end for

Deviation Detection. Throughout exploitation, the algorithm monitors each pair (i, j) for deviation from the learned model by tracking:

$$D_t(i, j) := \left| S_t^{\text{obs}}(i, j) - S_{T_0}^{\text{obs}}(i, j) - (N_t(i, j) - N_{T_0}(i, j)) \cdot \hat{P}_{T_0}(i, j) \right|. \tag{11}$$

A deviation triggers when $D_t(i, j) > \tau(\sqrt{N_t(i, j) - N_{T_0}(i, j)} + 1)$ for threshold $\tau = \sqrt{8 \log(K^2T^2/\delta)}$.

Adversarial Mode. Upon detecting deviation, the algorithm irreversibly switches to EXP3 with explicit uniform exploration following Gajane (2024). The sampling distribution mixes a Gibbs distribution over cumulative importance-weighted scores with uniform exploration:

$$q_t(i) = (1 - \gamma) \frac{\exp(\eta \tilde{S}_t(i))}{\sum_{j=1}^{K} \exp(\eta \tilde{S}_t(j))} + \frac{\gamma}{K}, \tag{12}$$

ensuring $q_t(i) \geq \gamma/K$ for bounded inverse probability weights. The importance-weighted Borda score update uses the transformed feedback g_t and accounts for the sampling probabilities of both arms in the pair.

5.2 Importance-Weighted Borda Estimation

In adversarial mode, we use importance-weighted estimators for the shifted Borda score.

Definition 5.1 (Importance-Weighted Borda Estimator)**.** *Given arms $x_t, y_t \sim q_t$ and transformed feedback g_t, the importance-weighted shifted Borda score estimate is:*

$$\hat{s}_t(i) := \frac{\mathbb{1}(i = x_t)}{K q_t(i)} \sum_{j \in [K]} \frac{\mathbb{1}(j = y_t) g(m, o_t, x_t)}{q_t(j)} \tag{13}$$

Lemma 5.2 (Unbiased Estimation)**.** *The estimator $\hat{s}_t(i)$ is unbiased:*

$$\mathbb{E}[\hat{s}_t(i) \mid q_t] = s_t(i) = \frac{1}{K} \sum_{j \in [K]} P_t(i, j). \tag{14}$$

Proof. Taking expectations over $x_t, y_t \overset{\text{i.i.d.}}{\sim} q_t$:

$$\mathbb{E}[\hat{s}_t(i) \mid q_t] = \sum_{x,y} q_t(x) q_t(y) \cdot \frac{\mathbb{1}(i = x) \cdot \mathbb{E}[g_t \mid x, y]}{K \cdot q_t(i) \cdot q_t(y)} = \frac{1}{K} \sum_{y} P_t(i, y) = s_t(i),$$

where we used $\mathbb{E}[g_t \mid x_t, y_t] = P_t(x_t, y_t)$ from the feedback transformation. □

5.3 Main Results: Borda Setting

We now state our main results for the Borda setting. The proofs rely on concentration arguments for the stochastic phase and EXP3 analysis for the adversarial phase.

Theorem 5.3 (Stochastic Regret Bound)**.** *In the stochastic setting ($P_t = P$ for all t), Algorithm 2 achieves:*

$$\mathbb{E}[R_T^{\mathrm{B}}] \leq O\left(K^2 \log KT + K \log^2 T + \sum_{i : \Delta_i^{\mathrm{B}} > 0} \frac{K \log KT}{(\Delta_i^{\mathrm{B}})^2} \right) \tag{15}$$

Proof Sketch. Conditioned on the high-probability event $\mathcal{E}$ where confidence bounds hold and no false adversarial switch triggers, we decompose the regret into four phases. The baseline exploration of T_0 rounds contributes $O(K^2 \log KT)$. During successive elimination, selecting a suboptimal arm i against a uniformly sampled opponent incurs $\Theta(1)$ instantaneous regret; by Lemma 5.10, arm i is eliminated after at most $O(K \log KT / (\Delta_i^{\mathrm{B}})^2)$ pulls, contributing $O\big(\sum_{i \neq i^\star} \frac{K \log KT}{(\Delta_i^{\mathrm{B}})^2}\big)$. Once only the optimal arm remains, exploitation yields exactly 0 regret. Finally, background monitoring forces uniform sampling with probability $p_t = \min(1, \frac{K \log t}{t})$, adding at most $\sum_{t=1}^{T} p_t \leq O(K \log^2 T)$. Adding these components yields the bound. See Appendix E.6 for full details. □

Theorem 5.4 (Adversarial Regret Bound)**.** *Under any oblivious adversarial sequence $P_1, \ldots, P_T$, Algorithm 2 achieves:*

$$\mathbb{E}[R_T^{\mathrm{B}}] \leq O\left(K \sqrt{T \log KT} + K^{1/3} T^{2/3} (\log K)^{1/3} \right) \tag{16}$$

Proof Sketch. We analyze two cases based on whether a mode switch occurs. If a switch occurs at round $t_{\text{sw}} \leq T$, the regret comprises the $O(K^2 \log KT)$ exploration overhead, the undetected pre-switch manipulation strictly bounded by the martingale detection threshold $\mathcal{R}_{\text{pre}} \leq O(K\sqrt{T \log KT})$, and the post-switch EXP3 exploration which contributes $\mathbb{E}[\mathcal{R}_{\text{post}}] \leq O(K^{1/3} T^{2/3} (\log K)^{1/3})$ (Lemma 5.12). Alternatively, if no switch occurs, the adversary's cumulative deviation is continuously bounded by the threshold, limiting regret to $O(K\sqrt{T \log KT})$. Full details are in Appendix E.6. □

Corollary 5.5 (Best-of-Both-Worlds for Borda Setting)**.** *Without prior knowledge of the environment type, Algorithm 2 simultaneously achieves:*

(i) ***Stochastic:*** $\mathbb{E}[R_T^{\mathrm{B}}] \leq O\left(K^2 \log KT + K \log^2 T + \sum_{i : \Delta_i^{\mathrm{B}} > 0} \frac{K \log KT}{(\Delta_i^{\mathrm{B}})^2} \right)$

(ii) ***Adversarial:*** $\mathbb{E}[R_T^{\mathrm{B}}] \leq O\big(K\sqrt{T \log KT} + K^{1/3} T^{2/3} (\log K)^{1/3} \big)$

5.4 Lower Bounds

Remark 5.6 (Optimality in the Borda Setting). *1.* ***Stochastic Setting:*** *Instance-dependent lower bounds imply an $\Omega\big(\sum_{i:\Delta_i^{\mathrm{B}}>0} \frac{\log T}{(\Delta_i^{\mathrm{B}})}\big)$ limit. This follows by constructing two instances that agree on all pairs except those involving arm i, with Borda gap Δ_i^B in one and 0 in the other. Consistency requires $\Omega(\log T/(\Delta_i^B)^2)$ pulls of i, since KL divergence scales as $(\Delta_i^B)^2$. Summing over all sub-optimal arm pulls yields $\Omega(\sum_{i:\Delta_i^{\mathrm{B}}>0} \log T/(\Delta_i^B))$. Our* `SA-MiDEX` *algorithm achieves gap-dependent rate of $\sum_i \log T/(\Delta_i^B)^2$ asymptotically, though it incurs an initial $O(K^2 \log KT)$ exploration overhead necessary to unbiasedly estimate global Borda scores.*

2. ***Adversarial Setting:*** *As proved in Gajane (2024, Theorem 2), the expected regret for any multi-dueling bandit algorithm measured against an adversarial Borda winner is lower bounded by $\Omega(K^{1/3}T^{2/3})$. Our adversarial guarantee given in Theorem 5.4 has two terms, $O(K\sqrt{T \log KT})$ and $O(K^{1/3}T^{2/3}(\log K)^{1/3})$; the former dominates for $T < K^4$ and the latter does for $T > K^4$. For $T > K^4$, which holds eventually in the standard asymptotic regime $T \to \infty$ with K fixed, our bound thus matches the theoretical limit up to logarithmic factors and matches the best-known result (Gajane, 2024, Theorem 1).*

5.5 Key Technical Lemmas

We state the key lemmas underlying our analysis. Full proofs are deferred to Appendix E.

Lemma 5.7 (Exploration Guarantees). *With $T_0 = 4K(K-1)\lceil \log(2K^2T^2/\delta) \rceil$, after exploration, with probability at least $1 - \delta/2$:*

(i) Each pair (i, j) is observed at least $n_0 = \Theta(\log T)$ times.

(ii) $|\hat{P}_{T_0}(i, j) - P(i, j)| \le O(\sqrt{\log T/n_0})$ for all pairs.

(iii) $|\hat{b}_{T_0}(i) - b(i)| \le O(\sqrt{\log T/n_0})$ for all arms.

Lemma 5.8 (No False Alarm in Stochastic Setting). *In the stochastic setting, with probability at least $1 - \delta/2$, the deviation detection never triggers throughout T rounds.*

Lemma 5.9 (Adversarial Detection Guarantee). *If the adversary's cumulative preference shift on pair (i, j) exceeds $3\gamma(\sqrt{n} + 1)$ where n is the post-exploration comparison count, then deviation is detected with probability at least $1 - \delta/(K^2T^2)$.*

Lemma 5.10 (Confidence Validity). *With probability at least $1 - \delta/4$, for all arms i and rounds t:*

$$|\hat{b}_t(i) - b(i)| \le \sqrt{\frac{2K \log(2K^2T/\delta)}{N_t(i)}}. \tag{17}$$

Lemma 5.11 (Elimination Sample Complexity). *Conditioned on confidence validity, each suboptimal arm j is eliminated from $\mathcal{C}$ after at most:*

$$N_T(j) \le \frac{32K \log(2K^2T/\delta)}{(\Delta_j^{\mathrm{B}})^2} + 1 \text{ pulls.} \tag{18}$$

Lemma 5.12 (EXP3 Regret in Adversarial Mode). *Starting from any round t_0 with fresh initialization, the EXP3 algorithm with explicit exploration parameters $\eta = \Theta(T^{-2/3})$ and $\gamma = \Theta(T^{-1/3})$ achieves:*

$$\mathbb{E}\left[\sum_{t=t_0}^{T}\left(s_t(i^\star) - \frac{s_t(x_t) + s_t(y_t)}{2}\right)\right] \le O\left(K^{1/3}(T - t_0 + 1)^{2/3}(\log K)^{1/3}\right). \tag{19}$$

6 Extensions

Our algorithmic framework flexibly accommodates several practical extensions.

Time-Varying Subset Sizes: Theorems 4.1, 5.3, and 5.4 extend immediately to the setting where the subset size $m_t \in \{2, \dots, K\}$ varies arbitrarily across rounds. As the rescaling factor β_{m_t} depends only on the current m_t and the feedback transformations remain unbiased pointwise, the regret bounds carry through with identical constants. See the formal statements and proofs in Appendix F.

Contextual, Batched, and High-Probability Settings: The black-box nature of the `MetaDueling` reduction allows it to effortlessly inherit capabilities for contexts, batched feedback, and high-probability bounds simply by swapping the base algorithm (e.g., `Versatile-DB`) for a variant designed for those settings. For the Borda setting, extending `SA-MiDEX` to contextual or high-probability domains requires modifying the global Borda score estimators and concentration thresholds, which presents an interesting direction for future work.

7 Discussion and Conclusion

We have presented the first best-of-both-worlds algorithms for multi-dueling bandits, addressing both Condorcet and Borda objectives.

The $O(K^2 \log KT)$ exploration phase in `SA-MiDEX` is required to learn all pairwise preferences. This overhead dominates for small T or large K. Structured preference models (e.g., low-rank, parametric) could potentially reduce this overhead. Both algorithms construct multisets containing only two distinct arms, potentially wasting the ability to compare $m > 2$ arms simultaneously. A more sophisticated reduction might extract further information from multi-way comparisons, though this would likely sacrifice the black-box property.

Our work opens several directions for future research. First, can we achieve best-of-both-worlds guarantees without assuming a Condorcet winner, requiring a regret notion that handles cyclic preferences while remaining sublinear in the adversarial case? Second, can multi-way comparisons ($m > 2$) be exploited more efficiently, potentially leveraging all $\binom{m}{2}$ pairwise comparisons to accelerate learning beyond the two-arm multiset approach? Third, can we compete with time-varying benchmarks, such as sequences of changing Condorcet winners, yielding dynamic regret bounds that scale with the number of changes or path length? Fourth, can structure in the preference matrix (e.g., low rank, Plackett-Luce, or feature-based models) be leveraged to reduce exploration or regret, particularly in the Borda setting with its $O(K^2)$ overhead? Finally, can our techniques extend beyond winner-only feedback to richer feedback models (rankings or pairwise margins) or sparser observations (e.g., only whether the top choice won)?

In conclusion, this paper provides the first comprehensive treatment of best-of-both-worlds guarantees for multi-dueling bandits under Condorcet and Borda objectives. Our `MetaDueling` reduction demonstrates that optimal Condorcet regret is achievable via a simple, modular transformation of existing dueling bandit algorithms. Our `SA-MiDEX` algorithm shows that gap-dependent stochastic rates and sublinear adversarial rates are simultaneously achievable for Borda objectives, despite the additional challenges of Borda score estimation. We complement our upper bounds with matching lower bounds for the Condorcet setting. For the Borda setting, our stochastic upper bound exceeds the lower bound by a factor of $O(K/\Delta_i^B)$ per suboptimal arm, while, for $T > K^4$, our adversarial upper bound matches the lower bound up to logarithmic factors and matches the best-known result in the literature. The modularity of our approach, particularly the black-box nature of `MetaDueling` ensures that future advances in dueling bandits will automatically benefit multi-dueling applications.

Acknowledgments

The work of Pratik Gajane was supported by the French National Research Agency (ANR) under Grant ANR-24-CPJ1-0088-01. The funders had no role in the conception of the study, the analysis, the conclusions drawn, or the writing of this manuscript.

References

Arpit Agarwal, Nicholas Johnson, and Shivani Agarwal. Choice bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 18399–18410, 2020.

Peter Auer and Chao-Kai Chiang. An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. In *Proceedings of the 29th Annual Conference on Learning Theory*, pp. 116–120. PMLR, 2016.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multi-armed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

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.

Brian Brost, Yevgeny Seldin, Ingemar J. Cox, and Christina Lioma. Multi-dueling bandits and their application to online ranker evaluation. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 2161–2166. ACM, 2016.

Sébastien Bubeck and Aleksandrs Slivkins. The best of both worlds: Stochastic and adversarial bandits. In *Proceedings of the 25th Annual Conference on Learning Theory*, Proceedings of Machine Learning Research, pp. 42.1–42.23. PMLR, 2012.

Miroslav Dudík, Katja Hofmann, Robert E. Schapire, Aleksandrs Slivkins, and Masrour Zoghi. Contextual dueling bandits. In *Proceedings of the 28th Conference on Learning Theory (COLT)*, pp. 563–587, 2015.

Moein Falahatgar, Yi Hao, Alon Orlitsky, Venkatadheeraj Pichapati, and Vaishakh Ravindrakumar. Maxing and ranking with few assumptions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Pratik Gajane. Adversarial multi-dueling bandits. In *ICML 2024 Workshop on Models of Human Feedback for AI Alignment*, 2024. https://icml.cc/virtual/2024/37748.

Kevin Jamieson, Sumeet Katariya, Atul Deshpande, and Robert Nowak. Sparse dueling bandits. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 416–424. PMLR, 2015.

Tiancheng Jin, Longbo Huang, and Haipeng Luo. The best of both worlds: stochastic and adversarial episodic mdps with unknown transition. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc.

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 (COLT)*, pp. 1141–1154. PMLR, 2015.

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 (ICML)*, pp. 1235–1244. PMLR, 2016.

Chung-Wei Lee, Haipeng Luo, Chen-Yu Wei, Mengxiao Zhang, and Xiaojin Zhang. Achieving near instance-optimality and minimax-optimality in stochastic and adversarial linear bandits simultaneously. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 6142–6151. PMLR, 2021.

R Duncan Luce. *Individual Choice Behavior: A Theoretical Analysis*. John Wiley, 1959.

Haipeng Luo, Chen-Yu Wei, Alekh Agarwal, and John Langford. Efficient contextual bandits in non-stationary worlds. In *Proceedings of the 31st Conference On Learning Theory*, pp. 1739–1776. PMLR, 2018.

Robin L Plackett. The analysis of permutations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 24(2):193–202, 1975.

Aadirupa Saha and Pierre Gaillard. Versatile dueling bandits: Best-of-both-world analyses for learning from relative preferences. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pp. 19011–19026. PMLR, 2022.

Aadirupa Saha and Aditya Gopalan. Battle of bandits. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 805–814. AUAI Press, 2018.

Aadirupa Saha and Aditya Gopalan. PAC battling bandits in the Plackett-Luce model. In *Proceedings of the 30th International Conference on Algorithmic Learning Theory (ALT)*, pp. 700–737. PMLR, 2019.

Aadirupa Saha, Tomer Koren, and Yishay Mansour. Adversarial dueling bandits. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 9235–9244. PMLR, 2021.

Yevgeny Seldin and Aleksandrs Slivkins. One practical algorithm for both stochastic and adversarial bandits. In *International Conference on Machine Learning (ICML)*, pp. 1287–1295. PMLR, 2014.

Hossein Azari Soufiani, David C. Parkes, and Lirong Xia. Random utility theory for social choice. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pp. 126–134, Red Hook, NY, USA, 2012. Curran Associates Inc.

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

Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The K-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012.

Julian Zimmert and Yevgeny Seldin. An optimal algorithm for stochastic and adversarial bandits. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 467–475. PMLR, 2019.

Julian Zimmert and Yevgeny Seldin. Tsallis-INF: An optimal algorithm for stochastic and adversarial bandits. In *Journal of Machine Learning Research*, volume 22, pp. 1–49, 2021.

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*, pp. 7683–7692. PMLR, 2019.

Masrour Zoghi, Shimon Whiteson, Rémi 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 (ICML)*, pp. 10–18. PMLR, 2014.

Masrour Zoghi, Zohar Karnin, Shimon Whiteson, and Maarten de Rijke. Copeland dueling bandits. In *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pp. 307–315, Cambridge, MA, USA, 2015. MIT Press.

Supplementary Materials

The following content was not necessarily subject to peer review.

A Extended Related Work

Dueling Bandits. The dueling bandit problem was introduced by Yue & Joachims (2009) for information retrieval applications. Early work focused on stochastic settings with Condorcet winners: Yue et al. (2012) proposed the Interleaved Filter algorithm, Zoghi et al. (2014) introduced RUCB with improved regret bounds, and Komiyama et al. (2015) established instance-optimal rates. Alternative notions of optimality have also been studied: Copeland winners (Zoghi et al., 2015; Komiyama et al., 2016), Borda winners (Jamieson et al., 2015; Falahatgar et al., 2017), and von Neumann winners (Dudík et al., 2015). Saha et al. (2021) initiated the study of adversarial dueling bandits, achieving $O(K^{1/3}T^{2/3})$ regret. Most relevant to our work, Saha & Gaillard (2022) achieved best-of-both-worlds guarantees for dueling bandits using Tsallis entropy regularization, obtaining $O(\sqrt{KT})$ adversarial regret and instance-optimal $O(\sum_{i\neq a^\star} \log T/\Delta_i)$ stochastic regret simultaneously. Our `MetaDueling` reduction leverages their `Versatile-DB` algorithm as a black-box.

Multi-Dueling Bandits. Multi-dueling bandits extend dueling bandits to settings where $m \geq 2$ arms are compared simultaneously. Brost et al. (2016) introduced the problem with subset-wise feedback, where the learner observes preferences among all pairs in the selected subset. Saha & Gopalan (2018) studied the more challenging *winner feedback* model, where only the identity of the winning arm is revealed, under stochastic preferences with Condorcet winners, achieving $O(K \log T/\Delta)$ regret. Agarwal et al. (2020) extended this to generalized Condorcet winners under multinomial logit choice models. Saha & Gopalan (2019) studied PAC identification in multi-dueling bandits. Gajane (2024) introduced adversarial multi-dueling bandits with Borda objectives, achieving $O((K \log K)^{1/3}T^{2/3})$ regret.

Our work bridges these lines: for the Condorcet setting, we provide the first best-of-both-worlds guarantee; for the Borda setting, we improve upon purely adversarial approaches by achieving gap-dependent stochastic regret while maintaining sublinear adversarial regret.

Best-of-Both-Worlds Bandits. The quest for algorithms that achieve optimal rates in both stochastic and adversarial settings without prior knowledge of the regime has a rich history. Bubeck & Slivkins (2012) first achieved best-of-both-worlds guarantees for multi-armed bandits, though with suboptimal constants. Seldin & Slivkins (2014) and Auer & Chiang (2016) refined these results. The breakthrough came with Zimmert & Seldin (2019; 2021), who showed that Tsallis entropy regularization (specifically, the Tsallis-INF algorithm) achieves optimal $O(\sqrt{KT})$ adversarial regret and $O(\sum_i \log T/\Delta_i)$ stochastic regret simultaneously. This approach has since been extended to linear bandits (Lee et al., 2021), combinatorial bandits (Zimmert et al., 2019), and dueling bandits (Saha & Gaillard, 2022).

Our work extends this line to multi-dueling bandits. For the Condorcet setting, we achieve this via a black-box reduction that preserves the best-of-both-worlds property of the base algorithm. For the Borda setting, we adopt a different approach based on the stochastic-and-adversarial paradigm.

Stochastic-and-Adversarial Learning. An alternative approach to best-of-both-worlds guarantees is the stochastic-and-adversarial paradigm, introduced by Auer & Chiang (2016). Rather than designing a single algorithm that simultaneously adapts to both regimes, this approach begins by assuming a stochastic environment and monitors for deviations; upon detecting adversarial behavior, the algorithm switches (potentially irreversibly) to a robust adversarial strategy. This paradigm has been applied to multi-armed bandits (Auer & Chiang, 2016), contextual bandits (Luo et al., 2018), and reinforcement learning (Jin et al., 2021).

Our SA-MiDEX algorithm for the Borda setting adopts this paradigm. The key challenge is designing a deviation detection mechanism that (i) never triggers false alarms in stochastic environments, ensuring gap-dependent regret, and (ii) detects adversarial shifts before they accumulate excessive regret. We achieve this using martingale concentration bounds calibrated to the multi-dueling feedback structure.

Choice Models and Preference Learning. The pairwise-subset choice model, in which our results hold, follows Saha & Gopalan (2018), where the probability of each arm winning is determined by averaging pairwise preferences. This generalizes the Bradley-Terry-Luce model (Bradley & Terry, 1952; Luce, 1959) and is related to Plackett-Luce models for ranking (Plackett, 1975). Alternative choice models include multinomial logit (Agarwal et al., 2020) and general random utility models (Soufiani et al., 2012).

B Relationship Between Condorcet and Borda Settings

The Condorcet and Borda criteria capture fundamentally different aspects of preference learning. A Condorcet winner $a^\star$ satisfies $P(a^\star, i) \geq 1/2$ for all $i \in [K]$. A Borda winner $i^\star$ maximizes $b(i) = \frac{1}{K-1}\sum_{j \neq i} P(i,j)$.

In arbitrary preference matrices, these two notions of optimality need not coincide. A Condorcet winner may secure narrow pairwise victories against all opponents, while another arm might lose narrowly to the Condorcet winner but dominate all other arms by a massive margin, thereby achieving a higher Borda score.

The two criteria are guaranteed to align only under structural assumptions on the preference matrix, such as the Bradley-Terry-Luce (BTL) model or Strong Stochastic Transitivity (SST).

Proposition B.1 (Condorcet and Borda Equivalence under SST)**.** *Assume the preference matrix P satisfies Strong Stochastic Transitivity (SST): for any $i, j, k \in [K]$, if $P(i,j) \geq 1/2$ and $P(j,k) \geq 1/2$, then $P(i,k) \geq \max\{P(i,j), P(j,k)\}$. If $a^\star$ is a Condorcet winner, then $a^\star$ is also Borda-optimal.*

Proof. Let $a^\star$ be the Condorcet winner, meaning $P(a^\star, i) \geq 1/2$ for all i. Let i be any other arm. We compare their Borda scores:

$$\begin{aligned} b(a^\star) - b(i) &= \frac{1}{K-1}\sum_{j \neq a^\star} P(a^\star, j) - \frac{1}{K-1}\sum_{j \neq i} P(i,j) \\ &= \frac{1}{K-1}\left[\sum_{j \neq a^\star, i} \big(P(a^\star, j) - P(i,j)\big) + P(a^\star, i) - P(i, a^\star)\right]. \end{aligned}$$

For any arm $j \neq a^\star, i$, there are two cases:

- **Case 1:** $P(i,j) \geq 1/2$**.** By the SST assumption, since $P(a^\star, i) \geq 1/2$ and $P(i,j) \geq 1/2$, it follows that $P(a^\star, j) \geq P(i,j)$. Thus, $P(a^\star, j) - P(i,j) \geq 0$.
- **Case 2:** $P(i,j) < 1/2$**.** Because $a^\star$ is the Condorcet winner, $P(a^\star, j) \geq 1/2$. Therefore, $P(a^\star, j) - P(i,j) > 0$.

In all cases, $P(a^\star, j) \geq P(i,j)$. Furthermore, since $P(a^\star, i) \geq 1/2$, we have $P(a^\star, i) - P(i, a^\star) = 2P(a^\star, i) - 1 \geq 0$.

Summing these non-negative terms yields $b(a^\star) - b(i) \geq 0$, proving that $a^\star$ maximizes the Borda score. □

Because such transitivity cannot be guaranteed in adversarial environments or complex user-preference datasets (where cyclic preferences naturally arise), the Borda objective remains the most robust benchmark for general multi-dueling problems.

C Validity of Rescaled Preferences

We now verify that the rescaled preference matrix $\widehat{P}_t$ is a valid preference matrix and preserves the Condorcet winner.

Lemma C.1 (Validity of Rescaled Preferences)**.** *The rescaled matrix $\widehat{P}_t$ satisfies:*

(i) ***Reciprocity:*** $\widehat{P}_t(i,j) + \widehat{P}_t(j,i) = 1$ *for all* $i, j \in [K]$.

(ii) ***Self-comparison:*** $\widehat{P}_t(i,i) = \frac{1}{2}$ *for all* $i \in [K]$.

(iii) ***Boundedness:*** $\widehat{P}_t(i,j) \in [0,1]$ *for all* $i, j \in [K]$.

(iv) ***Condorcet preservation:*** *If $a^\star$ is a Condorcet winner under P_t, then $a^\star$ is a Condorcet winner under $\widehat{P}_t$.*

Proof. *(i) Reciprocity:* By linearity of the rescaling:

$$\begin{aligned}\widehat{P}_t(i,j) + \widehat{P}_t(j,i) &= 2\alpha_m + \beta_m(P_t(i,j) + P_t(j,i)) \\ &= 2\alpha_m + \beta_m = (1-\beta_m) + \beta_m = 1,\end{aligned}$$

where we used $\alpha_m = (1-\beta_m)/2$ and $P_t(i,j) + P_t(j,i) = 1$.

(ii) Self-comparison: $\widehat{P}_t(i,i) = \alpha_m + \beta_m \cdot \frac{1}{2} = \frac{1-\beta_m}{2} + \frac{\beta_m}{2} = \frac{1}{2}$.

(iii) Boundedness: Since $P_t(i,j) \in [0,1]$, $\alpha_m \geq 0$, and $\beta_m > 0$:

$$\widehat{P}_t(i,j) \in [\alpha_m, \alpha_m + \beta_m] = [\alpha_m, 1-\alpha_m] \subseteq [0,1].$$

(iv) Condorcet preservation: If $P_t(a^\star, i) \geq \frac{1}{2}$ for all i, then:

$$\widehat{P}_t(a^\star, i) = \alpha_m + \beta_m P_t(a^\star, i) \geq \alpha_m + \frac{\beta_m}{2} = \frac{1}{2}. \qquad \square$$

D Proofs for Condorcet Setting

D.1 Proof of Lemma 3.3 (Unbiased Feedback)

Proof. We prove that $\mathbb{E}[o_t \mid x_t, y_t] = \widehat{P}_t(x_t, y_t) = \alpha_m + \beta_m P_t(x_t, y_t)$.

Let n_x denote the number of copies of x_t in $\mathcal{A}_t$ and $n_y = m - n_x$ the number of copies of y_t. By the pairwise-subset choice model (Definition 2), we compute the probability that some copy of x_t wins.

The multiset $\mathcal{A}_t$ contains n_x copies of x_t at positions $1, \ldots, n_x$ and n_y copies of y_t at positions $n_x + 1, \ldots, m$. The winning probability for position $i \in [n_x]$ (a copy of x_t) is:

$$W(i \mid \mathcal{A}_t, P_t) = \sum_{j \neq i} \frac{2P_t(\mathcal{A}_t(i), \mathcal{A}_t(j))}{m(m-1)}. \tag{20}$$

For position $i \leq n_x$:

- Comparisons against other copies of x_t (positions $j \leq n_x$, $j \neq i$): There are $n_x - 1$ such positions, each contributing $\frac{2 \cdot \frac{1}{2}}{m(m-1)} = \frac{1}{m(m-1)}$.
- Comparisons against copies of y_t (positions $j > n_x$): There are n_y such positions, each contributing $\frac{2P_t(x_t, y_t)}{m(m-1)}$.

Thus:

$$W(i \mid \mathcal{A}_t, P_t) = \frac{n_x - 1}{m(m-1)} + \frac{2n_y P_t(x_t, y_t)}{m(m-1)} \quad \text{for } i \le n_x. \tag{21}$$

The total probability that some copy of x_t wins is:

$$\begin{aligned}
\mathbb{P}[o_t = 1 \mid x_t, y_t, n_x] &= \sum_{i=1}^{n_x} W(i \mid \mathcal{A}_t, P_t) \\
&= n_x \left(\frac{n_x - 1}{m(m-1)} + \frac{2n_y P_t(x_t, y_t)}{m(m-1)} \right) \\
&= \frac{n_x(n_x - 1)}{m(m-1)} + \frac{2n_x n_y P_t(x_t, y_t)}{m(m-1)}.
\end{aligned} \tag{22}$$

Case 1: m **even.**

When m is even, $n_x = n_y = m/2$ deterministically. Substituting into (22):

$$\begin{aligned}
\mathbb{P}[o_t = 1 \mid x_t, y_t] &= \frac{(m/2)(m/2-1)}{m(m-1)} + \frac{2(m/2)(m/2)P_t(x_t, y_t)}{m(m-1)} \\
&= \frac{m(m-2)/4}{m(m-1)} + \frac{m^2 P_t(x_t, y_t)/2}{m(m-1)} \\
&= \frac{m-2}{4(m-1)} + \frac{mP_t(x_t, y_t)}{2(m-1)}.
\end{aligned} \tag{23}$$

Comparing with the rescaling constants for even m:

$$\beta_m = \frac{m}{2(m-1)}, \qquad \alpha_m = \frac{1 - \beta_m}{2} = \frac{1}{2} - \frac{m}{4(m-1)} = \frac{2(m-1) - m}{4(m-1)} = \frac{m-2}{4(m-1)}. \tag{24}$$

Thus (23) equals $\alpha_m + \beta_m P_t(x_t, y_t) = \widehat{P}_t(x_t, y_t)$.

Case 2: m **odd.**

When m is odd, with probability $1/2$ we have $(n_x, n_y) = (\frac{m+1}{2}, \frac{m-1}{2})$, and with probability $1/2$ we have $(n_x, n_y) = (\frac{m-1}{2}, \frac{m+1}{2})$.

For $(n_x, n_y) = (\frac{m+1}{2}, \frac{m-1}{2})$:

$$\begin{aligned}
\mathbb{P}[o_t = 1 \mid n_x = \tfrac{m+1}{2}] &= \frac{\frac{m+1}{2} \cdot \frac{m-1}{2}}{m(m-1)} + \frac{2 \cdot \frac{m+1}{2} \cdot \frac{m-1}{2} \cdot P_t(x_t, y_t)}{m(m-1)} \\
&= \frac{(m+1)(m-1)}{4m(m-1)} + \frac{(m+1)(m-1)P_t(x_t, y_t)}{2m(m-1)} \\
&= \frac{m+1}{4m} + \frac{(m+1)P_t(x_t, y_t)}{2m}.
\end{aligned} \tag{25}$$

For $(n_x, n_y) = (\frac{m-1}{2}, \frac{m+1}{2})$:

$$\begin{aligned}
\mathbb{P}[o_t = 1 \mid n_x = \tfrac{m-1}{2}] &= \frac{\frac{m-1}{2} \cdot \frac{m-3}{2}}{m(m-1)} + \frac{2 \cdot \frac{m-1}{2} \cdot \frac{m+1}{2} \cdot P_t(x_t, y_t)}{m(m-1)} \\
&= \frac{(m-1)(m-3)}{4m(m-1)} + \frac{(m-1)(m+1)P_t(x_t, y_t)}{2m(m-1)} \\
&= \frac{m-3}{4m} + \frac{(m+1)P_t(x_t, y_t)}{2m}.
\end{aligned} \tag{26}$$

Taking the average over the symmetrization:

$$\begin{aligned}\mathbb{P}[o_t = 1 \mid x_t, y_t] &= \frac{1}{2}\left(\frac{m+1}{4m} + \frac{m-3}{4m}\right) + \frac{(m+1)P_t(x_t, y_t)}{2m} \\ &= \frac{1}{2} \cdot \frac{2m-2}{4m} + \frac{(m+1)P_t(x_t, y_t)}{2m} \\ &= \frac{m-1}{4m} + \frac{m+1}{2m} P_t(x_t, y_t). \end{aligned} \tag{27}$$

Comparing with the rescaling constants for odd m:

$$\beta_m = \frac{m+1}{2m}, \qquad \alpha_m = \frac{1-\beta_m}{2} = \frac{1}{2} - \frac{m+1}{4m} = \frac{2m-m-1}{4m} = \frac{m-1}{4m}. \tag{28}$$

Thus (27) equals $\alpha_m + \beta_m P_t(x_t, y_t) = \widehat{P}_t(x_t, y_t)$.

In both cases, $\mathbb{E}[o_t \mid x_t, y_t] = \widehat{P}_t(x_t, y_t)$. □

D.2 Proofs for Gap Rescaling

Proof for Lemma 3.5

Proof. Using $\alpha_m = (1-\beta_m)/2$:

$$\begin{aligned}\widehat{\Delta}_t(i) &= \widehat{P}_t(a^\star, i) - \tfrac{1}{2} = \alpha_m + \beta_m P_t(a^\star, i) - \tfrac{1}{2} \\ &= \frac{1-\beta_m}{2} + \beta_m P_t(a^\star, i) - \tfrac{1}{2} = \beta_m \left(P_t(a^\star, i) - \tfrac{1}{2}\right) = \beta_m \, \Delta_t(i). \end{aligned}$$ □

Proof for Lemma 3.6

Proof. For even m: $\beta_m = \frac{m}{2(m-1)} = \frac{1}{2} \cdot \frac{m}{m-1} > \frac{1}{2}$. For odd m: $\beta_m = \frac{m+1}{2m} = \frac{1}{2} + \frac{1}{2m} > \frac{1}{2}$. In both cases, $\beta_m \geq \frac{1}{2}$, with $\beta_m \to \frac{1}{2}$ as $m \to \infty$. □

D.3 Regret Equivalence in Rescaled Preferences

Proof for Lemma 3.7

Proof. Define the instantaneous regrets:

$$\begin{aligned} r_t^{\mathrm{C}} &:= \frac{1}{m} \sum_{j=1}^{m} \Delta_t(\mathcal{A}_t(j)), \\ \hat{r}_t^{\mathrm{DB}} &:= \frac{1}{2}\left(\widehat{\Delta}_t(x_t) + \widehat{\Delta}_t(y_t)\right) = \frac{\beta_m}{2}\left(\Delta_t(x_t) + \Delta_t(y_t)\right). \end{aligned}$$

Case 1: m **even.** The multiset $\mathcal{A}_t$ contains $m/2$ copies each of x_t and y_t:

$$r_t^{\mathrm{C}} = \frac{1}{2}\Delta_t(x_t) + \frac{1}{2}\Delta_t(y_t) = \frac{1}{\beta_m}\hat{r}_t^{\mathrm{DB}}.$$

Case 2: m **odd.** Taking expectation over the symmetrization B_t:

$$\begin{aligned}\mathbb{E}[r_t^{\mathrm{C}} \mid x_t, y_t] &= \frac{1}{2} \cdot \frac{\lceil m/2 \rceil \Delta_t(x_t) + \lfloor m/2 \rfloor \Delta_t(y_t)}{m} + \frac{1}{2} \cdot \frac{\lfloor m/2 \rfloor \Delta_t(x_t) + \lceil m/2 \rceil \Delta_t(y_t)}{m} \\ &= \frac{1}{2}\Delta_t(x_t) + \frac{1}{2}\Delta_t(y_t) = \frac{1}{\beta_m}\hat{r}_t^{\mathrm{DB}}. \end{aligned}$$

Summing over $t = 1, \ldots, T$ and applying the tower property yields $\mathbb{E}[R_T^{\mathrm{C}}] = \frac{1}{\beta_m}\mathbb{E}[\widehat{R}_T^{\mathrm{DB}}]$. □

D.4 Proof for Theorem 4.1

Proof. By Lemma 3.3, the base algorithm receives feedback satisfying $\mathbb{E}[o_t \mid x_t, y_t] = \widehat{P}_t(x_t, y_t) \in [0, 1]$, which constitutes valid dueling bandit feedback for the rescaled preference matrix $\widehat{P}_t$ (Lemma C.1).

Part (i): Adversarial bound. The `Versatile-DB` algorithm achieves adversarial pseudo-regret

$$\mathbb{E}[\widehat{R}_T^{\mathrm{DB}}] \leq 4\sqrt{KT} + 1$$

on any sequence of valid preference matrices, including $\{\widehat{P}_t\}_{t=1}^T$ (Saha & Gaillard, 2022, Theorem 1). By Lemma 3.7:

$$\mathbb{E}[R_T^{\mathrm{C}}] = \frac{1}{\beta_m}\mathbb{E}[\widehat{R}_T^{\mathrm{DB}}] \leq \frac{1}{\beta_m}(4\sqrt{KT} + 1).$$

Since $\beta_m \geq 1/2$ for all $m \geq 2$ (Lemma 3.6), we have $1/\beta_m \leq 2$, establishing (4).

Part (ii): Bound for self-bounding preference matrices. The self-bounding condition (5) relates expected regret to the cumulative probability mass placed on suboptimal arms. When this condition holds, `Versatile-DB` satisfies (Saha & Gaillard, 2022, Theorem 2):

$$\mathbb{E}[\widehat{R}_T^{\mathrm{DB}}] \leq \sum_{i \neq a^\star} \frac{16 \log T + 48}{\widehat{\Delta}_i} + 4 \log T + 12 + 8\sqrt{K}.$$

By Lemma 3.5, $\widehat{\Delta}_i = \beta_m \Delta_i$, so:

$$\mathbb{E}[\widehat{R}_T^{\mathrm{DB}}] \leq \frac{1}{\beta_m} \sum_{i \neq a^\star} \frac{16 \log T + 48}{\Delta_i} + 4 \log T + 12 + 8\sqrt{K}.$$

Applying Lemma 3.7 (multiplication by $1/\beta_m$):

$$\mathbb{E}[R_T^{\mathrm{C}}] \leq \frac{1}{\beta_m^2} \sum_{i \neq a^\star} \frac{16 \log T + 48}{\Delta_i} + \frac{4 \log T + 12}{\beta_m} + \frac{8\sqrt{K}}{\beta_m}.$$

Using $1/\beta_m^2 \leq 4$ and $1/\beta_m \leq 2$ completes the proof. □

D.5 Proof of Lower Bounds

Proof of Theorem 4.4 (Adversarial Lower Bound). We prove the lower bound by reduction from adversarial dueling bandits.

Step 1: Reduction construction. Given any multi-dueling bandit algorithm $\mathcal{A}_{\mathrm{MDB}}$, we construct a dueling bandit algorithm $\mathcal{A}_{\mathrm{DB}}$ as follows. At each round t:

1. Run $\mathcal{A}_{\mathrm{MDB}}$ to obtain multiset $\mathcal{A}_t$ of size m.
2. Sample two distinct indices $i_t, j_t \in [m]$ uniformly without replacement.
3. Play the duel $(\mathcal{A}_t(i_t), \mathcal{A}_t(j_t))$ in the dueling environment.
4. Receive outcome $w_t \in \{i_t, j_t\}$ indicating the winner.
5. Construct a winner for the multi-dueling feedback: if $w_t = i_t$, return i_t as the winner index to $\mathcal{A}_{\mathrm{MDB}}$; otherwise return j_t.

This construction is valid: the feedback to $\mathcal{A}_{\mathrm{MDB}}$ is consistent with the pairwise-subset choice model restricted to the pair $(\mathcal{A}_t(i_t), \mathcal{A}_t(j_t))$.

Step 2: Regret equivalence. The instantaneous dueling regret for the constructed algorithm is:

$$r_t^{\mathrm{DB}} = \frac{1}{2}\left(\Delta_t(\mathcal{A}_t(i_t)) + \Delta_t(\mathcal{A}_t(j_t))\right). \tag{29}$$

Taking expectation over the uniform sampling of (i_t, j_t):

$$\begin{aligned}\mathbb{E}_{i_t,j_t}[r_t^{\mathrm{DB}}] &= \frac{1}{2}\cdot\frac{1}{\binom{m}{2}}\sum_{i<j}\left(\Delta_t(\mathcal{A}_t(i)) + \Delta_t(\mathcal{A}_t(j))\right)\\ &= \frac{1}{2}\cdot\frac{2}{m(m-1)}\cdot(m-1)\sum_{k=1}^{m}\Delta_t(\mathcal{A}_t(k))\\ &= \frac{1}{m}\sum_{k=1}^{m}\Delta_t(\mathcal{A}_t(k)) = r_t^{\mathrm{MDB}}.\end{aligned} \tag{30}$$

Summing over $t = 1, \ldots, T$:

$$\mathbb{E}[R_T^{\mathrm{DB}}] = \mathbb{E}[R_T^{\mathrm{C}}]. \tag{31}$$

Step 3: Apply adversarial dueling bandit lower bound. By the minimax lower bound for adversarial multi-armed bandits (Auer et al., 2002), for any algorithm there exists an oblivious adversary such that expected regret is at least $\Omega(\sqrt{KT})$.

Via the reduction of Saha et al. (2021), this transfers to dueling bandits: for any dueling bandit algorithm $\mathcal{A}_{\mathrm{DB}}$, there exists an adversarial sequence of preference matrices $P_1, \ldots, P_T$ (with a consistent Condorcet winner $a^\star$) such that:

$$\mathbb{E}[R_T^{\mathrm{DB}}] \geq c\sqrt{KT} \tag{32}$$

for some universal constant $c > 0$.

The hard instance can be constructed to admit a Condorcet winner: fix $a^\star = 1$ and set $P_t(1, i) = \frac{1}{2} + \epsilon_t(i)$ for adversarially chosen $\epsilon_t(i) \in [0, \frac{1}{2}]$. The adversary controls the gap magnitudes while maintaining the Condorcet structure.

Step 4: Transfer to multi-dueling. Suppose for contradiction that there exists a multi-dueling bandit algorithm $\mathcal{A}_{\mathrm{MDB}}$ achieving $\mathbb{E}[R_T^{\mathrm{C}}] = o(\sqrt{KT})$ on all adversarial Condorcet instances.

By Step 2, the constructed $\mathcal{A}_{\mathrm{DB}}$ would satisfy:

$$\mathbb{E}[R_T^{\mathrm{DB}}] = \mathbb{E}[R_T^{\mathrm{C}}] = o(\sqrt{KT}), \tag{33}$$

contradicting the lower bound of Step 3.

Therefore, for any multi-dueling bandit algorithm:

$$\mathbb{E}[R_T^{\mathrm{C}}] \geq \Omega(\sqrt{KT}). \tag{34}$$

□

Proof of Theorem 4.5 (Stochastic Lower Bound). We prove the lower bound by reduction from stochastic dueling bandits.

Step 1: Reduction construction. We use the same reduction as in Theorem 4.4: given $\mathcal{A}_{\mathrm{MDB}}$, construct $\mathcal{A}_{\mathrm{DB}}$ by sampling two indices uniformly and playing the corresponding duel.

Step 2: Regret equivalence. Since the preference matrix is now fixed ($P_t = P$ for all t), the same calculation as in (31) gives:

$$\mathbb{E}[R_T^{\mathrm{DB}}] = \mathbb{E}[R_T^{\mathrm{C}}]. \tag{35}$$

Step 3: Apply stochastic dueling bandit lower bound. Komiyama et al. (2015) established that for any stochastic dueling bandit algorithm and any instance with Condorcet winner $a^\star$:

$$\liminf_{T\to\infty} \frac{\mathbb{E}[R_T^{\mathrm{DB}}]}{\log T} \geq \sum_{i\neq a^\star} \frac{1}{2\cdot \mathrm{KL}(P(a^\star,i)\|\frac{1}{2})}, \tag{36}$$

where $\mathrm{KL}(p\|q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$ is the binary KL divergence.

Step 4: Simplify the KL term. For $P(a^\star,i) = \frac{1}{2}+\Delta_i$ with $\Delta_i > 0$, we compute:

$$\begin{aligned}\mathrm{KL}\left(\frac{1}{2}+\Delta_i \,\middle\|\, \frac{1}{2}\right) &= \left(\frac{1}{2}+\Delta_i\right)\log\frac{\frac{1}{2}+\Delta_i}{\frac{1}{2}} + \left(\frac{1}{2}-\Delta_i\right)\log\frac{\frac{1}{2}-\Delta_i}{\frac{1}{2}} \\ &= \left(\frac{1}{2}+\Delta_i\right)\log(1+2\Delta_i) + \left(\frac{1}{2}-\Delta_i\right)\log(1-2\Delta_i). \end{aligned}\tag{37}$$

Using the Taylor expansion $\log(1+x) = x - \frac{x^2}{2} + O(x^3)$ for small x:

$$\begin{aligned}\mathrm{KL}\left(\frac{1}{2}+\Delta_i \,\middle\|\, \frac{1}{2}\right) &= \left(\frac{1}{2}+\Delta_i\right)\left(2\Delta_i - 2\Delta_i^2 + O(\Delta_i^3)\right) \\ &\quad + \left(\frac{1}{2}-\Delta_i\right)\left(-2\Delta_i - 2\Delta_i^2 + O(\Delta_i^3)\right) \\ &= \Delta_i + 2\Delta_i^2 - \Delta_i^2 - 2\Delta_i^3 - \Delta_i + 2\Delta_i^2 + \Delta_i^2 - 2\Delta_i^3 + O(\Delta_i^3) \\ &= 2\Delta_i^2 + O(\Delta_i^3). \end{aligned}\tag{38}$$

Thus for small Δ_i:

$$\frac{1}{2\cdot \mathrm{KL}(\frac{1}{2}+\Delta_i\|\frac{1}{2})} = \frac{1}{4\Delta_i^2 + O(\Delta_i^3)} = \Theta\left(\frac{1}{\Delta_i^2}\right). \tag{39}$$

However, the regret contribution from arm i is Δ_i per pull. The lower bound (36) states that the expected number of pulls of suboptimal arm i must be at least $\Omega(\log T/\mathrm{KL}(\cdot))$. Since pulling arm i contributes Δ_i to regret, the total regret contribution from arm i is:

$$\Omega\left(\frac{\Delta_i\cdot\log T}{\mathrm{KL}(\frac{1}{2}+\Delta_i\|\frac{1}{2})}\right) = \Omega\left(\frac{\Delta_i\cdot\log T}{2\Delta_i^2}\right) = \Omega\left(\frac{\log T}{\Delta_i}\right). \tag{40}$$

Step 5: Transfer to multi-dueling. Summing over all suboptimal arms and applying the reduction:

$$\mathbb{E}[R_T^{\mathrm{C}}] = \mathbb{E}[R_T^{\mathrm{DB}}] \geq \Omega\left(\sum_{i\neq a^\star}\frac{\log T}{\Delta_i}\right). \tag{41}$$

□

E Proofs for Borda Setting

E.1 Transformed Feedback Properties

Lemma E.1 (Transformed Feedback Statistics)**.** *The transformed feedback* $g_t = (o_t - \alpha_m)/\beta_m$ *satisfies:*

(i) $\mathbb{E}[g_t \mid x_t, y_t] = P_t(x_t, y_t)$.

(ii) $g_t \in \left[-\frac{\alpha_m}{\beta_m}, \frac{1-\alpha_m}{\beta_m}\right] \subseteq [-1, 2]$.

(iii) $\mathrm{Var}(g_t \mid x_t, y_t) \le \frac{1}{4\beta_m^2} \le 1$.

Proof. *(i)* By Lemma 3.3, $\mathbb{E}[o_t \mid x_t, y_t] = \alpha_m + \beta_m P_t(x_t, y_t)$. Thus:

$$\mathbb{E}[g_t \mid x_t, y_t] = \frac{\mathbb{E}[o_t \mid x_t, y_t] - \alpha_m}{\beta_m} = \frac{\alpha_m + \beta_m P_t(x_t, y_t) - \alpha_m}{\beta_m} = P_t(x_t, y_t). \tag{42}$$

(ii) Since $o_t \in \{0, 1\}$:

$$g_t \in \left\{ \frac{0 - \alpha_m}{\beta_m}, \frac{1 - \alpha_m}{\beta_m} \right\} = \left\{ -\frac{\alpha_m}{\beta_m}, \frac{1 - \alpha_m}{\beta_m} \right\}. \tag{43}$$

For $m \ge 2$, we have $\beta_m \ge 1/2$ and $\alpha_m = (1 - \beta_m)/2 \le 1/4$. Thus:

$$-\frac{\alpha_m}{\beta_m} \ge -\frac{1/4}{1/2} = -\frac{1}{2} > -1, \qquad \frac{1 - \alpha_m}{\beta_m} \le \frac{1}{1/2} = 2. \tag{44}$$

(iii) The variance of a Bernoulli-derived random variable:

$$\mathrm{Var}(g_t \mid x_t, y_t) = \frac{\mathrm{Var}(o_t \mid x_t, y_t)}{\beta_m^2} \le \frac{1/4}{\beta_m^2} \le \frac{1/4}{1/4} = 1. \tag{45}$$

□

E.2 Exploration Phase Analysis

Proof of Lemma 5.7 (Exploration Guarantees). *(i) Sample counts.* The round-robin schedule cycles through all $K(K-1)$ ordered pairs. With $T_0 = 4K(K-1)n_0'$ where $n_0' = \lceil \log(2K^2T^2/\delta) \rceil$, each ordered pair (i, j) is observed exactly $4n_0' = n_0$ times.

(ii) Pairwise concentration. Fix pair (i, j). Let $\tau_1, \ldots, \tau_{n_0}$ be the rounds where this pair is compared. Define $Z_k = g_{\tau_k} - P(i, j)$, which satisfies:

- $\mathbb{E}[Z_k] = 0$ (by Lemma E.1(i)).
- $|Z_k| \le 3$ (since $g_{\tau_k} \in [-1, 2]$ and $P(i, j) \in [0, 1]$).
- $\mathrm{Var}(Z_k) \le 1$ (by Lemma E.1(iii)).

The empirical estimate is:

$$\hat{P}_{T_0}(i, j) = \frac{1}{n_0} \sum_{k=1}^{n_0} g_{\tau_k} = P(i, j) + \frac{1}{n_0} \sum_{k=1}^{n_0} Z_k. \tag{46}$$

By Hoeffding's inequality for bounded random variables:

$$\Pr\left[|\hat{P}_{T_0}(i, j) - P(i, j)| > \epsilon \right] = \Pr\left[\left| \frac{1}{n_0} \sum_{k=1}^{n_0} Z_k \right| > \epsilon \right] \le 2 \exp\left(-\frac{2n_0\epsilon^2}{9} \right). \tag{47}$$

Setting $\epsilon = \sqrt{\frac{9 \log(4K^2/\delta)}{2n_0}}$ and using $n_0 = 4\lceil \log(2K^2T^2/\delta) \rceil \ge 4 \log(4K^2/\delta)$ (for $T \ge 2$):

$$\Pr\left[|\hat{P}_{T_0}(i, j) - P(i, j)| > \sqrt{\frac{9 \log(4K^2/\delta)}{2n_0}} \right] \le 2 \exp(-\log(4K^2/\delta)) = \frac{\delta}{2K^2}. \tag{48}$$

By union bound over $K(K-1) < K^2$ pairs:

$$\Pr\left[\exists (i, j) : |\hat{P}_{T_0}(i, j) - P(i, j)| > \sqrt{\frac{9 \log(4K^2/\delta)}{2n_0}} \right] \le K^2 \cdot \frac{\delta}{2K^2} = \frac{\delta}{2}. \tag{49}$$

(iii) Borda score concentration. By definition:

$$|\hat{b}_{T_0}(i) - b(i)| = \left| \frac{1}{K-1} \sum_{j \neq i} (\hat{P}_{T_0}(i,j) - P(i,j)) \right| \leq \frac{1}{K-1} \sum_{j \neq i} |\hat{P}_{T_0}(i,j) - P(i,j)|. \tag{50}$$

On the event that all pairwise estimates are within ϵ of truth:

$$|\hat{b}_{T_0}(i) - b(i)| \leq \frac{1}{K-1} \cdot (K-1) \cdot \epsilon = \epsilon = \sqrt{\frac{9 \log(4K^2/\delta)}{2n_0}} = O\left(\sqrt{\frac{\log T}{n_0}}\right). \tag{51}$$

□

E.3 Deviation Detection Analysis

Proof of Lemma 5.8 (No False Alarm). We show that in the stochastic setting, the deviation detection never triggers with high probability.

Fix pair (i,j) with $i < j$. For rounds $t > T_0$ where this pair is compared, let $\tau_1, \tau_2, \ldots$ denote these rounds in order. Define the martingale:

$$M_n = \sum_{k=1}^{n} (g_{\tau_k} - P(i,j)). \tag{52}$$

Step 1: Martingale properties. The increments $X_k = g_{\tau_k} - P(i,j)$ satisfy:

- $\mathbb{E}[X_k \mid \mathcal{F}_{k-1}] = 0$ (martingale property).
- $|X_k| \leq 3$ (bounded increments).
- $\mathbb{E}[X_k^2 \mid \mathcal{F}_{k-1}] \leq 1$ (bounded conditional variance).

Step 2: Freedman's inequality. By Freedman's inequality, for a martingale (M_n) with $|M_n - M_{n-1}| \leq c$ and predictable quadratic variation $V_n = \sum_{k=1}^{n} \mathbb{E}[X_k^2 \mid \mathcal{F}_{k-1}]$:

$$\Pr\left[\exists n \leq N : |M_n| \geq x \text{ and } V_n \leq v\right] \leq 2 \exp\left(-\frac{x^2}{2v + 2cx/3}\right). \tag{53}$$

With $c = 3$, $v = n$ (since $V_n \leq n$), and $x = \tau(\sqrt{n}+1)$:

$$\Pr\left[|M_n| > \tau(\sqrt{n}+1)\right] \leq 2 \exp\left(-\frac{\tau^2(\sqrt{n}+1)^2}{2n + 2\tau(\sqrt{n}+1)}\right). \tag{54}$$

For $n \geq 1$ and $\tau = \sqrt{8 \log(K^2T^2/\delta)}$:

$$\frac{\tau^2(\sqrt{n}+1)^2}{2n + 2\tau(\sqrt{n}+1)} \geq \frac{\tau^2 n}{2n + 2\tau\sqrt{n} + 2\tau} \geq \frac{\tau^2}{4} \quad \text{for } n \geq \tau^2. \tag{55}$$

For $n < \tau^2$, we use a cruder bound. In either case:

$$\Pr\left[\exists n \leq T : |M_n| > \tau(\sqrt{n}+1)\right] \leq 2T \cdot \exp\left(-\frac{\tau^2}{8}\right) = 2T \cdot \exp(-\log(K^2T^2/\delta)) = \frac{2\delta}{K^2T}. \tag{56}$$

Step 3: Connect to detection statistic. The detection statistic is:

$$\begin{aligned} D_n &= |S_t^{\text{obs}}(i,j) - S_{T_0}^{\text{obs}}(i,j) - n \cdot \hat{P}_{T_0}(i,j)| \\ &= \left| \sum_{k=1}^{n} g_{\tau_k} - n \cdot \hat{P}_{T_0}(i,j) \right|. \end{aligned} \tag{57}$$

In the stochastic setting, $\hat{P}_{T_0}(i,j)$ estimates $P(i,j)$. By Lemma 5.7(ii):

$$|\hat{P}_{T_0}(i,j) - P(i,j)| \leq \epsilon_0 := O\left(\sqrt{\frac{\log T}{n_0}}\right). \tag{58}$$

Thus:

$$\begin{aligned} D_n &= \left|\sum_{k=1}^{n} g_{\tau_k} - n \cdot P(i,j) + n \cdot (P(i,j) - \hat{P}_{T_0}(i,j))\right| \\ &\leq |M_n| + n \cdot |\hat{P}_{T_0}(i,j) - P(i,j)| \\ &\leq |M_n| + n\epsilon_0. \end{aligned} \tag{59}$$

For detection to trigger, we need $D_n > \tau(\sqrt{n}+1)$. If $|M_n| \leq \tau(\sqrt{n}+1)/2$, then we need:

$$n\epsilon_0 > \frac{\tau(\sqrt{n}+1)}{2}. \tag{60}$$

For $n_0 = \Theta(\log T)$ and $\tau = \Theta(\sqrt{\log T})$, we have $\epsilon_0 = O(1/\sqrt{\log T})$. The condition becomes:

$$n \cdot O(1/\sqrt{\log T}) > \Omega(\sqrt{\log T} \cdot \sqrt{n}), \tag{61}$$

which simplifies to $\sqrt{n} > \Omega(\log T)$, i.e., $n > \Omega(\log^2 T)$.

For $n \leq T$ and sufficiently large T, the threshold is not exceeded when $|M_n| \leq \tau(\sqrt{n}+1)/2$.

Step 4: Union bound. By union bound over all $\binom{K}{2} < K^2/2$ pairs:

$$\Pr[\text{any false alarm}] \leq \frac{K^2}{2} \cdot \frac{2\delta}{K^2 T} = \frac{\delta}{T} \leq \frac{\delta}{2}. \tag{62}$$

□

Proof of Lemma 5.9 (Adversarial Detection). Suppose the adversary shifts preferences such that:

$$\left|\sum_{k=1}^{n} (P_{\tau_k}(i,j) - \hat{P}_{T_0}(i,j))\right| \geq 3\tau(\sqrt{n}+1). \tag{63}$$

Decompose the observed cumulative:

$$\sum_{k=1}^{n} g_{\tau_k} - n \cdot \hat{P}_{T_0}(i,j) = \underbrace{\sum_{k=1}^{n} (g_{\tau_k} - P_{\tau_k}(i,j))}_{=:M_n} + \underbrace{\sum_{k=1}^{n} (P_{\tau_k}(i,j) - \hat{P}_{T_0}(i,j))}_{\text{adversarial shift}}. \tag{64}$$

The term M_n is a martingale even under adversarial preferences, since $\mathbb{E}[g_{\tau_k} \mid \mathcal{F}_{k-1}, P_{\tau_k}] = P_{\tau_k}(i,j)$.

By the same Freedman argument as in Lemma 5.8:

$$\Pr[|M_n| > \tau(\sqrt{n}+1)] \leq \frac{2\delta}{K^2 T^2}. \tag{65}$$

Under condition (63), with probability at least $1 - \frac{2\delta}{K^2T^2}$:

$$\begin{aligned} D_n &= \left|\sum_{k=1}^{n} g_{\tau_k} - n \cdot \hat{P}_{T_0}(i,j)\right| \\ &\geq |\text{adversarial shift}| - |M_n| \\ &\geq 3\tau(\sqrt{n}+1) - \tau(\sqrt{n}+1) \\ &= 2\tau(\sqrt{n}+1) > \tau(\sqrt{n}+1). \end{aligned} \tag{66}$$

This exceeds the detection threshold, triggering a mode switch. □

E.4 Confidence and Elimination Analysis

Lemma E.2 (Borda Score Concentration)**.** *During the elimination phase, for any arm i and any $\epsilon > 0$, let $N_{\min}(i) = \min_{j \neq i} N_n(i,j)$ be the minimum number of comparisons arm i has against any specific opponent. Then:*

$$\Pr[|\hat{b}_n(i) - b(i)| > \epsilon] \leq 2K \exp\left(-\frac{N_{\min}(i)\epsilon^2}{5}\right). \tag{67}$$

Proof. The empirical Borda score is the unweighted average of pairwise estimates:

$$\hat{b}_n(i) = \frac{1}{K-1} \sum_{j \neq i} \hat{P}_n(i,j). \tag{68}$$

The error in the Borda score is bounded by the maximum pairwise error:

$$|\hat{b}_n(i) - b(i)| \leq \frac{1}{K-1} \sum_{j \neq i} |\hat{P}_n(i,j) - P(i,j)| \leq \max_{j \neq i} |\hat{P}_n(i,j) - P(i,j)|. \tag{69}$$

For any specific pair (i,j) observed $N_n(i,j)$ times, applying Hoeffding's inequality on the transformed feedback $g_t \in [-1,2]$ (range length 3) yields:

$$\Pr[|\hat{P}_n(i,j) - P(i,j)| > \epsilon] \leq 2\exp\left(-\frac{2N_n(i,j)\epsilon^2}{9}\right) \leq 2\exp\left(-\frac{N_{\min}(i)\epsilon^2}{5}\right). \tag{70}$$

Applying a union bound over all $K-1$ opponents $j \neq i$ ensures that the maximum error stays bounded:

$$\Pr\left[\max_{j \neq i} |\hat{P}_n(i,j) - P(i,j)| > \epsilon\right] \leq 2(K-1)\exp\left(-\frac{N_{\min}(i)\epsilon^2}{5}\right) \leq 2K\exp\left(-\frac{N_{\min}(i)\epsilon^2}{5}\right). \tag{71}$$

□

Proof of Lemma 5.10 (Confidence Validity). Define the confidence radius based on the total pulls of arm i:

$$\text{conf}_t(i) = \sqrt{\frac{20K\log(2K^2T/\delta)}{N_t(i)}}. \tag{72}$$

We show that $|\hat{b}_t(i) - b(i)| \leq \text{conf}_t(i)$ holds for all i and all $t > T_0$ with high probability.

Step 1: Relating $N_t(i)$ to $N_{\min}(i)$. Because the algorithm samples the opponent y_t uniformly from $[K] \setminus \{x_t\}$, the number of times arm i plays a specific opponent j follows a binomial distribution. By standard Chernoff bounds, with high probability, the minimum pairwise count concentrates heavily around its expectation: $N_{\min}(i) \geq \frac{N_t(i)}{2(K-1)} \geq \frac{N_t(i)}{2K}$. Substituting this into the denominator yields $\frac{1}{N_{\min}(i)} \leq \frac{2K}{N_t(i)}$.

Step 2: Fixed-time bound. For fixed n, let the confidence radius be $\text{conf}_n(i) = \sqrt{\frac{10\log(2K^2T/\delta)}{N_{\min}(i)}}$. By Lemma E.2 with $\epsilon = \text{conf}_n(i)$:

$$\Pr\left[|\hat{b}_n(i) - b(i)| > \text{conf}_n(i)\right] \leq 2K\exp\left(-\frac{N_{\min}(i)}{5} \cdot \frac{10\log(2K^2T/\delta)}{N_{\min}(i)}\right) = \frac{\delta}{KT}. \tag{73}$$

Using the relationship from Step 1, this radius is bounded by $\sqrt{\frac{20K\log(2K^2T/\delta)}{N_t(i)}}$.

Step 3: Time-uniform bound via peeling. Partition the range of $N_{\min}(i)$ into epochs:

$$E_\ell = \{t : 2^{\ell-1} \le N_{\min}(i) < 2^\ell\}, \quad \ell = 1, 2, \ldots, \lceil \log_2 T \rceil. \tag{74}$$

Step 4: Union bound. Summing over $O(\log T)$ epochs and K arms:

$$\Pr[\mathcal{E}_{\text{conf}}^c] \le K \cdot \log_2 T \cdot \frac{\delta}{KT} = \frac{\delta \log_2 T}{T} \le \frac{\delta}{4}. \tag{75}$$

□

Proof of Lemma 5.11 (Elimination Sample Complexity). Conditioned on the confidence event $\mathcal{E}_{\text{conf}}$ holding (Lemma 5.10), the true optimal arm $i^\star$ is never eliminated from the active set $\mathcal{C}$. This is because for any arm $j \in \mathcal{C}$, the following holds:

$$\hat{b}_t(i^\star) + \text{conf}_t(i^\star) \ge b(i^\star) \ge b(j) \ge \hat{b}_t(j) - \text{conf}_t(j). \tag{76}$$

Therefore, the strict elimination condition $\hat{b}_t(j) - \text{conf}_t(j) > \hat{b}_t(i^\star) + \text{conf}_t(i^\star)$ can never be satisfied against the optimal arm.

A suboptimal arm j is eliminated from $\mathcal{C}$ when there exists any arm $i \in \mathcal{C}$ such that:

$$\hat{b}_t(i) - \text{conf}_t(i) > \hat{b}_t(j) + \text{conf}_t(j). \tag{77}$$

Since the true Borda gap is $\Delta_j^{\text{B}} = b(i^\star) - b(j)$, and $i^\star$ remains in $\mathcal{C}$, this elimination condition is mathematically guaranteed to trigger against $i^\star$ once:

$$2\text{conf}_t(i^\star) + 2\text{conf}_t(j) < \Delta_j^{\text{B}}. \tag{78}$$

Because the algorithm selects $x_t \in \mathcal{C}$ using a round-robin schedule, all arms in the active set are pulled symmetrically, implying $\text{conf}_t(i^\star) \approx \text{conf}_t(j)$. Therefore, arm j is guaranteed to be eliminated once:

$$4\text{conf}_t(j) \le \Delta_j^{\text{B}}. \tag{79}$$

Substituting the definition of the confidence radius $\text{conf}_t(j) = \sqrt{\frac{20K\log(2K^2T/\delta)}{N_t(j)}}$:

$$4\sqrt{\frac{20K\log(2K^2T/\delta)}{N_t(j)}} \le \Delta_j^{\text{B}}. \tag{80}$$

Squaring both sides and solving for $N_t(j)$ yields the maximum number of times arm j can be pulled during the elimination phase before it is permanently removed:

$$N_T(j) \le \frac{320K\log(2K^2T/\delta)}{(\Delta_j^{\text{B}})^2} + 1. \tag{81}$$

□

E.5 Adversarial Mode Analysis

Proof of Lemma 5.12 (EXP3 Regret). Let $T' = T - t_0 + 1$ be the number of adversarial rounds. As established by Gajane (2024), the primary challenge in Borda estimation under winner-only feedback is that the variance of the importance-weighted estimator scales with $\sum_j 1/q_t(j)$.

The explicit exploration parameter γ guarantees that $q_t(j) \ge \gamma/K$ for all $j \in [K]$, strictly bounding the maximum variance. Applying the standard EXP3 regret analysis with the mixed distribution q_t (see Theorem 1 of Gajane (2024)) bounds the expected shifted Borda regret over T' rounds by:

$$\mathbb{E}[R_{T'}^s] \le O\left(K^{1/3}(T')^{2/3}(\log K)^{1/3}\right). \tag{82}$$

□

E.6 Main Theorem Proofs

Proof of Theorem 5.3 (Stochastic Regret). Define the high-probability event $\mathcal{E} = \mathcal{E}_{\text{conf}} \cap \mathcal{E}_{\text{no-switch}}$. By union bounds over the confidence intervals and the martingale detection thresholds, $\Pr[\mathcal{E}] \geq 1 - O(1/T)$, meaning $\Pr[\mathcal{E}^c] \leq O(1/T)$. Conditioned on $\mathcal{E}$, the algorithm never switches to adversarial mode, and the true optimal arm $i^\star$ is never eliminated from $\mathcal{C}$.

Step 1: Regret from Baseline Exploration Phase. During rounds $t \leq T_0$, the instantaneous Borda regret is at most 1. Thus, the initial exploration phase contributes:

$$\mathbb{E}[R_{T_0}^{\text{B}} \mid \mathcal{E}] \leq T_0 = 4K(K-1)\lceil \log(2K^2T^2/\delta) \rceil = O(K^2 \log KT). \tag{83}$$

Step 2: Regret from Elimination Phase. For any round $t > T_0$ where a suboptimal arm j is selected as $x_t \in \mathcal{C}$, the opponent y_t is drawn uniformly from $[K] \setminus \{x_t\}$. The expected instantaneous regret is:

$$\mathbb{E}[r_t \mid x_t = j, \mathcal{E}] = b(i^\star) - \left(\frac{b(j)}{2} + \frac{1}{K-1} \sum_{k \neq j} \frac{b(k)}{2} \right) = \frac{\Delta_j^{\text{B}}}{2} + \frac{1}{2(K-1)} \sum_{k \neq j} \Delta_k^{\text{B}} = \Theta(1). \tag{84}$$

By Lemma 5.11, arm j is pulled at most $N_T(j) \leq O(K \log T / (\Delta_j^{\text{B}})^2)$ times before elimination. The cumulative expected regret strictly from pulling suboptimal arms during the elimination phase is:

$$\sum_{j \neq i^\star} N_T(j) \cdot \mathbb{E}[r_t \mid x_t = j, \mathcal{E}] = O\left(\sum_{j \neq i^\star} \frac{K \log KT}{(\Delta_j^{\text{B}})^2} \right). \tag{85}$$

Step 3: Regret from Exploitation Phase. Once all suboptimal arms are eliminated, $\mathcal{C} = \{i^\star\}$. The algorithm plays $x_t = y_t = i^\star$, resulting in an expected instantaneous regret of exactly 0 for all subsequent rounds not subject to background monitoring.

Step 4: Regret from Background Monitoring. At any round $t > T_0$, the algorithm overrides the elimination logic with probability $p_t = \min(1, \frac{K \log t}{t})$ to randomly sample a pair and feed the deviation detection statistic. The maximum instantaneous regret of any pair is bounded by 1. The total expected regret from this monitoring is:

$$\begin{aligned}
\sum_{t=1}^{T} p_t \cdot 1 &= \sum_{t=1}^{T} \mathbb{1}\left(1 \leq \frac{K \log t}{t}\right) \min\left(1, \frac{K \log t}{t}\right) + \sum_{t=1}^{T} \mathbb{1}\left(1 > \frac{K \log t}{t}\right) \min\left(1, \frac{K \log t}{t}\right) \\
&= \sum_{t=1}^{T} \mathbb{1}\left(1 \leq \frac{K \log t}{t}\right) \cdot 1 + \sum_{t=1}^{T} \mathbb{1}\left(1 > \frac{K \log t}{t}\right) \frac{K \log t}{t} \\
&\leq \sum_{t=1}^{T} \frac{K \log t}{t} = O(K \log^2 T).
\end{aligned}$$

Step 5: Regret on the Failure Event. On the complement event $\mathcal{E}^c$, the algorithm's confidence bounds or detection thresholds fail. In the worst case, the algorithm suffers the maximum possible instantaneous regret of 1 for all T rounds, yielding a maximum total regret of T. Since $\Pr[\mathcal{E}^c] \leq O(1/T)$, the expected regret contribution from this failure case is strictly bounded:

$$\mathbb{E}[R_T^{\text{B}} \cdot \mathbb{1}[\mathcal{E}^c]] \leq T \cdot O\left(\frac{1}{T}\right) = O(1). \tag{86}$$

Step 6: Total Regret. Combining the expected regret conditioned on $\mathcal{E}$ (Steps 1, 2, and 4) with the expected regret on the failure event (Step 5), the total expected stochastic regret is:

$$\mathbb{E}[R_T^{\text{B}}] \leq O\left(K^2 \log KT + K \log^2 T + \sum_{i:\Delta_i^{\text{B}} > 0} \frac{K \log KT}{(\Delta_i^{\text{B}})^2} \right). \tag{87}$$

□

Proof of Theorem 5.4 (Adversarial Regret). We analyze two cases based on whether a mode switch occurs.

Case 1: Switch occurs at round $t_{\text{sw}} \leq T$.

Exploration regret:

$$R^{\text{B}}_{T_0} \leq T_0 = O(K^2 \log T). \tag{88}$$

Pre-switch regret ($T_0 < t < t_{sw}$): During this phase, the algorithm runs Successive Elimination and Background Monitoring. The adversary can manipulate preferences to cause regret while staying below the detection threshold.

For each pair (i, j), detection occurs when the cumulative deviation exceeds $\tau(\sqrt{n} + 1)$ where $n = N_t(i, j) - N_{T_0}(i, j)$. The maximum undetected deviation per pair is $O(\tau(\sqrt{T} + 1))$.

The total adversarial manipulation across all K^2 pairs before detection is:

$$\mathcal{R}_{\text{pre}} \leq \sum_{(i,j)} O(\tau\sqrt{N_{t_{\text{sw}}}(i,j)}) \leq O\left(\tau K \sqrt{\sum_{(i,j)} N_{t_{\text{sw}}}(i,j)}\right) = O(\tau K\sqrt{T}), \tag{89}$$

where we used Cauchy-Schwarz and $\sum_{(i,j)} N_t(i, j) \leq 2T$.

Since $\tau = O(\sqrt{\log KT})$:

$$\mathcal{R}_{\text{pre}} \leq O(K\sqrt{T \log KT}). \tag{90}$$

Post-switch regret ($t \geq t_{sw}$): After switching, the algorithm runs EXP3. By Lemma 5.12:

$$\mathbb{E}[\mathcal{R}_{\text{post}}] \leq O(K^{1/3}(T - t_{\text{sw}})^{2/3}(\log K)^{1/3}) \leq O(K^{1/3}T^{2/3}(\log K)^{1/3}). \tag{91}$$

Case 2: No switch occurs ($t_{\text{sw}} > T$).

If no switch occurs throughout T rounds, the adversary's cumulative deviation on each pair is bounded by the detection threshold. The total regret from adversarial deviations is:

$$\begin{aligned} R^{\text{adv}}_T &\leq \sum_{(i,j)} O(\tau(\sqrt{N_T(i,j)} + 1)) \\ &\leq O(\tau K^2) + O\left(\tau K\sqrt{T}\right) \\ &= O(K^2\sqrt{\log KT} + K\sqrt{T \log KT}). \end{aligned} \tag{92}$$

Combined bound. In both cases, the total adversarial regret is bounded by the sum of the exploration overhead, the maximum undetected pre-switch deviation, and the post-switch EXP3 regret:

$$\mathbb{E}[R^{\text{B}}_T] \leq O(K^2 \log KT) + O(K\sqrt{T \log KT}) + O(K^{1/3}T^{2/3}(\log K)^{1/3}). \tag{93}$$

For $T \geq K^2$, the initial exploration phase $O(K^2 \log KT)$ is strictly dominated by the pre-switch deviation term $O(K\sqrt{T \log KT})$.

However, neither the pre-switch deviation nor the post-switch EXP3 regret strictly dominates the other across all time horizons. The pre-switch detection delay dominates for moderate horizons ($T < K^4/\log^2 K$), while the EXP3 regret dominates for massive horizons ($T > K^4/\log^2 K$). Therefore, the final rigorous bound retains both primary terms:

$$\mathbb{E}[R^{\text{B}}_T] \leq O\left(K\sqrt{T \log KT} + K^{1/3}T^{2/3}(\log K)^{1/3}\right). \tag{94}$$

□

F Proofs for Extensions

We provide the formal corollaries and proofs for the time-varying subset size extension discussed in Section 6.

Corollary F.1 (Time-Varying Subset Size: Condorcet Setting)**.** *Theorem 4.1 and Corollary 4.2 hold when the subset size $m_t \in \{2, \dots, K\}$ varies arbitrarily across rounds.*

Proof. The proof of Lemma 3.7 holds for each round independently: at round t, the rescaling factor β_{m_t} depends only on the current subset size m_t. The key observations are:

(i) The feedback o_t remains an unbiased estimate of $\widehat{P}_t(x_t, y_t) = \alpha_{m_t} + \beta_{m_t} P_t(x_t, y_t)$.

(ii) The base algorithm operates on the rescaled preferences and is agnostic to the specific value of β_{m_t}.

(iii) The regret equivalence $\mathbb{E}[r_t^{\mathrm{C}}] = \frac{1}{\beta_{m_t}} \mathbb{E}[\hat{r}_t^{\mathrm{DB}}]$ holds pointwise.

Since $\beta_{m_t} \geq 1/2$ for all $m_t \geq 2$, the bounds carry through with the same constants. □

Corollary F.2 (Time-Varying Subset Size: Borda Setting)**.** *Theorems 5.3 and 5.4 hold when $m_t \in \{2, \dots, K\}$ varies arbitrarily across rounds, provided the transformed feedback g_t from* METADUELING *uses the current m_t.*

Proof. The transformed feedback $g_t = (o_t - \alpha_{m_t})/\beta_{m_t}$ remains an unbiased estimate of $P_t(x_t, y_t)$ for any m_t. The stochastic analysis (exploration, deviation detection) and adversarial analysis (EXP3) are unaffected by varying m_t. □

G Additional Technical Results

G.1 Concentration Inequalities

We state the concentration inequalities used throughout the proofs.

Lemma G.1 (Hoeffding's Inequality)**.** *Let $X_1, \dots, X_n$ be independent random variables with $X_i \in [a_i, b_i]$. Then for $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$:*

$$\Pr[|\bar{X} - \mathbb{E}[\bar{X}]| \geq t] \leq 2 \exp\left(-\frac{2n^2 t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right). \tag{95}$$

Lemma G.2 (Freedman's Inequality)**.** *Let $(M_n)_{n \geq 0}$ be a martingale with $M_0 = 0$ and $|M_n - M_{n-1}| \leq c$ almost surely. Let $V_n = \sum_{k=1}^n \mathbb{E}[(M_k - M_{k-1})^2 \mid \mathcal{F}_{k-1}]$ be the predictable quadratic variation. Then for any $x, v > 0$:*

$$\Pr[\exists n : M_n \geq x \text{ and } V_n \leq v] \leq \exp\left(-\frac{x^2}{2v + 2cx/3}\right). \tag{96}$$