IMPROVED ALGORITHMS FOR REPLICABLE BANDITS

Anonymous authors
Paper under double-blind review

ABSTRACT

This work is motivated by the growing demand for reproducible machine learning. We study the stochastic multi-armed bandit problem, where the algorithm's sequence of actions is, with a high probability, not affected by the randomness of the dataset. Existing algorithms require a regret scale of $O(K^3)$, which increases much faster than the number of actions (or "arms"), denoted as K. We introduce an algorithm with a distribution-dependent regret of $O(K)$ when the suboptimality gaps for each arm are within a constant factor. Furthermore, we propose another algorithm, which not only achieves a regret of $O(K)$ but also boasts a distribution-independent regret of $O(K^{1.5}\sqrt{T \log T})$. Additionally, we propose an algorithm for the linear bandit with regret of $O(d)$, which is linear in the dimension of associated features, denoted as d, and it is independent of K. For the analysis of these algorithms, we offer a principled approach to limiting the probability of non-replication, which clarifies the steps that existing research has implicitly followed.

1 INTRODUCTION

The *multi-armed bandit (MAB)* problem is one of the most well-known instances of sequential decision-making problems in uncertain environments, which can model various real-world scenarios. The problem involves conceptual entities called *arms*, of which there are a total of K. At each round $t = 1, 2, \ldots$, the forecaster selects one of the K arms and receives a corresponding reward. The forecaster's objective is to maximize the cumulative reward over these rounds. Maximizing this cumulative reward is equivalent to minimizing *regret*, the difference between the forecaster's cumulative reward and the reward of the best arm. The initial investigation of this problem took place within the field of statistics (Thompson, 1933; Robbins, 1952). In the past two decades, the machine learning community has conducted extensive research in this area, driven by numerous applications, including website optimization, A/B testing, and the formulation of meta-algorithms for algorithmic procedures (Auer et al., 2002; Li et al., 2010; Komiyama et al., 2015; Li et al., 2017).

Several algorithms have proven to be effective. Notably, the upper confidence bound (UCB, Lai & Robbins, 1985; Auer et al., 2002) and Thompson sampling (TS, Thompson, 1933) are widely recognized. Research has shown that these algorithms are asymptotically optimal (Cappé et al., 2013; Agrawal & Goyal, 2012; Kaufmann et al., 2012) in terms of their regret, meaning that these efficient algorithms exploit accumulated reward information to the fullest extent possible.

1.1 REPLICABILITY

One possible drawback of such efficiency is the algorithm's stability when dealing with small changes in the dataset, which can make replicating results challenging. To illustrate this, consider the following example:

Example 1. (Crowdsourcing (Abraham et al., 2013; Tran-Thanh et al., 2014)) Imagine a company conducting a crowd-based A/B testing with K items. In this scenario, each round t corresponds to a worker visiting their website, and each reward represents the feedback provided by the worker, such as a five-star rating. The key statistic of interest here is the mean score given by the workers. By using UCB or TS to allocate an item to each user, the system can quickly eliminate unpopular items from the candidate set. Although the company aims to publish the results, it is reluctant to disclose all the specific details of the setup. Therefore, it provides the experimental protocol along with summary statistics.

Table 1: Comparison of regret bounds in the K-armed and linear bandit problems. Means $\{\mu_i\}_{i\in[K]}$ are sorted in descending order, $\Delta_i = \mu_1 - \mu_i$ is the suboptimality gap of the arm i, and $\Delta = \Delta_2$. Algorithm 2 is referenced multiple times, implying that it has the smallest regret among the bounds. The lower bound is derived for the two-armed case. $\tilde{O}$ omits a polylog factor in d, K, T.

Problem	Esfandiari et al. (2023a)	This work
K-armed	$O\left(\sum_{i=2}^K \frac{1}{\Delta_i} \frac{K^2 \log T}{\rho^2}\right)$ (= $\boldsymbol{O(K^3)}$)	$O\left(\sum_{i=2}^K \frac{\Delta_i}{\Delta^2} \frac{\log T}{\rho^2}\right)$ (= $\boldsymbol{O(K)}$, REC (Theorem 3) and RSE (Theorem 4)) $O\left(\sum_{i=2}^K \frac{1}{\Delta_i} \frac{K^2 \log T}{\rho^2}\right)$ (RSE, Theorem 4) $O\left(\frac{K}{\rho}\sqrt{KT \log T}\right)$ (RSE, Theorem 4) $\Omega\left(\left(\frac{1}{\Delta}\right) \max\left(\log T, \frac{1}{\rho^2 \log((\rho\Delta)^{-1})}\right)\right)$ (Lower bound, Theorem 5)
Linear	$\tilde{O}\left(\frac{K^2\sqrt{dT}}{\rho^2}\right)$ (= $\boldsymbol{O(K^2)}$)	$O\left(\frac{d \log T}{\Delta^2 \rho^2}\right)$ (RLSE, Theorem 8) $O\left(\frac{K}{\rho}\sqrt{dKT \log T}\right)$ (RLSE, Theorem 8)

In Example 1, the original dataset is not disclosed, making it impossible for an external institution to perfectly replicate the experiment. To address this lack of guarantee due to the vague dataset specification, we focus on the algorithm's *stability*. Broadly speaking, a stable algorithm is robust against minor changes in the dataset. Notable categories of stability encompass differential privacy (Dwork et al., 2014), worst-case and average sensitivity (Varma & Yoshida, 2021), and pseudo-determinicity (Gat & Goldwasser, 2011). Among these notions of stability, we consider the *replicability* (Impagliazzo et al., 2022) in this work. In simple terms, a replicable algorithm exhibits almost identical behavior on two datasets sharing the same data-generating process. This concept aligns well with Example 1, where the data-generating process is clearly defined as a material method, whereas the dataset itself remains undisclosed.

Another advantage of promoting replicability in the context of sequential learning is its relation to statistical testing. It is well-known that the standard frequentist confidence interval no longer holds for the results of the multi-armed bandit problem because such an adaptive algorithm violates the assumption of statistical testing that the number of samples is fixed. In general, mean statistics derived from the multi-armed bandit algorithm are downward biased (Xu et al., 2013; Shin et al., 2019), and this bias persists even for a large sample scheme (Lai & Wei, 1982), rendering the use of standard confidence intervals ineffective even for asymptotics (Deshpande et al., 2018). Replicability is one of the best methods to address such bias because it forces the algorithm to exhibit identical behavior in multiple runs with different datasets sharing a common underlying data-generating process.

1.2 Our Contributions

The concept of replicability in learning was formalized by Impagliazzo et al. (2022). As far as we know, Esfandiari et al. (2023a) is the sole work that provides algorithms studying replicability for the multi-armed bandit problem. Roughly speaking, the regret of the algorithms therein is $O(K^2/\rho^2)$ times larger than that of non-replicable algorithms, which have $O(K)$ regret, such as UCB and TS. Here, K is the number of arms, and ρ is the probability of non-replication. Although the additional factor might appear necessary for the cost of replicability, it is somewhat disappointing because an $O(K^3)$-regret algorithm does not scale well with a moderate value of K. Upon closer examination of the problem, we discovered that K^2 factor can be eliminated, while the $1/\rho^2$ factor remains essential. The algorithmic contributions of this work are outlined as follows.

- We first introduce the general framework for bounding the probability of nonreplication.
- We introduce the Replicable Explore-then-Commit (REC) algorithm, which is the first replicable K-armed bandit algorithm with a regret bound of $O(K)$ when the suboptimality gaps for each arm are within a constant factor.

- While REC has an $O(K)$ bound, its distribution-dependent multiplicative factor may exceed that of existing algorithms in certain cases. To deal with this issue, we introduce the Replicable Successive Elimination (RSE) algorithm whose regret bound is the minimum of those of REC and the existing algorithms. Furthermore, we establish the distribution-independent regret bound for the replicable K-armed bandit problem.
- We derive the first lower bound for the replicable K-armed bandit problem, implying the necessity of the $1/\rho^2$ factor.
- Furthermore, we consider the linear bandit problem, in which each of the K arms is associated with $d < K$ features. We show that a straightforward modification of RSE yields an algorithm with a regret bound of $O(d)$, indepenent of K.

A comparison of existing algorithms and our algorithms is summarized in Table 1.

2 Problem Setup

We consider the finite-armed stochastic bandit problem with T rounds. At each round t, the forecaster who adopts an algorithm selects one of the arms $I_t \in [K] := \{1, 2, 3, \ldots, K\}$ and receives the corresponding reward r_t. Each arm $i \in [K]$ has an (unknown) mean parameter $\mu_i \in \mathbb{R}$. Here, $\mu_i \in [a, b]$ for some $a, b \in \mathbb{R}$ and let $S = b - a$. For ease of discussion, we assume $S = 1$, but all our results can be easily generalized[1] to any $S > 0$. The reward at round t is $r_t = \mu_{I_t} + \eta_t$, where η_t is a σ-subgaussian random variable that is independently drawn at each round.[2] The subgaussian assumption is quite general that is not limited to Gaussian random variables. Any bounded random variable is subgaussian, and thus, it is capable of representing binary events (Yes/No) and ordered choice (e.g., 5-star rating). For subgaussian random variables, the following inequality holds.

Lemma 1 (Concentration inequality)**.** *Let $X_1, X_2, \ldots, X_N$ be N independent (zero-mean) σ-subgaussian random variables, and $\hat{\mu}_N = (1/N)\sum_i X_i$ be the empirical mean. Then, we have*

$$\mathbb{P}\left[|\hat{\mu}_N| \geq s\right] \leq 2\exp\left(\frac{-s^2 N}{2\sigma^2}\right). \tag{1}$$

For ease of discussion, we assume the mean reward of each arm is distinct. In this case, we can assume $\mu_1 > \mu_2 > \cdots > \mu_K$ without loss of generality. Of course, an algorithm cannot exploit this ordering. A quantity called regret is defined as follows:

$$\mathrm{Regret}(T) := \sum_{t\in[T]} (\mu_1 - \mu_{I_t}) = \sum_{i\in[K]} \Delta_i N_i(T),$$

where $\Delta_i = \mu_1 - \mu_i$ and $N_i(T)$ is the number of draws on arm i during the T rounds. We also denote $\Delta = \min_{i\geq 2} \Delta_i = \Delta_2$. The performance of an algorithm is measured by the expected regret, where the expectation is taken over (hypothetical) multiple runs. Before discussing the replicability, we formalize the notion of dataset in a sequential learning problem because the reward r_t in the aforementioned procedure is drawn adaptively upon the choice of the arm I_t. The fact that each noise term η_t is drawn independently enables us to reformulate the problem as follows:

Definition 1. (Dataset) The process of the multi-armed bandit problem is equivalent to the following: First, draw a matrix $(r_{i,n})_{i\in[K], n\in[T]}$, where $r_{i,n} = \mu_i + \eta_{i,n}$ and $\eta_{i,n}$ is a σ-subgaussian random variable. Second, Run a multi-armed bandit problem. Here, r_t is the $(I_t, N_{I_t}(t))$-entry of the matrix. We call this matrix a *dataset* and denote it as $\mathcal{D}$. We call $(\mu_i)_{i\in[K]}$ a data-generating process or a model.

Following Esfandiari et al. (2023a), we consider the class of replicable algorithms that, with high probability, gives exactly the same sequence of selected arms for two independent runs.

Definition 2. (ρ-replicability, Impagliazzo et al. (2022); Esfandiari et al. (2023a)) For $\rho \in [0, 1]$, an algorithm is ρ-replicable if,

$$\mathbb{P}_{U,\mathcal{D}^{(1)},\mathcal{D}^{(2)}}\left[(I_1^{(1)}, I_2^{(1)}, \ldots, I_T^{(1)}) = (I_1^{(2)}, I_2^{(2)}, \ldots, I_T^{(2)})\right] \geq 1 - \rho, \tag{2}$$

[1] In particular, ϵ_p of Algorithms 1–3 should be replaced with $\epsilon_p = S2^{-p}$ and all the other parts remain the same.

[2] A random variable η is σ-subgaussian if $\mathbb{E}[\exp(\lambda\eta)] \leq \exp(\sigma^2\lambda^2/2)$ for any λ. For example, a zero-mean Gaussian random variable with variance σ^2 is σ-subgaussian.

where U represents the internal randomness, and $\mathcal{D}^{(1)}, \mathcal{D}^{(2)}$ are the two datasets that are drawn from the same data-generating process $\{\mu_i\}_{i\in[K]}$.

Here, we may consider U as a sequence of uniform random variables on $[0,1]$ that the algorithm can use to control its behavior. For an algorithm to be replicable, the use of such random variables is crucial. The value ρ corresponds to the probability of misreplication. The smaller ρ is, the more likely the sequence of actions is replicated. By definition, any algorithm is 1-replicable, and no nontrivial algorithm is 0-replicable.[3] In this paper, we consider $\rho \in (0,1)$ as an exogenous parameter, and our goal is to minimize the regret subject to the ρ-replicability.

3 General Bound of the Probability of Non-replication ρ

It is not very difficult to see that a standard bandit algorithm, such as UCB, lacks replicability. UCB, in each round, compares the UCB index of the arms, and thus, a minor change in the dataset can alter the sequence of draws $I_1, I_2, \dots, I_T$. Thus, designing a replicable algorithm must deviate significantly from standard bandit algorithms. This section presents a general framework for bounding non-replicable probability in the multi-armed bandit problem. We believe that this framework can be applied to many other sequential learning problems. First, a replicable algorithm should limit its flexibility by introducing phases.

Definition 3. A set of phases is a consecutive partition of rounds $[T]$. Namely, phase p is a consecutive subset of $[T]$, and the first round of phase $p+1$ follows the last round of phase p, and each round belongs to one of the phases. We define P to be the number of phases.

The sequence of draws $I_1, I_2, \dots, I_T$ is only allowed branch at the end of each phase, which we formalize in the following definition.

Definition 4. (Randomness) The randomness U consists of the one for each phase. Namely, $U = (U_1, U_2, \dots, U_P)$.

Definition 5. (Good events, decision variables, and decision points) We call the end of the final round of each phase a decision point, which we denote as T_p. For each $p \in [P]$, we consider the history $\mathcal{H}_p$ to be the set of all results up to the final round T_p of phase p. Namely,

$$\mathcal{H}_p = (I_1, r_1, I_2, r_2, \dots, I_{T_p}, r_{T_p}) \cup (U_1, U_2, \dots, U_p). \tag{3}$$

Each phase p is associated with good event $\mathcal{G}_p(\mathcal{H}_p)$, which is a binary function of $\mathcal{H}_p$. Each phase p is associated with random variables that are called *decision variables* d_p. Decision variables take discrete values and are functions of $\mathcal{H}_p$. Moreover, the sequence of draws on the next phase $\{I_{T_p+1}, I_{T_p+2}, \dots, I_{T_{p+1}}\}$ is uniquely determined by the decision variables $d_1, d_2, \dots, d_p$.

Intuitively speaking, the good events correspond to the concentration of statistics with its probability we can bound with concentration inequalities (by Lemma 1). The set of decision variables uniquely determines the sequence of draws. Note that each phase can be associated with more than one decision variable. To obtain intuition, we consider the following example.

Example 2. (A replicable elimination algorithm, Alg 2. in Esfandiari et al. (2023a)) At the end of each phase, the algorithm obtains an empirical estimate of μ_i for each arm. It tries to eliminate suboptimal arm i, and the corresponding decision variable is

$$d_{p,i} = \mathbf{1}[\max_j \mathrm{LCB}_j(p) \ge \mathrm{UCB}_i(p)], \tag{4}$$

where $\mathrm{UCB}_i(p), \mathrm{LCB}_i(p)$ are the (randomized) upper/lower confidence bounds of the arm i at phase p. Here, $\mathbf{1}[\mathcal{E}]$ is 1 if event $\mathcal{E}$ holds or 0 otherwise. Under good events, by randomizing the confidence bounds with U_p, it bounds the probability of non-replication of each decision variable.

In the following, we defined the non-replication probability for each component.

Definition 6. (Probability of bad event) Let $\rho_p^G = \mathbb{P}\left[\mathcal{G}_p^c\right]$, where $\mathcal{G}^c$ is a complement event of $\mathcal{G}$.

[3] A 0-replicable algorithm draws an identical sequence of arms for almost all fixed U.

Definition 7. (Non-replication probability of a decision variable) Let $d_{p,i}$ be the i-th decision variable at phase p. Its non-replication probability $\rho_{p,i}$ is defined as

$$\rho^{(p,i)} := \mathbb{P}_{U,\mathcal{D}^{(1)},\mathcal{D}^{(2)}}\left[d_p^{(1)} \neq d_p^{(2)} \,\middle|\, \bigcap_{p'=1}^{p-1}\left\{d_{p'}^{(1)} = d_{p'}^{(2)}, \mathcal{G}_{p'}^{(1)}, \mathcal{G}_{p'}^{(2)}\right\}, \mathcal{G}_p^{(1)}, \mathcal{G}_p^{(2)},\right] \tag{5}$$

where we use superscripts (1) and (2) for the corresponding variables on the two datasets $\mathcal{D}^{(1)}, \mathcal{D}^{(2)}$.

Theorem 2. (Replicability of an algorithm) *An algorithm is ρ-replicabile with*

$$\rho \leq 2\sum_p \rho_p^G + \sum_{p,i} \rho^{(p,i)}. \tag{6}$$

In summary, Theorem 2 enables us to decompose the non-replication probability ρ into the sum of the non-replication probabilities due to the bad events ($\mathcal{G}^c$) and decision variables ($\rho^{(p,i)}$).

3.1 Comparision of Algorithms in view of Decision Variables

The smaller the non-replication probability of each decision variable is, the higher the cost the algorithm must pay to guarantee it. Assuming all $\rho^{(p,i)}$ are equal, $\rho^{(p,i)} \sim \rho/|\{\rho^{(p,i)}\}|$, where $|\{\rho^{(p,i)}\}|$ is the number of decision variables. Algorithm 2 in Karbasi et al. (2023) uses the decision variables for eliminating each suboptimal arm. Therefore, it has $\tilde{O}(K)$ decision variables, implying that each of the non-replication probability must be $\tilde{O}(\rho/K)$. As a result, these algorithms has an $O(K^2/\rho^2)$ factor in the leading term of the regret. In our Replicable Explore-the-Commit (REC) algorithm, we use decision variables representing whether or not to finish the exploration process, which means that we only need $O(1)$ (in fact, only one of them is effective!) decision variables with its non-replication level $O(\rho)$, and as a result, it has $O(1/\rho^2)$ factor in the leading term of the regret, which dramatically reduces the dependence on K.

4 An $O(K)$-regret Algorithm for K-armed Bandit Problem

This section introduces the Replicable Explore-then-Commit (REC, Algorithm 1), an $O(K)$-regret algorithm for the K-armed bandit problem. This algorithm consists of multiple exploration phases and an exploitation period. The last round of each phase is a decision point, where the algorithm decides whether it terminates the exploration period or not. For this aim, it utilizes the minimum suboptimality gap estimator $\hat{\Delta}(p) = \max_i \hat{\mu}_i(p) - \max_i^{(2)} \hat{\mu}_i(p)$, where $\max_i^{(2)}$ denotes the second largest element. This algorithm involves a single uniform random variable $U_p \sim \mathrm{Unif}(0,1)$ for each phase p.

Let $\mathrm{Conf}(p) := \epsilon_p/C_\rho$ and $\epsilon_p = 2^{-p}$. Here, the universal constant C_ρ is clarified later in Theorem 3. At phase p, if the algorithm is in the exploration period, we draw each arm up to

$$N_p := 4\sigma^2 \frac{\log(KPT)}{(\mathrm{Conf}(p))^2} = O\left(4^p \log T\right) \tag{7}$$

times, where $P = \min_p\{N_p \geq T\} = O(\log T)$ is the maximum number of phases. Lemma 1 with $s = \mathrm{Conf}(p)$ implies that, with probability at least $1 - 1/PT$, we have $|\Delta_{ij}(p) - \Delta_{ij}| \leq \mathrm{Conf}(p)$, for each gap $\Delta_{ij} = |\mu_i - \mu_j|$ and its empirical estimator $\hat{\Delta}_{ij}(p) = |\hat{\mu}_i(p) - \hat{\mu}_j(p)|$.

Algorithm 1: Replicable Explore-then-Commit (REC)

```
// Exploration period
for p = 1, 2, ..., P do
    ε_p = 2^{-p} ;
    Draw shared random variable U_p ~ Unif(0, 1);
    Draw each arm for N_p times;
    If Δ̂(p) ≥ (2+U_p)ε_p / ρ, then fix the estimated best arm î* ∈ arg max_i μ̂_i(p) and break the
      loop. Note that Δ̂(p) > 0 implies |A_{p+1}| = 1.
// Exploitation period
Draw arm î* for the rest of the rounds.
```

The following theorem guarantees the replicability and performance of Algorithm 1.

Theorem 3. *Let $C_\rho \geq 9/4$. Assume that $\rho \leq 1/2$ and $T \geq 36K/\rho$. Then, Algorithm 1 is ρ-replicable and the following regret bound holds:*

$$\mathbb{E}[\mathrm{Regret}(T)] = O\left(\sum_{i=2}^{K} \frac{\Delta_i}{\Delta^2} \frac{\log T}{\rho^2}\right). \tag{8}$$

Remark 1. (Use of explore-then-commit) Esfandiari et al. (2023a) briefly remarked the possibility of the use of the explore-then-commit strategy and sketched an $O(T \sum_i \Delta_i)$ regret, from which we significantly improved the regret as well as getting rid of the assumption of known Δ.

5 A Generalized Algorithm for K-armed Bandit Problem

Algorithm 2: Replicable Successive Elimination (RSE)

1 Initialize the candidate set $\mathcal{A}_1 = [K]$;
2 **for** $p = 1, 2, \ldots, P$ **do**
3 $\quad \epsilon_p = 2^{-p}$;
4 $\quad$ Draw shared random variables $U_{p,i} \sim \mathrm{Unif}(0,1)$ for $i = 0, 1, 2, \ldots, K$;
5 $\quad$ Draw each arm in $\mathcal{A}_p$ up to N_p times;
6 $\quad \mathcal{A}_{p+1} \leftarrow \mathcal{A}_p$;
7 $\quad$ **if** $\hat{\Delta}(p) \geq \frac{(2+U_{p,0})\epsilon_p}{\rho_a}$ **then**
8 $\quad\quad \mathcal{A}_{p+1} = \{\arg\max_i \hat{\mu}_i(p)\}$; $\quad \triangleright$ Eliminate all arms except for one. By definition of $\hat{\Delta}(p) > 0$, $|\mathcal{A}_{p+1}| = 1$ holds.
9 $\quad$ **for** $i \in \mathcal{A}_{p+1}$ **do**
10 $\quad\quad$ **if** $\hat{\Delta}_i(p) \geq \frac{(2+U_{p,i})\epsilon_p}{\rho_e}$ **then**
11 $\quad\quad\quad \mathcal{A}_{p+1} \leftarrow \mathcal{A}_{p+1} \setminus \{i\}$; $\quad \triangleright$ Eliminate arm i.

Although Algorithm 1 improves the existing $O(K^3)$ regret bound in terms of the dependency to K, there are two potential drawbacks: First, when we further compare the ratio between REC's regret bound with the existing regret bound of

$$O\left(\sum_{i=2}^{K} \frac{1}{\Delta_i} \frac{K^2 \log T}{\rho^2}\right) \quad \text{(Alg.2 in Esfandiari et al. (2023a))}, \tag{9}$$

the ratio between the two distribution-dependent bounds is bounded as (equation 8/equation 9) $\leq ((\Delta_K)/(K\Delta))^2$, which implies that REC may be inferior if the minimum suboptimality gap Δ is extremely small (i.e., $\Delta \ll (\Delta_K)/K$). Second, it does not have a distribution-independent regret bound. To address these issues, we introduce Algorithm 2, which generalizes Algorithm 1. Unlike Algorithm 1, it keeps the list of remaining arms $\mathcal{A}_p$ that it draws. At the end of each phase, it attempts to eliminate all but one arm (Line 7). If that fails, it attempts to eliminate each arm i (Line 10). Here, $\hat{\Delta}_i(p) = \max_j \hat{\mu}_j(p) - \hat{\mu}_i(p)$ be the estimated suboptimality gap. Here, the hyperparameters ρ_a, ρ_e therein determine the confidence level for elimination. One can confirm that Algorithm 1 is a specialized version of Algorithm 2 where $(\rho_a, \rho_e) = (\rho, 0)$, where $\rho_e = 0$ implies the corresponding elimination never occurs. Here, eliminating all but one arm is equivalent to switching to the exploration period. However, when $\rho_e > 0$, it attempts to eliminate each arm as well. The following theorem guarantees the replicability and the regret of Algorithm 2.

Theorem 4. *Let $C_\rho \geq 9/4$, $\rho_a = \rho/2$, and $\rho_e = \rho/(2\max(K-2, 2))$. Assume that $\rho \leq 1/2$ and $T \geq 36K/\rho$. Then, Algorithm 2 is ρ-replicable. Moreover, the following three regret bounds hold:*

$$\mathbb{E}[\mathrm{Regret}(T)] = O\left(\sum_{i=2}^{K} \frac{\Delta_i}{\Delta^2} \frac{\log T}{\rho^2}\right), \quad \textit{(same as equation 8)} \tag{10}$$

$$\mathbb{E}[\mathrm{Regret}(T)] = O\left(\sum_{i=2}^{K} \frac{K^2 \log T}{\Delta_i \rho^2}\right), \quad \textit{(same as equation 9)} \tag{11}$$

$$\mathbb{E}[\text{Regret}(T)] = O\left(\frac{K}{\rho}\sqrt{KT\log T}\right). \quad \textit{(distribution-independent regret)} \tag{12}$$

In other words, Algorithm 2 has the best bound of Algorithm 1 and existing algorithms. Moreover, this algorithm is the first replicable algorithm that has a distribution-independent regret bound in the K-armed bandit problem.

6 Regret Lower Bound for Replicable Algorithms

This section provides the regret lower bound for K-armed bandit algorithms. Following the literature, we consider the class of uniformly good algorithms. Intuitively speaking, an algorithm is uniformly good if it works with any model $(\mu_1, \mu_2, \dots, \mu_K)$.

Definition 8 (Uniformly good, Lai & Robbins (1985))**.** An algorithm is *uniformly good*, if for any $a > 0$ and for any model $(\mu_1, \mu_2, \dots, \mu_K)$, there exists a function $R(T) = o(T^a)$ such that

$$\mathbb{E}[\text{Regret}(T)] \leq R(T). \tag{13}$$

Theorem 5. *Consider a two-armed bandit problem where reward is drawn from* $\text{Bernoulli}(\mu_i)$ *for each arm* $i = 1, 2$ *with mean parameters* μ_1, μ_2*. Consider an algorithm that is uniformly good and* ρ*-replicable. Then, for any* $\Delta > 0$*, there exists an instance* (μ_1, μ_2) *with* $\Delta = |\mu_1 - \mu_2|$ *such that the regret of any* ρ*-replicable bandit algorithm is lower-bounded as*

$$\mathbb{E}[\text{Regret}(T)] = \Omega\left(\frac{1}{\rho^2 \Delta \log((\rho\Delta)^{-1})}\right). \tag{14}$$

This bound implies that REC, RSE, and the algorithms in Esfandiari et al. (2023a) are optimal up to a polylogarithmic factor for two-armed bandit problem.

Remark 2. It is well-known that another lower bound

$$\mathbb{E}[\text{Regret}(T)] = \Omega\left(\frac{\log T}{\Delta}\right) \tag{15}$$

holds for a uniformly good algorithm (c.f., Theorem 1 in Lai & Robbins (1985)). Therefore, a lower bound for a ρ-replicable uniformly good algorithm is the maximum of equation 14 and equation 15.

The absence of $\log T$ in equation 14 appears to be essential. The factor $\log T$ is derived from the uniformly good property of a bandit algorithm. However, the cost of replicability and the cost of uniform goodness are not necessarily compounded. Any ρ-replicable algorithm (if it is not uniformly good) that frequently selects the best arm should maintain the lower bound of equation 14.

7 An Algorithm for Linear Bandit Problem

Next, we consider the linear bandit problem, a special version of the K-armed bandit problem where associated information is available. In this problem, each arm $i \in [K]$ is associated with a d-dimensional feature vector $\boldsymbol{x}_i \in \mathbb{R}^d$ and the reward r_t of choosing an arm I_t is $\boldsymbol{x}_{I_t}^\top \boldsymbol{\theta} + \eta_t$, where $\boldsymbol{\theta}$ is (unknown) shared parameter vector, and η_t is a σ-subgaussian random variable. Namely, the mean $\mu_i = \boldsymbol{x}_i^\top \boldsymbol{\theta}$ can be estimated via known feature $\boldsymbol{x}_i$ and unknown shared coefficients $\boldsymbol{\theta}$. Without loss of generality, we assume $\text{span}(\{x_i\}_{i=1}^K) = \mathbb{R}^d$.

We introduce the replicable linear successive elimination (RLSE). Similarly to RSE (Algorithm 2), this algorithm is elimination-based. The main innovation here is to use the G-optimal design that explores all dimensions in an efficient way. Namely,

Definition 9. (G-optimal design) For $\mathcal{A}_p \subseteq [K]$, let π be a distribution over $\mathcal{A}_p$. Let

$$\boldsymbol{V}(\pi) = \sum_{i \in \mathcal{A}_p} \pi(\boldsymbol{x}_i)\boldsymbol{x}_i\boldsymbol{x}_i^\top, \quad g(\pi) = \max_{i \in \mathcal{A}_p} ||\boldsymbol{x}_i||^2_{\boldsymbol{V}(\pi)^{-1}}. \tag{16}$$

A distribution π^* is called a G-optimal design if it minimizes g, i.e., $\pi^* \in \arg\min_\pi g(\pi)$.

We use the following well-known results for G-optimal designs (See, e.g., Section 21 of Lattimore & Szepesvári (2020)).

Lemma 6 (Kiefer-Wolfowitz)**.** *A G-optimal design π^* satisfies $g(\pi^*) = d$.*

In this paper, we assume the availability of a constant approximation of optimal design $\pi^{*,\mathrm{app}} = \pi^{*,\mathrm{app}}(\mathcal{A}_p)$ with $g(\pi^{*,\mathrm{app}}) \leq 2d$. An explicit construction of such an approximated G-optimal design is found in the literature (e.g., Lemma 7 of (Esfandiari et al., 2023a)). Given an oracle for an approximated G-optimal design, we define the allocation at phase p to be $N_i^{\mathrm{lin}}(p) = \lceil N^{\mathrm{lin}}(p)\pi_i^{*,\mathrm{app}} \rceil$,where

$$N^{\mathrm{lin}}(p) := \frac{16\sigma^2 d \log(|\mathcal{A}_p|PT)}{(\mathrm{Conf}(p))^2}. \tag{17}$$

Note that $\sum_i N_i^{\mathrm{lin}}(p) \leq N^{\mathrm{lin}}(p) + K$. We use the following lemma for the confidence bound (see e.g., Section 21.1 of Lattimore & Szepesvári (2020)):

Lemma 7 (Fixed-sample bound)**.** *Consider the estimator $\hat{\boldsymbol{\theta}}_p$ at the end of phase p. Then, with probability at least $1 - \frac{2}{PT}$, the following bound holds uniformly for any $i \in \mathcal{A}_p$:*

$$|\boldsymbol{x}_i^\top(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_p)| \leq \frac{\mathrm{Conf}(p)}{2}. \tag{18}$$

Letting $\hat{\mu}_i = \boldsymbol{x}_i^\top \hat{\boldsymbol{\theta}}_p$ and $\hat{\Delta}_{ij} = |\hat{\mu}_i - \hat{\mu}_j|$, equation 18 implies

$$|\Delta_{ij} - \hat{\Delta}_{ij}| \leq \mathrm{Conf}(p). \tag{19}$$

Apart from applying approximated G-optimal exploration, the algorithm closely mirrors the steps of RSE. A comprehensive description of RLSE can be found in Appendix A. The subsequent theorem assures both the replicability and regret of RLSE.

Theorem 8. *Let $C_\rho \geq 9/4$, $\rho_a = \rho/2$, and $\rho_e = \rho/(2\max(K-2, 2))$. Assume that $\rho \leq 1/2$ and $T \geq 36K/\rho$. Then, RLSE is ρ-replicable. Moreover, the following two regret bounds hold:*

$$\mathbb{E}[\mathrm{Regret}(T)] = O\left(\frac{d \log T}{\Delta^2 \rho^2}\right), \quad \textit{(an } O(d) \textit{ distribution-dependent bound)} \tag{20}$$

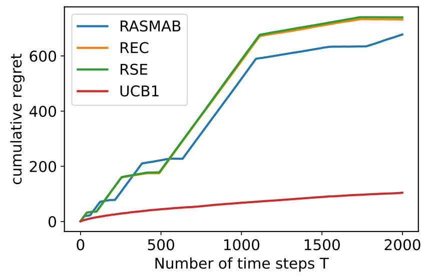
$$\mathbb{E}[\mathrm{Regret}(T)] = O\left(\frac{K}{\rho}\sqrt{dKT \log T}\right). \quad \textit{(distribution-independent bound)} \tag{21}$$

The first bound depends on the suboptimality gap Δ and is independent of K. The second bound is distribution-independent and is smaller than the existing bound by a $\sqrt{K}/\rho$ factor (c.f., Table 1).

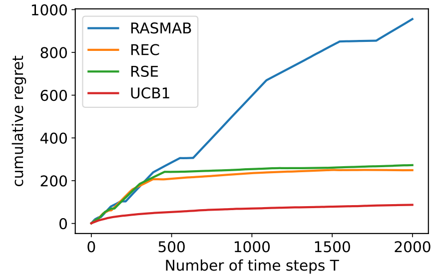
8 Simulation

We compared our REC (Algorithm 1) and RSE (Algorithm 2) with RASMAB (Algorithm 2 of Esfandiari et al. (2023a), "Replicable Algorithm for Stochastic Multi-Armed Bandits"). We did not include Algorithm 1 of Esfandiari et al. (2023a) because its regret bound is always inferior to RASMAB. Three models of K-armed Gaussian bandit problems were considered. To ensure fair comparison, as RASMAB relies on the Hoeffding inequality, we standardized the variance of the arms at 0.5. The results were averaged over 100 runs. The algorithms can be characterized as follows: REC eliminates all arms simultaneously, RASMAB eliminates each arm independently, and RSE incorporates both strategies. Theoretical results suggest that REC outperforms RASMAB provided that $\Delta_K/\Delta = o(K)$. We optimize C_ρ in REC and RSE and β of RASMAB for $\hat{\rho} = 0.3$ by using a grid search. Here, the empirical nonreplication probability $\hat{\rho}$ is obtained by bootstrapping. Namely, assume that the algorithm results in S different sequences of draws, where the corresponding number of occurrences for each sequence are $N_{(1)}, N_{(2)}, N_{(3)}, \dots, N_{(S)}$. By definition, $\sum_s N_{(s)} = 100$. Then, $\hat{\rho} := 1 - \sum_s (N_{(s)}/100)^2$.

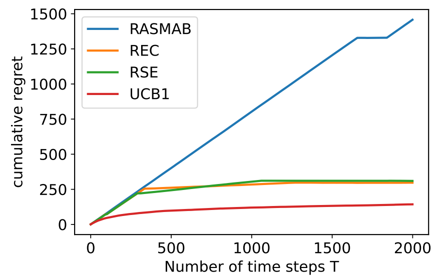
We set the mean parameters as follows: $\boldsymbol{\mu} = (0.1, 0.1, 0.8, 0.8, 0.9)$ for Model 1, $\boldsymbol{\mu} = (0.1, 0.1, 0.5, 0.5, 0.9)$ for Model 2, and $\boldsymbol{\mu} = (0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.9)$ for Model 3. The amount of regret is depicted in Figure 1. A lower regret signifies superior performance. As



Model 1



Model 2



Model 3

Figure 1: Regret of algorithms. The horizontal axis indicates the number of rounds t from 1 to T, whereas the vertical axis indicates $\mathrm{Regret}(t)$. Results of REC and RSE in Model 1 are very similar.

a non-replicable algorithm, UCB1 naturally outperforms all other replicable algorithms. Model 1, with a large $\Delta/\Delta_2 = 0.8/0.1 = 8$, is designed to favor RASMAB while Model 3, having $\Delta/\Delta_2 = 1$, is tailored to favor REC. In Models 2 and 3, REC and RSE significantly surpass RASMAB, indicating their success in replicably selecting the optimal arm. In Model 1, RASMAB marginally outperforms REC and RSE. As RASMAB is designed to eliminate each arm independently, the early elimination of clearly suboptimal arms (specifically, arms 1 and 2) decreases cumulative regret. These findings align with our theoretical results. Additional simulations employing theoretically-chosen hyperparameters are in the appendix.

9 Related Work

Replicability was introduced by Impagliazzo et al. (2022) and they designed replicable algorithms for answering statistical queries, identifying heavy hitters, finding median, and learning halfspaces. Since then, replicable algorithms have been studied for bandit problems (Karbasi et al., 2023), reinforcement learning (Eaton et al., 2023), and clustering (Esfandiari et al., 2023b). The equivalence of various stability notions, including replicability and differential privacy (Dwork et al., 2014) was shown for a broad class of statistical problems (Bun et al., 2023). However, the equivalence therein does not necessarily guarantee an efficient conversion. Kalavasis et al. (2023) considered a relaxed notion of replicability. Note also that there are several relevant works Dixon et al. (2023); Chase et al. (2023) that study a different notion of applicability.

Prior to the introduction of the replicable bandit algorithm, the batched bandit problem was considered (Auer & Ortner, 2010; Cesa-Bianchi et al., 2013; Komiyama et al., 2013; Perchet et al., 2016; Gao et al., 2019; Esfandiari et al., 2021). In this problem, the algorithm needs to determine the sequence of draws at the beginning of each batch. Existing replicable bandit algorithms in Esfandiari et al. (2023a), as well as our algorithms, adopt phased approaches, and one can find similarities in the algorithmic design. In particular, Perchet et al. (2016) considered the two-armed batched bandit problem. They utilized the fact that the termination of the exploration phases in EtC only occurs in a fixed number of rounds, a concept that we also utilize in the proof of our algorithms. However, their algorithm does not guarantee ρ-replicability for $\rho < 1/2$. Our REtC extends their results by introducing a randomized confidence level to guarantee a further level of applicability. Furthermore, our RSE generalizes both EtC and successive elimination (Gao et al., 2019; Esfandiari et al., 2023a) in a replicable way, and we recover essentially the same performance towards a large value of ρ.

Several different notions of stability have been explored in the context of sequential learning. For instance, robustness against corrupted distributions has been examined in the multi-armed bandit problem (Kim & Lim, 2016; Gajane et al., 2018; Kapoor et al., 2019; Basu et al., 2022). Differential privacy has also been considered in this context (Shariff & Sheffet, 2018; Basu et al., 2019; Hu & Hegde, 2022). Differential privacy considers the change of decision against the change of a single data point, whereas in the replicable bandits, we have more than one change of data points between two datasets that are generated from the identical data-generating process. Recent work (Dong & Yoshida, 2023) showed that an algorithm with a low average sensitivity (Varma & Yoshida, 2021) can be transformed to an online learning algorithm with low regret and inconsistency in the random-order setting, and hence in the stochastic setting.

REFERENCES

Ittai Abraham, Omar Alonso, Vasilis Kandylas, and Aleksandrs Slivkins. Adaptive crowdsourcing algorithms for the bandit survey problem. In *Proceedings of the 26th Annual Conference on Learning Theory (COLT)*, pp. 882–910, 2013.

Shipra Agrawal and Navin Goyal. Analysis of Thompson sampling for the multi-armed bandit problem. In *Proceedings of the 25th Conference on Learning Theory (COLT)*, pp. 39.1–39.26, 2012.

Peter Auer and Ronald Ortner. UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Period. Math. Hung.*, 61(1-2):55–65, 2010. doi: 10.1007/S10998-010-3055-6. URL `https://doi.org/10.1007/s10998-010-3055-6`.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.

Debabrota Basu, Christos Dimitrakakis, and Aristide C. Y. Tossou. Differential privacy for multi-armed bandits: What is it and what is its cost? *CoRR*, abs/1905.12298, 2019.

Debabrota Basu, Odalric-Ambrym Maillard, and Timothée Mathieu. Bandits corrupted by nature: Lower bounds on regret and robust optimistic algorithm. *CoRR*, abs/2203.03186, 2022.

Mark Bun, Marco Gaboardi, Max Hopkins, Russell Impagliazzo, Rex Lei, Toniann Pitassi, Satchit Sivakumar, and Jessica Sorrell. Stability is stable: Connections between replicability, privacy, and adaptive generalization. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing (STOC)*, pp. 520–527, 2023.

Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. Kullback–Leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, 41(3):1516 – 1541, 2013.

Nicolò Cesa-Bianchi, Ofer Dekel, and Ohad Shamir. Online learning with switching costs and other adaptive adversaries. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pp. 1160–1168, 2013. URL `https://proceedings.neurips.cc/paper/2013/hash/9cf81d8026a9018052c429cc4e56739b-Abs`

Zachary Chase, Shay Moran, and Amir Yehudayoff. Replicability and stability in learning, 2023.

Yash Deshpande, Lester W. Mackey, Vasilis Syrgkanis, and Matt Taddy. Accurate inference for adaptive linear models. In Jennifer G. Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 1202–1211, 2018.

Peter Dixon, A. Pavan, Jason Vander Woude, and N. V. Vinodchandran. List and certificate complexities in replicable learning, 2023.

Jing Dong and Yuichi Yoshida. General transformation for consistent online approximation algorithms, 2023.

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Eric Eaton, Marcel Hussing, Michael Kearns, and Jessica Sorrell. Replicable reinforcement learning. *CoRR*, abs/2305.15284, 2023. doi: 10.48550/ARXIV.2305.15284. URL `https://doi.org/10.48550/arXiv.2305.15284`.

Hossein Esfandiari, Amin Karbasi, Abbas Mehrabian, and Vahab S. Mirrokni. Regret bounds for batched bandits. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pp. 7340–7348. AAAI Press, 2021. doi: 10.1609/AAAI.V35I8.16901. URL `https://doi.org/10.1609/aaai.v35i8.16901`.

Hossein Esfandiari, Alkis Kalavasis, Amin Karbasi, Andreas Krause, Vahab Mirrokni, and Grigoris Velegkas. Replicable bandits. In *The Eleventh International Conference on Learning Representations*, 2023a.

Hossein Esfandiari, Amin Karbasi, Vahab Mirrokni, Grigoris Velegkas, and Felix Zhou. Replicable clustering. *CoRR*, abs/2302.10359, 2023b.

Pratik Gajane, Tanguy Urvoy, and Emilie Kaufmann. Corrupt bandits for preserving local privacy. In *Proceedings of the 29th International Conference on Algorithmic Learning Theory (ALT)*, pp. 387–412, 2018.

Zijun Gao, Yanjun Han, Zhimei Ren, and Zhengqing Zhou. Batched multi-armed bandits problem. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 501–511, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/20f07591c6fcb220ffe637cda29bb3f6-Abs

Eran Gat and Shafi Goldwasser. Probabilistic search algorithms with unique answers and their cryptographic applications. In *Electronic Colloquium on Computational Complexity (ECCC)*, volume 18, pp. 136, 2011.

Bingshan Hu and Nidhi Hegde. Near-optimal thompson sampling-based algorithms for differentially private stochastic bandits. In James Cussens and Kun Zhang (eds.), *Uncertainty in Artificial Intelligence, Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence, UAI 2022, 1-5 August 2022, Eindhoven, The Netherlands*, volume 180 of *Proceedings of Machine Learning Research*, pp. 844–852. PMLR, 2022. URL https://proceedings.mlr.press/v180/hu22a.html.

Russell Impagliazzo, Rex Lei, Toniann Pitassi, and Jessica Sorrell. Reproducibility in learning. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, pp. 818–831, 2022.

Alkis Kalavasis, Amin Karbasi, Shay Moran, and Grigoris Velegkas. Statistical indistinguishability of learning algorithms. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Sayash Kapoor, Kumar Kshitij Patel, and Purushottam Kar. Corruption-tolerant bandit learning. *Machine Learning*, 108(4):687–715, 2019.

Amin Karbasi, Grigoris Velegkas, Lin F. Yang, and Felix Zhou. Replicability in reinforcement learning. *CoRR*, abs/2305.19562, 2023.

Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *Proceedings of the 23rd International Conference on Algorithmic Learning Theory (ALT)*, pp. 199–213, 2012.

Michael Jong Kim and Andrew E. B. Lim. Robust multiarmed bandit problems. *Management Science*, 62(1):264–285, 2016.

Junpei Komiyama, Issei Sato, and Hiroshi Nakagawa. Multi-armed bandit problem with lock-up periods. In Cheng Soon Ong and Tu Bao Ho (eds.), *Asian Conference on Machine Learning, ACML 2013, Canberra, ACT, Australia, November 13-15, 2013*, volume 29 of *JMLR Workshop and Conference Proceedings*, pp. 116–132. JMLR.org, 2013. URL http://proceedings.mlr.press/v29/Komiyama13.html.

Junpei Komiyama, Junya Honda, and Hiroshi Nakagawa. Optimal regret analysis of Thompson sampling in stochastic multi-armed bandit problem with multiple plays. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pp. 1152–1161, 2015.

T.L Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

Tze Leung Lai and C. Z. Wei. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *Annals of Statistics*, 10:154–166, 1982.

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

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web (WWW)*, pp. 661–670, 2010.

Lisha Li, Kevin G. Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, 18:185:1–185:52, 2017.

Vianney Perchet, Philippe Rigollet, Sylvain Chassang, and Erik Snowberg. Batched bandit problems. *The Annals of Statistics*, 44(2):660 – 681, 2016. doi: 10.1214/15-AOS1381. URL `https://doi.org/10.1214/15-AOS1381`.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527 – 535, 1952.

Roshan Shariff and Or Sheffet. Differentially private contextual linear bandits. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 4301–4311, 2018.

Jaehyeok Shin, Aaditya Ramdas, and Alessandro Rinaldo. Are sample means in multi-armed bandits positively or negatively biased? In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 7102–7111, 2019.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

Long Tran-Thanh, Trung Dong Huynh, Avi Rosenfeld, Sarvapali D. Ramchurn, and Nicholas R. Jennings. Budgetfix: budget limited crowdsourcing for interdependent task allocation with quality guarantees. In *International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, pp. 477–484, 2014.

Nithin Varma and Yuichi Yoshida. Average sensitivity of graph algorithms. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 684–703, 2021.

Min Xu, Tao Qin, and Tie-Yan Liu. Estimation bias in multi-armed bandit algorithms for search advertising. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 2400–2408, 2013.

A REPLICABLE LINEAR SUCCESSIVE ELIMINATION

The Replicable Linear Successive Elimination (RLSE) algorithm is described in Algorithm 3.

Algorithm 3: Replicable Linear Successive Elimination (RLSE)

Initialize the candidate set $\mathcal{A}_1 = [K]$;
while $p = 1, 2, \ldots, P$ **do**
  $\epsilon_p = 2^{-p}$;
  Draw shared random variables $U_{p,i} \sim \mathrm{Unif}(0,1)$ for each $i = 0, 1, 2, \ldots, K$;
  Draw each arm for $N_i^{\mathrm{lin}}(p)$ times; ▷ Approximated G-optimal exploration
  $\mathcal{A}_{p+1} \leftarrow \mathcal{A}_p$;
  if $\hat{\Delta}(p) \geq \frac{(2+U_{p,0})\epsilon_p}{\rho_a}$ **then**
    $\mathcal{A}_{p+1} = \{\arg\max_i \hat{\mu}_i(p)\}$; ▷ Eliminate all arms except for one. Note that $\hat{\Delta}(p) > 0$ implies $|\mathcal{A}_{p+1}| = 1$.
  for $i \in \mathcal{A}_{p+1}$ **do**
    if $\hat{\Delta}_i(p) \geq \frac{(2+U_{p,i})\epsilon_p}{\rho_e}$ **then**
      $\mathcal{A}_{p+1} \leftarrow \mathcal{A}_{p+1} \setminus \{i\}$; ▷ Eliminate arm i.

B PROOFS ON GENERAL BOUND

Proof of Theorem 2. Let $\mathcal{G} = \bigcap_p \mathcal{G}_p$. We have

$$
\begin{aligned}
\rho &:= \mathbb{P}_{U,\mathcal{D}^{(1)},\mathcal{D}^{(2)}}\left[(I_1^{(1)}, I_2^{(1)}, \ldots, I_T^{(1)}) \neq (I_1^{(2)}, I_2^{(2)}, \ldots, I_T^{(2)})\right] && (22)\\
&\leq \mathbb{P}[(I_1^{(1)}, I_2^{(1)}, \ldots, I_T^{(1)}) \neq (I_1^{(2)}, I_2^{(2)}, \ldots, I_T^{(2)}), \mathcal{G}^{(1)}, \mathcal{G}^{(2)}] + \mathbb{P}_{\mathcal{D}^{(1)}}\left[\cup_p \mathcal{G}_p^c\right] + \mathbb{P}_{\mathcal{D}^{(2)}}\left[\cup_p \mathcal{G}_p^c\right] && (23)\\
&\leq \mathbb{P}[(I_1^{(1)}, I_2^{(1)}, \ldots, I_T^{(1)}) \neq (I_1^{(2)}, I_2^{(2)}, \ldots, I_T^{(2)}), \mathcal{G}^{(1)}, \mathcal{G}^{(2)}] + 2\sum_p \rho_p^G && (24)\\
&\quad \text{(by Definition 6)} && (25)\\
&\leq \mathbb{P}\left[(d_p^{(1)})_{p=1}^P \neq (d_p^{(2)})_{p=1}^P, \mathcal{G}^{(1)}, \mathcal{G}^{(2)}\right] + 2\sum_p \rho_p^G && (26)\\
&\quad \text{(by definition of decision variables)} && (27)\\
&\leq \sum_p \mathbb{P}\left[d_p^{(1)} \neq d_p^{(2)}, \cap_{p'=1}^{p-1}\left\{d_{p'}^{(1)} = d_{p'}^{(2)}\right\}, \mathcal{G}^{(1)}, \mathcal{G}^{(2)}\right] + 2\sum_p \rho_p^G && (28)\\
&\leq \sum_p \mathbb{P}\left[d_p^{(1)} \neq d_p^{(2)}, \cap_{p'=1}^{p-1}\left\{d_{p'}^{(1)} = d_{p'}^{(2)}, \mathcal{G}_{p'}^{(1)}, \mathcal{G}_{p'}^{(2)}\right\}, \mathcal{G}_p^{(1)}, \mathcal{G}_p^{(2)}\right] + 2\sum_p \rho_p^G && (29)\\
&\leq \sum_p \mathbb{P}\left[d_p^{(1)} \neq d_p^{(2)} \,\middle|\, \cap_{p'=1}^{p-1}\left\{d_{p'}^{(1)} = d_{p'}^{(2)}, \mathcal{G}_{p'}^{(1)}, \mathcal{G}_{p'}^{(2)}\right\}, \mathcal{G}_p^{(1)}, \mathcal{G}_p^{(2)}\right] + 2\sum_p \rho_p^G && (30)\\
&\leq \sum_{p,i} \rho^{(p,i)} + 2\sum_p \rho_p^G && (31)\\
&\quad \text{(by Definition 7).} && (32)
\end{aligned}
$$

□

C Proofs on Algorithm 1

C.1 Replicability of Algorithm 1

This section bounds the probability of non-replication of Theorem 3. We use the general bound of Section 3. Let the good event (Definition 6) be

$$\mathcal{G} = \bigwedge_{p \in [P]} \mathcal{G}_p \tag{33}$$

$$\mathcal{G}_p = \bigwedge_{i,j \in [K]} \left\{ |\Delta_{ij} - \hat{\Delta}_{ij}(p)| \le \mathrm{Conf}(p) \right\}. \tag{34}$$

Event $\mathcal{G}$ states that all estimators lie in the confidence region.

Lemma 9. *Event $\mathcal{G}$ holds with probability at least $1 - \rho/18$.*

It is easy to see that the only decision variable at phase p is

$$d_{(p,0)} := \mathbf{1}\left[\hat{\Delta}(p) \ge \frac{(2 + U_p)\epsilon_p}{\rho}\right].$$

Lemma 10. *Let*

$$p_{\mathrm{s}} = \min_p \left\{ \epsilon_p \le \frac{10\Delta}{17\rho} \right\}.$$

Under $\mathcal{G}$, the misreplication probability is $\rho^{(p,0)} = 0$ for any $p \ne p_{\mathrm{s}}$.

Lemma 10 states that the only effective decision variable is that of phase p_{s}.

Lemma 11. *Under $\mathcal{G}$, the probability of non-replication at each decision point is at most $\rho^{(p,0)} = 8\rho/9$ for $p = p_{\mathrm{s}}$.*

Proof of non-replicability part of Theorem 3. Theorem 2 and Lemmas 9–11 imply that the probability of misidentification is at most $2 \times \rho/18 + 8\rho/9 = \rho$, which completes the proof. □

In the following, we derive Lemmas 9–11.

Proof of Lemma 9. Since $\Delta_{ij} = \mu_i - \mu_j$, $\hat{\Delta}_{ij} - \Delta_{ij}$ is estimating the sum of two σ-subgaussian random variables, which is a 2σ-subgaussian random variable. By using Lemma 1 and taking a union bound over all possible $K(K-1)/2$ pairs of ij and phases $1, \dots, P$, Event $\mathcal{G}$ holds with high probability:

$$\mathbb{P}[\mathcal{G}] \ge 1 - \frac{2K}{T}, \tag{35}$$

By assumption, $2K/T \le \rho/18$, which completes the proof. □

Proof of Lemma 10. First, we show that there are at most 2 phases where the break from the loop (i.e., $\min_p d_{(p,1)} = 1$) occurs. We first show that a break never occurs if

$$\Delta \le \frac{17}{10}\frac{\epsilon_p}{\rho}.$$

This holds because

$$\begin{aligned}
\hat{\Delta}(p) &\le \Delta + \mathrm{Conf}(p) && (36)\\
&\le \frac{17}{10}\frac{\epsilon_p}{\rho} + \mathrm{Conf}(p) && (37)\\
&\le \frac{2\epsilon_p}{\rho} \quad \left(\text{by } \frac{20-17}{10}\frac{\epsilon_p}{\rho} \ge \frac{6}{10}\frac{\epsilon_p}{1} \ge \frac{\epsilon_p}{C_\rho} = \mathrm{Conf}(p)\right) && (38)\\
&\le \frac{(2+U_p)\epsilon_p}{\rho}, && (39)
\end{aligned}$$

We next show that

$$\Delta \geq \frac{17}{5} \frac{\epsilon_p}{\rho}$$

implies a break. This holds since

$$\begin{aligned}
\hat{\Delta}(p) &\geq \Delta - \mathrm{Conf}(p) \quad \text{(by } \mathcal{G}\text{)} && (40)\\
&\geq \frac{17}{5}\frac{\epsilon_p}{\rho} - \mathrm{Conf}(p) && (41)\\
&\geq \frac{3\epsilon_p}{\rho} \quad \left(\text{by } \frac{(17-15)\epsilon_p}{5\rho} \geq \frac{4\epsilon_p}{5} \geq \frac{\epsilon_p}{C_\rho} \geq \mathrm{Conf}(p)\right) && (42)\\
&\geq \frac{(2+U_p)\epsilon_p}{\rho}. && (43)
\end{aligned}$$

The above results, combined with the fact that ϵ_p is halved at each phase, implying that the only decision points where the decision variable can take both of $\{0, 1\}$ are $p_\mathrm{s}, p_\mathrm{s} + 1$. Therefore, if the decision variable at phase p_s matches, the decision variable at $p_\mathrm{s} + 1$ and subsequent match. $\square$

Proof of Lemma 11. For each phase $p \in \{p_s, p_s + 1\}$, we bound the probability of non-replication. At the end of phase p, it utilizes the randomness $U_p \sim \mathrm{Unif}(0, 1)$, and the random variable $\frac{(2+U_p)\epsilon_p}{\rho}$ is uniformly distributed on a region on size ϵ_p/ρ. Meanwhile, event $\mathcal{G}$ implies

$$\left| \left(\hat{\Delta}^{(1)}(p) - \frac{(2+U_p)\epsilon_p}{\rho} \right) - \left(\hat{\Delta}^{(2)}(p) - \frac{(2+U_p)\epsilon_p}{\rho} \right) \right| \leq 2\mathrm{Conf}(p). \quad (44)$$

This implication suggests that within a region of at most width $2\mathrm{Conf}(p)$, the expressions $\left(\hat{\Delta}^{(1)}(p) - \frac{(2+U_p)\epsilon_p}{\rho}\right)$ and $\left(\hat{\Delta}^{(2)}(p) - \frac{(2+U_p)\epsilon_p}{\rho}\right)$ can have different signs. Therefore, the probability of non-replication is at most

$$\frac{2\mathrm{Conf}(p)\rho}{\epsilon_p} \leq 8\rho/9,$$

where the last inequality follows from the assumption $C_\rho \geq 9/4$. $\square$

C.2 Regret bound of Algorithm 1

This section derives the regret bound in Theorem 3.

Assume that $\mathcal{G}$ holds. Lemma 10 implies that the break occurs by the end of phase p_s or $p_\mathrm{s} + 1$. The regret up to this phase is at most

$$\sum_i \Delta_i \times N_{p_\mathrm{s}+1}.$$

By using equation 7, we have

$$N_{p_\mathrm{s}+1} = O\left(\log T \times 4^{p_\mathrm{s}}\right) = O\left(\log T \times \frac{1}{\Delta^2\rho^2}\right).$$

Moreover, assume that $\hat{\mu}_i(p) \geq \hat{\mu}_1(p)$ when a break occurs at $p \in \{p_\mathrm{s}, p_\mathrm{s} + 1\}$. Then,

$$\begin{aligned}
\hat{\mu}_i(p) - \hat{\mu}_1(p) &\geq \frac{2\epsilon_{p_\mathrm{s}}}{\rho} \quad \text{(by the fact that break occurs)} && (45)\\
&\geq 2\epsilon_{p_\mathrm{s}} && (46)
\end{aligned}$$

Meanwhile, $\mathcal{G}$ implies for all phase p

$$\hat{\mu}_i(p) - \hat{\mu}_1(p) \leq 2\mathrm{Conf}(p). \quad (47)$$

Since $\mathrm{Conf}(p) < \epsilon_p$ for all p, this contradicts. By proof of contradiction. $\hat{\mu}_1(p) > \hat{\mu}_i(p)$. Therefore, the empirically best arm is always the true best arm, and we have zero regret during the exploitation period under $\mathcal{G}$. In summary,

$$\mathbb{E}[\mathrm{Regret}(T)] \leq \mathbb{E}[\mathbf{1}[\mathcal{G}] \cdot \mathrm{Regret}(T)] + O(1) \quad \text{(equation 35 implies } \Pr[\mathcal{G}^c] \text{ is } K/T = O(1/T)\text{)}$$

$$(48)$$

$$\leq \underbrace{O\left(\sum_i \Delta_i N_{P_c}\right)}_{\text{Regret during exploration, due to Lemma 10}} + \underbrace{0}_{\text{Regret during exploitation}} + \underbrace{O(1)}_{\text{Regret in the case of } \mathcal{G}^c} \tag{49}$$

$$\leq O\left(\sum_i \Delta_i \frac{\log T}{\Delta^2 \rho^2}\right) + O(1) \tag{50}$$

$$= O\left(\sum_i \Delta_i \frac{\log T}{\Delta^2 \rho^2}\right), \tag{51}$$

which completes the proof.

D Proofs on Algorithm 2

D.1 Replicability of Algorithm 2

Similar to that of Theorem 3, we utilize Theorem 2. We use the event $\mathcal{G}$ defined in equation 33 during the proof.

Lemma 12. *Event $\mathcal{G}$ holds with probability at least $1 - \rho/6$. Moreover, under $\mathcal{G}$, arm 1 (= best arm) is never eliminated (i.e., $1 \in \mathcal{A}_p$ for all p).*

There are $K + 1$ binary decision variables at phase p. The first decision variable is

$$d_{(p,0)} := \mathbf{1}\left[\hat{\Delta}(p) > \frac{(2 + U_{p,0})\epsilon_p}{\rho_a}\right], \tag{52}$$

which corresponds to Line 7. The other K decision variables are

$$d_{(p,i)} := \mathbf{1}\left[\hat{\Delta}_i(p) > \frac{(2 + U_{p,i})\epsilon_p}{\rho_e}\right] \tag{53}$$

for each $i = 1, 2, \ldots, K$, which corresponds to Line 10. If all decision variables are identical between two runs then the sequence of draws is identical between them.

Lemma 13. *Let*

$$p_{\mathrm{s},0} = \min_p \left\{\epsilon_p \leq \frac{10\Delta}{17\rho_a}\right\}.$$

Under $\mathcal{G}$, we have $\rho^{(p,0)} = 0$ for any $p \neq p_{\mathrm{s},0}$.

Lemma 14. *Let*

$$p_{\mathrm{s},i} = \min_p \left\{\epsilon_p \leq \frac{10\Delta_i}{17\rho_e}\right\},$$

for each $i \in [K]$. Under $\mathcal{G}$, we have $\rho^{(p,1)}, \rho^{(p,2)} = 0$ for all p. Moreover, $\rho^{(p,i)} = 0$ for all $p \neq p_{\mathrm{s},i}$.

Lemma 15. *Under $\mathcal{G}$, the probability of non-replication at each decision point is at most $\rho^{(p,0)} = 8\rho_a/9$ or $\rho^{(p,i)} = 8\rho_e/9$ for $i \in \{3, \ldots, K\}$.*

Proof of non-replicability part of Theorem 4. Theorem 2 and Lemmas 12–15 imply that the probability of misidentification is at most $2 \times \rho/18 + 8(\rho_a + (K - 2)\rho_e)/9 = \rho$, which completes the proof. □

Proof of Lemma 12. Lemma 9 implies that event $\mathcal{G}$ holds with probability at least $1 - \rho/6$.

In the following, we show that arm 1 is never eliminated under $\mathcal{G}$. Elimination of arm 1 at phase p at Line 7 implies that there exists a suboptimal arm $i \neq 1$ such that

$$\hat{\mu}_i(p) - \mu_1(p) \geq \frac{(2 + U_{p,0})\epsilon_p}{\rho_a} \geq \frac{2\epsilon_p}{\rho_a} \geq 2\epsilon_p,$$

which never occurs under $\mathcal{G}$, since $\mathcal{G}$ implies

$$|\mu_1 - \hat{\mu}_1(p)|, |\mu_i - \hat{\mu}_i(p)| \leq \mathrm{Conf}(p) < \epsilon_p/2$$

and $\mu_1 > \mu_i$ by definition. The same discussion goes for elimination at Line 10. □

Proof of Lemma 13. We first show arm 2 is never eliminated by Line 10 before phase $p_{\mathrm{s},0}+1$. For $p \le p_{\mathrm{s},0}+1$, we have

$$\begin{aligned}
\hat{\Delta}_2(p) - \frac{(2+U_{p,l})\epsilon_p}{\rho_e} &\le \hat{\Delta}_2(p) - \frac{2\epsilon_p}{\rho_e} && (54)\\
&\le \Delta + 2\mathrm{Conf}(p) - \frac{2\epsilon_p}{\rho_e} \quad (\text{by } \mathcal{G}) && (55)\\
&= \Delta + \frac{2\epsilon_p}{C_\rho} - \frac{2\epsilon_p}{\rho_e} && (56)\\
&\le \frac{17\epsilon_p}{5\rho_a} + \frac{2\epsilon_p}{C_\rho} - \frac{2\epsilon_p}{\rho_e} \quad (\text{by } p \le p_{\mathrm{s},0}) && (57)\\
&\le \frac{17\epsilon_p}{5\rho_a} + \frac{2\epsilon_p}{3} - \frac{2\epsilon_p}{\rho_e} && (58)\\
&\le 0, \quad (\text{by } 1/2 \ge \rho_a \ge 2\rho_e) && (59)
\end{aligned}$$

which implies $d_{(p,2)} = 0$.

Given arm 1 is never eliminated and arm 2 remains in $\mathcal{A}_{p_{\mathrm{s},0}+1}$ unless $d_{(p,0)} = 1$ occurs, the minimum gap $\Delta = \mu_1 - \mu_2$ is the same as the minimum gap among $\mathcal{A}_{p_{\mathrm{s},0}+1}$. The rest of this Lemma, which shows $d_{(p,0)} = 1$ for the first phase at $p = p_{\mathrm{s},0}$ or $p_{\mathrm{s},0}+1$, is very similar to that of Lemma 10, and thus we omit it. The only non-zero decision variable is $\rho^{(p,0)}$ at $p = p_{\mathrm{s},0}$ by using the same discussion. □

Proof of Lemma 14. We omit the proof because it is very similar to Lemma 10. □

Proof of Lemma 15. We omit the proof because the proof for each $i = 0, 1, 2, \ldots, K$ is very similar to Lemma 11.

□

D.2 Regret bound of Algorithm 2

(A) The $O(K)$ regret bound: We first derive the same bound as Algorithm 1. This part is very identical to that of Algorithm 1 because, under $\mathcal{G}$, all but arm 1 is eliminated by phase $p_{\mathrm{s},0}+1$.

(B) The other two regret bounds: We first derive the distribution-dependent bound. Lemma 15 states that, under $\mathcal{G}$, each arm i is eliminated by phase $p_{\mathrm{s},i}+1$. Therefore, each arm i is drawn at most

$$N_{i,\max} := O\left(4^{p_{\mathrm{s},i}+1}\log T\right) = O\left(\frac{K^2}{\rho^2\Delta_i^2}\log T\right) \tag{60}$$

times, and thus regret is bounded as

$$\begin{aligned}
\mathbb{E}[\mathrm{Regret}(T)] &\le \mathbb{E}[\mathbf{1}[\mathcal{G}]\cdot\mathrm{Regret}(T)] + O(1) && (61)\\
&\le \underbrace{\sum_i \Delta_i N_{i,\max}}_{\text{Regret during exploration}} + \underbrace{0}_{\text{Regret during exploitation, which is 0 by Lemma 12}} + \underbrace{O(1)}_{\text{Regret in the case of } \mathcal{G}^c} && (62)\\
&\le O\left(\sum_i \frac{K^2\log T}{\Delta_i\rho^2}\right), && (63)
\end{aligned}$$

which is the second regret bound of Theorem 4.

We finally derive the distribution-independent regret bound. Letting $N_i(T)$ be the number of draws of arm i in the T rounds, we have

$$\mathrm{Regret}(T)\mathbf{1}[\mathcal{G}] \le \sum_i \Delta_i N_i(T) + O(1) \tag{64}$$

$$= \sum_i \Delta_i \sqrt{N_{i,\max}} \sqrt{N_i(T)} + O(1), \tag{65}$$

and

$$\sum_i \Delta_i \sqrt{N_{i,\max}} \sqrt{N_i(T)} \le O(1) \times \sum_i \Delta_i \sqrt{\frac{K^2}{\rho^2 \Delta_i^2} \log T} \sqrt{N_i(T)} \quad \text{(by equation 60)} \tag{66}$$

$$\le O(1) \times \sum_i \sqrt{\frac{K^2}{\rho^2} \log T} \sqrt{N_i(T)} \tag{67}$$

$$\le O(1) \times \sqrt{\frac{K^2}{\rho^2} \log T} \sqrt{KT} \quad \text{(by Cauchy-Schwarz and } \sum_i N_i(T) = T\text{)}, \tag{68}$$

and thus

$$\mathbb{E}[\text{Regret}(T)] \le \mathbb{E}[\text{Regret}(T)\mathbf{1}[\mathcal{G}]] + O(1) \tag{69}$$

$$= O\left(\sqrt{\frac{K^2}{\rho^2} \log T} \sqrt{KT} \right) \quad \text{(by equation 106)} \tag{70}$$

$$= O\left(\frac{K}{\rho} \sqrt{KT \log T} \right), \tag{71}$$

which is the third regret bound of Theorem 4.

E Proofs on the Lower Bound

In the following, we derive Theorem 5. The proof is inspired by Theorem 7.2 of Impagliazzo et al. (2022) but is significantly more challenging due to the adaptiveness of sampling. In particular, Lemma 16 utilizes the change-of-measure argument and works even if the number of draws $N_2(T)$ on arm 2 is a random variable.

Proof of Theorem 5. The goal of the proof here is to derive the inequality:

$$\mathbb{E}[\text{Regret}(T)] = \Omega\left(\frac{1}{\rho^2 \Delta \log((\rho\Delta)^{-1})} \right). \tag{72}$$

We consider the set of models $\mathcal{P}$, where $\mu_1 = 1/2$ is fixed and $\mu_2 \in [1/2 - \Delta, 1/2 + \Delta]$. With a slight abuse of notation, we specify a model in $\mathcal{P}$ by $\mu_2 - 1/2$. We also denote U to the internal randomness For example,

$$\mathbb{P}_{-\Delta,U}[\mathcal{X}]$$

be the probability that event $\mathcal{X}$ occurs under the corresponding model $(\mu_1, \mu_2) = (1/2, 1/2 - \Delta)$ and randomness U.

Markov's inequality on the regret and the assumption of uniformly goodness implies that, there exists a sublinear function $R(T)$ such that, at least 3/4 of the choice of the randomness U, we have $\mathbb{E}_{-\Delta,U}[\text{Regret}(T)] \le R(T)$. Similarly, at least 3/4 of the choice of the randomness U, we have $\mathbb{E}_{\Delta,U}[\text{Regret}(T)] \le R(T)$. By taking a union bound, at least half $(= 1 - (1/4 + 1/4))$ over the choice of U, we have

$$\max\{\mathbb{E}_{-\Delta,U}[\text{Regret}(T)], \mathbb{E}_{\Delta,U}[\text{Regret}(T)]\} \le R(T). \tag{73}$$

Let T_0 be such that

$$\forall T \ge T_0 \; R(T) \le \frac{T\Delta}{8}. \tag{74}$$

In the following, we consider an arbitrary $T \ge T_0$. In parts (A) and (B), we fix U such that equation 73 holds. Part (C) marginalizes it over U.

(A) Fix the randomness U and consider behavior of algorithm for different models:

Let N_C be the smallest among the integer n such that

$$\mathbb{P}_{-\Delta,U}\left[N_2(T) \geq n\right] \leq \frac{1}{4} \text{ and} \tag{75}$$

$$\mathbb{P}_{\Delta,U}\left[N_2(T) \geq n\right] \geq \frac{3}{4}. \tag{76}$$

At least one such an integer exists; $n = T/2$ satisfies this condition because otherwise

$$\max\{\mathbb{E}_{-\Delta,U}[\mathrm{Regret}(T)], \mathbb{E}_{\Delta,U}[\mathrm{Regret}(T)]\} \tag{77}$$

$$\geq \max\left\{\mathbb{P}_{-\Delta,U}\left[N_2(T) \geq \frac{T}{2}\right]\frac{T\Delta}{2}, \left(1 - \mathbb{P}_{\Delta,U}\left[N_2(T) \geq \frac{T}{2}\right]\right)\frac{T\Delta}{2}\right\} \tag{78}$$

$$> \frac{T\Delta}{8} \tag{79}$$

$$\text{(by } \mathbb{P}_{-\Delta,U}\left[N_2(T) \geq \frac{T}{2}\right] > \frac{1}{4} \text{ or } \mathbb{P}_{\Delta,U}\left[N_2(T) \geq \frac{T}{2}\right] < \frac{3}{4}\text{)}, \tag{80}$$

which violates equation 74.

By continuity, there exists $\nu = \nu(U) \in (-\Delta, +\Delta)$ such that $\mathbb{P}_{\nu,U}[N_2(T) \geq N_C] = 1/2$. By Lemma 16, there exists $C_I = \Theta(1)$ such that, for any $\nu' \in [\nu - C_I/\sqrt{\log(N_C)N_C}, \nu + C_I/\sqrt{\log(N_C)N_C}] =: \mathcal{I}(U)$ we have $\mathbb{P}_{\nu',U}[N_2(T) \geq N_C] \in (1/3, 2/3)$.

(B) Marginalize it over models: Assume that we first draw a model uniformly random from $[-\Delta, \Delta]$, and then run the algorithm. Conditioned on the shared randomness U, with probability at least $C_I/(\sqrt{\log(N_C)N_C}\Delta)$, we draw model in $\mathcal{I}(U)$. For a model in $\mathcal{I}(U)$, there is $2 \times (1/3) \times (2/3) = 4/9$ probability of non-replicability.

(C) Marginalize it over shared random variable U: equation 73 implies that the discussions in (A) and (B) hold at least half of the random variable U. Marginalize it over the distribution on U, we have the probability of non-replicability at least

$$\frac{1}{2} \times \frac{4}{9}\frac{C_I}{\sqrt{\log(N_C)N_C}\Delta}. \tag{81}$$

Since the algorithm is ρ-replicable, we have

$$\frac{2}{9}\frac{C_I}{\sqrt{\log(N_C)N_C}\Delta} \leq \rho,$$

which implies

$$N_C = \Omega\left(\frac{1}{(\rho\Delta)^2 \log((\rho\Delta)^{-1})}\right). \tag{82}$$

The regret is lower-bounded as

$$\max\{\mathbb{E}_{-\Delta}[\mathrm{Regret}(T)], \mathbb{E}_{\Delta}[\mathrm{Regret}(T)]\} \tag{83}$$

$$\geq \max\left\{\mathbb{P}_{-\Delta}[N_2(T) \geq N_C - 1]\Delta(N_C - 1), (1 - \mathbb{P}_{\Delta}[N_2(T) \geq N_C - 1])\,\Delta(T - N_C + 1)\right\} \tag{84}$$

$$\geq \max\left\{\mathbb{P}_{-\Delta}[N_2(T) \geq N_C - 1]\Delta(N_C - 1), (1 - \mathbb{P}_{\Delta}[N_2(T) \geq N_C - 1])\,\Delta\left(\frac{T}{2} + 1\right)\right\} \tag{85}$$

$$\text{(by } N_C \leq T/2\text{)} \tag{86}$$

$$\geq \frac{1}{4} \times \Delta(N_C - 1) \tag{87}$$

$$\text{(by } N_C - 1 \text{ violates equation 75 or equation 76, and } N_C \leq T/2\text{)} \tag{88}$$

$$= \Omega\left(\frac{1}{\rho^2\Delta\log((\rho\Delta)^{-1})}\right). \quad \text{(by equation 82)} \tag{89}$$

□

E.1 Lemmas for Regret Lower Bound

The following lemma is used to bound the gradient of the probability of occurrences. This lemma corresponds to the derivative of the acceptance function[4] of Lemma 7.2 in Impagliazzo et al. (2022), but more technical due to the fact that $N_2(T)$ is a random variable.

Lemma 16. (Likelihood ratio) *Let*

$$\mathcal{E} = \{N_2(T) \le N_C\} \tag{90}$$

and ν be such that $\mathbb{P}_{\nu,U}[\mathcal{E}] = 1/2$. There exists a value $C_I = \Theta(1)$ that does not depend on the shared random variable U such that, for any model $\nu' \in [\nu - C_I/\sqrt{\log(N_C)N_C}, \nu + C_I/\sqrt{\log(N_C)N_C}]$, we have

$$\mathbb{P}_{\nu',U}[\mathcal{E}] \in \left(\frac{1}{3}, \frac{2}{3}\right). \tag{91}$$

Proof of Lemma 16. In the following, we bound $\mathbb{P}_{\nu',U}[\mathcal{E}]$ by using the change-of-measure argument.

Let the log-likelihood ratio between from models ν, ν' be[5]

$$L_t = \sum_{n=1}^{N_2(t)} \log\left(\frac{X_{2,n}\left(\frac{1}{2}+\nu\right) + (1-X_{2,n})\left(\frac{1}{2}-\nu\right)}{X_{2,n}\left(\frac{1}{2}+\nu'\right) + (1-X_{2,n})\left(\frac{1}{2}-\nu'\right)}\right), \tag{92}$$

where $X_{2,n}$ is the n-th reward from arm 2 and $N_2(t)$ be the number of draws on arm 2 during the first t rounds.

We have

$$\mathbb{P}_{\nu',U}[\mathcal{E}] = \mathbb{E}_{\nu,U}\left[\mathbf{1}[\mathcal{E}]e^{-L_T}\right]. \quad \text{(change-of-measure)} \tag{93}$$

Note that, under ν the random variable

$$\log\left(\frac{X_{2,n}\left(\frac{1}{2}+\nu\right) + (1-X_{2,n})\left(\frac{1}{2}-\nu\right)}{X_{2,n}\left(\frac{1}{2}+\nu'\right) + (1-X_{2,n})\left(\frac{1}{2}-\nu'\right)}\right)$$

is mean $d_{\mathrm{KL}}(\nu,\nu')$ and bounded by

$$R = \left|\log\left(\frac{2+\nu}{2+\nu'}\right), \log\left(\frac{2-\nu}{2-\nu'}\right)\right| = O(|\nu-\nu'|) = O\left(\frac{C_I}{\sqrt{\log(N_C)N_C}}\right).$$

Under $\mathcal{E}$, $N_2(T) \le N_C$ and L_T is bounded as the max of random variables

$$L_T \le \max_{N \le N_C}\left(\sum_{n \le N} Z_n\right),$$

where

$$Z_n := \log\left(\frac{X_{2,n}\left(\frac{1}{2}+\nu\right) + (1-X_{2,n})\left(\frac{1}{2}-\nu\right)}{X_{2,n}\left(\frac{1}{2}+\nu'\right) + (1-X_{2,n})\left(\frac{1}{2}-\nu'\right)}\right)$$

is a random variable with its mean $d_{\mathrm{KL}}(\nu,\nu')$ and radius R. Hoeffding inequality and union bound over $N = 1, 2, \dots, N_C$ implies that, with probability at least $1 - 1/12$ we have

$$|L_T| \le N_C d_{\mathrm{KL}}(1/2+\nu, 1/2+\nu') + R\sqrt{\log(2 \times 12 \times N_C)N_C/2} \tag{94}$$

$$= O\left(N_C \times \frac{C_I^2}{\log(N_C)N_C}\right) + O\left(\frac{C_I}{\sqrt{\log(N_C)N_C}} \times \sqrt{N_C \log(N_C)}\right) = O(C_I) \tag{95}$$

and by setting an appropriate width $C_I = \Theta(1)$ guarantees that $e^{-L_T} \in [1 - 1/6, 1 + 1/6]$ with probability at least $1 - 1/12$, which, together with the change-of-measure implies equation 91. □

[4] Namely, $\mathrm{ACC}(p)$ therein.

[5] On these models, $\mu_1 = 1/2$, $\mu_2 = 1/2 + \nu$ or $1/2 + \nu'$.

F PROOFS ON ALGORITHM 3

F.1 REPLICABILITY OF ALGORITHM 3

We omit the derivation of the non-replicability bound of Algorithm 3 because it is very similar to that of Algorithm 2. The only difference is that the amount of exploration is based on a G-optimal design, but its confidence bound of equation 19 suffices to derive the good event that is identical to equation 33.

F.2 REGRET BOUND OF ALGORITHM 3

This section shows the regret bounds of Theorem 8. The main difference from Algorithm 2 is that the number of samples for each arm is $N_i^{\mathrm{lin}}(p)$ that satisfies $\sum_i N_i^{\mathrm{lin}}(p) \le N^{\mathrm{lin}}(p) + K$.

(A) The $O(d)$ regret bound: We derive the first regret bound in Theorem 8. We first derive the distribution-dependent bound. Similar discussion as Lemma 15 states that, under $\mathcal{G}$, Line 7 in Algorithm 3 eliminates all but best arm by phase $p_{\mathrm{s},0}+1$, where $p_{\mathrm{s},0}$ is identical to that of Algorithm 3. The regret is bounded as

$$\begin{aligned}
\mathbb{E}[\mathrm{Regret}(T)] &\le \mathbb{E}[\mathbf{1}[\mathcal{G}]\mathrm{Regret}(T)] + O(1) && (96)\\
&\le \underbrace{O\left(\sum_i N_i^{\mathrm{lin}}(p_{\mathrm{s},i}+1)\right)}_{\text{Regret during exploration}} + \underbrace{0}_{\text{Regret during exploitation}} + \underbrace{O(1)}_{\text{Regret in the case of } \mathcal{G}^c} && (97)\\
&= O\left(\frac{d\log T}{\Delta^2\rho^2}\right) + o(\log T), && (98)\\
&\quad \left(\text{by } \sum_i N_i^{\mathrm{lin}}(p_{\mathrm{s},i}+1) = N^{\mathrm{lin}}(p_{\mathrm{s},i}+1) \le \frac{16\sigma^2 d\log(|\mathcal{A}_p|PT)}{(\mathrm{Conf}(p_{\mathrm{s},0}))^2} = O\left(\frac{d\log T}{\Delta^2\rho^2}\right)\right) && (99)
\end{aligned}$$

which is the first regret bound of Theorem 8.

(B) The distribution-independent regret bound:

Similar discussion as Algorithm 2 states that, under $\mathcal{G}$, arm i is eliminated by $p_{\mathrm{s},i}+1$. The regret is bounded as

$$\begin{aligned}
\mathrm{Regret}(T)\mathbf{1}[\mathcal{G}] &\le \sum_i \Delta_i N_i(T) + O(1) && (100)\\
&= \sum_i \Delta_i \sqrt{\sum_{p\le p_{\mathrm{s},i}+1} N_i^{\mathrm{lin}}(p)}\sqrt{N_i(T)} + O(1), && (101)
\end{aligned}$$

and

$$\begin{aligned}
&\sum_i \Delta_i \sqrt{\sum_{p\le p_{\mathrm{s},i}+1} N_i^{\mathrm{lin}}(p)}\sqrt{N_i(T)} && (102)\\
&\le O(1)\times \sum_i \Delta_i \sqrt{\frac{dK^2}{\rho^2\Delta_i^2}\log T}\sqrt{N_i(T)} \quad \text{(by equation 17)} && (103)\\
&\le O(1)\times \sum_i \sqrt{\frac{dK^2}{\rho^2}\log T}\sqrt{N_i(T)} && (104)\\
&\le O(1)\times \sqrt{\frac{dK^2}{\rho^2}\log T}\sqrt{KT} \quad \left(\text{by Cauchy-Schwarz and } \sum_i N_i(T) = T\right), && (105)\\
& && (106)
\end{aligned}$$

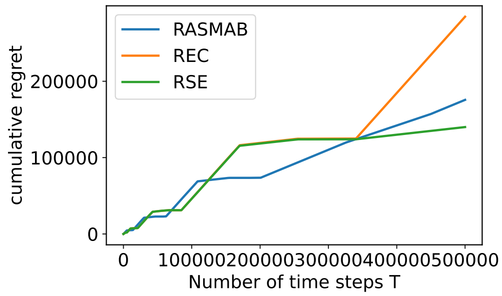
and thus

$$\mathbb{E}[\mathrm{Regret}(T)] \le \mathbb{E}[\mathrm{Regret}(T)\mathbf{1}[\mathcal{G}]] + O(1) \qquad (107)$$

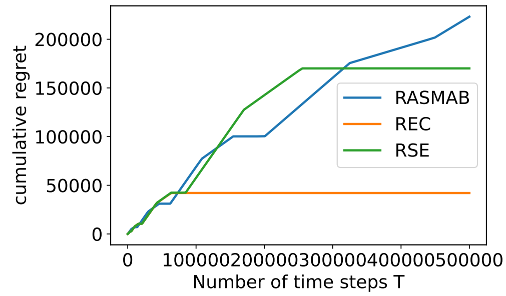
$$= O\left(\sqrt{\frac{dK^2}{\rho^2}\log T}\sqrt{KT}\right) \quad \text{(by equation 106)} \tag{108}$$

$$= O\left(\frac{K}{\rho}\sqrt{dKT\log T}\right), \tag{109}$$

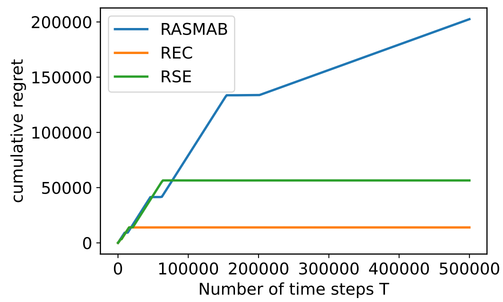
which is the second regret bound of Theorem 8.



Model 4



Model 5



Model 6

Figure 2: Regret of algorithms with theoretical parameters.

G ADDITIONAL SIMULATION

In this section, we report simulation results in which the algorithmic hyperparameters are selected based on their replicability bounds. Specifically, we set $C_\rho = 9/4$ for REC and RSE, and we adhere to the original version of Algorithm 2 as outlined in Esfandiari et al. (2023a). It should be noted that RASMAB is not explicitly designed to meet a non-replication level of ρ, owing to its disregard of the multiplication factor on the number of effective decision variables. This fact suggests that this comparison offers a substantial advantage to RASMAB . Notably, Appendix B therein specifies that there are 6 possible bad events. This implies that, for large T, their algorithm would need to choose an amount of exploration at least $(6-1)^2$ times greater.

We have set the models (i.e., mean parameters) as follows: $\boldsymbol{\mu} = (0.1, 0.9, 1.0)$ for Model 4, $\boldsymbol{\mu} = (0.1, 0.5, 1.0)$ for Model 5, and $\boldsymbol{\mu} = (0.0, 0.0, 1.0)$ for Model 6. Additionally, we set $\rho = 0.5$. Model 4 is more advantageous for RASMAB and RSE, while Model 6 favors REC.

The results for the Theoretical configuration are presented in Figure 2. All algorithms were confirmed to be ρ-replicable. However, due to the conservative selection of theoretical hyperparameters, the required exploration amount exceeded that of previous simulations. As a result, all algorithms incurred a larger regret compared to the simulations in the main paper.

Among the three algorithms, REC showed superior performance in Models 5 and 6, while RASMAB and RSE had some advantages in Model 4. These results align with our theoretical findings, which suggest that REC and RSE have a significant advantage over RASMAB when the value of Δ_K/Δ is moderate.