The Learnability of the Multiplayer Adversarial Bandit Problem

Anonymous authors
Paper under double-blind review

Abstract

Following the work of Chang et al. (2022), we consider the information asymmetric multiplayer adversarial bandits where at each round, the players each have their own set of actions to select from, contributing to a joint action. The players are not allowed to communicate during learning; however, they are allowed to agree on a strategy beforehand. We show that when the players pull simultaneously, there always exists an adaptive adversary that can incur linear regret. We modify the setting so that the pulls are successive instead, and show near optimal regret bounds in the case of successive pulls, both with and without reward asymmetry.

1 Introduction

The adversarial bandit problem has emerged as a foundational framework for decision-making across multiple time steps under uncertainty when faced with a potentially malicious environment. Unlike stochastic bandits, where rewards are drawn from fixed but unknown distributions, adversarial bandits consider scenarios where rewards may be chosen by an adaptive adversary who actively attempts to minimize the decision-maker's cumulative reward. This model has proven invaluable in applications ranging from online advertising and recommendation systems to network routing and security, where unpredictable or strategic environments naturally arise. The problem has been extensively studied since the seminal work of Auer et al. (1995), which established the EXP3 and Hedge algorithms achieving near-optimal regret bounds of $O(\sqrt{KT \log K})$ for K arms over T rounds. Subsequent research has further refined these bounds (examples include Audibert & Bubeck (2010) and Orabona & Pál (2016)), as well as developed algorithms with improved performance guarantees under various constraints and assumptions about the adversary's capabilities and the feedback mechanisms available to the learner (Bubeck & Slivkins (2012)).

In a different vein, many real-world applications inherently involve multiple decision-makers operating in shared environments. Numerous recent works have extended bandit frameworks in general to multiplayer settings Chang et al. (2022); Chang & Lu (2024); Chang & Karthik (2025); Chang & Lu (2025), where each player selects actions from their own action set, collectively generating joint actions and corresponding rewards. While the objective of multiplayer bandits remains similar to the single-player case, these settings introduce a fundamental challenge: *information asymmetry.* Depending on the problem formulation, players may observe joint actions but not other players' rewards (action asymmetry), receive different i.i.d. observations of the same underlying events (reward asymmetry), or face both action and reward asymmetries. Addressing these various forms of information asymmetry represents the central challenge in multiplayer bandits research.

In this work, we are interested in analyzing the intersection of adversarial bandits in a multi-player sense. As such, we consider a version of multiplayer adversarial bandits (MAB) where each agent has a distinct set of individual actions (marginal arms), and the resulting joint outcome is determined by the combination of arms selected by all players. This formulation naturally models distributed decision-making scenarios where agents have different capabilities or action spaces, such as in multi-robot coordination or decentralized resource allocation problems. Our approach extends the classical adversarial framework established by Auer et al. (1995) to the multiplayer domain, while building upon the current literature of multiplayer bandits as

established by Chang et al. (2022). This setting also covers information asymmetry challenges, as players must coordinate despite having limited visibility into others' action spaces and observations.

Our contribution In Chang et al. (2022); Chang & Lu (2024; 2025); Chang & Karthik (2025), the authors were able to prove near optimal regret bounds for the multiplayer information asymmetry bandit problem. Remarkably, we are able to show that in the same setting, when the rewards are adversarial, a strategic adversary can enforce linear regret. To address this, we modify the multiplayer setting so that actions occur sequentially, allowing each player to observe the choices of the previous one. In the case where all players experience the same losses, we propose `mHedge-A`, an algorithm that achieves a regret bound of $\tilde{O}(\sqrt{TM})$, where T represents the number of rounds and M the number of players. This has practical relevance for situations where communication is costly in terms of time or resources. Lastly, we tackle reward asymmetry by introducing `mHedge-B`, an extension of `mHedge-A`, which attains a regret bound of $\tilde{O}(M^2\sqrt{T})$. This result highlights how performance deteriorates as noise affects the observed rewards while still near optimal regret growth with respect to T.

Related Works

The adversarial bandits landscape has evolved substantially since the pioneering work of Auer et al. (1995), which introduced the problem of prediction with expert advice under adversarial rewards. Building on this foundation, Auer et al. (2002) developed the seminal EXP3 algorithm that addressed the more challenging bandit feedback setting. This momentum continued with Audibert & Bubeck (2010), who refined these approaches through their Implicitly Normalized Forecaster, establishing tighter minimax optimal regret bounds that significantly improved performance. Bubeck & Slivkins (2012) further crafted sophisticated algorithms that leverage consistency conditions to achieve polylog(T) regret in stochastic environments while maintaining $\tilde{O}(T)$ guarantees for adversarial settings. The challenge of developing algorithms that perform well without prior knowledge of the environment type was tackled by Seldin & Slivkins (2014) and later Auer & Chiang (2016), who enhanced EXP3 to create robust algorithms that adapt to the environment without requiring prior knowledge of whether it is stochastic or adversarial. This line of research culminated in the work of Zimmert & Seldin (2019), who developed theoretically optimal algorithms achieving the best possible regret bounds across arbitrary environments, effectively bridging the gap between stochastic and adversarial formulations.

Many other variations of adversarial models address practical challenges and connect to popular frameworks in the bandit literature. Contextual adversarial bandits, introduced by Auer (2002) and further developed by Beygelzimer et al. (2011) and Agarwal et al. (2014), incorporate side information available before decision-making, allowing algorithms to adapt their strategies based on observed contexts. This approach has been refined by Syrgkanis et al. (2016) with variance-based improvements. Combinatorial adversarial bandits Cesa-Bianchi & Lugosi (2012) significantly expand the action space by enabling learners to select subsets of base actions simultaneously, with Neu (2015) and Kveton et al. (2015) establishing tighter bounds for semi-bandit feedback. The bandits with knapsacks framework, pioneered by Badanidiyuru et al. (2013), extended by Agrawal & Devanur (2016) and Rangi et al. (2018), and finally incorporated with adversarial patterns by Immorlica et al. (2022), incorporates resource constraints where actions consume from limited budgets, modeling real-world scenarios with capacity limitations or cost considerations. Dueling bandits Yue et al. (2012) represent a fundamental shift in feedback structure, focusing on preference-based learning where only relative comparisons between actions are observable rather than absolute rewards, with Zoghi et al. (2014), Komiyama et al. (2015) and Saha et al. (2021) developing efficient algorithms for both stochastic and adversarial variants. Finally, corrupted adversarial models, as shown in Hajiesmaili et al. (2020) and Ma & Zhou (2023), consider corruption budgets for corrupted non-stochastic bandits to ensure robustness against adversaries modifying limited rewards or contexts.

Our work also intersects with the expanding field of multi-player bandits, first introduced by Awerbuch & Kleinberg (2008) for collective action selection. This area has evolved through diverse algorithmic approaches, including ϵ-greedy variations Szorenyi et al. (2013); Jin et al. (2023), distributed methods like gossip UCB Landgren et al. (2016); Martínez-Rubio et al. (2019), distinctive actions through time iterations (Uchiya et al. (2010)) and the leader-follower DPE1 framework Wang et al. (2020). Other researchers have explored various coordination mechanisms, which include observation constraints limited to neighboring players Cesa-Bianchi

et al. (2016), information exchange through graph networks Xu & Klabjan (2024), asynchronous activation of player subsets Bonnefoi et al. (2017); Cesa-Bianchi et al. (2020), collective decision-making through voting systems Shahrampour et al. (2017), and communicating agents with personalized action sets to optimize cumulative group performance Dubey et al. (2020). Note that in these works, communication typically happens during the learning process where they are able to (for some cost) communicate feedback from the environment. In our setting, by enforcing communication only at the start of learning, the only information they are given at that moment is the size of the environment and the number of learning rounds.

At the heart of this intersection lies adversarial, multi-player bandits, which has also garnered some interest. Alatur et al. (2020) pioneered work in this area by introducing a collision penalty framework motivated by radio networks, where multiple agents compete for limited channel resources. Mahesh et al. (2024) further used adversarial algorithms to achieve better results than simple algorithms in radar network applications, particularly in contested electromagnetic environments. Similarly, Tong et al. (2021) extended these concepts to CSMA networks and developed specialized metrics to accomodate for the challenges that distributed wireless access in adversarial settings bring. Finally, Shi & Shen (2021) proposed a new orthogonal hardness metric called attackability that characterizes vulnerability to strategic manipulation, investigating fundamental tradeoffs between robustness and performance. While these works have been crucial in expanding adversarial, multi-player bandits, as far as we are aware, none of the literature covers the restrictions that information asymmetry gives.

2 Preliminary

Single-Player Adversarial Bandits We first define the single-player adversarial setting. Let K be the number of arms identified with the set $[K] = \{1, 2, \ldots, K\}$ and let T be the time horizon. The problem proceeds in rounds $t = 1, 2, \ldots, T$ as follows:

1. First, an adversary selects a loss vector $\ell_t \in [0, 1]^K$, where $\ell_t(a)$ represents the loss for arm a in round t. The adversary may choose these losses adaptively based on the learner's previous actions.
2. Then, without observing ℓ_t, the learner selects an arm $a_t \in [K]$.
3. Finally, the learner observes the full loss vector ℓ_t and suffers the loss $\ell_t(a_t)$.

Let $\mathcal{A}_{t-1}$ denote all information available to the learner before round t. As mentioned above, since the adversary is adaptive, they may choose losses based on $\mathcal{A}_{t-1}$, and may also anticipate the learner's action a_t in the current round.

The goal is to minimize the regret with respect to the best fixed arm in hindsight, defined below:

$$R_T = \sum_{t=1}^{T} \ell_t(a_t) - \min_{a \in [K]} \sum_{t=1}^{T} \ell_t(a) \tag{1}$$

Note: The problem is often equivalently formulated with rewards $r_t(a) = 1 - \ell_t(a)$ instead of losses, in which case the goal is to maximize rewards and the regret definition changes accordingly.

Oblivious vs Adaptive Adversary It's worth noting that various characterizations of adversarial knowledge appear in the literature (refer to Related Works above), but this distinction is largely theoretical in our scenario. Under our approach of adversarial bandits, the essence of the problem lies in preparing for worst-case scenarios rather than average-case performance, as the adversarial framework makes no probabilistic assumptions about the environment. As such, we will assume below that our adversaries are adaptive, and we must deal with the omniscient adversary's decisions as needed. This also ensures that our results are generalizable towards all adversaries.

Multi-Player Extension of Adversarial Bandits

We extend this definition to the multi-player setting, noting that multiple players pull arms that combine to form a joint arm. As in the adversarial setting, the players are given the entire loss vector for all possible joint arms.

To formally define this, let each of M players have K arms (represented by $[K]$), for a total of K^M joint arms. Each joint arm $\mathbf{a} \in [K]^M$ has a dimension in the loss vector ℓ_t.. We shall borrow notation from computer science and define $\boldsymbol{a}[1:i]$ to be the first $i-1$ components of the joint arm $\boldsymbol{a}$. The problem proceeds in rounds $t = 1, 2, \ldots, T$ as follows:

1. First, an adversary selects a loss vector $\ell_t \in [0,1]^{K^M}$, where $\ell_t(\mathbf{a})$ represents the loss for joint arm $\mathbf{a}$ in round t. As clarified above, the adversary may now choose these losses adaptively based on the learner's previous actions.
2. Then, without observing ℓ_t, each learner $m \in [M]$ selects their marginal arm a_m to form the joint arm $\mathbf{a}_t = [a_1, \ldots, a_M] \in [K]^M$.
3. Finally, all learners observes the full loss vector ℓ_t and suffers the loss $\ell_t(\mathbf{a_t})$ defined above.

Note that in the above setting the players are not allowed to communicate during the learning, however they are able to coordinate a strategy prior to the start of learning. Regret is defined similarly to the single-player setting:

$$R_T = \sum_{t=1}^{T} \ell_t(\mathbf{a}_t) - \min_{\mathbf{a} \in [K]^M} \sum_{t=1}^{T} \ell_t(\mathbf{a}) \tag{2}$$

Note that we can generalize this setting to variable action count settings simply by setting some marginal actions to have regret of 1. In particular, if player P_i has K_{P_i} arms, and $K = \max_i K_{P_i}$, we can set $\ell_t(\mathbf{a}) = 1$ for any $\ni \notin \prod_{i=1}^{M}[1, K_{P_i}]$. This forces all arms outside of $\prod_{i=1}^{M}[1, K_{P_i}]$ to be sub-optimal regardless of strategy (as long as one joint arm has loss less than 1), and ensures that only joint arms within $\prod_{i=1}^{M}[1, K_{P_i}]$ can be selected.

Information Asymmetry In contrast to the stochastic framework of Chang et al. (2022), the adversarial setting eliminates action asymmetry since players observe complete loss vectors for all arms, including those not selected. Note that this is necessary, as our definition of regret can be ill-defined otherwise. This fundamentally changes the information structure compared to stochastic settings, and forces us to concentrate on modeling reward asymmetry through a modified version of Problem B defined in Chang et al. (2022), where each player receives the shared adversarial loss corrupted by independent identically distributed 0-mean sub-Gaussian noise. More specifically, we define

- No asymmetry: each player receives the same loss
- Reward asymmetry: each player recives the same loss but corrupted by some identically distributed 0-mean sub-Gaussian noise. They do not get to see the corrupted copies of the other players.

3 A Lower Bound on Simultaneous (Skull Adversary)

In Chang et al. (2022); Chang & Lu (2024; 2025); Chang & Karthik (2025) they were able to prove near optimal regret bounds for the multiplayer information asymmetry bandit problem. In this section we construct an adaptive adversary (the skull adversary) that causes the learner to experience linear regret. Recall the multi-player adversarial bandit setting defined above, where players may employ stochastic strategies, and decide their marginal arms simultaneously. As mentioned above, we assume each player also has exactly K arms.

Since the players are not allowed to communicate, they can only decide on a random policy on their own action space (rather than the joint actions space). Therefore, the players' strategy space has $(K-1)M$

degrees of freedom, as each player can independently assign probabilities to $K-1$ of their actions (with the probability of the K-th action being determined by the constraint that probabilities sum to 1). However, the joint action space contains K^M possible action combinations. This dimensional discrepancy implies that there exist numerous distributions over the joint action space that cannot be realized by any combination of independent player strategies, which allows any space-exploitative adversary to force higher regret bounds, assuming no communication is allowed.

Using that bit of wisdom, we now demonstrate that an adaptive adversary can force the players to incur linear regret in this setting.

Theorem 1 (Skull Adversary Theorem) *In a two-player adversarial bandit game where each player has two actions, there exists an adaptive adversary strategy (the Skull Adversary) that forces players to incur $\Omega(T)$ regret.*

Proof: Consider a scenario with two players ($M = 2$), each having two actions ($K = 2$). Let the joint action space be denoted by $a_{1,1}, a_{1,2}, a_{2,1}, a_{2,2}$, where $a_{i,j}$ represents player 1 selecting action i and player 2 selecting action j. We define the Skull Adversary strategy as follows:

1. Assign loss $\ell_S(a_{1,2}) = \ell_S(a_{2,1}) = 1$ for all rounds.
2. Let $p_t(a_{i,j})$ denote the probability of joint action $a_{i,j}$ under the players' strategy profile at round t. For the remaining two actions, assign:
   (a) $\ell_S(a_{i,j}) = 0$ to the joint action $a_{i,j} \in a_{1,1}, a_{2,2}$ with $\min\{p_t(a_{1,1}), p_t(a_{2,2})\}$
   (b) $\ell_S(a_{i,j}) = 1/2$ to the other joint action in $a_{1,1}, a_{2,2}$
   (c) In the case where $p_t(a_{1,1}) = p_t(a_{2,2})$, assign $\ell_S(a_{1,1}) = 1/2$ and $\ell_S(a_{2,2}) = 0$

Let $p_{i,j}$ and $\ell_{i,j}$ be the probability and resulting loss, respectively, of player 1 picking arm i and player 2 picking arm j. The MMAB is then described as

	$a_{1,\cdot}$	$a_{2,\cdot}$
$a_{\cdot,1}$	x	1
$a_{\cdot,2}$	1	$0.5 - x$

where $x = 0.5 \times \mathbf{1}(p_{1,1} \geq p_{2,2})$ with $\mathbf{1}$ being the commonly defined indicator function.

To show that the Skull Adversary ensures linear regret in T, let us first consider a lower bound for the loss of the overall best joint arm (in hindsight). Note that arms $a_{1,2}$ and $a_{2,1}$ both induce a total loss of T. For the remaining arms, note that

$$\sum_{t=1}^{T} \ell_{1,1} + \sum_{t=1}^{T} \ell_{2,2} = \sum_{t=1}^{T} (\ell_{1,1} + \ell_{2,2}) = \sum_{t=1}^{T} \frac{1}{2} = T/2,$$

which implies that either $\sum_{t=1}^{T} \ell_{a_{1,1}}$ or $\sum_{t=1}^{T} \ell_{a_{2,2}}$ must be at most $T/4$.

Now we consider the optimized loss of any non-communicative strategy. WLOG, assume that $a_{1,1} = 0.5$. This implies $p_{1,1} \geq p_{2,2}$. Let $q = p_{1,\cdot}$ and $r = p_{\cdot,1}$. The loss in round t is then

$$\begin{aligned}
\langle \ell, p \rangle_t &= \frac{1}{2} \times p_{1,1} + 1 \times p_{1,2} + 1 \times p_{2,1} + 0 \times p_{2,2} \\
&= \frac{1}{2} qr + q(1-r) + r(1-q) \\
&= -\frac{3}{2} qr + q + r \\
&= \frac{4 - (3q-2)(3r-2)}{6}.
\end{aligned}$$

We cannot have both $q < 1/2$ and $r < 1/2$ since that would mean $qr < (1-q)(1-r) \iff p_{1,1} < p_{2,2}$.

Thus, WLOG assume $q \geq 1/2$. Since $3q - 2 \in [-1/2, 1]$ and $3r - 2 \in [-2, 1]$, we have $(3q-2)(3r-2) \leq 1$. Thus,

$$\frac{4 - (3q-2)(3r-2)}{6} \geq \frac{4-1}{6} = \frac{1}{2}$$

so the loss for each round is at least $\frac{1}{2}$. Hence, for any given simultaneous policy,

$$\mathrm{Reg}_T = \sum_{t=1}^{T} \langle \ell, p \rangle_t - \min_{\boldsymbol{a} \in [\boldsymbol{a}]} \sum_{t=1}^{T} \ell_{\boldsymbol{a}} \geq \frac{T}{2} - \frac{T}{4} = \frac{T}{4}, \tag{3}$$

which implies that the Skull Adversary must always induce linear regret.

□

It suffices to generalize the Skull Adversary to the general case, which will be done below.

Theorem 2 (Generalized Skull Adversary Theorem) *Let $K \geq 2$ be the number of players and $M \geq 2$ be the number of actions per player. For any set of independent stochastic strategies $\{\mathcal{P}_i \in \Delta(M)\}_{i=1}^{K}$, there exists an adaptive adversarial strategy that forces the players to incur $\Omega(T)$ loss.*

Proof: We proceed by constructing an adversarial strategy that extends from Theorem 1.

Define the loss function $\ell^* : [M]^K \to [0, 1]$ as:

$$\ell^*(a_1, a_2, \ldots, a_K) = \begin{cases} 1 & \text{if } (a_1, a_2, \ldots, a_K) \notin \{1,2\}^2 \times \{1\}^{K-2} \\ \ell_S(a_1, a_2) & \text{otherwise} \end{cases},$$

where ℓ_S is the Skull Adversary strategy established in Theorem 1. This forces player 1 and 2 to only pick arms 1 and 2, and player 3 onwards to pick arm 1, as any other set of actions would lead to loss 1. The problem of optimizing loss then simplifies to the two-player, two-armed problem shown in Theorem 1, and we're done.

□

4 Successive Pulls Full Information

As no policy can achieve sublinear regret in the multiplayer information-asymmetric setting as proposed in Chang et al. (2022), we now investigate an alternative formulation using successive pulls. In this model, players act in a predetermined sequence where each player observes all preceding players' actions before making their own decision. This approach effectively addresses the communication constraints that led to our impossibility result, and results for Problem A (no information asymmetry) and Problem B (reward asymmetry) are shown below.

4.1 Problem A: No Information Asymmetry

In this section, we propose `mHedge-A`, motivated by the single player hedge in Freund & Schapire (1997), where the i-th player will observe the previous $i-1$-players actions $(a_1, \ldots, a_{i-1})$, and apply hedge on all the joint actions such that the first $i-1$ components of that action match $(a_1, \ldots, a_{i-1})$. In other words they are applying a distribution over K^{M-i+1} actions. Note that despite sampling across the joint action space, they can only pull their corresponding action. With an application of Bayes' rule, we are able to show that the resulting probability of sampling a joint action $\boldsymbol{a}$ is actually equal to the probability that a single player applying hedge across the entire joint action space would pick $\boldsymbol{a}$. Thus we are able to apply their well known regret bound. The Psuedocode is given in Algorithm 1.

Algorithm 1 `mHedge-A`

Input: learning rate $\eta > 0$ ▷ a.k.a. step size / temperature
Initialization: For each player $i \in \{1, \dots, M\}$, set $L_0[i] \leftarrow \mathbf{0}$ (zero vector of length K^M).
for $t = 1$ **to** T **do**
  for $i = 1$ **to** M **do**
    Restrict $L_{t-1}[i]$ to the subvector corresponding to joint arms with prefix $\mathbf{a}[1{:}i] = (a_1, \dots, a_{i-1})$.
    $p_t^i(\mathbf{a}) \leftarrow \dfrac{\exp\big(-\eta\, L_{t-1}[i](\mathbf{a})\big)}{\sum_{\mathbf{b}} \exp\big(-\eta\, L_{t-1}[i](\mathbf{b})\big)}$ ▷ softmax over $\Delta(K^{M-i+1})$
  end for
  Play p_t and observe loss vector $\ell_t[i] \in [0,1]^{K^M}$ for each player i.
  for $i = 1$ **to** M **do**
    $L_t[i] \leftarrow L_{t-1}[i] + \ell_t[i]$
  end for
end for

We now present the regret bound for `mHedge-A`.

Theorem 3 *The regret for* `mHedge-A` *in Algorithm 1 satisfies*

$$R_T \leq 2\sqrt{TM\log(K)}. \tag{4}$$

Proof: Given a joint action $(a_1, \dots, a_M)$, consider the probability that it is pulled. By Bayes' rule, this is given by

$$\mathbb{P}(\boldsymbol{a} = (a_1, \dots, a_M)) = \prod_{i=1}^{M} P\left(\boldsymbol{a}[i] = a_i \,\middle|\, \bigwedge_{j=1}^{i-1} \boldsymbol{a}[j] = a_j\right).$$

(For clarity, summing over some condition like $\boldsymbol{a}[1] = a_i$ is to sum over all $\boldsymbol{a} \in [K]^M$ that satisfy the condition. List slicing of the form $[a : b]$ denotes the a through $(b-1)$st elements.)

Note that for player 1, the probability that they pull action a_1 is

$$\mathbb{P}(\boldsymbol{a}[1] = a_1) = \frac{\sum_{\boldsymbol{a}[1]=a_i} \exp(-\eta L_{t-1}(\boldsymbol{a}))}{\sum_{\boldsymbol{a}} \exp(-\eta L_{t-1}(\boldsymbol{a}))}.$$

By induction, we can show that for player i, the probability that they pull action a_i given the previous $i-1$ actions is

$$\mathbb{P}(\boldsymbol{a}[i] = a_i | \boldsymbol{a}[1] = a_1, \dots, \boldsymbol{a}[i-1] = a_{i-1}) = \frac{\sum_{\boldsymbol{a}[1:i+1]=(a_1,a_2,\dots,a_i)} \exp(-\eta L_{t-1}(\boldsymbol{a}))}{\sum_{\boldsymbol{a}[1:i]=(a_1,a_2,\dots,a_{i-1})} \exp(-\eta L_{t-1}(\boldsymbol{a}))}. \tag{5}$$

Note that in the numerator we are summing across actions where the first i components align with $(a_1, \dots, a_M)$, while in the denominator, we are summing across actions where only the first $i-1$ components align with $(a_1, \dots, a_M)$. Therefore, based on Bayes' rule above, we have

$$\begin{aligned}\mathbb{P}(\boldsymbol{a} = (a_1, \dots, a_M)) &= \prod_{i=1}^{M} \frac{\sum_{\boldsymbol{a}[1:i+1]=(a_1,a_2,\dots,a_i)} \exp(-\eta L_{t-1}(\boldsymbol{a}))}{\sum_{\boldsymbol{a}[1:i]=(a_1,a_2,\dots,a_{i-1})} \exp(-\eta L_{t-1}(\boldsymbol{a}))} \\ &= \frac{\exp(-\eta L_{t-1}((a_1, \dots, a_M)))}{\sum_{\boldsymbol{a}} \exp(-\eta L_{t-1}(\boldsymbol{a}))}\end{aligned}$$

where the last equality uses the fact that the product telescopes.

However, note that this would be the probability that $(a_1, \ldots, a_M)$ is sampled if there were only one player who played according to Hedge. Therefore we can apply the results from the regret analysis for the single-player Hedge algorithm in Luo (2022) to obtain the desired regret bound. In particular,

$$R_T \leq \frac{\log(K^M)}{\eta} + T\eta \leq 2\sqrt{TM\log(K)},$$

with $\eta = \sqrt{\frac{M \log K}{T}}$. □

4.2 Problem B: Information Asymmetry in Rewards

In this section, we propose `mHedge-B` which deals with the successive pulls in the case where the players receive an adversarial loss plus some stochastic subgaussian loss with 0 mean. Formally, at each round t, player i observes the modified loss vector $\ell_t + X_t^i$, where $X_t^i \sim \text{SubG}(1)$. Remarkably, we can apply the same principle used in `mHedge-A` in the previous section to obtain a nearly optimal regret bound for this setting as well. More specifically, the i-th player will observe the previous $(i-1)$-players actions $(a_1, \ldots, a_{i-1})$, and apply hedge on all the joint actions such that the first $(i-1)$ components of that action match $(a_1, \ldots, a_{i-1})$. However, when we apply Bayes rule, the stochastic losses prevent the probabilities of sampling a joint action from being equal to the standard hedge algorithm. However, we can upper bound these probabilities via using the standard concentration properties of sub-Gaussian distributions, characterized by the following well-known "Corollary 5.5" inequality.

Lemma 4 (Corollary 5.5 of Lattimore & Szepesvári (2020)) *Assume that $X_i - \mu$ are independent, σ-subgaussian random variables. Then for any $\varepsilon \geq 0$,*

$$\mathbb{P}(\hat{\mu} \geq \mu + \varepsilon) \leq \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right)$$

$$\textit{and} \quad \mathbb{P}(\hat{\mu} \leq \mu - \varepsilon) \leq \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right)$$

where $\hat{\mu} = \frac{1}{n}\sum_{t=1}^{n} X_t$.

Since the reward is 1-subgaussian, we can construct the following interval, which contains the empirical mean of the noise with high probability

$$G_{\boldsymbol{a}}(t) = \left\{ \widehat{X}_{\boldsymbol{a}}(t) \in \left(-\sqrt{\frac{4\log(TK^M)}{t-1}}, \sqrt{\frac{4\log(TK^M)}{t-1}} \right) \right\}$$

Let the "good event" be defined as

$$G = \bigcap_{t=1}^{T} \bigcap_{\boldsymbol{a}} G_{\boldsymbol{a}}(t). \tag{6}$$

the extra K^M factor will allow us to remove the exponential dependence on the action space in the regret bound. We have the following lemma, which upper bounds the probability of the complement of G.

Lemma 5 *We have*

$$\mathbb{P}(G^c) \leq \frac{1}{T}$$

Proof: Using DeMorgan's laws and the probability union bound we have

$$\begin{aligned}
\mathbb{P}(G^c) &= P\left(\bigcup_{t=1}^{T}\bigcup_{\boldsymbol{a}} G_{\boldsymbol{a}}(t)^c\right) \\
&\leq \sum_{t=1}^{T}\sum_{\boldsymbol{a}} \mathbb{P}(G_{\boldsymbol{a}}(t)^c) \\
&\leq \sum_{t=1}^{T}\sum_{\boldsymbol{a}} \exp\left(-\frac{t-1}{2}\left(\sqrt{\frac{4\log(K^M T)}{t-1}}\right)^2\right) \\
&\leq \sum_{t=1}^{T}\sum_{\boldsymbol{a}} \frac{1}{(K^M T)^2} \\
&\leq \frac{1}{T}
\end{aligned}$$

where in the second inequality, we use Lemma 4 with $n = t-1$, since at round t, we have $t-1$ samples of the stochastic noise. □

Players take turns pulling successively and condition their pulls on the previous players' pulls.

Algorithm 2 `mHedge-B`

Input: learning rate $\eta > 0$
Initialization: For each player $i \in \{1, \ldots, M\}$, set $L_0[i] \leftarrow \mathbf{0}$ (a zero vector of length K^M).
for $t = 1$ **to** T **do**
  for $i = 1$ **to** M **do**
    Restrict $L_{t-1}[i]$ to the subvector for joint arms with prefix $\mathbf{a}[1{:}i] = (a_1, \ldots, a_{i-1})$.
    $p_t^i(\mathbf{a}) \leftarrow \dfrac{\exp\big(-\eta\, L_{t-1}[i](\mathbf{a})\big)}{\sum_{\mathbf{b}} \exp\big(-\eta\, L_{t-1}[i](\mathbf{b})\big)}$ ▷ softmax over $\Delta(K^{M-i+1})$
  end for
  Play p_t and observe $\ell_t + X_t[i] \in [0,1]^{K^M}$ with $X_t[i](\mathbf{a})$ 1-subgaussian.
  for $i = 1$ **to** M **do**
    $L_t[i] \leftarrow L_{t-1}[i] + \ell_t + X_t[i]$
  end for
end for

Theorem 6 *`mHedgeB` gives a regret bound of*

$$R_T \leq 4M^2 \ln K \sqrt{2T \log T} + 1. \tag{7}$$

Comparing this to the regret bound in Theorem 3. It is the same order in T, but contains an extra factor of M as a result of the union bound over the players.

Proof: Note that we can decompose the regret as

$$R_T = \sum_{t=1}^{T} \mathbb{E}[\ell_t(a_t)] - \min_{a\in[K]} \sum_{t=1}^{T} \ell_t(a) = \sum_{t=1}^{T} \mathbb{E}[\ell_t(a_t) - \ell_t(a^*)]$$

where a^* is the best arm in hindsight and the expectation is over the policy as well as the stochastic noise. Using the total law of expectation we obtain

$$
\begin{aligned}
R_T &= \sum_{t=1}^{T} \mathbb{E}[\ell_t(a_t) - \ell_t(a^*)|G]\mathbb{P}(G) + \sum_{t=1}^{T} \mathbb{E}[\ell_t(a_t) - \ell_t(a^*)|G^c]\mathbb{P}(G^c) \\
&\leq \sum_{t=1}^{T} \mathbb{E}[\ell_t(a_t) - \ell_t(a^*)|G] + T\mathbb{P}(G^c) \\
&\leq \sum_{t=1}^{T} \mathbb{E}[\ell_t(a_t) - \ell_t(a^*)|G] + 1.
\end{aligned}
$$

In the first inequality, we use the fact that the adversarial part of each loss is in $[0, 1]$, and that the noise has averages out to 0. Therefore, the regret (under the condition of the complement of the good event) is upper bounded by T. The last inequality uses Lemma 5.

We now bound the regret under good event in equation (6). Given a joint action $(a_1, \ldots, a_M)$, we calculate the probability that it is pulled. By Bayes' rule, this is given by

$$
\mathbb{P}(\boldsymbol{a} = (a_1, \ldots, a_M)) = \prod_{i=1}^{M} P\left(\boldsymbol{a}[i] = a_i \,\middle|\, \bigwedge_{j=1}^{i-1} \boldsymbol{a}[j] = a_j\right).
$$

Note that for player 1, the probability that they pull action a_1 is

$$
\mathbb{P}(\boldsymbol{a}[1] = a_1) = \frac{\sum\limits_{\boldsymbol{a}[1]=a_1} \exp\big(-\eta L_{t-1}^1(\boldsymbol{a})\big)}{\sum_{\boldsymbol{a}} \exp\big(-\eta L_{t-1}^1(\boldsymbol{a})\big)} \frac{\sum\limits_{\boldsymbol{a}[1]=a_1} \exp(-\eta L_{t-1}(\boldsymbol{a})) \exp\Big(-\eta(t-1)\widehat{X}_{\boldsymbol{a}}^1(t-1)\Big)}{\sum_{\boldsymbol{a}} \exp(-\eta L_{t-1}(\boldsymbol{a})) \exp\Big(-\eta(t-1)\widehat{X}_{\boldsymbol{a}}^1(t-1)\Big)},
$$

where we have split the losses in the adversarial part as well as the stochastic part. Note the extra factor of $t-1$ since we use the sum of the noises. For player i, the probability that they pull action a_i given the previous $i-1$ actions is

$$
\begin{aligned}
&\mathbb{P}(\boldsymbol{a}[i] = a_i | \boldsymbol{a}[1] = a_1, \ldots, \boldsymbol{a}[i-1] = a_{i-1}) \\
&= \frac{\sum\limits_{\boldsymbol{a}[1:i+1]=(a_1,\ldots,a_i)} \exp\big(-\eta L_{t-1}^i(\boldsymbol{a})\big)}{\sum\limits_{\boldsymbol{a}[1:i]=(a_1,\ldots,a_{i-1})} \exp\big(-\eta L_{t-1}^i(\boldsymbol{a})\big)} \\
&= \frac{\sum\limits_{\boldsymbol{a}[1:i+1]=(a_1,\ldots,a_i)} \exp(-\eta L_{t-1}(\boldsymbol{a}))}{\sum\limits_{\boldsymbol{a}[1:i]=(a_1,\ldots,a_{i-1})} \exp(-\eta L_{t-1}(\boldsymbol{a}))} \cdot \frac{\exp\Big(-\eta(t-1)\widehat{X}_{\boldsymbol{a}}^i(t-1)\Big)}{\exp\Big(-\eta(t-1)\widehat{X}_{\boldsymbol{a}}^i(t-1)\Big)}
\end{aligned}
$$

Note that in the numerator we are summing across actions where the first i components align with $(a_1, \ldots, a_M)$, while in the denominator, we are summing across actions where only the first $i-1$ components align with $(a_1, \ldots, a_M)$. Therefore, based on Bayes' rule above, we have

$$
\mathbb{P}(\boldsymbol{a} = (a_1, \ldots, a_M)) = \prod_{i=1}^{M} \frac{\sum\limits_{\boldsymbol{a}[1:i+1]=(a_1,\ldots,a_i)} \exp(-\eta L_{t-1}(\boldsymbol{a}))}{\sum\limits_{\boldsymbol{a}[1:i]=(a_1,\ldots,a_{i-1})} \exp(-\eta L_{t-1}(\boldsymbol{a}))} \cdot \frac{\exp\Big(-\eta(t-1)\widehat{X}_{\boldsymbol{a}}^i(t-1)\Big)}{\exp(-\eta L_{t-1}(\boldsymbol{a})) \exp\Big(-\eta(t-1)\widehat{X}_{\boldsymbol{a}}^i(t-1)\Big)}
$$

Note that even for actions that were not taken, we obtain a loss for that round. Therefore the "number of pulls" of each joint arm up to and not including time t is $t-1$. Under the good event (6), we have

$\widehat{X_a} \in (-\sqrt{2\log(TK^M)/(t-1)}, \sqrt{2\log(TK^M)/(t-1)})$. Therefore,

$$\mathbb{P}(\boldsymbol{a} = (a_1, \ldots, a_M)) \leq \prod_{i=1}^{M} \frac{\sum_{\boldsymbol{a}[1:i+1]=(a_1,\ldots,a_i)} \exp(-\eta L_{t-1}(\boldsymbol{a}))}{\sum_{\boldsymbol{a}[1:i]=(a_1,\ldots,a_{i-1})} \exp(-\eta L_{t-1}(\boldsymbol{a}))} \cdot \exp\left(2\eta\sqrt{2(t-1)\log(TK^M)}\right)$$
$$\leq \exp\left(2M\eta\sqrt{2(t-1)\log(TK^M)}\right) \cdot \frac{\exp(-\eta L_{t-1}((a_1, \ldots, a_M)))}{\sum_{\boldsymbol{a}} \exp(-\eta L_{t-1}(\boldsymbol{a}))}$$

where in the equality we used the fact that the adversarial part of the loss telescopes in the product. Note that for $x \in [0, 1]$, we have $e^x \leq 1 + 2x$. Therefore, the above can be bounded by

$$\mathbb{P}(\boldsymbol{a} = (a_1, \ldots, a_M)) \leq \left(1 + 4M\eta\sqrt{2(t-1)\log(TK^M)}\right) \cdot \frac{\exp(-\eta L_{t-1}((a_1, \ldots, a_M)))}{\sum_{\boldsymbol{a}} \cdot \exp(-\eta L_{t-1}(\boldsymbol{a}))}$$
$$\leq \left(1 + 4M\eta\sqrt{2T\log(TK^M)}\right) \cdot \frac{\exp(-\eta L_{t-1}((a_1, \ldots, a_M)))}{\sum_{\boldsymbol{a}} \exp(-\eta L_{t-1}(\boldsymbol{a}))}.$$

Note that the term in blue is the probability that $(a_1, \ldots, a_M)$ is sampled if there were only one player that played according to Hedge. Therefore, we can apply the regret analysis for the single-player Hedge algorithm in Luo (2022) and above, to deduce that

$$\sum_{t=1}^{T} \mathbb{E}[\ell_t(a_t) - \ell_t(a^*)|G] \leq \left(1 + 4M\eta\sqrt{2T\log(TK^M)}\right)\left(\frac{\log K^M}{\eta} + T\eta\right)$$

for the good event. Setting $\eta = \sqrt{\frac{K\log M}{T}}$ and adding back in the bad event term, we obtain the regret bound

$$R_T \leq 8M^2 \log K\sqrt{2TK\log T} + 1.$$

□

5 Conclusions

In this paper, we have extended the multiplayer bandit framework of Chang et al. (2022) to accommodate adversarial environments. Our analysis reveals a fundamental result: with simultaneous arm pulls, no strategy can achieve sublinear regret against an adaptive adversary, as demonstrated by our Skull Adversary construction. This impossibility result highlights the critical role of coordination in adversarial multiplayer settings.

To overcome this limitation, we proposed a sequential decision-making approach where players act in a predetermined order, with each player observing the actions of preceding players. For the setting with identical losses across players (Problem A), we developed `mHedge-A`, which achieves a regret bound of $\tilde{O}(\sqrt{TM})$. This result has significant practical implications for distributed systems where communication incurs substantial costs in time or resources, as it demonstrates that one-way communication along a predetermined ordering suffices for sublinear regret.

We further addressed environments with reward asymmetry (Problem B) through our algorithm `mHedge-B`, achieving a regret bound of $\tilde{O}(M^2\sqrt{T})$. This bound quantifies the performance degradation introduced by observation noise while maintaining asymptotic sublinearity in T. Together, these results provide a comprehensive understanding of the learnability of multiplayer adversarial bandits under various information structures, as well as establish the theoretical foundations for designing effective learning algorithms in these extremely challenging environments.

References

Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pp. 1638–1646, 2014.

Shipra Agrawal and Nikhil R Devanur. Efficient algorithms for the dynamic knapsack problem. In *Algorithmica*, volume 74, pp. 19–46, 2016.

Pragnya Alatur, Kfir Y. Levy, and Andreas Krause. Multi-player bandits: The adversarial case. *Journal of Machine Learning Research*, 21(77):1–23, 2020. URL `http://jmlr.org/papers/v21/19-912.html`.

Jean-Yves Audibert and Sébastien Bubeck. Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, 11:2785–2836, 2010.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. In *Advances in Neural Information Processing Systems*, pp. 107–114, 2002.

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

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE 36th annual foundations of computer science*, pp. 322–331. IEEE, 1995.

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

Baruch Awerbuch and Robert Kleinberg. Competitive collaborative learning. In *Conference on Learning Theory*, pp. 233–244, 2008.

Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. In *54th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pp. 207–216. IEEE, 2013.

Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 19–26. JMLR Workshop and Conference Proceedings, 2011.

Rémi Bonnefoi, Lilian Besson, Christophe Moy, Emilie Kaufmann, and Jacques Palicot. Multi-armed bandit learning in iot networks: Learning helps even in non-stationary settings. In *International Conference on Cognitive Radio Oriented Wireless Networks*, pp. 173–185. Springer, 2017.

Sébastien Bubeck and Aleksandrs Slivkins. The best of both worlds: stochastic and adversarial bandits. In *Conference on Learning Theory*, pp. 42.1–42.23. JMLR Workshop and Conference Proceedings, 2012.

Nicolò Cesa-Bianchi and Gábor Lugosi. Combinatorial bandits. In *Journal of Computer and System Sciences*, volume 78, pp. 1404–1422, 2012.

Nicolò Cesa-Bianchi, Tommaso Cesari, and Claire Monteleoni. Cooperative online learning: Keeping your neighbors updated. In *Algorithmic learning theory*, pp. 234–250. PMLR, 2020.

Nicol'o Cesa-Bianchi, Claudio Gentile, Yishay Mansour, and Alberto Minora. Delay and cooperation in nonstochastic bandits. In *Conference on Learning Theory*, pp. 605–622. PMLR, 2016.

William Chang and Aditi Karthik. Multiplayer information asymmetric bandits in metric spaces. *arXiv preprint arXiv:2503.08004*, 2025.

William Chang and Yuanhao Lu. Optimal cooperative multiplayer learning bandits with noisy rewards and no communication. In *2024 IEEE 63rd Conference on Decision and Control (CDC)*, pp. 823–830. IEEE, 2024.

William Chang and Yuanhao Lu. Multiplayer information asymmetric contextual bandits. *arXiv preprint arXiv:2503.08961*, 2025.

William Chang, Mehdi Jafarnia-Jahromi, and Rahul Jain. Online learning for cooperative multi-player multi-armed bandits. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pp. 7248–7253. IEEE, 2022.

Abhimanyu Dubey et al. Cooperative multi-agent bandits with heavy tails. In *International conference on machine learning*, pp. 2730–2739. PMLR, 2020.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

Mohammad Hajiesmaili, Mohammad Sadegh Talebi, John Lui, Wing Shing Wong, et al. Adversarial bandits with corruptions: Regret lower bound and no-regret algorithm. *Advances in Neural Information Processing Systems*, 33:19943–19952, 2020.

Nicole Immorlica, Karthik Sankararaman, Robert Schapire, and Aleksandrs Slivkins. Adversarial bandits with knapsacks. *Journal of the ACM*, 69(6):1–47, 2022.

Tianyuan Jin, Hao-Lun Hsu, William Chang, and Pan Xu. Finite-time frequentist regret bounds of multi-agent thompson sampling on sparse hypergraphs, 2023. URL `https://arxiv.org/abs/2312.15549`.

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

Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvari. Tight regret bounds for stochastic combinatorial semi-bandits. In *Artificial Intelligence and Statistics*, pp. 535–543, 2015.

Peter Landgren, Vaibhav Srivastava, and Naomi Ehrich Leonard. On distributed cooperative decision-making in multiarmed bandits. In *2016 European Control Conference (ECC)*, pp. 243–248. IEEE, 2016.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Haipeng Luo. Lecture 1: Introduction to bandit algorithms. Course materials for CSCI 659: Advanced Topics in Machine Learning and Artificial Intelligence, 2022. URL `https://haipeng-luo.net/courses/CSCI659/2022_fall/lectures/lecture1.pdf`. University of Southern California.

Yuzhe Ma and Zhijin Zhou. Adversarial attacks on adversarial bandits. *arXiv preprint arXiv:2301.12595*, 2023.

Shivakumar Mahesh, Anshuka Rangi, Haifeng Xu, and Long Tran-Thanh. Attacking multi-player bandits and how to robustify them. In *Proceedings of 23rd Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024)*. ACM; International Foundation for Autonomous Agents and Multiagent Systems . . . , 2024.

David Martínez-Rubio, Varun Kanade, and Patrick Rebeschini. Decentralized cooperative stochastic bandits. *Advances in Neural Information Processing Systems*, 32, 2019.

Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *Advances in Neural Information Processing Systems*, pp. 3168–3176, 2015.

Francesco Orabona and Dávid Pál. Optimal non-asymptotic lower bound on the minimax regret of learning with expert advice. *arXiv:1511.02176*, 2016.

Anshuka Rangi, Massimo Franceschetti, and Long Tran-Thanh. Unifying the stochastic and the adversarial bandits with knapsack. *arXiv preprint arXiv:1811.12253*, 2018.

Aadirupa Saha, Tomer Koren, and Yishay Mansour. Adversarial dueling bandits. In *International Conference on Machine Learning*, 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*, pp. 1287–1295. PMLR, 2014.

Shahin Shahrampour, Alexander Rakhlin, and Ali Jadbabaie. Multi-armed bandits in multi-agent networks. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2786–2790. IEEE, 2017.

Chengshuai Shi and Cong Shen. An attackability perspective on no-sensing adversarial multi-player multi-armed bandits. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pp. 533–538. IEEE, 2021.

Vasilis Syrgkanis, Akshay Krishnamurthy, and Robert Schapire. Efficient bandit algorithms for online multiclass prediction. In *International Conference on Machine Learning*, pp. 273–281, 2016.

Balazs Szorenyi, Róbert Busa-Fekete, István Hegedus, Róbert Ormándi, Márk Jelasity, and Balázs Kégl. Gossip-based distributed stochastic bandit algorithms. In *International conference on machine learning*, pp. 19–27. PMLR, 2013.

Jingwen Tong, Liqun Fu, and Zhu Han. Throughput enhancement of full-duplex csma networks using multiplayer bandits. *IEEE Internet of Things Journal*, 8(15):11807–11821, 2021.

Taishi Uchiya, Atsuyoshi Nakamura, and Mineichi Kudo. Algorithms for adversarial bandit problems with multiple plays. In *International Conference on Algorithmic Learning Theory*, pp. 375–389. Springer, 2010.

Po-An Wang, Alexandre Proutiere, Kaito Ariu, Yassir Jedra, and Alessio Russo. Optimal algorithms for multiplayer multi-armed bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 4120–4129. PMLR, 2020.

Mengfan Xu and Diego Klabjan. Decentralized randomly distributed multi-agent multi-armed bandit with heterogeneous rewards. *Advances in Neural Information Processing Systems*, 36, 2024.

Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. K-armed bandit problems with commodity structure. In *Journal of Machine Learning Research*, volume 13, pp. 2643–2675, 2012.

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

Masrour Zoghi, Shimon Whiteson, Rémi Munos, and Maarten Rijke. Relative upper confidence bound for the k-armed dueling bandit problem. In *International Conference on Machine Learning*, pp. 10–18, 2014.