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 upper bounds are near-optimal and match the best-known results 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 and 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 and 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 and Seldin, 2021) and dueling bandits (Saha and Gaillard, 2022), but remain unexplored for multi-dueling bandits. Existing work addresses these regimes separately: Saha and 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.

Contributions.

1. **Condorcet Setting: The `MetaDueling` Reduction.** We propose `MetaDueling` (Algorithm 1), 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 9). Instantiating this with `Versatile-DB` (Saha and Gaillard, 2022) yields the first best-of-both-worlds multi-dueling guarantee (Theorem 5): $O(\sqrt{KT})$ 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` (Algorithm 2) for Borda winners. `SA-MiDEX` initially assumes stochastic preferences using successive elimination to obtain unbiased estimates, monitors for adversarial deviations via martingale concentration, and upon detection irreversibly switches to adversarial mode. This yields $O\Big(K^2 \log KT + K \log^2 T + \sum_{i:\Delta_i^{\mathrm{B}}>0} \frac{K \log KT}{(\Delta_i^{\mathrm{B}})^2}\Big)$ stochastic regret and $O\big(K\sqrt{T \log KT} + K^{1/3} T^{2/3} (\log K)^{1/3}\big)$ adversarial regret (Theorems 6 and 7).

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, establishing the optimality of our Condorcet guarantees. For the Borda setting, lower bounds of $\Omega(K^{1/3} T^{2/3})$ (adversarial) show our upper bounds are near-optimal.

Related Work. Multi-dueling bandits (Brost et al., 2016) generalize the standard dueling framework (Yue and Joachims, 2009) to simultaneous m-way comparisons. The quest for best-of-both-worlds algorithms has seen significant breakthroughs via Tsallis entropy regularization (Zimmert and Seldin, 2019, 2021), extended to dueling bandits by Saha and Gaillard (2022). Our `MetaDueling` reduction directly leverages their `Versatile-DB` algorithm. For the Borda setting, we draw upon the stochastic-and-adversarial paradigm of Auer and Chiang (2016), which actively monitors for adversarial deviations and permanently switches strategies upon detection.

2. Preliminaries

We consider online learning over K arms $[K]$ for T rounds. At each round t, the environment 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 ($P_t(i,i) = 1/2$). The learner selects a multiset $\mathcal{A}_t$ of size m and observes a winning arm. Following Saha and Gopalan (2018), the probability arm i wins is $W(i \mid \mathcal{A}, P) := \sum_{j \neq i} \frac{2P(\mathcal{A}(i), \mathcal{A}(j))}{m(m-1)}$.

Assumption 1 (Condorcet Winner and Gap). *There exists a Condorcet winner $a^\star \in [K]$ such that $P_t(a^\star, i) \geq 1/2$ for all $i \in [K]$, $t \in [T]$. The sub-optimality gap is $\Delta_t(i) := P_t(a^\star, i) - 1/2$, written $\Delta(i)$ in the stochastic setting, with $\Delta_{\min} := \min_{i \neq a^\star} \Delta(i) > 0$.*

Cumulative Condorcet pseudo-regret: $R_T^{\mathrm{C}} := \sum_{t=1}^{T} \frac{1}{m} \sum_{j=1}^{m} \Delta_t(\mathcal{A}_t(j))$.

Borda Setting. The Borda score of arm i is $b_t(i) := \frac{1}{K-1} \sum_{j \neq i} P_t(i,j)$, i.e. the probability it beats a uniformly random opponent. Writing $i_t^\star := \arg\max_i b_t(i)$ for the (per-round) Borda winner, the Borda gap of arm i is $\Delta_t^{\mathrm{B}}(i) := b_t(i_t^\star) - b_t(i)$, abbreviated $\Delta_i^{\mathrm{B}} := b(i^\star) - b(i)$ in the stochastic setting.

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**
  Let $(q_t^{(1)}, q_t^{(2)})$ be the distributions maintained by $\mathcal{D}$ ▷ **Query base alg. for arm pair**
  Sample $x_t \sim q_t^{(1)}$ and $y_t \sim q_t^{(2)}$ independently
  Sample $B_t \sim \text{Bernoulli}(1/2)$
  if m is even **then** ▷ **Construct multiset with symmetrization**
    $(n_x, n_y) \leftarrow (m/2, m/2)$
  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 $\mathcal{A}_t$ and observe winner index $I_t \in [m]$ ▷ **Play and observe feedback**
  Set $o_t \leftarrow \mathbb{1}[\mathcal{A}_t(I_t) = x_t]$
  Update $\mathcal{D}$ with outcome o_t for pair (x_t, y_t) ▷ **Feed outcome to base algorithm**
end for

The cumulative Borda pseudo-regret compares the average Borda score of selected arms against the best fixed arm in hindsight:

$$R_T^{\mathrm{B}} := \max_{a \in [K]} \sum_{t=1}^{T} \frac{1}{m} \sum_{j=1}^{m} \big(b_t(a) - b_t(\mathcal{A}_t(j)) \big),$$

which in the stochastic setting reduces to $\sum_t \frac{1}{m} \sum_j \Delta^{\mathrm{B}}(\mathcal{A}_t(j))$. Unlike a Condorcet winner, a Borda winner always exists and is more robust to cyclic preferences (Jamieson et al., 2015b).

3. The MetaDueling Reduction

We present `MetaDueling` (Algorithm 1), a black-box reduction that decouples the multi-way feedback mechanism from the core learning strategy. 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, $n_x = n_y = m/2$. When m is odd, we randomize symmetrically with probability $1/2$ to ensure unbiased feedback in expectation.

Rescaled Preferences. The multiset construction introduces bias from ties between identical arms. We characterize this precisely.

Definition 2 (Rescaling Constants). *Define:*

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

Definition 3 (Rescaled Preference Matrix). *The rescaled preference matrix is* $\widehat{P}_t(i,j) := \alpha_m + \beta_m P_t(i,j)$.

The rescaled matrix $\widehat{P}_t$ is a valid preference matrix and preserves the Condorcet winner (Lemma 9, proved in Appendix C). The key lemma shows that o_t is an unbiased estimate of $\widehat{P}_t$:

Lemma 4 (Unbiased Feedback). *Under Algorithm 1:* $\mathbb{E}[o_t \mid x_t, y_t] = \widehat{P}_t(x_t, y_t) = \alpha_m + \beta_m P_t(x_t, y_t)$.

Since $\beta_m \geq 1/2$ for all $m \geq 2$ (Lemma 11), the rescaled gaps satisfy $\widehat{\Delta}_t(i) = \beta_m \Delta_t(i)$ (Lemma 10), and regret equivalence gives $\mathbb{E}[R_T^{\mathrm{C}}] = \frac{1}{\beta_m}\mathbb{E}[\widehat{R}_T^{\mathrm{DB}}]$ (Lemma 12), where $\widehat{R}_T^{\mathrm{DB}}$ is the base algorithm's pseudo-regret under $\widehat{P}_t$. No modification to $\mathcal{D}$'s internals is needed.

4. Main Results

4.1. Condorcet Setting

Theorem 5 (Best-of-Both-Worlds for Condorcet Multi-Dueling). *Under Assumption 1,* `MetaDueling` *instantiated with* `Versatile-DB` *satisfies simultaneously for any* $T \geq 1$*:*

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

(ii) Stochastic: *In the stochastic setting with fixed* $P_t = P$ *and Condorcet winner* $a^\star$*:*

$$\mathbb{E}[R_T^{\mathrm{C}}] \leq \sum_{i \neq a^\star} \frac{64 \log T + 192}{\Delta_i} + 8 \log T + 24 + 16\sqrt{K} = O\left(\sum_{i \neq a^\star} \frac{\log T}{\Delta_i}\right). \tag{2}$$

Both bounds hold simultaneously without prior knowledge of the regime.

Our bound $O(\sum_{i \neq a^\star} \log T/\Delta_i)$ improves upon Saha and Gopalan (2018)'s $O(K \log T/\Delta_{\min})$ by depending on individual gaps. Moreover, our algorithm simultaneously achieves $O(\sqrt{KT})$ adversarial regret, which Saha and Gopalan (2018) do not address.

Lower Bounds. For any multi-dueling bandit algorithm: $\mathbb{E}[R_T^{\mathrm{C}}] \geq \Omega(\sqrt{KT})$ for adversarial instances, and $\mathbb{E}[R_T^{\mathrm{C}}] \geq \Omega\left(\sum_{i \neq a^\star} \frac{\log T}{\Delta_i}\right)$ for stochastic instances (Theorems 13–14), establishing optimality.

4.2. Borda Setting

We propose `SA-MiDEX` (Algorithm 2), using the `MetaDueling` reduction with transformed feedback $g_t = (o_t - \alpha_m)/\beta_m$ satisfying $\mathbb{E}[g_t \mid x_t, y_t] = P_t(x_t, y_t)$. It combines decaying uniform opponent sampling, a two-phase exploration-exploitation scheme, martingale-based deviation detection, and an EXP3 fallback.

The algorithm tracks, for each pair (i, j), the cumulative deviation of post-baseline feedback from the locked baseline $\hat{P}_{T_0}$,

$$D_t(i,j) := \left| S_t^{\mathrm{obs}}(i,j) - S_{T_0}^{\mathrm{obs}}(i,j) - \big(N_t(i,j) - N_{T_0}(i,j)\big)\, \hat{P}_{T_0}(i,j) \right|,$$

and switches mode when $D_t(i,j) > \tau(\sqrt{N_t(i,j) - N_{T_0}(i,j)} + 1)$, $\tau = \sqrt{8 \log(K^2 T^2/\delta)}$.

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:

$$q_t(i) = (1-\gamma)\frac{\exp(\eta \tilde{S}_t(i))}{\sum_{j=1}^{K} \exp(\eta \tilde{S}_t(j))} + \frac{\gamma}{K}, \tag{3}$$

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.

Here $\tilde{S}_t(i) := \sum_{\tau \leq t} \hat{s}_\tau(i)$ accumulates the per-round importance-weighted estimator

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

which is unbiased for the shifted Borda score $s_t(i) := \frac{1}{K} \sum_j P_t(i,j)$ (Lemma 16). Since $s_t(i) = \frac{1}{K}\big((K-1)b_t(i) + \frac{1}{2}\big)$, it preserves the Borda ranking.

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. (3)**
  end if
  $o_t \leftarrow$ `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**
    for all $i \in [K]$: $\tilde{S}(i) \leftarrow \tilde{S}(i) + \hat{s}_t(i)$
  end if
end for

Theorem 6 (Stochastic Regret). *In the stochastic setting ($P_t = P$ for all t), Algorithm 2 achieves*

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

Theorem 7 (Adversarial Regret). *Under any oblivious adversarial sequence $P_1, \ldots, P_T$, Algorithm 2 achieves*

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

The adversarial guarantee matches the $\Omega(K^{1/3}T^{2/3})$ lower bound of Gajane (2024) up to logarithmic factors. The stochastic guarantee comes with an $O(K^2 \log KT)$ exploration overhead necessary to estimate global Borda scores.

5. Extensions

Our algorithmic framework flexibly accommodates several practical extensions.

Time-Varying Subset Sizes: Theorems 5, 6, and 7 extend immediately to the setting where the subset size $m_t \in \{2, \ldots, K\}$ varies arbitrarily across rounds. Because 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. Formal statements and proofs are deferred to 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.

6. Discussion and Conclusion

We have presented the first best-of-both-worlds algorithms for multi-dueling bandits, addressing both Condorcet and Borda objectives. The `MetaDueling` reduction demonstrates that optimal Condorcet regret is achievable via a simple, modular transformation of existing dueling bandit algorithms. `SA-MiDEX` shows that gap-dependent stochastic rates and sublinear adversarial rates are simultaneously achievable for Borda objectives. The modularity of our approach ensures that future advances in dueling bandits will automatically benefit multi-dueling applications.

Open directions include: best-of-both-worlds guarantees without a Condorcet winner; exploiting all $\binom{m}{2}$ pairwise comparisons rather than two-arm multisets; dynamic regret against time-varying benchmarks; and leveraging preference structure (e.g., low rank, Plackett-Luce) to reduce the $O(K^2)$ Borda exploration overhead.

References

Arpit Agarwal, Nicholas Johnson, and Shivani Agarwal. Choice bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 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*, pages 116–120. PMLR, 2016.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed 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)*, pages 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, pages 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)*, pages 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. arXiv:2406.12475.

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)*, pages 416–424. PMLR, 2015a.

Kevin Jamieson, Sumeet Katariya, Atul Deshpande, and Robert Nowak. Sparse Dueling Bandits. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38, pages 416–424, San Diego, California, USA, 2015b. PMLR.

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)*, pages 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)*, pages 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*, pages 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*, pages 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)*, pages 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)*, pages 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)*, pages 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)*, pages 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)*, pages 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, page 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)*, pages 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)*, pages 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, pages 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*, pages 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)*, pages 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, page 307–315, Cambridge, MA, USA, 2015. MIT Press.

A. Extended Related Work

Dueling Bandits. The dueling bandit problem was introduced by Yue and 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., 2015a; 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 and 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 and 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 and 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 and Slivkins (2012) first achieved best-of-both-worlds guarantees for multi-armed bandits, though with suboptimal constants. Seldin and Slivkins (2014) and Auer and Chiang (2016) refined these results. The breakthrough came with Zimmert and 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 and Gaillard, 2022).

Stochastic-and-Adversarial Learning. An alternative approach to best-of-both-worlds guarantees is the stochastic-and-adversarial paradigm, introduced by Auer and 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 and 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.

Choice Models and Preference Learning. The pairwise-subset choice model follows Saha and Gopalan (2018), where the probability of each arm winning is determined by averaging pairwise preferences. This generalizes the Bradley-Terry-Luce model (Bradley and 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. The two criteria are guaranteed to align only under structural assumptions such as the Bradley-Terry-Luce (BTL) model or Strong Stochastic Transitivity (SST).

Proposition 8 (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

Lemma 9 (Validity of Rescaled Preferences). *The rescaled matrix $\widehat{P}_t(i, j) = \alpha_m + \beta_m P_t(i, j)$ 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:

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

D. Proofs for Condorcet Setting

D.1. Proof of Lemma 4 (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, the winning probability for position $i \in [n_x]$ (a copy of x_t) is:

$$W(i \mid \mathcal{A}_t, P_t) = \sum_{j \neq i} \frac{2P_t(\mathcal{A}_t(i), \mathcal{A}_t(j))}{m(m-1)}. \tag{4}$$

For position $i \leq n_x$:

- Comparisons against other copies of x_t ($j \le n_x$, $j \ne i$): $n_x - 1$ positions, each contributing $\frac{1}{m(m-1)}$.
- Comparisons against copies of y_t ($j > n_x$): n_y positions, each contributing $\frac{2P_t(x_t,y_t)}{m(m-1)}$.

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

The total probability that some copy of x_t wins is:

$$\mathbb{P}[o_t = 1 \mid x_t, y_t, n_x] = \sum_{i=1}^{n_x} W(i \mid \mathcal{A}_t, P_t) = \frac{n_x(n_x-1)}{m(m-1)} + \frac{2n_x n_y P_t(x_t,y_t)}{m(m-1)}. \tag{5}$$

Case 1: m even. When m is even, $n_x = n_y = m/2$ deterministically. Substituting into (5):

$$\begin{aligned}\mathbb{P}[o_t = 1 \mid x_t, y_t] &= \frac{(m/2)(m/2-1)}{m(m-1)} + \frac{2(m/2)(m/2)P_t(x_t,y_t)}{m(m-1)} \\ &= \frac{m-2}{4(m-1)} + \frac{mP_t(x_t,y_t)}{2(m-1)}. \end{aligned} \tag{6}$$

Comparing with $\beta_m = \frac{m}{2(m-1)}$ and $\alpha_m = \frac{m-2}{4(m-1)}$, we have (6) $= \alpha_m + \beta_m P_t(x_t,y_t)$.

Case 2: m odd. With probability 1/2: $(n_x, n_y) = (\frac{m+1}{2}, \frac{m-1}{2})$; with probability 1/2: $(n_x, n_y) = (\frac{m-1}{2}, \frac{m+1}{2})$.

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

$$\begin{aligned}\mathbb{P}[o_t = 1 \mid n_x = \tfrac{m+1}{2}] &= \frac{(m+1)(m-1)}{4m(m-1)} + \frac{(m+1)(m-1)P_t(x_t,y_t)}{2m(m-1)} \\ &= \frac{m+1}{4m} + \frac{(m+1)P_t(x_t,y_t)}{2m}. \end{aligned} \tag{7}$$

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

$$\begin{aligned}\mathbb{P}[o_t = 1 \mid n_x = \tfrac{m-1}{2}] &= \frac{(m-1)(m-3)}{4m(m-1)} + \frac{(m-1)(m+1)P_t(x_t,y_t)}{2m(m-1)} \\ &= \frac{m-3}{4m} + \frac{(m+1)P_t(x_t,y_t)}{2m}. \end{aligned} \tag{8}$$

Taking the average over the symmetrization:

$$\begin{aligned}\mathbb{P}[o_t = 1 \mid x_t, y_t] &= \frac{1}{2}\left(\frac{m+1}{4m} + \frac{m-3}{4m}\right) + \frac{(m+1)P_t(x_t,y_t)}{2m} \\ &= \frac{m-1}{4m} + \frac{m+1}{2m}P_t(x_t,y_t). \end{aligned} \tag{9}$$

Comparing with $\beta_m = \frac{m+1}{2m}$ and $\alpha_m = \frac{m-1}{4m}$, we have (9) $= \alpha_m + \beta_m P_t(x_t,y_t)$.

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

Lemma 10 (Gap Rescaling). *For any arm $i \in [K]$, the rescaled gap satisfies: $\widehat{\Delta}_t(i) := \widehat{P}_t(a^\star, i) - \frac{1}{2} = \beta_m \Delta_t(i)$.*

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

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

□

Lemma 11 (Bound on Rescaling Factor). *For all $m \geq 2$, the rescaling factor satisfies $\beta_m \geq \frac{1}{2}$, with equality as $m \to \infty$.*

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

Lemma 12 (Regret Equivalence). *For any sequence of preference matrices $P_1, \ldots, P_T$ satisfying Assumption 1:*

$$\mathbb{E}[R_T^{\mathrm{C}}] = \frac{1}{\beta_m} \, \mathbb{E}[\widehat{R}_T^{\mathrm{DB}}], \tag{10}$$

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 $\widehat{P}_t$.

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 of Theorem 5

Proof. By Lemma 4, 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 9).

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 (Saha and Gaillard, 2022, Theorem 1). By Lemma 12:

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

Since $\beta_m \geq 1/2$ (Lemma 11), $1/\beta_m \leq 2$, establishing the adversarial bound.

Part (ii): Stochastic bound. The self-bounding condition relates expected regret to the cumulative probability mass placed on suboptimal arms. When this condition holds (which is automatic in the stochastic setting), `Versatile-DB` satisfies (Saha and 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 10, $\widehat{\Delta}_i = \beta_m \Delta_i$, 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 12 (multiplication by $1/\beta_m$) and using $1/\beta_m^2 \leq 4$, $1/\beta_m \leq 2$ completes the proof. □

D.5. Proof of Lower Bounds

Theorem 13 (Adversarial Lower Bound). *For any multi-dueling bandit algorithm and any $T \geq K \geq 4$, there exists an adversarial instance satisfying Assumption 1 such that $\mathbb{E}[R_T^{\mathrm{C}}] \geq \Omega(\sqrt{KT})$.*

Proof. 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. 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 is $r_t^{\mathrm{DB}} = \frac{1}{2}\left(\Delta_t(\mathcal{A}_t(i_t)) + \Delta_t(\mathcal{A}_t(j_t))\right)$. Taking expectation over the uniform sampling of (i_t, j_t):

$$\begin{aligned}
\mathbb{E}_{i_t, j_t}[r_t^{\mathrm{DB}}] &= \frac{1}{2} \cdot \frac{1}{\binom{m}{2}} \sum_{i<j} \left(\Delta_t(\mathcal{A}_t(i)) + \Delta_t(\mathcal{A}_t(j))\right) \\
&= \frac{1}{2} \cdot \frac{2}{m(m-1)} \cdot (m-1) \sum_{k=1}^{m} \Delta_t(\mathcal{A}_t(k)) = \frac{1}{m} \sum_{k=1}^{m} \Delta_t(\mathcal{A}_t(k)) = r_t^{\mathrm{MDB}}. \qquad (11)
\end{aligned}$$

Summing over $t = 1, \ldots, T$: $\mathbb{E}[R_T^{\mathrm{DB}}] = \mathbb{E}[R_T^{\mathrm{C}}]$.

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 $a^\star = 1$, with $P_t(1, i) = \frac{1}{2} + \epsilon_t(i)$ for adversarially chosen $\epsilon_t(i) \in [0, \frac{1}{2}]$) such that $\mathbb{E}[R_T^{\mathrm{DB}}] \geq c\sqrt{KT}$ for some universal constant $c > 0$.

Step 4: Transfer to multi-dueling. Suppose for contradiction that $\mathcal{A}_{\mathrm{MDB}}$ achieves $\mathbb{E}[R_T^{\mathrm{C}}] = o(\sqrt{KT})$ on all adversarial Condorcet instances. Then the constructed $\mathcal{A}_{\mathrm{DB}}$ satisfies $\mathbb{E}[R_T^{\mathrm{DB}}] = \mathbb{E}[R_T^{\mathrm{C}}] = o(\sqrt{KT})$, contradicting Step 3.

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

Theorem 14 (Stochastic Lower Bound). *For any multi-dueling bandit algorithm, there exists a stochastic instance with Condorcet winner $a^\star$ such that $\mathbb{E}[R_T^{\mathrm{C}}] \geq \Omega\left(\sum_{i \neq a^\star} \frac{\log T}{\Delta_i}\right)$.*

Proof. **Steps 1–2** (same reduction as Theorem 13): $\mathbb{E}[R_T^{\mathrm{DB}}] = \mathbb{E}[R_T^{\mathrm{C}}]$.

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 $a^\star$:

$$\liminf_{T \to \infty} \frac{\mathbb{E}[R_T^{\mathrm{DB}}]}{\log T} \geq \sum_{i \neq a^\star} \frac{1}{2 \cdot \mathrm{KL}(P(a^\star, i) \| \frac{1}{2})}, \tag{12}$$

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 $\Delta_i > 0$:

$$\mathrm{KL}\left(\tfrac{1}{2} + \Delta_i \middle\| \tfrac{1}{2}\right) = \left(\tfrac{1}{2} + \Delta_i\right) \log(1 + 2\Delta_i) + \left(\tfrac{1}{2} - \Delta_i\right) \log(1 - 2\Delta_i). \tag{13}$$

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

$$\begin{aligned} \mathrm{KL}\left(\tfrac{1}{2} + \Delta_i \middle\| \tfrac{1}{2}\right) &= \left(\tfrac{1}{2} + \Delta_i\right)\left(2\Delta_i - 2\Delta_i^2 + O(\Delta_i^3)\right) + \left(\tfrac{1}{2} - \Delta_i\right)\left(-2\Delta_i - 2\Delta_i^2 + O(\Delta_i^3)\right) \\ &= 2\Delta_i^2 + O(\Delta_i^3). \end{aligned} \tag{14}$$

Hence $\frac{1}{2\,\mathrm{KL}(\frac{1}{2}+\Delta_i \| \frac{1}{2})} = \Theta(1/\Delta_i^2)$. The lower bound (12) states that arm i must be pulled at least $\Omega(\log T / \mathrm{KL}(\cdot))$ times. Since each pull contributes Δ_i to regret:

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

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

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

E. Proofs for Borda Setting

E.1. Transformed Feedback Properties

Lemma 15 (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) $\mathrm{Var}(g_t \mid x_t, y_t) \leq \frac{1}{4\beta_m^2} \leq 1$.

Proof. (i) By Lemma 4, $\mathbb{E}[o_t \mid x_t, y_t] = \alpha_m + \beta_m P_t(x_t, y_t)$. Thus:

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

(ii) Since $o_t \in \{0, 1\}$, $g_t \in \{-\alpha_m/\beta_m, (1-\alpha_m)/\beta_m\}$. For $m \geq 2$, $\beta_m \geq 1/2$ and $\alpha_m \leq 1/4$, so $-\alpha_m/\beta_m \geq -1/2 > -1$ and $(1-\alpha_m)/\beta_m \leq 2$.

(iii)

$$\mathrm{Var}(g_t \mid x_t, y_t) = \frac{\mathrm{Var}(o_t \mid x_t, y_t)}{\beta_m^2} \leq \frac{1/4}{\beta_m^2} \leq \frac{1/4}{1/4} = 1. \qquad \square$$

Lemma 16 (Unbiased Importance-Weighted Estimation). *For $x_t, y_t \overset{i.i.d.}{\sim} q_t$, the estimator $\hat{s}_t(i) = \frac{\mathbb{1}(x_t=i)}{K\, q_t(i)} \sum_{j\in[K]} \frac{\mathbb{1}(y_t=j)\, g_t}{q_t(j)}$ is unbiased for the shifted Borda score: $\mathbb{E}[\hat{s}_t(i) \mid q_t] = s_t(i) = \frac{1}{K}\sum_{j\in[K]} P_t(i,j)$.*

Proof. Taking expectations over $x_t, y_t \overset{\text{i.i.d.}}{\sim} q_t$ and using $\mathbb{E}[g_t \mid x_t, y_t] = P_t(x_t, y_t)$ (Lemma 15(i)):

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

E.2. Exploration Phase Analysis

Lemma 17 (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) $|\hat{P}_{T_0}(i,j) - P(i,j)| \leq O(\sqrt{\log T/n_0})$ for all pairs.

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

Proof. (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 $\mathbb{E}[Z_k] = 0$, $|Z_k| \leq 3$ (since $g_{\tau_k} \in [-1, 2]$ and $P(i,j) \in [0,1]$), and $\mathrm{Var}(Z_k) \leq 1$. By Hoeffding's inequality:

$$\Pr\Big[|\hat{P}_{T_0}(i,j) - P(i,j)| > \epsilon\Big] \leq 2\exp\left(-\frac{2n_0\epsilon^2}{9}\right).$$

Setting $\epsilon = \sqrt{\frac{9\log(4K^2/\delta)}{2n_0}}$ and using $n_0 \geq 4\log(4K^2/\delta)$:

$$\Pr\Big[|\hat{P}_{T_0}(i,j) - P(i,j)| > \epsilon\Big] \leq 2\exp(-\log(4K^2/\delta)) = \frac{\delta}{2K^2}.$$

Union bound over $K(K-1) < K^2$ pairs gives failure probability $\leq \delta/2$.

(iii) Borda score concentration.

$$|\hat{b}_{T_0}(i) - b(i)| \leq \frac{1}{K-1}\sum_{j\neq i} |\hat{P}_{T_0}(i,j) - P(i,j)| \leq \epsilon = O\left(\sqrt{\frac{\log T}{n_0}}\right). \qquad \square$$

E.3. Deviation Detection Analysis

Lemma 18 (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.*

Proof. 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:

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

Step 1: Martingale properties. The increments $X_k = g_{\tau_k} - P(i, j)$ satisfy: $\mathbb{E}[X_k \mid \mathcal{F}_{k-1}] = 0$; $|X_k| \leq 3$; $\mathbb{E}[X_k^2 \mid \mathcal{F}_{k-1}] \leq 1$.

Step 2: Freedman's inequality. With $c = 3$, $V_n \leq n$, and $x = \tau(\sqrt{n} + 1)$ where $\tau = \sqrt{8 \log(K^2 T^2/\delta)}$:

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

For $n \geq \tau^2$, $\frac{\tau^2(\sqrt{n}+1)^2}{2n+2\tau(\sqrt{n}+1)} \geq \frac{\tau^2}{4}$. In either case:

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

Step 3: Connect to detection statistic.

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

where $\epsilon_0 = O(\sqrt{\log T/n_0})$ from Lemma 17(ii). For the threshold $\tau(\sqrt{n}+1)$ to be spuriously exceeded, we would need $n\epsilon_0 > \tau(\sqrt{n} + 1)/2$. For $n_0 = \Theta(\log T)$ and $\tau = \Theta(\sqrt{\log T})$, $\epsilon_0 = O(1/\sqrt{\log T})$, and this simplifies to $\sqrt{n} > \Omega(\log T)$, i.e., $n > \Omega(\log^2 T)$—which requires $|M_n| > \tau(\sqrt{n} + 1)/2$ to also hold. The probability of this is bounded above.

Step 4: Union bound.

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

□

Lemma 19 (Adversarial Detection Guarantee). *If the adversary's cumulative preference shift on pair (i, j) exceeds $3\tau(\sqrt{n} + 1)$ where n is the post-exploration comparison count, then deviation is detected with probability at least $1 - \frac{2\delta}{K^2 T^2}$.*

Proof. Suppose the adversary shifts preferences such that:

$$\left|\sum_{k=1}^{n} (P_{\tau_k}(i, j) - \hat{P}_{T_0}(i, j))\right| \geq 3\tau(\sqrt{n} + 1). \tag{15}$$

Decompose:

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

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

Under condition (15), with probability at least $1 - \frac{2\delta}{K^2T^2}$:

$$D_n \geq |\text{adversarial shift}| - |M_n| \geq 3\tau(\sqrt{n}+1) - \tau(\sqrt{n}+1) = 2\tau(\sqrt{n}+1) > \tau(\sqrt{n}+1),$$

triggering a mode switch. □

E.4. Confidence and Elimination Analysis

Lemma 20 (Borda Score Concentration). *During the elimination phase, for any arm i, $\epsilon > 0$, and $N_{\min}(i) = \min_{j \neq i} N_n(i,j)$:*

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

Proof. The error satisfies:

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

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

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

Union bound over all $K-1$ opponents $j \neq i$:

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

Lemma 21 (Confidence Validity). *With probability at least $1 - \delta/4$, for all arms i and rounds $t > T_0$:*

$$|\hat{b}_t(i) - b(i)| \leq \mathit{conf}_t(i) := \sqrt{\frac{20K\log(2K^2T/\delta)}{N_t(i)}}.$$

Proof. **Step 1: Relating $N_t(i)$ to $N_{\min}(i)$.** Because the algorithm samples y_t uniformly from $[K] \setminus \{x_t\}$, by standard Chernoff bounds with high probability: $N_{\min}(i) \geq \frac{N_t(i)}{2(K-1)} \geq \frac{N_t(i)}{2K}$, so $\frac{1}{N_{\min}(i)} \leq \frac{2K}{N_t(i)}$.

Step 2: Fixed-time bound. For fixed n, let $\mathrm{conf}_n(i) = \sqrt{\frac{10\log(2K^2T/\delta)}{N_{\min}(i)}}$. By Lemma 20 with $\epsilon = \mathrm{conf}_n(i)$:

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

Using 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 $E_\ell = \{t : 2^{\ell-1} \leq N_{\min}(i) < 2^\ell\}$ for $\ell = 1, \ldots, \lceil \log_2 T \rceil$.

Step 4: Union bound.

$$\Pr[\mathcal{E}_{\mathrm{conf}}^c] \leq K \cdot \log_2 T \cdot \frac{\delta}{KT} = \frac{\delta \log_2 T}{T} \leq \frac{\delta}{4}. \quad \square$$

Lemma 22 (Elimination Sample Complexity). *Conditioned on confidence validity, each suboptimal arm j is eliminated from $\mathcal{C}$ after at most $N_T(j) \leq \frac{320K\log(2K^2T/\delta)}{(\Delta_j^{\mathrm{B}})^2} + 1$ pulls.*

Proof. Conditioned on $\mathcal{E}_{\text{conf}}$, the true optimal arm $i^\star$ is never eliminated: for any $j \in \mathcal{C}$:

$$\hat{b}_t(i^\star) + \text{conf}_t(i^\star) \geq b(i^\star) \geq b(j) \geq \hat{b}_t(j) - \text{conf}_t(j),$$

so the elimination condition $\hat{b}_t(j) - \text{conf}_t(j) > \hat{b}_t(i^\star) + \text{conf}_t(i^\star)$ can never hold against $i^\star$.

A suboptimal arm j is eliminated when there exists $i \in \mathcal{C}$ with $\hat{b}_t(i) - \text{conf}_t(i) > \hat{b}_t(j) + \text{conf}_t(j)$. Since $i^\star$ remains in $\mathcal{C}$ and $\Delta_j^{\mathrm{B}} = b(i^\star) - b(j)$, this condition is guaranteed to trigger against $i^\star$ once $2\text{conf}_t(i^\star) + 2\text{conf}_t(j) < \Delta_j^{\mathrm{B}}$. Because the algorithm uses round-robin over $\mathcal{C}$, arms are pulled symmetrically, so $\text{conf}_t(i^\star) \approx \text{conf}_t(j)$, and arm j is eliminated once $4\,\text{conf}_t(j) \leq \Delta_j^{\mathrm{B}}$, i.e.,

$$4\sqrt{\frac{20K\log(2K^2T/\delta)}{N_t(j)}} \leq \Delta_j^{\mathrm{B}}.$$

Squaring both sides:

$$N_T(j) \leq \frac{320K\log(2K^2T/\delta)}{(\Delta_j^{\mathrm{B}})^2} + 1.$$

□

E.5. Adversarial Mode Analysis

Lemma 23 (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:*

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

Proof. 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 $q_t(j) \geq \gamma/K$ 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:

$$\mathbb{E}[R_{T'}^s] \leq O\Big(K^{1/3}(T')^{2/3}(\log K)^{1/3}\Big).$$

□

E.6. Main Theorem Proofs

Proof of Theorem 6 (Stochastic Regret). Define the high-probability event $\mathcal{E} = \mathcal{E}_{\text{conf}} \cap \mathcal{E}_{\text{no-switch}}$. By union bounds, $\Pr[\mathcal{E}^c] \leq O(1/T)$. Conditioned on $\mathcal{E}$, the algorithm never switches to adversarial mode, and $i^\star$ is never eliminated.

Step 1: Regret from Baseline Exploration. During rounds $t \leq T_0$, instantaneous Borda regret is at most 1. Thus:

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

Step 2: Regret from Elimination Phase. For $t > T_0$ where suboptimal arm j is selected as $x_t \in \mathcal{C}$, opponent $y_t \sim \text{Uniform}([K] \setminus \{x_t\})$. The expected instantaneous regret is:

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

By Lemma 22, arm j is pulled at most $N_T(j) \leq O(K \log KT/(\Delta_j^{\mathrm{B}})^2)$ times. The cumulative elimination regret is:

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

Step 3: Regret from Exploitation. Once $\mathcal{C} = \{i^\star\}$, the algorithm plays $x_t = y_t = i^\star$ with instantaneous regret 0.

Step 4: Regret from Background Monitoring. The algorithm overrides with probability $p_t = \min(1, \frac{K \log t}{t})$; instantaneous regret is at most 1. The total monitoring regret satisfies:

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

Step 5: Regret on the Failure Event. $\mathbb{E}[R_T^{\mathrm{B}} \cdot \mathbb{1}[\mathcal{E}^c]] \leq T \cdot O(1/T) = O(1)$.

Step 6: Total Regret.

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

Proof of Theorem 7 (Adversarial Regret). We analyze two cases based on whether a mode switch occurs.

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

Exploration regret: $R_{T_0}^{\mathrm{B}} \leq T_0 = O(K^2 \log KT)$.

Pre-switch regret ($T_0 < t < t_{sw}$): For each pair (i, j), detection occurs when 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:

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

where we used Cauchy-Schwarz and $\sum_{(i,j)} N_t(i, j) \leq 2T$. Since $\tau = O(\sqrt{\log KT})$: $\mathcal{R}_{\mathrm{pre}} \leq O(K \sqrt{T \log KT})$.

Post-switch regret ($t \geq t_{sw}$): By Lemma 23: $\mathbb{E}[\mathcal{R}_{\mathrm{post}}] \leq O(K^{1/3}(T - t_{\mathrm{sw}})^{2/3} (\log K)^{1/3}) \leq O(K^{1/3} T^{2/3} (\log K)^{1/3})$.

Lemma 23 bounds the shifted Borda regret $R_T^s := \sum_t \big(s_t(i^\star) - \frac{1}{2}(s_t(x_t) + s_t(y_t)) \big)$; since $s_t(i) = \frac{K-1}{K} b_t(i) + \frac{1}{2K}$, the Borda and shifted regrets satisfy $R_T^{\mathrm{B}} = \frac{K}{K-1} R_T^s \leq 2 R_T^s$ for $K \geq 2$, so the same rate holds for R_T^{B} up to this constant.

Case 2: No switch occurs ($t_{\mathbf{sw}} > T$). The adversary's cumulative deviation on each pair is bounded by the detection threshold. The total regret from adversarial deviations satisfies:

$$R_T^{\text{adv}} \leq \sum_{(i,j)} O(\tau(\sqrt{N_T(i,j)}+1)) \leq O(\tau K^2) + O\Big(\tau K\sqrt{T}\Big) = O(K^2\sqrt{\log KT} + K\sqrt{T\log KT}). \tag{16}$$

Combined bound. In both cases:

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

For $T \geq K^2$, $O(K^2 \log KT)$ is dominated by $O(K\sqrt{T\log KT})$. Since neither the pre-switch nor post-switch term dominates universally across all horizons, we retain both:

$$\mathbb{E}[R_T^{\text{B}}] \leq O\Big(K\sqrt{T\log KT} + K^{1/3}T^{2/3}(\log K)^{1/3}\Big).$$ □

F. Proofs for Extensions

Corollary 24 (Time-Varying Subset Size: Condorcet Setting). *Theorem 5 holds when the subset size $m_t \in \{2,\dots,K\}$ varies arbitrarily across rounds.*

Proof. The proof of Lemma 12 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 $\mathbb{E}[r_t^{\text{C}}] = \frac{1}{\beta_{m_t}}\mathbb{E}[\hat{r}_t^{\text{DB}}]$ holds pointwise.

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

Corollary 25 (Time-Varying Subset Size: Borda Setting). *Theorems 6 and 7 hold when $m_t \in \{2,\dots,K\}$ varies arbitrarily across rounds, provided the transformed feedback g_t 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 (by Lemma 15(i) applied pointwise). The stochastic analysis (exploration, deviation detection) and adversarial analysis (EXP3) depend only on the unbiasedness and boundedness of g_t, both of which hold uniformly over $m_t \geq 2$. □

G. Additional Technical Results

G.1. Concentration Inequalities

Lemma 26 (Hoeffding's Inequality). *Let $X_1,\dots,X_n$ be independent random variables with $X_i \in [a_i,b_i]$. Then for $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$:*

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

Lemma 27 (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 $x, v > 0$:*

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