

# Factored Bandits

**Julian Zimmert**  
University of Copenhagen  
zimmert@di.ku.dk

**Yevgeny Seldin**  
University of Copenhagen  
seldin@di.ku.dk

## Abstract

We introduce the factored bandits model, which is a framework for learning with limited (bandit) feedback, where actions can be decomposed into a Cartesian product of atomic actions. Factored bandits incorporate rank-1 bandits as a special case, but significantly relax the assumptions on the form of the reward function. We provide an anytime algorithm for stochastic factored bandits and up to constants matching upper and lower regret bounds for the problem. Furthermore, we show how a slight modification enables the proposed algorithm to be applied to utility-based dueling bandits. We obtain an improvement in the additive terms of the regret bound compared to state-of-the-art algorithms (the additive terms are dominating up to time horizons that are exponential in the number of arms).

## 1 Introduction

We introduce *factored bandits*, which is a bandit learning model, where actions can be decomposed into a Cartesian product of atomic actions. As an example, consider an advertising task, where the actions can be decomposed into (1) selection of an advertisement from a pool of advertisements and (2) selection of a location on a web page out of a set of locations, where it can be presented. The probability of a click is then a function of the quality of the two actions, the attractiveness of the advertisement and the visibility of the location it was placed at. In order to maximize the reward the learner has to maximize the quality of actions along each dimension of the problem. Factored bandits generalize the above example to an arbitrary number of atomic actions and arbitrary reward functions satisfying some mild assumptions.

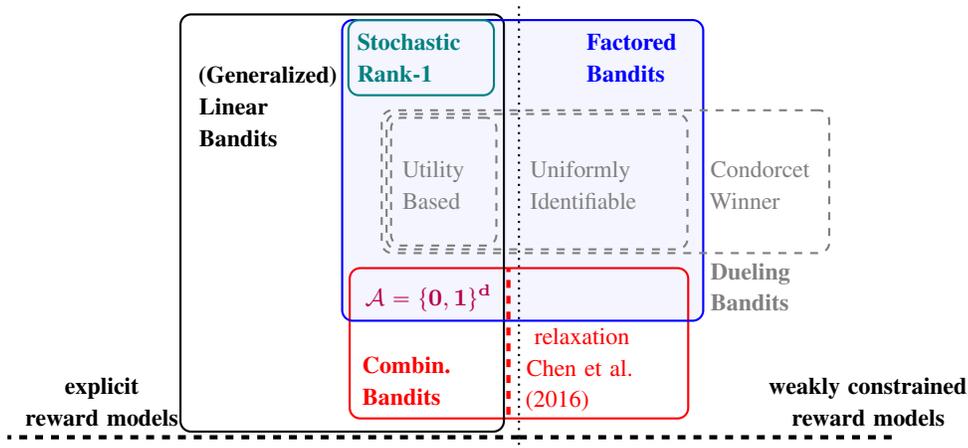


Figure 1: Relations between factored bandits and other bandit models.

In a nutshell, at every round of a factored bandit game the player selects  $L$  atomic actions,  $a_1, \dots, a_L$ , each from a corresponding finite set  $\mathcal{A}_\ell$  of size  $|\mathcal{A}_\ell|$  of possible actions. The player then observes a reward, which is an arbitrary function of  $a_1, \dots, a_L$  satisfying some mild assumptions. For example, it can be a sum of the quality of atomic actions, a product of the qualities, or something else that does not necessarily need to have an analytical expression. The learner does not have to know the form of the reward function.

Our way of dealing with combinatorial complexity of the problem is through introduction of *unique identifiability* assumption, by which the best action along each dimension is uniquely identifiable. A bit more precisely, when looking at a given dimension we call the collection of actions along all other dimensions a *reference set*. The unique identifiability assumption states that in expectation the best action along a dimension outperforms any other action along the same dimension by a certain margin when both are played with the same reference set, irrespective of the composition of the reference set. This assumption is satisfied, for example, by the reward structure in linear and generalized linear bandits, but it is much weaker than the linearity assumption.

In Figure 1, we sketch the relations between factored bandits and other bandit models. We distinguish between bandits with explicit reward models, such as linear and generalized linear bandits, and bandits with weakly constrained reward models, including factored bandits and some relaxations of combinatorial bandits. A special case of factored bandits are rank-1 bandits [7]. In rank-1 bandits the player selects two actions and the reward is the product of their qualities. Factored bandits generalize this to an arbitrary number of actions and significantly relax the assumption on the form of the reward function.

The relation with other bandit models is a bit more involved. There is an overlap between factored bandits and (generalized) linear bandits [1; 6], but neither is a special case of the other. When actions are represented by unit vectors, then for (generalized) linear reward functions the models coincide. However, the (generalized) linear bandits allow a continuum of actions, whereas factored bandits relax the (generalized) linearity assumption on the reward structure to uniform identifiability.

There is a partial overlap between factored bandits and combinatorial bandits [3]. The action set in combinatorial bandits is a subset of  $\{0, 1\}^d$ . If the action set is unrestricted, i.e.  $\mathcal{A} = \{0, 1\}^d$ , then combinatorial bandits can be seen as factored bandits with just two actions along each of the  $d$  dimensions. However, typically in combinatorial bandits the action set is a strict subset of  $\{0, 1\}^d$  and one of the parameters of interest is the permitted number of non-zero elements. This setting is not covered by factored bandits. While in the classical combinatorial bandits setting the reward structure is linear, there exist relaxations of the model, e.g. Chen et al. [4].

Dueling bandits are not directly related to factored bandits and, therefore, we depict them with faded dashed blocks in Figure 1. While the action set in dueling bandits can be decomposed into a product of the basic action set with itself (one for the first and one for the second action in the duel), the observations in dueling bandits are the identities of the winners rather than rewards. Nevertheless, we show that the proposed algorithm for factored bandits can be applied to utility-based dueling bandits.

The main contributions of the paper can be summarized as follows:

1. We introduce factored bandits and the uniform identifiability assumption.
2. Factored bandits with uniformly identifiable actions are a generalization of rank-1 bandits.
3. We provide an anytime algorithm for playing factored bandits under uniform identifiability assumption in stochastic environments and analyze its regret. We also provide a lower bound matching up to constants.
4. Unlike the majority of bandit models, our approach does not require explicit specification or knowledge of the form of the reward function (as long as the uniform identifiability assumption is satisfied). For example, it can be a weighted sum of the qualities of atomic actions (as in linear bandits), a product thereof, or any other function not necessarily known to the algorithm.
5. We show that the algorithm can also be applied to utility-based dueling bandits, where the additive factor in the regret bound is reduced by a multiplicative factor of  $K$  compared to state-of-the-art (where  $K$  is the number of actions). It should be emphasized that in state-of-the-art regret bounds for utility-based dueling bandits the additive factor is dominating

for time horizons below  $\Omega(\exp(K))$ , whereas in the new result it is only dominant for time horizons up to  $\mathcal{O}(K)$ .

6. Our work provides a unified treatment of two distinct bandit models: rank-1 bandits and utility-based dueling bandits.

The paper is organized in the following way. In Section 2 we introduce the factored bandit model and uniform identifiability assumption. In Section 3 we provide algorithms for factored bandits and dueling bandits. In Section 4 we analyze the regret of our algorithm and provide matching upper and lower regret bounds. In Section 5 we compare our work empirically and theoretically with prior work. We finish with a discussion in Section 6.

## 2 Problem Setting

### 2.1 Factored bandits

We define the game in the following way. We assume that the set of actions  $\mathcal{A}$  can be represented as a Cartesian product of atomic actions,  $\mathcal{A} = \bigotimes_{\ell=1}^L \mathcal{A}^\ell$ . We call the elements of  $\mathcal{A}^\ell$  *atomic arms*. For rounds  $t = 1, 2, \dots$  the player chooses an action  $\mathbf{A}_t \in \mathcal{A}$  and observes a reward  $r_t$  drawn according to an unknown probability distribution  $p_{\mathbf{A}_t}$  (i.e., the game is “stochastic”). We assume that the mean rewards  $\mu(\mathbf{a}) = \mathbb{E}[r_t | \mathbf{A}_t = \mathbf{a}]$  are bounded in  $[-1, 1]$  and that the noise  $\eta_t = r_t - \mu(\mathbf{A}_t)$  is conditionally 1-sub-Gaussian. Formally, this means that

$$\forall \lambda \in \mathbb{R} \quad \mathbb{E} \left[ e^{\lambda \eta_t} | \mathcal{F}_{t-1} \right] \leq \exp \left( \frac{\lambda^2}{2} \right),$$

where  $\mathcal{F}_t := \{\mathbf{A}_1, r_1, \mathbf{A}_2, r_2, \dots, \mathbf{A}_t, r_t\}$  is the filtration defined by the history of the game up to and including round  $t$ . We denote  $\mathbf{a}^* = (a_1^*, a_2^*, \dots, a_L^*) = \operatorname{argmax}_{\mathbf{a} \in \mathcal{A}} \mu(\mathbf{a})$ .

**Definition 1** (uniform identifiability). *An atomic set  $\mathcal{A}^k$  has a uniformly identifiable best arm  $a_k^*$  if and only if*

$$\forall a \in \mathcal{A}^k \setminus \{a_k^*\} : \Delta_k(a) := \min_{\mathbf{b} \in \bigotimes_{\ell \neq k} \mathcal{A}^\ell} \mu(a_k^*, \mathbf{b}) - \mu(a, \mathbf{b}) > 0. \quad (1)$$

We assume that all atomic sets have uniformly identifiable best arms. The goal is to minimize the pseudo-regret, which is defined as

$$\operatorname{Reg}_T = \mathbb{E} \left[ \sum_{t=1}^T \mu(\mathbf{a}^*) - \mu(\mathbf{A}_t) \right].$$

Due to generality of the uniform identifiability assumption we cannot upper bound the instantaneous regret  $\mu(\mathbf{a}^*) - \mu(\mathbf{A}_t)$  in terms of the gaps  $\Delta_\ell(a_\ell)$ . However, a sequential application of (1) provides a lower bound

$$\begin{aligned} \mu(\mathbf{a}^*) - \mu(\mathbf{a}) &= \mu(\mathbf{a}^*) - \mu(a_1, a_2^*, \dots, a_L^*) + \mu(a_1, a_2^*, \dots, a_L^*) - \mu(\mathbf{a}) \\ &\geq \Delta_1(a_1) + \mu(a_1, a_2^*, \dots, a_L^*) - \mu(\mathbf{a}) \geq \dots \geq \sum_{\ell=1}^L \Delta_\ell(a_\ell). \end{aligned} \quad (2)$$

For the upper bound let  $\kappa$  be a problem dependent constant, such that  $\mu(\mathbf{a}^*) - \mu(\mathbf{a}) \leq \kappa \sum_{\ell=1}^L \Delta_\ell(a_\ell)$  holds for all  $\mathbf{a}$ . Since the mean rewards are in  $[-1, 1]$ , the condition is always satisfied by  $\kappa = \min_{\mathbf{a}, \ell} 2\Delta_\ell^{-1}(a_\ell)$  and by equation (2)  $\kappa$  is always larger than 1. The constant  $\kappa$  appears in the regret bounds. In the extreme case when  $\kappa = \min_{\mathbf{a}, \ell} 2\Delta_\ell^{-1}(a_\ell)$  the regret guarantees are fairly weak. However, in many specific cases mentioned in the previous section,  $\kappa$  is typically small or even 1. We emphasize that algorithms proposed in the paper do not require the knowledge of  $\kappa$ . Thus, the dependence of the regret bounds on  $\kappa$  is not a limitation and the algorithms automatically adapt to more favorable environments.

## 2.2 Dueling bandits

The set of actions in dueling bandits is factored into  $\mathcal{A} \times \mathcal{A}$ . However, strictly speaking the problem is not a factored bandit problem, because the observations in dueling bandits are not the rewards.<sup>1</sup> When playing two arms,  $a$  and  $b$ , we observe the identity of the winning arm, but the regret is typically defined via average relative quality of  $a$  and  $b$  with respect to a “best” arm in  $\mathcal{A}$ .

The literature distinguishes between different dueling bandit settings. We focus on *utility-based dueling bandits* [14] and show that they satisfy the uniform identifiability assumption.

In utility-based dueling bandits, it is assumed that each arm has a utility  $u(a)$  and that the winning probabilities are defined by  $\mathbb{P}[a \text{ wins against } b] = \nu(u(a) - u(b))$  for a monotonously increasing link function  $\nu$ . Let  $w(a, b)$  be 1 if  $a$  wins against  $b$  and 0 if  $b$  wins against  $a$ . Let  $a^* := \operatorname{argmax}_{a \in \mathcal{A}} u(a)$  denote the best arm. Then for any arm  $b \in \mathcal{A}$  and any  $a \in \mathcal{A} \setminus a^*$ , it holds that  $\mathbb{E}[w(a^*, b)] - \mathbb{E}[w(a, b)] = \nu(u(a^*) - u(b)) - \nu(u(a) - u(b)) > 0$ , which satisfies the uniform identifiability assumption. For the rest of the paper we consider the linear link function  $\nu(x) = \frac{1+x}{2}$ . The regret is then defined by

$$\operatorname{Reg}_T = \mathbb{E} \left[ \sum_{t=1}^T \frac{u(a^*) - u(A_t)}{2} + \frac{u(a^*) - u(B_t)}{2} \right]. \quad (3)$$

## 3 Algorithms

Although in theory an asymptotically optimal algorithm for any structured bandit problem was presented in [5], for factored bandits this algorithm does not only require solving an intractable semi-infinite linear program at every round, but it also suffers from additive constants which are exponential in the number of atomic actions  $L$ . An alternative naive approach could be an adaptation of sparring [16], where each factor runs an independent  $K$ -armed bandit algorithm and does not observe the atomic arm choices of other factors. The downside of sparring algorithms, both theoretically and practically, is that each algorithm operates under limited information and the rewards become non i.i.d. from the perspective of each individual factor.

Our Temporary Elimination Algorithm (TEA, Algorithm 1) avoids these downsides. It runs independent instances of the Temporary Elimination Module (TEM, Algorithm 3) in parallel, one per each factor of the problem. Each TEM operates on a single atomic set. The TEA is responsible for the synchronization of TEM instances. Two main ingredients ensure information efficiency. First, we use relative comparisons between arms instead of comparing absolute mean rewards. This cancels out the effect of non-stationary means. The second idea is to use local randomization in order to obtain unbiased estimates of the relative performance without having to actually play each atomic arm with the same reference, which would have led to prohibitive time complexity.

```

1  $\forall \ell : \text{TEM}^\ell \leftarrow \text{new TEM}(\mathcal{A}^\ell)$ 
2  $t \leftarrow 1$ 
3 for  $s = 1, 2, \dots$  do
4    $M_s \leftarrow$ 
   |  $\operatorname{argmax}_\ell |\text{TEM}^\ell . \text{getActiveSet}(f(t)^{-1})|$ 
5    $T_s \leftarrow (t, t + 1, \dots, t + M_s - 1)$ 
6   for  $\ell \in \{1, \dots, L\}$  in parallel do
7     |  $\text{TEM}^\ell . \text{scheduleNext}(T_s)$ 
8   for  $t \in T_s$  do
9     |  $r_t \leftarrow \text{play}((\text{TEM}^\ell . A_t)_{\ell=1, \dots, L})$ 
10  for  $\ell \in \{1, \dots, L\}$  in parallel do
11    |  $\text{TEM}^\ell . \text{feedback}((r_{t'})_{t' \in T_s})$ 
12   $t \leftarrow t + |T_s|$ 

```

**Algorithm 1:** Factored Bandit TEA

```

1 TEM  $\leftarrow$  new TEM( $\mathcal{A}$ )
2  $t \leftarrow 1$ 
3 for  $s = 1, 2, \dots$  do
4    $\mathcal{A}_s \leftarrow \text{TEM} . \text{getActiveSet}(f(t)^{-1})$ 
5    $T_s \leftarrow (t, t + 1, \dots, t + |\mathcal{A}_s| - 1)$ 
6   TEM . scheduleNext( $T_s$ )
7   for  $b \in \mathcal{A}_s$  do
8     |  $r_t \leftarrow \text{play}(\text{TEM} . A_t, b)$ 
9     |  $t \leftarrow t + 1$ 
10  TEM . feedback( $((r_{t'})_{t' \in T_s})$ )

```

**Algorithm 2:** Dueling Bandit TEA

<sup>1</sup>In principle, it is possible to formulate a more general problem that would incorporate both factored bandits and dueling bandits. But such a definition becomes too general and hard to work with. For the sake of clarity we have avoided this path.

The TEM instances run in parallel in externally synchronized phases. Each module selects active arms in  $getActiveSet(\delta)$ , such that the optimal arm is included with high probability. The length of a phase is chosen such that each module can play each potentially optimal arm at least once in every phase. All modules schedule all arms for the phase in  $scheduleNext$ . This is done by choosing arms in a round robin fashion (random choices if not all arms can be played equally often) and ordering them randomly. All scheduled plays are executed and the modules update their statistics through the call of  $feedback$  routine. The modules use slowly increasing lower confidence bounds for the gaps in order to temporarily eliminate arms that are with high probability suboptimal. In all algorithms, we use  $f(t) := (t + 1) \log^2(t + 1)$ .

**Dueling bandits** For dueling bandits we only use a single instance of TEM. In each phase the algorithm generates two random permutations of the active set and plays the corresponding actions from the two lists against each other. (The first permutation is generated in Line 6 and the second in Line 7 of Algorithm 2.)

### 3.1 TEM

The TEM tracks empirical differences between rewards of all arms  $a_i$  and  $a_j$  in  $D_{i,j}$ . Based on these differences, it computes lower confidence bounds for all gaps. The set  $\mathcal{K}^*$  contains those arms where all LCB gaps are zero. Additionally the algorithm keeps track of arms that were never removed from  $\mathcal{B}$ . During a phase, each arm from  $\mathcal{K}^*$  is played at least once, but only arms in  $\mathcal{B}$  can be played more than once. This is necessary to keep the additive constants at  $M \log(K)$  instead of  $MK$ .

```

19 Function scheduleNext ( $\mathcal{T}$ )
20   for  $a \in \mathcal{K}^*$  do
21      $\tilde{t} \leftarrow$  random unassigned index in  $\mathcal{T}$ 
22      $A_{\tilde{t}} \leftarrow a$ 
23   while not all  $A_{t_s}, \dots, A_{t_s+|\mathcal{T}|-1}$  assigned
24     do
25       for  $a \in \mathcal{B}$  do
26          $\tilde{t} \leftarrow$  random unassigned index in  $\mathcal{T}$ 
27          $A_{\tilde{t}} \leftarrow a$ 
28 Function feedback ( $\{R_t\}_{t_s, \dots, t_s+M_s-1}$ )
29    $\forall a_i : N_s^i, R_s^i \leftarrow 0, 0$ 
30   for  $t = t_s, \dots, t_s + M_s - 1$  do
31      $R_s^{A_t} \leftarrow R_s^{A_t} + R_t$ 
32      $N_s^{A_t} \leftarrow N_s^{A_t} + 1$ 
33   for  $a_i, a_j \in \mathcal{K}^*$  do
34      $D_{i,j} \leftarrow D_{i,j} + \min\{N_i^s, N_j^s\} \left( \frac{R_s^i}{N_s^i} - \frac{R_s^j}{N_s^j} \right)$ 
35      $N_{i,j} \leftarrow N_{i,j} + \min\{N_i^s, N_j^s\}$ 

```

```

global  $N_{i,j}, D_{i,j}, \mathcal{K}^*, \mathcal{B}$ 
1 Function initialize ( $\mathcal{K}$ )
2    $\forall a_i, a_j \in \mathcal{K} : N_{i,j}, D_{i,j} \leftarrow 0, 0$ 
3    $\mathcal{B} \leftarrow \mathcal{K}$ 
4
5 Function getActiveSet ( $\delta$ )
6   if  $\exists N_{i,j} = 0$  then
7      $\mathcal{K}^* \leftarrow \mathcal{K}$ 
8   else
9     for  $a_i \in \mathcal{K}$  do
10       $\hat{\Delta}^{LCB}(a_i) \leftarrow \max_{a_j \neq a_i} \frac{D_{j,i}}{N_{j,i}} -$ 
11         $\sqrt{\frac{12 \log(2Kf(N_{j,i})\delta^{-1})}{N_{j,i}}}$ 
12       $\mathcal{K}^* \leftarrow \{a_i \in \mathcal{K} | \hat{\Delta}^{LCB}(a_i) \leq 0\}$ 
13      if  $|\mathcal{K}^*| = 0$  then
14         $\mathcal{K}^* \leftarrow \mathcal{K}$ 
15       $\mathcal{B} \leftarrow \mathcal{B} \cap \mathcal{K}^*$ 
16      if  $|\mathcal{B}| = 0$  then
17         $\mathcal{B} \leftarrow \mathcal{K}^*$ 
18   return  $\mathcal{K}^*$ 

```

**Algorithm 3:** Temporary Elimination Module (TEM) Implementation

## 4 Analysis

We start this section with the main theorem, which bounds the number of times the TEM pulls sub-optimal arms. Then we prove upper bounds on the regret for our main algorithms. Finally, we prove a lower bound for factored bandits that shows that our regret bound is tight up to constants.

### 4.1 Upper bound for the number of sub-optimal pulls by TEM

**Theorem 1.** *For any TEM submodule  $TEM^\ell$  with an arm set of size  $K = |\mathcal{A}^\ell|$ , running in the TEA algorithm with  $M := \max_\ell |\mathcal{A}^\ell|$  and any suboptimal atomic arm  $a \neq a^*$ , let  $N_t(a)$  denote the number of times TEM has played the arm  $a$  up to time  $t$ . Then there exist constants  $C(a) \leq M$  for*

$a \neq a^*$ , such that

$$\mathbb{E}[N_t(a)] \leq \frac{120}{\Delta(a)^2} \left( \log(2Kt \log^2(t)) + 4 \log \left( \frac{48 \log(2Kt \log^2(t))}{\Delta(a)^2} \right) \right) + C(a),$$

where  $\sum_{a \neq a^*} C(a) \leq M \log(K) + \frac{5}{2}K$  in the case of factored bandits and  $C(a) \leq \frac{5}{2}$  for dueling bandits.

*Proof sketch.* [The complete proof is provided in the Appendix.]

**Step 1** We show that the confidence intervals are constructed in such a way that the probability of all confidence intervals holding at all epochs up from  $s'$  is at least  $1 - \max_{s \geq s'} f(t_s)^{-1}$ . This requires a novel concentration inequality (Lemma 3) for a sum of conditionally  $\sigma_s$ -sub-gaussian random variables, where  $\sigma_s$  can be dependent on the history. This technique might be useful for other problems as well.

**Step 2** We split the number of pulls into pulls that happen in rounds where the confidence intervals hold and those where they fail:  $N_t(a) = N_t^{conf}(a) + N_t^{\overline{conf}}(a)$ .

We can bound the expectation of  $N_t^{\overline{conf}}(a)$  based on the failure probabilities given by  $\mathbb{P}[\text{conf failure at round } s] \leq \frac{1}{f(t_s)}$ .

**Step 3** We define  $s'$  as the last round in which the confidence intervals held and  $a$  was not eliminated. We can split  $N_t^{conf}(a) = N_{t_{s'}}^{conf}(a) + C(a)$  and use the confidence intervals to upper bound  $N_{t_{s'}}^{conf}(a)$ . The upper bound on  $\sum_a C(a)$  requires special handling of arms that were eliminated once and carefully separating the cases where confidence intervals never fail and those where they might fail.  $\square$

## 4.2 Regret Upper bound for Dueling Bandit TEA

A regret bound for the Factored Bandit TEA algorithm, Algorithm 1, is provided in the following theorem.

**Theorem 2.** *The pseudo-regret of Algorithm 1 at any time  $T$  is bounded by*

$$\begin{aligned} \text{Reg}_T \leq & \kappa \left( \sum_{\ell=1}^L \sum_{a_\ell \neq a_\ell^*} \frac{120}{\Delta_\ell(a_\ell)} \left( \log(2|\mathcal{A}^\ell|t \log^2(t)) + 4 \log \left( \frac{48 \log(2|\mathcal{A}^\ell|t \log^2(t))}{\Delta_\ell(a_\ell)} \right) \right) \right) \\ & + \max_{\ell} |\mathcal{A}^\ell| \sum_{\ell} \log(|\mathcal{A}^\ell|) + \sum_{\ell} \frac{5}{2} |\mathcal{A}^\ell|. \end{aligned}$$

*Proof.* The design of TEA allows application of Theorem 1 to each instance of TEM. Using  $\mu(\mathbf{a}_*) - \mu(\mathbf{a}) \leq \kappa \sum_{\ell=1}^L \Delta_\ell(a_\ell)$ , we have that

$$\text{Reg}_T = \mathbb{E} \left[ \sum_{t=1}^T \mu(\mathbf{a}^*) - \mu(\mathbf{a}_t) \right] \leq \kappa \sum_{\ell=1}^L \sum_{a_\ell \neq a_\ell^*} \mathbb{E}[N_T(a_\ell)] \Delta_\ell(a_\ell).$$

Applying Theorem 1 to the expected number of pulls and bounding the sums  $\sum_a C(a) \Delta(a) \leq \sum_a C(a)$  completes the proof.  $\square$

## 4.3 Dueling bandits

A regret bound for the Dueling Bandit TEA algorithm (DBTEA), Algorithm 2, is provided in the following theorem.

**Theorem 3.** *The pseudo-regret of Algorithm 2 for any utility-based dueling bandit problem at any time  $T$  (defined in equation (3)) satisfies  $\text{Reg}_T \leq \mathcal{O} \left( \sum_{a \neq a^*} \frac{\log(T)}{\Delta(a)} \right) + \mathcal{O}(K)$ .*

*Proof.* At every round, each arm in the active set is played once in position  $A$  and once in position  $B$  in  $\text{play}(A, B)$ . Denote by  $N_t^A(a)$  the number of plays of an arm  $a$  in the first position,  $N_t^B(a)$  the number of plays in the second position, and  $N_t(a)$  the total number of plays of the arm. We have

$$\text{Reg}_T = \sum_{a \neq a_*} \mathbb{E}[N_t(a)]\Delta(a) = \sum_{a \neq a_*} \mathbb{E}[N_t^A(a) + N_t^B(a)]\Delta(a) = \sum_{a \neq a_*} 2\mathbb{E}[N_t^A(a)]\Delta(a).$$

The proof is completed by applying Theorem 1 to bound  $\mathbb{E}[N_t^A(a)]$ .  $\square$

#### 4.4 Lower bound

We show that without additional assumptions the regret bound cannot be improved. The lower bound is based on the following construction. The mean reward of every arm is given by  $\mu(\mathbf{a}) = \mu(\mathbf{a}^*) - \sum_{\ell} \Delta_{\ell}(a_{\ell})$ . The noise is Gaussian with variance 1. In this problem, the regret can be decomposed into a sum over atomic arms of the regret induced by pulling these arms,  $\text{Reg}_T = \sum_{\ell} \sum_{a_{\ell} \in \mathcal{A}^{\ell}} \mathbb{E}[N_T(a_{\ell})]\Delta_{\ell}(a_{\ell})$ . Assume that we only want to minimize the regret induced by a single atomic set  $\mathcal{A}^{\ell}$ . Further, assume that  $\Delta_k(a)$  for all  $k \neq \ell$  are given. Then the problem is reduced to a regular  $K$ -armed bandit problem. The asymptotic lower bound for  $K$ -armed bandit under 1-Gaussian noise goes back to [10]: For any consistent strategy  $\theta$ , the asymptotic regret is lower bounded by  $\liminf_{T \rightarrow \infty} \frac{\text{Reg}_T^{\theta}}{\log(T)} \geq \sum_{a \neq a_*} \frac{2}{\Delta(a)}$ . Due to regret decomposition, we can apply this bound to every atomic set separately. Therefore, the asymptotic regret in the factored bandit problem is

$$\liminf_{T \rightarrow \infty} \frac{\text{Reg}_T^{\theta}}{\log(T)} \geq \sum_{\ell=1}^L \sum_{a^{\ell} \neq a_*^{\ell}} \frac{2}{\Delta^{\ell}(a^{\ell})}.$$

This shows that our general upper bound is asymptotically tight up to leading constants and  $\kappa$ .

**$\kappa$ -gap** We note that there is a problem-dependent gap of  $\kappa$  between our upper and lower bounds. Currently we believe that this gap stems from the difference between information and computational complexity of the problem. Our algorithm operates on each factor of the problem independently of other factors and is based on the ‘‘optimism in the face of uncertainty’’ principle. It is possible to construct examples in which the optimal strategy requires playing surely sub-optimal arms for the sake of information gain. For example, this kind of constructions were used by Lattimore and Szepesvari [11] to show suboptimality of optimism-based algorithms. Therefore, we believe that removing  $\kappa$  from the upper bound is possible, but requires a fundamentally different algorithm design. What is not clear is whether it is possible to remove  $\kappa$  without significant sacrifice of the computational complexity.

## 5 Comparison to Prior Work

### 5.1 Stochastic rank-1 bandits

Stochastic rank-1 bandits introduced by Katariya et al. [7] are a special case of factored bandits. The authors published a refined algorithm for Bernoulli rank-1 bandits using KL confidence sets in Katariya et al. [8]. We compare our theoretical results with the first paper because it matches our problem assumptions. In our experiments, we provide a comparison to both the original algorithm and the KL version.

In the stochastic rank-1 problem there are only 2 atomic sets of size  $K_1$  and  $K_2$ . The matrix of expected rewards for each pair of arms is of rank 1. It means that for each  $u \in \mathcal{A}^1$  and  $v \in \mathcal{A}^2$ , there exist  $\bar{u}, \bar{v} \in [0, 1]$  such that  $\mathbb{E}[r(u, v)] = \bar{u} \cdot \bar{v}$ . The proposed Stochastic rank-1 Elimination algorithm introduced by Katariya et al. is a typical elimination style algorithm. It requires knowledge of the time horizon and uses phases that increase exponentially in length. In each phase, all arms are played uniformly. At the end of a phase, all arms that are sub-optimal with high probability are eliminated.

**Theoretical comparison** It is hard to make a fair comparison of the theoretical bounds because TEA operates under much weaker assumptions. Both algorithms have a regret bound of  $\mathcal{O}\left(\left(\sum_{u \in \mathcal{A}^1 \setminus u_*} \frac{1}{\Delta_1(u)} + \sum_{v \in \mathcal{A}^2 \setminus v_*} \frac{1}{\Delta_2(v)}\right) \log(t)\right)$ . The problem independent multiplicative factors

hidden under  $\mathcal{O}$  are smaller for TEA, even without considering that rank-1 Elimination requires a doubling trick for anytime applications. However, the problem dependent factors are in favor of rank-1 Elimination, where the gaps correspond to the mean difference under uniform sampling  $(\bar{u}^* - \bar{u}) \sum_{v \in \mathcal{A}^2} \bar{v}/K_2$ . In factored bandits, the gaps are defined as  $(\bar{u}^* - \bar{u}) \min_{v \in \mathcal{A}^2} \bar{v}$ , which is naturally smaller. The difference stems from different problem assumptions. Stronger assumptions of rank-1 bandits make elimination easier as the number of eliminated suboptimal arms increases. The TEA analysis holds in cases where it becomes harder to identify suboptimal arms after removal of bad arms. This may happen when highly suboptimal atomic actions in one factor provide more discriminative information on atomic actions in other factors than close to optimal atomic actions in the same factor (this follows the spirit of illustration of suboptimality of optimistic algorithms in [11]). We leave it to future work to improve the upper bound of TEA under stronger model assumptions.

In terms of memory and computational complexity, TEA is inferior to regular elimination style algorithms, because we need to keep track of relative performances of the arms. That means both computational and memory complexities are  $\mathcal{O}(\sum_{\ell} |\mathcal{A}^{\ell}|^2)$  per round in the worst case, as opposed to rank-1 Elimination that only requires  $\mathcal{O}(|\mathcal{A}^1| + |\mathcal{A}^2|)$ .

**Empirical comparison** The number of arms is set to 16 in both sets. We always fix  $\bar{u}^* - \bar{u} = v^* - \bar{v} = 0.2$ . We vary the absolute value of  $u^*v^*$ . As expected, rank1ElimKL has an advantage when the Bernoulli random variables are strongly biased towards one side. When the bias is close to  $\frac{1}{2}$ , we clearly see the better constants of TEA. In the evaluation we clearly outperform rank-1 Elimination

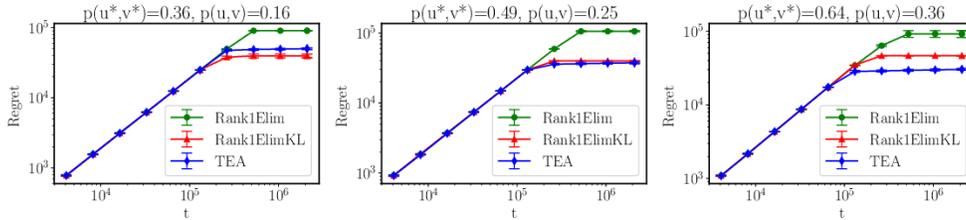


Figure 2: Comparison of Rank1Elim, Rank1ElimKL, and TEA for  $K = L = 16$ . The results are averaged over 20 repetitions of the experiment.

over different parameter settings and even beat the KL optimized version if the means are not too close to zero or one. This supports that our algorithm does not only provide a more practical anytime version of elimination, but also improves on constant factors in the regret. We believe that our algorithm design can be used to improve other elimination style algorithms as well.

## 5.2 Dueling Bandits: Related Work

To the best of our knowledge, the proposed Dueling Bandit TEA is the first algorithm that satisfies the following three criteria simultaneously for utility-based dueling bandits:

- It requires no prior knowledge of the time horizon (nor uses the doubling trick or restarts).
- Its pseudo-regret is bounded by  $\mathcal{O}(\sum_{a \neq a^*} \frac{\log(t)}{\Delta(a)})$ .
- There are no additive constants that dominate the regret for time horizons  $T > \mathcal{O}(K)$ .

We want to stress the importance of the last point. For all state-of-the-art algorithms known to us, when the number of actions  $K$  is moderately large, the additive term is dominating for any realistic time horizon  $T$ . In particular, Ailon et al. [2] introduces three algorithms for the utility-based dueling bandit problem. The regret of Doubler scales with  $\mathcal{O}(\log^2(t))$ . The regret of MultiSBM has an additive term of order  $\sum_{a \neq a^*} \frac{K}{\Delta(a)}$  that is dominating for  $T < \Omega(\exp(K))$ . The last algorithm, Sparring, has no theoretical analysis.

Algorithms based on the weaker Condorcet winner assumption apply to utility-based setting, but they all suffer from equally large or even larger additive terms. The RUCB algorithm introduced by Zoghi et al. [17] has an additive term in the bound that is defined as  $2D\Delta_{max} \log(2D)$ , for

$\Delta_{max} = \max_{a \neq a^*} \Delta(a)$  and  $D > \frac{1}{2} \sum_{a_i \neq a^*} \sum_{a_j \neq a_i} \frac{4\alpha}{\min\{\Delta(a_i)^2, \Delta(a_j)^2\}}$ . By unwrapping these definitions, we see that the RUCB regret bound has an additive term of order  $2D\Delta_{max} \geq \sum_{a \neq a^*} \frac{K}{\Delta(a)}$ . This is again the dominating term for time horizons  $T \leq \Omega(\exp(K))$ . The same applies to the RMED algorithm introduced by Komiyama et al. [9], which has an additive term of  $\mathcal{O}(K^2)$ . (The dependencies on the gaps are hidden behind the  $\mathcal{O}$ -notation.) The D-TS algorithm by Wu and Liu [13] based on Thompson Sampling shows one of the best empirical performances, but its regret bound includes an additive constant of order  $\mathcal{O}(K^3)$ .

Other algorithms known to us, Interleaved Filter [16], Beat the Mean [15], and SAVAGE [12], all require knowledge of the time horizon  $T$  in advance.

**Empirical comparison** We have used the framework provided by Komiyama et al. [9]. We use the same utility for all sub-optimal arms. In Figure 3, the winning probability of the optimal arm over suboptimal arms is always set to 0.7, we run the experiment for different number of arms  $K$ . TEA outperforms all algorithms besides RMED variants, as long as the number of arms are sufficiently big. To show that there also exists a regime where the improved constants gain an advantage over RMED, we conducted a second experiment in Figure 4 (in the Appendix), where we set the winning probability to 0.95<sup>2</sup> and significantly increase the number of arms. The evaluation shows that the additive terms are indeed non-negligible and that Dueling Bandit TEA outperforms all baseline algorithms when the number of arms is sufficiently large.

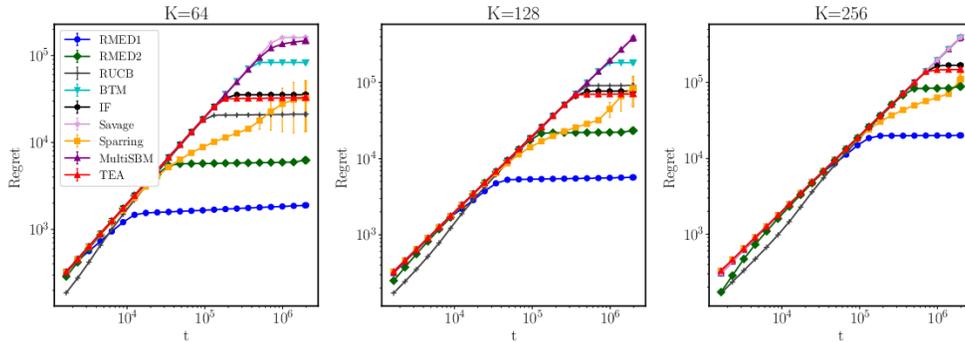


Figure 3: Comparison of Dueling Bandits algorithms with identical gaps of 0.4. The results are averaged over 20 repetitions of the experiment.

## 6 Discussion

We have presented the factored bandits model and uniform identifiability assumption, which requires no knowledge of the reward model. We presented an algorithm for playing stochastic factored bandits with uniformly identifiable actions and provided matching upper and lower bounds for the problem up to constant factors. Our algorithm and proofs might serve as a template to turn other elimination style algorithms into improved anytime algorithms.

Factored bandits with uniformly identifiable actions generalize rank-1 bandits. We have also provided a unified framework for the analysis of factored bandits and utility-based dueling bandits. Furthermore, we improve the additive constants in the regret bound compared to state-of-the-art algorithms for utility-based dueling bandits.

There are multiple potential directions for future research. One example mentioned in the text is the possibility of improving the regret bound when additional restrictions on the form of the reward function are introduced or improvements of the lower bound when algorithms are restricted in computational or memory complexity. Another example is the adversarial version of the problem.

<sup>2</sup>Smaller gaps show the same behavior but require more arms and more timesteps.

## References

- [1] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.
- [2] N. Ailon, Z. Karnin, and T. Joachims. Reducing dueling bandits to cardinal bandits. In *International Conference on Machine Learning*, pages 856–864, 2014.
- [3] N. Cesa-Bianchi and G. Lugosi. Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404–1422, 2012.
- [4] W. Chen, Y. Wang, Y. Yuan, and Q. Wang. Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *The Journal of Machine Learning Research*, 17(1):1746–1778, 2016.
- [5] R. Combes, S. Magureanu, and A. Proutiere. Minimal exploration in structured stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 1761–1769, 2017.
- [6] S. Filippi, O. Cappé, A. Garivier, and C. Szepesvári. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*, pages 586–594, 2010.
- [7] S. Katariya, B. Kveton, C. Szepesvári, C. Vernade, and Z. Wen. Stochastic rank-1 bandits (long version). In *AISTATS*, volume 54 of *PMLR*, pages 392–401, April 2017.
- [8] S. Katariya, B. Kveton, C. Szepesvári, C. Vernade, and Z. Wen. Bernoulli rank-1 bandits for click feedback. *International Joint Conference on Artificial Intelligence*, 2017.
- [9] J. Komiyama, J. Honda, H. Kashima, and H. Nakagawa. Regret lower bound and optimal algorithm in dueling bandit problem. In *Conference on Learning Theory*, pages 1141–1154, 2015.
- [10] T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.
- [11] T. Lattimore and C. Szepesvári. The end of optimism? An asymptotic analysis of finite-armed linear bandits (long version). In *AISTATS*, volume 54 of *PMLR*, pages 728–737, April 2017.
- [12] T. Urvoy, F. Clerot, R. Féraud, and S. Naamane. Generic exploration and k-armed voting bandits. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 91–99, 2013.
- [13] H. Wu and X. Liu. Double thompson sampling for dueling bandits. In *Advances in Neural Information Processing Systems*, pages 649–657, 2016.
- [14] Y. Yue and T. Joachims. Interactively optimizing information retrieval systems as a dueling bandits problem. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1201–1208. ACM, 2009.
- [15] Y. Yue and T. Joachims. Beat the mean bandit. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 241–248, 2011.
- [16] Y. Yue, J. Broder, R. Kleinberg, and T. Joachims. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012.
- [17] M. Zoghi, S. Whiteson, R. Munos, and M. Rijke. Relative upper confidence bound for the k-armed dueling bandit problem. In *International Conference on Machine Learning*, pages 10–18, 2014.

## A Auxiliary Lemmas

**Lemma 1.** *Given positive real numbers  $\sigma_1, \sigma_2, \dots, \sigma_n$ . If  $(X_i)_{i=1, \dots, n}$  is a sequence of random variables such that  $X_i$  conditioned on  $X_{i-1}, X_{i-2}, \dots$  is  $\sigma_i$ -sub-Gaussian. Then  $Z = \sum_{i=1}^n X_i$  is  $\sqrt{\sum_{i=1}^n \sigma_i^2}$ -sub-Gaussian.*

We believe this is a standard result, however we only found references for independent sub-Gaussian random variables.

*Proof of Lemma 1.* For  $t = 1, \dots, n$  define  $M_{s,t} = \exp(s \sum_{i=1}^t X_i - \frac{1}{2} \sum_{i=1}^t s^2 \sigma_i^2)$ . We claim  $M_{s,t}$  is a super-martingale. Given that  $X_i$  are conditionally sub-Gaussian, we have  $\mathbb{E}[\exp(sX_{t+1}) | X_t, X_{t-1}, \dots] \leq \exp(\frac{s^2 \sigma_{t+1}^2}{2})$ . So

$$\begin{aligned} \mathbb{E}[M_{s,t+1} | M_{s,t}] &= \mathbb{E}[\exp(sX_{t+1} - \frac{1}{2} s^2 \sigma_{t+1}^2) M_{s,t} | M_{s,t}] \\ &= \mathbb{E}[\exp(sX_{t+1} - \frac{1}{2} s^2 \sigma_{t+1}^2) | M_{s,t}] M_{s,t} \leq M_{s,t}. \end{aligned}$$

Additionally by definition of sub-Gaussian  $\mathbb{E}[M_{s,1}] \leq 1$ . Therefore  $\mathbb{E}[M_{s,n}] \leq 1$ . Finally we get that  $\mathbb{E}[\exp(sZ)] = \mathbb{E}[M_{s,n} \cdot \exp(\sum_{i=1}^n \frac{s^2 \sigma_i^2}{2})] \leq \exp(\sum_{i=1}^n \frac{s^2 \sigma_i^2}{2})$ . So  $Z$  is  $\sqrt{\sum_{i=1}^n \sigma_i^2}$ -sub-Gaussian.  $\square$

**Lemma 2.** *Let  $y \geq 1, z \geq 10$ , then for any  $x > zy + 4z \log(zy)$ :*

$$\frac{z(\log(f(x)) + y)}{x} < 1.$$

*Proof.* We can reparameterize  $x = zy + \alpha z \log(zy)$  for  $\alpha > 4$ . Then

$$\begin{aligned} \frac{zy + z \log(f(zy + \alpha z \log(zy)))}{zy + \alpha z \log(zy)} &< 1 \\ \Leftrightarrow \frac{\log(f(zy + \alpha z \log(zy)))}{\alpha \log(zy)} &< 1 \\ \Leftrightarrow f(zy + \alpha z \log(zy)) &< (zy)^\alpha \\ \Leftrightarrow f(zy + \alpha zy \log(zy)) &< (zy)^\alpha. \end{aligned}$$

Using  $\log(x) \leq \sqrt{x} - \frac{1}{2}$  and  $\alpha > 4$ , we have that

$$f(zy + \alpha zy \log(zy)) < f\left(zy + \alpha zy \left(\sqrt{zy} - \frac{1}{2}\right)\right) < f(\alpha(zy)^{\frac{3}{2}} - 1) = \alpha(zy)^{\frac{3}{2}} \log^2(\alpha(zy)^{\frac{3}{2}}).$$

It is therefore sufficient to prove that for all  $\tilde{x} > 10$  and  $\alpha > 4$ :

$$\begin{aligned} \alpha \log^2(\alpha \tilde{x}^{\frac{3}{2}}) &< \tilde{x}^{\alpha - \frac{3}{2}} \\ \Leftrightarrow \alpha(\sqrt{\alpha} + \tilde{x}^{\frac{3}{4}})^2 &< \tilde{x}^{\alpha - \frac{3}{2}} \\ \Leftrightarrow \sqrt{\alpha}(\sqrt{\alpha} \tilde{x}^{\frac{3}{4} - \frac{\alpha}{2}} + \tilde{x}^{\frac{3}{2} - \frac{\alpha}{2}}) &< 1. \end{aligned}$$

The minimum on the left hand side is obtained for  $\alpha = 4$  and  $\tilde{x} = 10$  for with it holds true.  $\square$

**Lemma 3.** *Let  $\sigma \in \mathbb{R}$  and  $X_1, X_2, \dots$  be a sequence of sub-Gaussian random variables adapted to the filtration  $\mathcal{F}_1, \mathcal{F}_2, \dots$ , i.e.  $\mathbb{E}[e^{sX_t} | X_1, X_2, \dots, X_{t-1}] \leq e^{-\frac{\sigma_t^2 s^2}{2}}$ . Assume for all  $t$ :  $\sum_{i=1}^t \sigma_i^2 = n_t \sigma^2$ , with  $n_t \in \mathbb{N}$  almost surely. Then*

$$\mathbb{P}\left[\exists t \in \mathbb{N} : \sum_{i=1}^t X_i \geq \sqrt{2\sigma^2 n_t \log\left(\frac{f(n_t)}{\delta}\right)}\right] \leq \delta,$$

where  $f(n_t) = 2(1 + n_t) \log^2(1 + n_t)$ .

Note that unlike in Lemma 1, we do not require  $\sigma_t$  to be independent of  $X_1, \dots, X_{t-1}$ .

*Proof.* The proof follows closely the arguments presented in the proofs of Lemma 8 in Abbasi-Yadkori et al. [1] and Lemma 14 in Lattimore and Szepesvári [11]. For  $\psi \in \mathbb{R}$  define

$$M_{t,\psi} = \exp \left( \sum_{s=1}^t \psi X_s - \frac{\psi^2 \sigma_s^2}{2} \right).$$

If  $t_0 \leq \tau \leq t$  is a stopping time with respect to  $\mathcal{F}$ , then as in the proof of Abbasi-Yadkori et al. [1, Lemma 8] we have  $\mathbb{E}[M_{\tau,\psi}] \leq 1$ . By Markov's inequality, we have

$$\mathbb{P}[M_{\tau,\psi} \geq 1/\delta] \leq \delta \quad \Leftrightarrow \quad \mathbb{P} \left[ \sum_{s=1}^{\tau} X_s \geq \frac{\log(\delta^{-1})}{\psi} + \frac{\psi n_{\tau} \sigma^2}{2} \right] \leq \delta.$$

An optimal choice of  $\psi$  would be  $\psi = \sqrt{2 \frac{\log(1/\delta)}{n_{\tau} \sigma^2}}$ , however  $\psi X_t$  would not be  $\mathcal{F}_t$ -measurable for  $t \leq \tau$  and  $M_{t,\psi}$  would not be well defined. Instead, for  $k \geq 1$  we define

$$\psi_k := \sqrt{\frac{2 \log(f(k) \delta^{-1})}{k \sigma^2}}.$$

With a union bound, we get that

$$\mathbb{P} \left[ \exists k \geq 1 : \sum_{s=1}^{\tau} X_s \geq \frac{\log(f(k) \delta^{-1})}{\psi_k} + \frac{\psi_k n_{\tau} \sigma^2}{2} \right] \leq \sum_{k=1}^{\infty} \frac{\delta}{f(k)} \leq \delta.$$

Using now  $k = n_{\tau}$ , for which this also holds, we get that

$$\mathbb{P} \left[ \sum_{s=1}^{\tau} X_s \geq \sqrt{2 \sigma^2 n_{\tau} \log \left( \frac{f(n_{\tau})}{\delta} \right)} \right] \leq \delta.$$

The proof is completed by choosing a stopping time  $\tau$ :

$$\tau = \min \left( \infty \cup \left\{ t \geq 1 : \sum_{s=1}^t X_s \geq \sqrt{2 n_t \sigma^2 \log \left( \frac{f(n_t)}{\delta} \right)} \right\} \right).$$

□

**Lemma 4.** Given  $X_1, X_2, \dots, X_n$  random variables with means  $p_1, p_2, \dots, p_n \in [-1, 1]$ , such that all  $X_i - p_i$  are 1-sub-Gaussian. (e.g. Bernoulli random variables) Given further two sample sizes  $m, k \geq 1$ , such that  $m + k \leq n$ . Then for  $I_m : |I_m| = m$  and  $I_k : |I_k| = k$  disjoint uniform samples of indices in  $(1, 2, \dots, n)$  without replacement, the random variable

$$Z = \frac{1}{m} \sum_{i \in I_m} X_i - \frac{1}{k} \sum_{i \in I_k} X_i,$$

is  $\sqrt{\frac{3(m+k)}{mk}}$ -sub-Gaussian.

*Proof.* Without loss of generality, we set  $m \leq k$ . By definition, the random variables  $X_i$  can be decomposed into  $X_i = p_i + \eta_i$ , where  $\eta_i$  are conditionally independent 1-sub-Gaussian random variables. Decomposing  $Z$  gives:

$$Z = \frac{1}{m} \sum_{i \in I_m} p_i - \frac{1}{k} \sum_{i \in I_k} p_i + \frac{1}{m} \sum_{i \in I_m} \eta_i - \frac{1}{k} \sum_{i \in I_k} \eta_i.$$

We define  $\bar{I} = \{1, \dots, n\} \setminus (I_m \cup I_k)$ , the indices of remaining  $X_i$ 's and  $\bar{p} = \frac{1}{n} \sum_{i=1}^n p_i$  the mean of means. In order to show that  $\frac{1}{m} \sum_{i \in I_m} p_i - \frac{1}{k} \sum_{i \in I_k} p_i$  is sub-Gaussian, we first draw the elements in

$(p_i)_{i \in \bar{I}} = (\bar{P}_i)_{i=1, \dots, n-m-k}$  and then the set  $(p_i)_{i \in I_m} = (P_i^m)_{i=1, \dots, m}$ . Drawing the first element  $\bar{P}_1$  can be written as  $\bar{P}_1 = \bar{p} + \zeta_1$ , where  $\zeta_1$  is sub-Gaussian. With continuous drawings, it holds that

$$\begin{aligned}
\mathbb{E}[\bar{P}_2 | \bar{P}_1] &= \bar{p} - \frac{1}{n-1} \zeta_1 \\
\bar{P}_2 &= \bar{p} - \frac{1}{n-1} \zeta_1 + \zeta_2 \\
\mathbb{E}[\bar{P}_3 | \bar{P}_1, \bar{P}_2] &= \bar{p} - \frac{1}{n-1} \zeta_1 - \frac{1}{n-2} \zeta_2 \\
\bar{P}_3 &= \bar{p} - \frac{1}{n-1} \zeta_1 - \frac{1}{n-2} \zeta_2 + \zeta_3 \\
&\dots \\
\mathbb{E}[\bar{P}_{n-m-k} | \bar{P}_1, \dots, \bar{P}_{n-m-k-1}] &= \bar{p} - \sum_{i=1}^{n-m-k-1} \frac{1}{n-i} \zeta_i \\
\bar{P}_{n-m-k} &= \bar{p} - \sum_{i=1}^{n-m-k-1} \frac{1}{n-i} \zeta_i + \zeta_{n-m-k} \\
\sum_{i=1}^{n-m-k} \bar{P}_i &= (n-m-k)\bar{p} + \sum_{i=1}^{n-m-k} \frac{m+k}{n-i} \zeta_i
\end{aligned}$$

The noise variables  $\zeta_i$  are all conditionally independent and 1-sub-Gaussian.

We continue with  $P_i^m$  in the same fashion:

$$\begin{aligned}
\mathbb{E}[P_1^m | \bar{P}] &= \bar{p} - \sum_{i=1}^{n-m-k} \frac{1}{n-i} \zeta_i \\
P_1^m &= \bar{p} - \sum_{i=1}^{n-m-k} \frac{1}{n-i} \zeta_i + \zeta_{n-k-m+1} \\
\mathbb{E}[P_2^m | \bar{P}, P_1^m] &= \bar{p} - \sum_{i=1}^{n-m-k+1} \frac{1}{n-i} \zeta_i \\
P_2^m &= \bar{p} - \sum_{i=1}^{n-m-k} \frac{1}{n-i} \zeta_i + \zeta_{n-k-m+2} \\
&\dots \\
\mathbb{E}[P_m^m | \bar{P}, P_1^m, \dots, P_{m-1}^m] &= \bar{p} - \sum_{i=1}^{n-k-1} \frac{1}{n-i} \zeta_i \\
P_m^m &= \bar{p} - \sum_{i=1}^{n-k-1} \frac{1}{n-i} \zeta_i + \zeta_{n-k} \\
\sum_{i=1}^m P_m^m &= (n-k)\bar{p} + \sum_{i=1}^{n-k} \frac{k}{n-i} \zeta_i - \sum_{i=1}^{n-m-k} \bar{P}_i \\
&= m\bar{p} - \sum_{i=1}^{n-m-k} \frac{m}{n-i} \zeta_i + \sum_{i=n-m-k+1}^{n-k} \frac{k}{n-i} \zeta_i.
\end{aligned}$$

We can now use

$$\frac{1}{k} \sum_{i \in I_k} p_i = \frac{1}{k} \left( n\bar{p} - \sum_{i=1}^{n-m-k} \bar{P}_i - \sum_{i=1}^m P_i^m \right),$$

to substitute

$$\begin{aligned}
\frac{1}{m} \sum_{i \in I_m} p_i - \frac{1}{k} \sum_{i \in I_k} p_i &= \frac{1}{m} \sum_{i=1}^m P_i^m - \frac{1}{k} \left( n\bar{p} - \sum_{i=1}^{n-m-k} \bar{P}_i - \sum_{i=1}^m P_i^m \right) \\
&= \frac{m+k}{mk} \sum_{i=1}^m P_i^m + \frac{1}{k} \sum_{i=1}^{n-m-k} \bar{P}_i - \frac{n}{k} \bar{p} \\
&= \frac{m+k}{mk} \left( m\bar{p} - \sum_{i=1}^{n-m-k} \frac{m}{n-i} \zeta_i + \sum_{i=n-m-k+1}^{n-k} \frac{k}{n-i} \zeta_i \right) \\
&\quad + \frac{1}{k} \left( (n-m-k)\bar{p} + \sum_{i=1}^{n-m-k} \frac{m+k}{n-i} \zeta_i \right) - \frac{n}{k} \bar{p} \\
&= \sum_{i=n-m-k+1}^{n-k} \frac{m+k}{m(n-i)} \zeta_i \\
&= \sum_{i=0}^{m-1} \frac{m+k}{m(k+i)} \zeta_{n-k-i}.
\end{aligned}$$

With these substitutions  $Z$  can be written as a weighted sum of conditionally independent sub-Gaussian random variables:

$$Z = \sum_{i=0}^{m-1} \frac{m+k}{m(k+i)} \zeta_{n-k-i} + \frac{1}{m} \sum_{i \in I_m} \eta_i - \frac{1}{k} \sum_{i \in I_k} \eta_i.$$

Therefore  $Z$  is according to Lemma 1 at least

$$\sqrt{\sum_{i=0}^{m-1} \left( \frac{m+k}{m(k+i)} \right)^2 + \sum_{i=1}^m \frac{1}{m^2} + \sum_{i=1}^k \frac{1}{k^2}} \leq \sqrt{\frac{3(m+k)}{mk}}$$

-sub-Gaussian.

The last step uses the inequality

$$\begin{aligned}
\sum_{i=0}^{m-1} \frac{1}{(k+i)^2} &= \int_0^m \frac{1}{(k+x)^2} dx + \sum_{i=0}^{m-1} \left( \frac{1}{(k+i)^2} - \int_{x=i}^{i+1} \frac{1}{(k+x)} dx \right) \\
&= \frac{m}{(k+m)k} + \sum_{i=0}^{m-1} \frac{1}{(k+i)^2(k+i+1)} \\
&\leq \frac{m}{(k+m)k} + \frac{1}{k+1} \sum_{i=0}^{m-1} \frac{1}{(k+i)^2} \\
&\leq \frac{m(k+1)}{(k+m)k^2} \\
&\leq \frac{2m}{(k+m)k}.
\end{aligned}$$

□

## B Proof of Theorem 1

With the Lemmas from the previous section, we can proof our main theorem.

*Proof of Theorem 1.* We follow the steps from the sketch.

**Step 1** We define the following shifted random variables.

$$\begin{aligned}\tilde{R}_t &:= R_t + \mu_t(a_*) - \mu_t(A_t) \\ \tilde{R}_s^i &:= \sum_{t \in T_s} \mathbb{I}\{A_t = a_i\} \tilde{R}_t \\ \Delta \tilde{D}_s^i &:= \frac{\tilde{R}_s^*}{N_s^*} - \frac{\tilde{R}_s^i}{N_s^i} \\ \tilde{D}_s(a_i) &:= \sum_{k=1}^s \min\{N_s^i, N_s^*\} \Delta \tilde{D}_k^i \\ \tilde{\Delta}_s(a_i) &:= \frac{\tilde{D}_s(a_i)}{N_{*,i}(s)}.\end{aligned}$$

The reward functions satisfy  $\mu_t(a_*) - \mu_t(a_t) > \Delta(a_t)$  for all  $a_t$ . Therefore  $R_t > \tilde{R}_t - \Delta(A_t)$ . So we can bound  $\frac{D_{*,i}^*}{N_{*,i}^*} > \Delta(a_i) + \tilde{\Delta}_s(a_i)$  and  $\frac{D_{i,*}^i}{N_{i,*}^i} < -\Delta(a_i) - \tilde{\Delta}_s(a_i)$ .

Define the events

$$\mathcal{E}_s := \left\{ \forall i : |\tilde{\Delta}_s(a_i)| \leq \sqrt{\frac{12 \log(2K f(N_{*,i}) \delta_s^{-1})}{N_{*,i}}} \right\}, \quad \mathcal{F} := \bigcap_{s \geq 2} \mathcal{E}_s$$

and their complements  $\bar{\mathcal{E}}_s, \bar{\mathcal{F}}$ .

According to lemma 1,  $\Delta \tilde{D}_s^i$  is  $\sqrt{\frac{6}{\min\{N_s^*, N_s^i\}}}$ -sub-Gaussian. So  $\tilde{D}_s(a_i)$  is a sum of conditionally  $\sigma_i$ -sub-Gaussian random variables, such that  $\sum_{i=1}^s \sigma_i^2 = 6N_{*,i}(s)$ , Therefore we can apply Lemma 3. For both cases  $\delta_s = \frac{1}{f(t_s)}$  and  $\delta_s = \delta$ , the probability never increases in time.

$$\begin{aligned}\mathbb{P} \left[ \exists s' \geq s : \tilde{\Delta}_{s'}(a_i) \geq \sqrt{\frac{12 \log(2K f(\delta_{s'}) N_{*,i})}{\delta_{s'}}} \right] \\ \leq \mathbb{P} \left[ \exists s' \geq s : \tilde{D}_{s'}(a_i) \geq N_{*,i} \sqrt{\frac{12 \log(2K f(N_{*,i}) \delta_s)}{N_{*,i}}} \right] \leq \frac{\delta_s}{2K}.\end{aligned}$$

Using a union bound over  $\pm \tilde{D}_s(a_i)$  for  $a_i \in \mathcal{A}$ , we get

$$\mathbb{P}[\bar{\mathcal{E}}_s] \leq \delta_s \text{ and } \mathbb{P}[\bar{\mathcal{F}}] \leq \delta_2.$$

**step 2** We split the number of pulls in two categories: those that appear in rounds where the confidence intervals hold, and those that appear in rounds where they fail:  $N_t^{\mathcal{E}}(a_i) = \sum_{s'=1}^{s(t)} \mathbb{I}\{\mathcal{E}_s\} N_s^i$ ,  $N_t^{\bar{\mathcal{E}}}(a_i) = \sum_{s'=1}^{s(t)} \mathbb{I}\{\bar{\mathcal{E}}_s\} N_s^i$ .

$$\begin{aligned}N_t(a_i) &\leq N_t^{\mathcal{E}}(a_i) + N_t^{\bar{\mathcal{E}}}(a_i) \\ \mathbb{E}[N_t^{\mathcal{E}}(a_i)] &= \mathbb{P}[\bar{\mathcal{F}}] \mathbb{E}[N_t^{\mathcal{E}}(a_i) | \bar{\mathcal{F}}] + \mathbb{P}[\mathcal{F}] \mathbb{E}[N_t^{\mathcal{E}}(a_i) | \mathcal{F}].\end{aligned}$$

In the high probability case, we are with probability  $1 - \delta$  in the event  $\mathcal{F}$  and  $N_t^{\bar{\mathcal{E}}}(a_i)$  is 0. In the setting of  $\delta_s = f(t_s)^{-1}$ , we can exclude the first round and start with  $s = 2$  and  $t_2 = M + 1$ . This is because we do not use the confidence intervals in the first round.

$$\begin{aligned}\mathbb{E}[N_t^{\bar{\mathcal{E}}}(a_i)] &\leq \sum_{s=2}^{\infty} \frac{t_{s+1} - t_s}{f(t_s)} \leq \sum_{s=1}^{\infty} \frac{M}{f(Ms)} \\ &\leq \frac{M}{f(M)} + \sum_{s=2}^{\infty} \frac{M}{f(Ms)} \leq \frac{1}{2} + \sum_{s=1}^{\infty} \frac{1}{f(s)} \leq \frac{3}{2}\end{aligned}$$

We use the fact that  $\frac{1}{f(t_s)}$  is monotonically decreasing, so the expression gets minimized if all rounds are maximally long.

**Step 3:** bounding  $\mathbb{E}[N_t^{\mathcal{E}}(a_i)|\mathcal{F}], \mathbb{E}[N_t^{\mathcal{E}}(a_i)|\overline{\mathcal{F}}]$

Let  $s'$  be the last round at which the arm  $a_i$  is not eliminated. We claim that  $N_{i,*}$  at the beginning of round  $s'$  must be surely smaller or equal to  $\frac{48}{\Delta(a_i)^2} \left( \log(2K\delta_{s'}^{-1}) + 4 \log\left(\frac{48 \log(2K\delta_{s'}^{-1})}{\Delta(a_i)^2}\right) \right)$ . Assume the opposite holds, then according to Lemma 2 with  $z = \frac{48}{\Delta(a_i)^2}$  and  $y = \log(2K\delta_{s'}^{-1})$ :

$$\frac{\frac{48}{\Delta(a_i)^2} (\log(f(N_{i,*}(s'))) + \log(2K\delta_{s'}^{-1}))}{N_{i,*}(s')} < 1 \quad \Leftrightarrow \quad \sqrt{\frac{12 \log(2Kf(N_{*,i})\delta_{s'}^{-1})}{N_{*,i}}} < \frac{1}{2}\Delta(a_i).$$

So we have that

$$\hat{\Delta}_{s'}^{LCB}(a_i) \geq \Delta(a_i) - 2\sqrt{\frac{12 \log(2Kf(N_{*,i})\delta_{s'}^{-1})}{N_{*,i}}} > 0,$$

and  $a_i$  would have been excluded at the beginning of round  $s'$ , which is a contradiction.

Let  $C(a_i)$  denote the number of plays of  $a_i$  in round  $s'$ . Then for the different cases we have:

$$N_t^{\mathcal{E}}(a_i) - C(a_i) \leq \begin{cases} M \cdot N_{i,*}(s'), & \text{under the event } \overline{\mathcal{F}} \\ 2 \cdot N_{i,*}(s'), & \text{under the event } \mathcal{F} \\ N_{i,*}(s'), & \text{if } M_s = |\mathcal{A}_A| \end{cases}$$

$$\sum_{a \neq a_*} C(a) \leq \begin{cases} MK, & \text{under the event } \overline{\mathcal{F}} \\ M \log(K) + K, & \text{under the event } \mathcal{F} \\ K & \text{if } M_s = |\mathcal{A}_A| \end{cases}$$

The first case is trivial because each arm can only be played  $M$  times in a single round and  $\min\{N_s^i, N_s^*\} \geq 1$  in rounds with  $\mathcal{E}_s$ . The second case follows from the fact that  $a_*$  is always in set  $\mathcal{B}$  under the event  $\mathcal{F}$ . So  $N_s^* \geq \max\{1, N_s^i - 1\}$  and  $\min\{N_s^i, N_s^*\} \geq \frac{N_s^i}{2}$ . The amount of pulls in a single round is naturally bounded by  $\lceil \frac{M}{|\mathcal{B}|} \rceil \leq M$ . Given that under the event  $\mathcal{F}$ , the set  $\mathcal{B}$  never resets and the set  $\mathcal{B}$  only decreases if an arm is eliminated, we can bound

$$\sum_{a_i \neq a_*} C(a_i) \leq \sum_{i=2}^K \lceil \frac{M}{i} \rceil \leq M \log(K) + K.$$

Finally the last case follows trivially because in the case of  $M_s = |\mathcal{A}_A|$ , we have  $N_s^i = N_s^* = C(a_i) = 1$ .

**Step 4:** combining everything

In the high probability case, we have with probability at least  $1 - \delta$ :

$$\begin{aligned} N_t(a_i) &\leq N_t^{\mathcal{E}}(a_i) + N_t^{\overline{\mathcal{E}}}(a_i) \\ &\leq 2N_{i,*}(s') + C(a_i) \\ &\leq \frac{96}{\Delta(a)^2} \left( \log(2K\delta^{-1}) + 4 \log\left(\frac{48 \log(2K\delta^{-1})}{\Delta(a)^2}\right) \right) + C(a_i) \end{aligned}$$

and also

$$\sum_{a \neq a_*} C(a) \leq M \log(K) + K.$$

If additionally  $M_s = |\mathcal{A}_A|$ , then the bound improves to

$$\begin{aligned} N_t(a_i) &\leq N_t^{\mathcal{E}}(a_i) + N_t^{\overline{\mathcal{E}}}(a_i) \\ &\leq N_{i,*}(s') + 1 \\ &\leq \frac{48}{\Delta(a)^2} \left( \log(2K\delta^{-1}) + 4 \log\left(\frac{48 \log(2K\delta^{-1})}{\Delta(a)^2}\right) \right) + 1. \end{aligned}$$

In the setting of  $\delta_s = f(t_s)^{-1}$ , we have

$$\begin{aligned} \mathbb{E}[N_t^{\mathcal{E}}(a_i) - C(a_i)] &\leq 2N_{i,*}(s') + \frac{1}{f(M)}MN_{i,*}(s') \\ &\leq \frac{120}{\Delta(a)^2} \left( \log(2Kt \log^2(t)) + 4 \log \left( \frac{48 \log(2Kt \log^2(t))}{\Delta(a)^2} \right) \right). \end{aligned}$$

So

$$\mathbb{E}[N_t(a_i)] \leq \mathbb{E}[C(a) + N_t^{\bar{\mathcal{E}}}(a_i)] + \frac{120}{\Delta(a)^2} \left( \log(2Kt \log^2(t)) + 4 \log \left( \frac{48 \log(2Kt \log^2(t))}{\Delta(a)^2} \right) \right).$$

where

$$\begin{aligned} \sum_{a \neq a_*} \mathbb{E}[C(a) + N_t^{\bar{\mathcal{E}}}(a_i)] &\leq M \log(K) + K + \frac{1}{f(M)}MK + \frac{3}{2}K \\ &\leq M \log(K) + \frac{5}{2}K. \end{aligned}$$

Finally if additionally  $M_s = |\mathcal{A}_A|$ , this bound improves to

$$\begin{aligned} \mathbb{E}[N_t(a_i)] &\leq \mathbb{E}[N_t^{\bar{\mathcal{E}}}(a_i)] + N_{*,i}(s') + 1 \\ &\leq \frac{5}{3} + \frac{48}{\Delta(a)^2} \left( \log(2Kt \log^2(t)) + 4 \log \left( \frac{48 \log(2Kt \log^2(t))}{\Delta(a)^2} \right) \right). \end{aligned}$$

□

## C Additional experiment

The winning probability is set to 0.95. All sub-optimal arms are identical

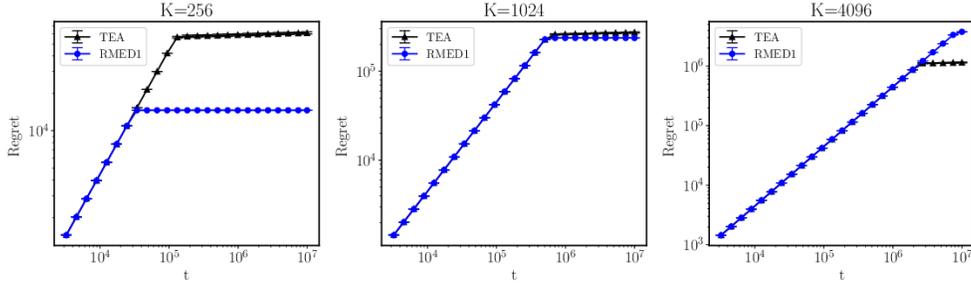


Figure 4: Comparison with identical gaps of 0.9. The results are averaged over 20 repetitions of the experiment.