BANDITS WITH MULTIPLE OPTIMAL ARMS: MINIMAX REGRET AND NON-ADAPTIVITY

Anonymous authors
Paper under double-blind review

ABSTRACT

We study multi-armed bandits (MAB) with multiple optimal arms, motivated by the fact that many practical decision making problems admit multiple correct answers. For K-armed bandits with A optimal arms, we first provide a sharper analysis of previous sub-sampling algorithms (De Heide et al., 2021; Zhu & Nowak, 2020), establishing a $\widetilde{O}\left(\frac{K-A}{\sqrt{KA}}\sqrt{T}\right)$ minimax regret, where T is the total number of interactions and $\widetilde{O}(\cdot)$ drops all constant and logarithmic factors, improving the previous $\widetilde{O}(\sqrt{KT/A})$ regret. We then provide a matching lower bound up to logarithmic factors, indicating that our established rate is nearly minimax optimal. We further show that the knowledge of A up to $\widetilde{O}(1)$ factors is necessary to achieve near optimal regret, as near optimal algorithms for one number of optimal arms must incur substantially larger regret than optimal regret for a smaller number. Overall, our results provide a comprehensive minimax characterization of K-armed bandits with A over the entire range of $1 \le A \le K-1$.

1 INTRODUCTION

Multi-armed bandits (MAB) provide a basic model of sequential decision making under uncertainty, whose implication covers a wide range of practical problems (Bouneffouf et al., 2020; Pryzant et al., 2023; Panda et al., 2025). In an MAB, a learner repeatedly selects an action from an action set, observes a noisy reward, and aims at maximizing the cumulative reward via balancing exploration and exploitation. Classical results have characterized the difficulty of the problem by a worst case regret $\Theta(\sqrt{KT})$ when $T \ge K$, where K is the number of arms and T is the total number of interactions (Auer et al., 2002; Audibert & Bubeck, 2009; Lattimore, 2018).

However, in practice, many decision making problems, including code generation (Jimenez et al., 2024; Jain et al., 2025), math problem solving (Hendrycks et al., 2021) and good recommendation (Kano et al., 2019), admit multiple correct or satisfactory solutions. Such problems motivate the study of bandit learning with multiple correct answers (Degenne & Koolen, 2019), on which the benefit of the abundance of optimal arms is not captured by the classical $\Theta(\sqrt{KT})$ regret. Prior work exploits this benefit through sub-sampling: select approximately K/A arms and run a standard bandit algorithm on the selected subset (Zhu & Nowak, 2020; De Heide et al., 2021), yielding an expected regret of $\widetilde{O}(\sqrt{KT/A})$. However, this rate does not fully capture the benefit when A is close to K. Indeed, for $A = K-1$, if we randomly sample two arms, with probability only $O(1/K)$ one of the two arms is sub-optimal, leading to a regret of $O(\sqrt{T}/K)$. This discrepancy raises the question:

What is the minimax regret of multi-armed bandit with multiple optimal arms?

In this paper, we solve this question up to logarithmic factors. We provide a refined analysis of previous sub-sampling algorithms, and a novel near-matching lower bound for the regime $A > K/4$ that was not covered by prior work. Formally, writing the class of K-armed bandit with A optimal arms as $\Xi_{K,A}$, our upper bound and lower bound together gives

$$\inf_{\texttt{Alg}} \sup_{\nu \in \Xi_{K,A}} \mathrm{Regret}_{\texttt{Alg},\nu}(T) = \widetilde{\Theta}\left(\frac{K-A}{\sqrt{KA}}\sqrt{T}\right),$$

where $\mathrm{Regret}_{\texttt{Alg},\nu}(T)$ denotes expected regret under algorithm `Alg` and bandit instance ν. This characterization recovers the previous rate when $K - A = \Omega(K)$ and sharpens the previous regret upper bound by a factor of $\widetilde{\Theta}((K-A)/K)$ when $K - A = o(K)$.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The minimax characterization also raises an adaptation question that whether a learner can attain the appropriate rate without knowing A (Zhu & Nowak, 2020; De Heide et al., 2021). De Heide et al. (2021) showed that, for well separated $B = o(A)$, if $A + B \leq K$, then no strategy can simultaneously achieve minimax optimal regret on both $\Xi_{K,A}$ and $\Xi_{K,B}$. In this paper, we provide a novel analysis that removes the constraint on $A + B \leq K$, showing that any strategy achieving near-optimal regret on $\Xi_{K,A}$ must pay an additional multiplicative factor of $\widetilde{\Omega}(\sqrt{A/B})$ for regret on $\Xi_{K,B}$, thus providing a near-complete picture of non-adaptivity. Our contributions are summarized as follows

- We provide a refined analysis of previous algorithms for bandit with multiple optimal arms. Together with our novel lower bound for the regime $A > K/4$, we establish the minimax regret for K-armed bandit with A optimal arms, up to logarithmic factors, for all $K, T \geq K$ and $A \in [1, K-1]$.
- We show that, for any well separated $A > B$, any strategy achieving near-optimal regret on $\Xi_{K,A}$ will demonstrate a multiplicative loss of $\widetilde{\Omega}(\sqrt{A/B})$ compared to the minimax optimal rate on $\Xi_{K,B}$, necessitating the knowledge of A to achieve the near-optimal regret.
- Technically, standard two-point lower-bound arguments based on disjoint sets of optimal arms become ineffective when the number of optimal arms is close to K. We overcome this obstacle by considering a larger collection of symmetric instances and comparing them with a common all-equal reference instance. Together with a multiplicative change-of-measure argument, this approach enables us to establish worst-case hardness results even when the instances necessarily share optimal arms.

Notation. The set $\mathcal{A}$ are assumed to be finite throughout the paper. For nonnegative sequences $\{x_n\}$ and $\{y_n\}$, we write $x_n = O(y_n)$ if $\limsup_{n\to\infty} x_n/y_n < \infty$, $x_n = o(y_n)$ if $\limsup_{n\to\infty} x_n/y_n = 0$, $y_n = \Omega(x_n)$ (interchangeably written as $y_n \gtrsim x_n$) if $x_n = O(y_n)$, and $y_n = \Theta(x_n)$ if $x_n = O(y_n)$ and $x_n = \Omega(y_n)$. We further employ $\widetilde{O}(\cdot), \widetilde{\Omega}(\cdot)$, and $\widetilde{\Theta}(\cdot)$ to hide $\mathrm{polylog}$ factors. For a pair of probability measures $P \ll Q$ on the same space, we use $\mathrm{TV}(\mathbb{P}, \mathbb{Q})$ to denote their total variation, and $\mathsf{KL}(P\|Q) \coloneqq \int \log(\mathrm{d}P/\mathrm{d}Q)\,\mathrm{d}P$ to denote their KL divergence. We denote $[N] \coloneqq \{1, \cdots, N\}$ for any positive integer N. Boldfaced lowercase letters, like $\mathbf{x}$ are reserved for vectors, whose i-th entry is represented as x_i.

2 Related Work

Bandits with Multiple Best Arms. MABs with multiple optimal arms has been widely investigated in both regret minimization and pure exploration settings. In the pure exploration setting, the existence of multiple correct answer largely reduces the sample complexity (De Heide et al., 2021; Katz-Samuels & Jamieson, 2020), and the asymptotical optimal instance-dependent rate for fixed confidence best arm identification was established when number of optimal arm A is unknown (Degenne & Koolen, 2019) and known (Truong, 2026). In the regret minimization setting, when restricting to uniformly fast convergent (Garivier et al., 2019, Definition 1), the instance dependent regret grows with $\sum_{a\in\mathcal{A}\setminus\mathcal{A}^*} \Delta_a^{-1} \log T$ asymptotically as $T \to \infty$, linearly with the number of suboptimal arms $K - A$ (Lai & Robbins, 1985; Burnetas & Katehakis, 1996). The uniform convergence requirement is relaxed in a subsequent line of work that considered bandits with large or even infinite number of arms (Chaudhuri & Kalyanakrishnan, 2017; 2018; Aziz et al., 2018; De Heide et al., 2021; Zhu & Nowak, 2020). Specializing back to our setting, their results give an instance-dependent regret of $K \log T \log(\Delta^{-1})(A\Delta)^{-1}$ (De Heide et al., 2021) and correspondingly, $\widetilde{O}(\sqrt{KT/A})$ worst case regret, which is near-optimal when $K - A = \Omega(K)$ (Zhu & Nowak, 2020), via an algorithmic sub-sampling (Wang et al., 2008; Bayati et al., 2020; Kalvit & Zeevi, 2020). Despite these advances, the minimax characterization of regret is restricted to the regime $K - A = \Omega(K)$, since the standard lower bound techniques like Le Cam's method are ineffective at $K - A = o(K)$. In offline setting, Ji et al. (2026) provided a near-optimal characterization for all $A \in (0, K)$, when restricting to uniform behavior policy, through a multiplicative change-of-measure appeared in Degenne & Koolen (2019); Garivier & Kaufmann (2021). However, the corresponding result in online setting is not yet clear.

Adaptivity in Bandits. Adaptivity asks whether a learner can match the regret attainable when a problem parameter is known, without receiving that parameter as input. Some examples of affirmative results of adaptivity include adapting to unknown misspecification in contextual bandit (Foster et al., 2020), unknown weak corruption budgets (Wei et al., 2022; Liu et al., 2024), and unknown noise variance (Zhao et al., 2023). Despite these results, other adaptation problems demonstrate intrinsic

hardness, leading to a number of hardness results. Examples in this category include adapting to unknown corruption tolerance in strong corruption (Bogunovic et al., 2021; He et al., 2022), unknown smoothness classes in continuum-armed bandit (Locatelli & Carpentier, 2018; Hadiji, 2019) and nonparametric contextual bandit (Gur et al., 2022), unknown reward range (Hadiji & Stoltz, 2023), unknown reward noise distribution (Ashutosh et al., 2021; Genalti et al., 2024), and unknown smallest realizable function class within a nested function class (Zhu & Nowak, 2022; Marinov & Zimmert, 2021). For the proportion of optimal arms that is studied in this paper, Zhu & Nowak (2020) prove non-adaptivity and characterize Pareto-optimal gap-independent rates, and De Heide et al. (2021) establish a gap-dependent non-adaptivity lower bound in the reservoir model. Compared to their proof, our hardness result also cover regimes with $A + B > K$, where A and B are the respective numbers of optimal arms in two K-armed instance classes. Any two instances, one from each class, then share an optimal arm, so the earlier constructions in Zhu & Nowak (2020); De Heide et al. (2021) based on disjoint optima cannot establish non-adaptivity in this regime.

3 Problem Setup

We consider K-armed multi-armed bandit (MAB) with A optimal arms and denote it with a tuple $(\mathcal{A}, r, T)$, where $K := |\mathcal{A}| < \infty$ is the number of actions, $r : \mathcal{A} \to [0, 1]$ is the reward function unknown to the learner, and $T \geq 1$ is the total number of interactions. At each round $t \in [T]$, the learner selects an action $a_t \in \mathcal{A}$ according to a some strategy, and observes a noisy reward $r_t = r(a_t) + \varepsilon_t$, where ε_t is 1-sub-Gaussian (Lattimore & Szepesvári, 2020, Definition 5.2). The learner's goal is to minimize the regret:

$$\mathrm{Regret}(T) = \sum_{t=1}^{T} r^* - \sum_{t=1}^{T} \mathbb{E}[r(a_t)],$$

where $r^* = \max_{a \in \mathcal{A}} r(a)$ and the expectation is taken with respect to the algorithm and the noises. In this paper, we consider the case that the set of optimal arm $\mathcal{A}^{\text{opt}} := \{a \in \mathcal{A} : r(a) = \max_{a' \in \mathcal{A}} r(a')\}$ has a known cardinality $A \in [1, K-1]$. We denote $\mathcal{A}^{\text{sub}} = \mathcal{A} \setminus \mathcal{A}^{\text{opt}}$. Our goal is to characterize the minimax optimality of the instance class

$$\inf_{\texttt{Alg}} \sup_{\nu \in \Xi_{K,A}} \mathrm{Regret}_{\texttt{Alg},\nu}(T), \tag{3.1}$$

for given K, A, T, where $\Xi_{K,A} := \{\nu : |\mathcal{A}^{\text{opt}}(\nu)| = A\}$.

4 Algorithms and Upper Bounds

In this section, we provide the algorithm for learning a K-armed bandit with A optimal arms. The core algorithm design, which first samples a portion of arm and then run a bandit algorithm on it, follows the previous algorithm design for bandit with many optimal arms (Zhu & Nowak, 2020; De Heide et al., 2021). In this paper, we provide a sharper analysis, achieving a near-optimal rate on the entire range $A \in [1, K-1]$. We first summarize the algorithm in Algorithm 1. At a high level, the algorithm first randomly samples a subset $\mathcal{B} \subseteq \mathcal{A}$ with cardinality

$$|\mathcal{B}| = L = \begin{cases} \min\left\{K - A + 1, \left\lceil \frac{\log(2T)}{\log(K(K-A)^{-1})} \right\rceil\right\} & \text{if } A \leq K/2 \\ \min\{K - A + 1, \lceil \log_2 T + 1 \rceil\} & \text{otherwise} \end{cases}. \tag{4.1}$$

Then Algorithm 1 runs a minimax-optimal algorithm on $\mathcal{B}$. The following theorem characterizes the regret upper bound of Algorithm 1.

Theorem 4.1. The regret of Algorithm 1 with L set in (4.1) on any K-armed bandit instance with horizon $T \geq K$ and exactly $A \in [1, K-1]$ optimal arms admits an upper bound

$$\mathrm{Regret}(T) = \widetilde{O}\left(\frac{K - A}{\sqrt{KA}}\sqrt{T}\right).$$

Although we analyze the same algorithmic framework as Zhu & Nowak (2020), we establish a strictly superior regret bound via a sharper theoretical analysis. Specifically, our result shaves off a factor of $(K - A)/K$ from the previous $\widetilde{O}(\sqrt{KT/A})$ bound. This improvement is particularly pronounced in bandit instances with a large proportion of optimal arms.

Algorithm 1 SMOSS (Zhu & Nowak, 2020; De Heide et al., 2021)

Require: Action set $\mathcal{A}$, number of total rounds T, number of optimal arm A if $A \leq K/2$.

1: Set L as in (4.1) and sample a subset $\mathcal{B} \subseteq \mathcal{A}$ of L distinct arms uniformly without replacement.
2: Run any minimax-optimal L-armed bandit algorithm, such as MOSS (Audibert & Bubeck, 2009), on $\mathcal{B}$ for T rounds.

Remark 4.2. Theorem 4.1 provides an improved guarantee solely in expectation. One may naturally ask whether this improvement can also be attained with high probability. In Appendix D, we show that such an extension is impossible. We prove that any algorithm satisfying the regret bound of Theorem 4.1 must suffer $\widetilde{\Omega}(\sqrt{T})$ regret with non-vanishing probability. This finding highlights an inherent separation between expected and high-probability regret in the setting of bandits with multiple optima.

4.1 Proof Overview of Theorem 4.1

We outline the proof of Theorem 4.1, deferring the full details to Appendix A. We distinguish three cases according to the composition of the randomly selected subset $\mathcal{B}$:

1. The subset $\mathcal{B}$ contains no optimal arms, corresponding to the event $\mathcal{E}_{\text{miss}} := \{\mathcal{B} \cap \mathcal{A}^{\text{opt}} = \varnothing\}$. In this case, we use the trivial regret bound of T.
2. The subset $\mathcal{B}$ contains both optimal and suboptimal arms, denoted by event $\mathcal{E}_{\text{sub}} := \{\mathcal{B} \cap \mathcal{A}^{\text{sub}} \neq \varnothing\} \cap \{\mathcal{B} \cap \mathcal{A}^{\text{opt}} \neq \varnothing\}$. In this case, running MOSS for T rounds incurs expected regret $O(\sqrt{LT})$.
3. The subset $\mathcal{B}$ contains only optimal arms, corresponding to the event $\mathcal{E}_{\text{opt}} := \{\mathcal{B} \subseteq \mathcal{A}^{\text{opt}}\}$. In this case, the regret is zero.

Previous Analysis. Previously, De Heide et al. (2021); Zhu & Nowak (2020) distinguish between missing all optimal arms and including at least one. Choosing L of order K/A, up to suitable logarithmic factors, makes $\mathbb{P}[\mathcal{E}_{\text{miss}}]$ negligible. On the complementary event $\mathcal{E}_{\text{sub}} \cup \mathcal{E}_{\text{opt}}$, the expected regret is bounded by $O(\sqrt{LT})$, yielding an overall bound of $\widetilde{O}(\sqrt{KT/A})$. This rate is optimal up to logarithmic factors when $K - A = \Omega(K)$.

Our Approach. The previous analysis is not sharp when A is close to K, because it does not exploit the zero regret incurred on $\mathcal{E}_{\text{opt}}$. In this regime, each sampled arm is suboptimal with probability only $(K - A)/K$, and the subset size is $L = \widetilde{O}(1)$. A union bound therefore gives

$$\mathbb{P}[\mathcal{E}_{\text{sub}}] \leq \mathbb{P}[\mathcal{E}_{\text{miss}} \cup \mathcal{E}_{\text{sub}}] \leq L\frac{K-A}{K} = \widetilde{O}\left(\frac{K-A}{K}\right).$$

Combining this probability bound with the conditional expected regret bound $\widetilde{O}(\sqrt{T})$ on $\mathcal{E}_{\text{sub}}$ yields

$$\text{Regret}(T) \lesssim \mathbb{P}[\mathcal{E}_{\text{miss}}]T + \mathbb{P}[\mathcal{E}_{\text{sub}}]\sqrt{LT} = \widetilde{O}\left(\frac{K-A}{K}\sqrt{T}\right),$$

where the contribution from $\mathcal{E}_{\text{miss}}$ is made negligible by the choice of L. This establishes the desired rate when A is close to K.

5 Lower Bounds

In this section, we provide a lower bound of (3.1). We begin with the regime in which the number of optimal arms is small relative to K. The lower bound in this regime is suggested by previous work (Zhu & Nowak, 2020) and follows from a standard Le Cam two-point argument (see, e.g., Lattimore & Szepesvári 2020, Theorem 15.2).

Theorem 5.1. Given any K, $1 \leq A \leq K/4$ and $T \geq (K - A)/A$, for any algorithm, there exists a K-armed bandit with A optimal arms, on which the algorithm suffers from $\widetilde{\Omega}(\sqrt{A^{-1}KT})$ regret.

We next consider the regime in which A is comparable to K, where the proofs technique of previous lower bounds (De Heide et al., 2021; Zhu & Nowak, 2020) is not effective. The main difficulty arises when A is close to K: the overlap between the sets of optimal arms makes a standard two-point argument ineffective. The lower bound for this regime is stated in the following theorem.

Theorem 5.2. Given any K, $A > K/4$ and $T \geq 4/\log K$, for any algorithm, there exists a K-armed bandit with A optimal arms, on which the algorithm suffers from $\widetilde{\Omega}((K - A)K^{-1}\sqrt{T})$ regret.

Together, Theorems 5.1 and 5.2, combined with Theorem 4.1, yield

$$\inf_{\texttt{Alg}} \sup_{\nu \in \Xi_{K,A}} \mathrm{Regret}_{\texttt{Alg},\nu}(T) = \widetilde{\Theta}\left(\frac{K-A}{\sqrt{KA}}\sqrt{T}\right),$$

for all K, $1 \le A \le K-1$ and $T \ge K$. This characterizes the minimax regret of K-armed bandit with A optimal arms up to logarithmic factors across the full range of A.

5.1 Proof Overview of Theorem 5.1

In this section, we provide a quick overview of the proof of Theorem 5.1 and defer the full proof to Appendix B.1. Starts from here, unless stated otherwise, we design the noise of rewards to be standard Gaussian.

Fix a constant $\Delta > 0$ to be specified later. For the first instance, we consider $\mathcal{S}_0 \subset \mathcal{A}$ such that $|\mathcal{S}_0| = A$ and define the instance ν_0 whose reward given by $r_0(a) = 1/2 + \Delta$ for $a \in \mathcal{S}_0$ and $r_0(a) = 1/2$ otherwise. For any algorithm $\texttt{Alg}$, let $N_{\mathcal{S}}(T)$ be the number of pulls within $\mathcal{S}$ up to time T. By the pigeonhole principle, there exists some $\mathcal{S}_1 \subset \mathcal{A}$ such that $|\mathcal{S}_1| = A$ and

$$\mathbb{E}_{\nu_0,\mathsf{Alg}}[N_{\mathcal{S}_1}(T)] \le \frac{TA}{K-A} \lesssim \frac{TA}{K},$$

where the expectation is taken over the distribution jointly given by instance r_0 and Alg. Now we construct the second instance ν_1 as $r_1(a) = 1/2 + 2\Delta$ for all $a \in \mathcal{S}_1$ and $r_1(a) = r_0(a)$ otherwise. Under this construction, the distribution induced by the two instances is close, as the KL-divergence between the trajectory distributions $\mathbb{P}_0$ and $\mathbb{P}_1$ is given by

$$\sum_{a \in \mathcal{A}} \mathbb{E}_0[N_a(T)]\, \mathrm{KL}\left(\mathcal{N}(r_0(a),1) \| \mathcal{N}(r_1(a),1)\right) \lesssim \frac{TA}{K}\Delta^2, \tag{5.1}$$

where we adopt the shorthand $\mathbb{P}_i := \mathbb{P}_{\nu_i,\mathsf{Alg}}$ to denote the probability distributions over trajectories induced by the interaction between algorithm Alg and instances ν_i for $i \in \{0,1\}$. Now we pick $\Delta \sim \sqrt{K/(AT)}$, yielding $\mathsf{KL}\left(\mathbb{P}_0 \| \mathbb{P}_1\right) = O(1)$, indicating that under algorithm Alg, it is hard to distinguish ν_1 from ν_0[1].

However, we see that the optimal actions of ν_0 and ν_1 are disjoint. In other words, at each step, the learner incurs a regret of Δ on either ν_0 or ν_1 if the learner cannot distinguish ν_0 and ν_1. Since under our construction, the learner cannot reliably distinguish ν_0 and ν_1 until time step T, it must suffer from a regret of $\Omega(T\Delta) = \Omega(\sqrt{KT/A})$ on one of the instances.

5.2 Proof Overview of Theorem 5.2

The preceding proof of Theorem 5.1 relies on two instances whose trajectory distributions are difficult to distinguish but whose *sets of optimal arms are disjoint*. Disjointness is essential: if the instances share an optimal arm, an algorithm can repeatedly pull that arm and incur zero regret on both instances without distinguishing between them.

When $A > K/2$, however, any two instances in $\Xi_{K,A}$ share at least one optimal arm. An argument based solely on such a pair therefore cannot establish the desired lower bound. To overcome this obstacle, we consider a larger collection of instances, where each instance is given by subtracting the reward by a small gap Δ on some subset of $\mathcal{A}$ with cardinality $K - A$. Formally, we state the instances as indexed by

$$\mu \in \mathcal{V}_A := \left\{ \boldsymbol{\mu} \in \{1/2, 1/2-\Delta\}^K \,\middle|\, \sum_{i=1}^{K} \mathbb{1}(\mu_i = 1/2) = A \right\},$$

where $\Delta = \widetilde{\Theta}(1/\sqrt{T})$. Given any $\mu, \nu \in \mathcal{V}_A$, the reward distribution on any of the arms differs by only Δ, making any algorithm unable to distinguish between any individual instance among $\mathcal{V}_A$. To formally state this, we introduce an auxiliary instance θ whose reward vector is given by $\boldsymbol{\theta} = (1/2, \cdots, 1/2)^\top \in \mathbb{R}^K$, and let $\mathcal{V} = \mathcal{V}_A \cup \{\theta\}$.

[1] In general, the KL-divergence $\mathsf{KL}\left(\mathbb{P} \| \mathbb{Q}\right)$ between two distributions $\mathbb{P}$ and $\mathbb{Q}$ being a constant indicates that $\mathbb{P}$ and $\mathbb{Q}$ cannot be reliably distinguished.

We first show that, for any $\mu \in \mathcal{V}_A$, the induced distribution is sufficiently close to θ. We do this by computing the KL-divergence between the two distribution $\mathbb{P}_\mu$ and $\mathbb{P}_\theta$, induced by a fixed algorithm $\texttt{Alg}$, and μ, θ correspondingly. By the chain rule of KL-divergence, we see that

$$\mathsf{KL}\left(\mathbb{P}_\theta \| \mathbb{P}_\mu\right) = \frac{\Delta^2}{2}\mathbb{E}_{a\in\mathcal{A}^{\text{sub}}(\mu)}[N_a(T)] \le T\Delta^2 = O(1),$$

where $N_a(T)$ denotes the number of pulls on arm a up to T interactions, and the inequality holds due to $\mathbb{E}_{a\in\mathcal{A}^{\text{sub}}(\mu)}[N_a(T)] \le T$. Recall that, an $O(1)$ KL-divergences indicates that $\mathbb{P}_\mu$ and $\mathbb{P}_\theta$ cannot be reliably distinguished. This implies that, $\texttt{Alg}$ cannot behaves too distinctively on μ and θ. To characterize this, we need the following multiplicative change-of-measure proposition. See Remark B.2 for its previous applications in proving bandit lower bounds.

Proposition 5.3. Let $\mathbb{P}$ and $\mathbb{Q}$ be probability measures satisfying $\mathbb{P} \ll \mathbb{Q}$ and define the log-likelihood ratio $L = \log(\,\mathrm{d}\mathbb{P}/\,\mathrm{d}\mathbb{Q})$. Then, for every bounded random variable $X \in [0, b]$ and every $\gamma \in \mathbb{R}$,

$$\mathbb{E}_{\mathbb{P}}[X] \le e^{\gamma}\mathbb{E}_{\mathbb{Q}}[X] + b\mathbb{P}[L > \gamma]. \tag{5.2}$$

We apply Proposition 5.3 with $\mathbb{Q} = \mathbb{P}_\mu$, $\mathbb{P} = \mathbb{P}_\theta$, and $\gamma \approx \mathsf{KL}\left(\mathbb{P}_\theta \| \mathbb{P}_\mu\right)$. Since $\mathbb{E}_\theta[L] = \mathsf{KL}\left(\mathbb{P}_\theta \| \mathbb{P}_\mu\right)$, a careful selection of γ actually allows $\mathbb{P}[L > \gamma]$ negligible. The application of Proposition 5.3 results in the following change-of-measure inequality

$$\mathbb{E}_\mu[X] \gtrsim \mathbb{E}_\theta[X] - o\left(\frac{T}{K}\right), \tag{5.3}$$

for any random variable $X \in [0, T]$.

However, since on the instance θ, the rewards on all the arms are equal, the algorithm $\texttt{Alg}$ must allocate $T(K-A)/K$ pulls on one of the subset $\mathcal{S} \subset \mathcal{A}$ with $|\mathcal{S}| = K - A$ by symmetry. Now we consider the instance $\mu \in \mathcal{V}_A$ such that $\mathcal{A}^{\text{sub}}(\mu) = \mathcal{S}$. We transfer the behavior of $\texttt{Alg}$ on θ to μ via (5.3). In particular, substitute X with $N_{\mathcal{A}^{\text{sub}}(\mu)}(T) = \sum_{t=1}^{T} \mathbb{1}\{a_t \in \mathcal{A}^{\text{sub}}(\mu)\}$, we obtain

$$\mathbb{E}_\mu[N_{\mathcal{A}^{\text{sub}}(\mu)}(T)] \gtrsim \mathbb{E}_\theta[N_{\mathcal{A}^{\text{sub}}(\mu)}(T)] - o\left(\frac{T}{K}\right) \ge \frac{K-A}{K}T - o\left(\frac{T}{K}\right) = \Omega\left(\frac{K-A}{K}T\right).$$

Finally, recall that $\mathrm{Regret}_{\mu,\texttt{Alg}}(T) = \Delta\mathbb{E}_\mu[N_{\mathcal{A}^{\text{sub}}(\mu)}(T)]$, we see that

$$\mathrm{Regret}_{\mu,\texttt{Alg}}(T) \gtrsim \Omega\left(\frac{K-A}{K}T\Delta\right) = \widetilde{\Omega}\left(\frac{K-A}{K}\sqrt{T}\right),$$

where the last equation is obtained from $\Delta = \Theta(1/\sqrt{T})$. This gives the desired lower bound.

6 Impossibility to Adapt to Number of Optimal Arms

When $A = o(K)$, Algorithm 1 requires A to know the sample size L to achieve the near-minimax rate. In this section, we show that such knowledge of the A is necessary. In particular, we show that, any algorithm cannot achieve (near)-minimax rate simultaneously on $\Xi_{K,A}$ and $\Xi_{K,B}$, if at least one of A or B is $o(K)$. From a technical perspective, we split the results into two disjoint regimes: $A + B \le K$ and $A + B > K$. We first consider the simpler case $B \le K - A$.

Theorem 6.1. Let $1 \le A \le K - 1$, and consider any algorithm Alg satisfying that there exists some $\Gamma_{K,T} = \mathrm{polylog}(K, T)$ such that for all $T > 0$,

$$\sup_{\nu:\,|\mathcal{A}^{\text{opt}}(\nu)|=A} \mathrm{Regret}_{\nu,\mathsf{Alg}}(T) \le \Gamma_{K,T}\frac{K-A}{\sqrt{KA}}\sqrt{T}.$$

Then $\forall\ B \le \min\{A/(8\Gamma_{K,T}), K - A\}$ and $T \ge \max\{256B^2\Gamma_{K,T}^2/(KA), KA/(16B^2\Gamma_{K,T}^2)\}$, there exists a K-armed bandit instance ν with exactly B optimal arms such that

$$\mathrm{Regret}_{\nu,\mathsf{Alg}}(T) \ge \Omega\left(\frac{\sqrt{KA}}{B\Gamma_{K,T}}\sqrt{T}\right). \tag{6.1}$$

Theorem 6.1, in terms of both technique and results statement, follows De Heide et al. (2021, Theorem 3). Thus, it also admits constraints to the range of B similar to De Heide et al. (2021). The constraint

$B = \widetilde{O}(A)$ is mild, as if $B = \widetilde{\Omega}(A)$, then running SMOSS with B also gives a near-optimal rate on $\Xi_{K,A}$ up to a factor of $O(A/B) = \widetilde{O}(1)$. The second constraint, $B \leq K - A$, however, is intrinsic to the proof technique. Relaxing this constraint requires fundamentally different approach. We provide an overview of the proof technique of Theorem 6.1, along with its limitation, in Section 6.1.

We next state the results in the regime $B \geq K - A$ as follows,

Theorem 6.2. Let $K/2 \leq A \leq K-1$, and consider any algorithm Alg achieving near-optimal regret on $\Xi_{K,A}$, that is, for all $T > 0$ and some fixed $\Gamma_{K,T} = \mathrm{polylog}(K, T)$

$$\sup_{\nu:\, |\mathcal{A}^{\mathrm{opt}}(\nu)|=A} \mathrm{Regret}_{\nu,\mathsf{Alg}}(T) \leq \Gamma_{K,T} \frac{K-A}{K}\sqrt{T}.$$

Then, for any B satisfying $K - A \leq B \leq K/2$ and $B\sqrt{\log(32B/(K-A))} \leq K/(256\Gamma_{K,T})$ and any $T \geq \max\{K^2\big(B^2\Gamma_{K,T}^2 \log(32B/(K-A))\big)^{-1}, 4\}$, there exists a K-armed bandit instance ν with exactly B optimal arms such that

$$\mathrm{Regret}_{\nu,\mathsf{Alg}}(T) \geq \widetilde{\Omega}\left(\frac{\sqrt{KA}}{B\Gamma_{K,T}}\sqrt{T}\right). \tag{6.2}$$

Ignoring all logarithmic factor, Theorem 6.1 and Theorem 6.2 together show that, when $B = O(A)$, any algorithm that is near-minimax optimal on $\Xi_{K,A}$ must suffer a regret of at least $\widetilde{\Omega}(\sqrt{KAT}/B)$, which demonstrates a multiplicative loss of $\widetilde{\Omega}(\sqrt{A/B})$ compared to the minimax optimal rate $\widetilde{O}(\sqrt{KT/B})$ on $\Xi_{K,B}$. These results indicates that, algorithm agnostic to the number of optimal arms cannot simultaneously attain the near-optimal regret on $\Xi_{K,A}$ and $\Xi_{K,B}$ if $B = o(A)$, necessitating the knowledge of the number of optimal arms up to logarithmic factors.

Remark 6.3. Theorems 6.1 and 6.2 together show that, when $B = o(A)$, an algorithm cannot be simultaneously optimal on both $\Xi_{K,A}$ and $\Xi_{K,B}$. This tradeoff motivates the notion of a Pareto-optimal algorithm (Zhu & Nowak, 2020): one for which no other algorithm achieves better performance pointwise over all $1 \leq A \leq K-1$. Designing a Pareto-optimal algorithm for our setting remains an interesting direction for future work.

6.1 Proof Overview of Theorem 6.1

In this section, we provide an overview of the proof of Theorem 6.1. At a high level, the proof follows De Heide et al. (2021) and considers the two-point-type instances, and then shows that algorithm cannot reliably distinguish these two instances through a TV-based change-of-measure argument.

For simplicity, we constraint us to the case $K - A = \Omega(K)$. At a high level, our goal is to consider two instance $\nu_0 \in \Xi_{K,A}$ and $\nu_1 \in \Xi_{K,B}$, such that they are similar enough, yet any algorithm must suffer from sufficiently large regret on at least one of them. To do this, we consider partition $\mathcal{A}$ as follows, where $\mathcal{S}_0$ is fixed and $\mathcal{S}_1$ determined later

$$\mathcal{A} = \mathcal{S}_0 \cup \mathcal{S}_1 \cup \mathcal{S}_2, \quad \text{where } |\mathcal{S}_0| = A,\ |\mathcal{S}_1| = B, \text{ and },|\mathcal{S}_2| = K - A - B.$$

We use r_0 for the reward mean of ν_0 and r_1 for the reward of ν_1, which are given by

$$r_0(a) = \begin{cases} 1/2, & a \in \mathcal{S}_0 \\ 1/2 - \Delta, & a \in \mathcal{S}_1 \\ 1/2 - \Delta, & a \in \mathcal{S}_2 \end{cases}, \qquad r_1(a) = \begin{cases} 1/2, & a \in \mathcal{S}_0 \\ 1/2 + \Delta, & a \in \mathcal{S}_1 \\ 1/2 - \Delta, & a \in \mathcal{S}_2 \end{cases}.$$

It is easy to verify that $\nu_0 \in \Xi_{K,A}$ and $\nu_1 \in \Xi_{K,B}$. The key observation is that, since arms in $\mathcal{S}_0^{\complement}$ is sub-optimal under ν_0, any algorithm `Alg` optimal on $\Xi_{K,A}$ cannot allocate too much pulls on $\mathcal{S}_0^{\complement}$, therefore one of its subset $\mathcal{S}_1$. In particular, fix such algorithm and drop the subscripts of `Alg`, if $\mathrm{Regret}(T) \leq R_A$, we must observe

$$\mathbb{E}_0[N_{\mathcal{S}_0^{\complement}}(T)] \leq \frac{R_A}{\Delta},$$

where $\mathbb{E}_0$ is a short hand for $\mathbb{E}_{\nu_0}$ and similar abbreviation applies to ν_1. By symmetry, there exist a subset $\mathcal{S} \subseteq \mathcal{S}_0^{\complement}$ such that $|\mathcal{S}| = B$ and

$$\mathbb{E}_0[N_{\mathcal{S}}(T)] \leq \frac{BR_A}{(K-A)\Delta}.$$

Without loss of generality, we pick $\mathcal{S}_1 = \mathcal{S}$. With this construction, we see that (i) since `Alg` does not allocate sufficient pulls on $\mathcal{S}_1$, `Alg` cannot reliably distinguish ν_1 from ν_0; and (ii) the optimal actions of ν_0 and ν_1 are disjoint, therefore `Alg` cannot be optimal simultaneously on them.

Formally, we first show that `Alg` cannot reliably distinguish ν_1 from ν_0, through a computation of the KL-divergence between there distribution. Taking $\Delta = \Theta((K-A)/(BR_A))$

$$\mathsf{KL}\left(\mathbb{P}_0 \| \mathbb{P}_1\right) = \mathbb{E}_0[N_{\mathcal{S}_1}(T)] \cdot (2\Delta)^2 = O\bigg(\frac{BR_A\Delta}{K-A}\bigg) = O(1).$$

In this case, the distribution $\mathbb{P}_0$ and $\mathbb{P}_1$ is close enough, therefore the random variable $N_{\mathcal{S}_1}(T)$ behaves similarly under these two distribution. In particular, a change-of-measure leads to

$$\mathbb{E}_1[N_{\mathcal{S}_1}(T)] \le \mathbb{E}_0[N_{\mathcal{S}_1}(T)] + T\mathrm{TV}(\mathbb{P}_0, \mathbb{P}_1) \le \mathbb{E}_0[N_{\mathcal{S}_1}(T)] + T\sqrt{\frac{1}{2}\mathsf{KL}\left(\mathbb{P}_0 \| \mathbb{P}_1\right)},$$

where the last inequality holds due to Pinsker's inequality. Plugging in $\Delta = \Theta((K-A)/(BR_A))$ into $\mathbb{E}_0[N_{\mathcal{S}_1}(T)] \le BR_A/((K-A)\Delta)$ and $\mathsf{KL}\left(\mathbb{P}_0 \| \mathbb{P}_1\right) = O(1)$ gives

$$\mathbb{E}_1[N_{\mathcal{S}_1}(T)] \le \frac{B^2R_A^2}{(K-A)^2} + O(T) = \widetilde{O}\bigg(\frac{B^2}{(K-A)^2}\frac{(K-A)^2T}{KA} + T\bigg) = \widetilde{O}(T),$$

where the first equation holds by plugging in the rate in Theorem 4.1. Finally, a careful selection of constants enables $T - \mathbb{E}_1[N_{\mathcal{S}_1}(T)] = \Omega(T)$, thus

$$\mathrm{Regret}_{\nu_1}(T) \ge (T - \mathbb{E}_1[N_{\mathcal{S}_1}(T)])\Delta = \Omega(T\Delta) = \Omega\bigg(\frac{T(K-A)}{BR_A}\bigg) = \widetilde{\Omega}\bigg(\frac{\sqrt{KA}}{B}\sqrt{T}\bigg),$$

which gives the desired rate.

6.2 Proof Overview of Theorem 6.2

Similar to Theorem 5.1, the proof of Theorem 6.1 relies on the two-point construction that requires *the two instances have disjoint sets of optimal arms*. However, in the case $A + B > K$, any instance $\nu_0 \in \Xi_{K,A}$ and $\nu_1 \in \Xi_{K,B}$ must share at least one optimal arm $a \in \mathcal{A}^{\mathrm{opt}}(\nu_0) \cap \mathcal{A}^{\mathrm{opt}}(\nu_1)$, preventing an application in this regime.

We overcome this obstacle by showing that any near-optimal algorithm must exhibit a non-uniform exploration profile akin to that of Algorithm 1, therefore ruling out algorithms that concentrate their pulls on a fixed arm or set of arms. Yet even under a fixed algorithm, different instances in $\Xi_{K,A}$ induce different trajectory distributions, complicating a unified characterization of its behavior. We therefore characterize the algorithm's exploration profile on a common reference instance θ, in which all arms have the same mean reward. Such uniformed characterization is enabled by the change-of-measure in Proposition 5.3, if we select the instances in $\Xi_{K,A}$ close enough to θ. Formally, we have the following lemma.

Lemma 6.4. Let $A = \widetilde{\Omega}(K)$, and suppose that `Alg` is minimax-optimal on $\Xi_{K,A}$ up to logarithmic factors, namely

$$\sup_{\nu:\, |\mathcal{A}^{\mathrm{opt}}(\nu)|=A} \mathrm{Regret}_{\nu,\mathtt{Alg}}(T) \le \Gamma_{K,T}\frac{K-A}{\sqrt{KA}}\sqrt{T},$$

where $\Gamma_{K,T} = \mathrm{polylog}(K,T)$. Let θ denote the all-equal instance. Then, ignoring logarithmic factors, for every $m \in [T]$, and every $\mathcal{S} \subset \mathcal{A}$ with $|\mathcal{S}| = K - A$,

$$\mathbb{P}_\theta\big[N_{\mathcal{S}}(T) \ge m\big] \lesssim \widetilde{O}\bigg(\frac{K-A}{K}\sqrt{\frac{T}{m}}\bigg).$$

Moreover, for any $\mathcal{U} \subseteq \mathcal{A}$ such that $|\mathcal{U}| = B \ge K - A$, we have

$$\mathbb{P}_\theta\big[N_{\mathcal{U}}(T) \ge m\big] \lesssim \widetilde{O}\bigg(\frac{B}{K}\sqrt{\frac{T}{m}}\bigg). \tag{6.3}$$

Lemma 6.4 reveals a mismatch between the exploration profiles required for near-optimality on $\Xi_{K,A}$ and $\Xi_{K,B}$ in the regime $K - A \le B \ll K$. To illustrate this mismatch, suppose for simplicity that B

divides K, and partition the arms of θ into K/B groups of size B, treating each group as a super-arm. We compare the exploration levels that algorithms $\texttt{Alg}_A$ and $\texttt{Alg}_B$, near-optimal on their respective instance classes, can attain with high probability on each fixed super-arm.

For $\texttt{Alg}_B$, we can prove that for every fixed $\mathcal{U} \subseteq \mathcal{A}$ with $|\mathcal{U}| = B$, there is a threshold

$$m_B = \widetilde{\Omega}\left(\frac{TB}{K}\right), \text{ such that } N_{\mathcal{U}}(T) \geq m_B,$$

with high probability under θ. Thus, at the level of super-arms, $\texttt{Alg}_B$ must explore relatively uniformly: each fixed super-arm receives at least $\widetilde{\Omega}(TB/K)$ pulls with high probability[2].

In contrast, for $\texttt{Alg}_A$, inequality (6.3) implies that one can choose a threshold

$$m_A = \widetilde{O}\left(\frac{TB^2}{K^2}\right), \text{ for which, } \mathbb{P}_{\theta,\texttt{Alg}_A}[N_{\mathcal{U}}(T) \geq m_A] \leq 1/2,$$

provided this threshold lies within the horizon. Consequently, $\texttt{Alg}_A$ cannot ensure more than $\widetilde{O}(TB^2/K^2)$ pulls of any fixed super-arm with high probability. When $B \ll K$, this scale is smaller than the exploration level required of $\texttt{Alg}_B$, up to logarithmic factors. This incompatibility yields the desired non-adaptivity result: the exploration profile imposed by near-optimality on $\Xi_{K,A}$ precludes the exploration needed to handle the hardest instances in $\Xi_{K,B}$.

Remark 6.5. Lemma 6.4 shows that, when A is close to K, near-optimal algorithms must exhibit non-uniform exploration behavior, consistent with the sub-sampling approach in Section 4. A more precise characterization of "sub-sampling-like" behavior, and whether such behavior is necessary for near-optimality, remains an interesting direction for future work.

Formally, we begin with any fixed $\mathcal{U} \subseteq \mathcal{A}$ such that $|\mathcal{U}| = B$, Lemma 6.4 shows that only an exploration of level $\widetilde{O}(TB^2/K^2)$ is guaranteed. Therefore, taking $m = \widetilde{\Theta}(TB^2/K^2)$ with properly selected constants, we can ensure that

$$\mathbb{P}_\theta\big[N_{\mathcal{U}}(T) \geq m\big] \lesssim \widetilde{O}\left(\frac{B}{K}\sqrt{\frac{T}{m}}\right) < 1 - \Omega(1).$$

On the other side, an exploration of level $\widetilde{O}(TB^2/K^2)$ is not sufficient for identifying hard-enough instance. Particularly, we consider the instance ν with reward function

$$r_\nu(a) = \begin{cases} 1/2 + \Delta, & a \in \mathcal{U} \\ 1/2, & a \in \mathcal{U}^{\complement} \end{cases},$$

where $\Delta = \widetilde{\Theta}(1/\sqrt{m})$. Once again, under the event $\{N_{\mathcal{U}}(T) \leq m\}$, algorithm allocates only at most m to $\mathcal{U}$, leading to $\mathsf{KL}\left(\mathbb{P}_\theta \| \mathbb{P}_\nu\right) = O(1)$, therefore the learner are not able to reliably distinguish θ and ν in this case. Formally, under the event $\{N_{\mathcal{U}}(T) \leq m\}$, the distribution of $\mathbb{P}_\nu$ and $\mathbb{P}_\theta$ is close enough, so change of measure is at a cost of only multiplicative constant:

$$\mathbb{P}_\nu\big[N_{\mathcal{U}}(T) \geq m\big] \lesssim \mathbb{P}_\theta\big[N_{\mathcal{U}}(T) \geq m\big] < 1 - \Omega(1).$$

Recall that on ν, the regret is given by $\Delta\big(T - \mathbb{E}_\nu[N_{\mathcal{U}}(T)]\big)$, therefore,

$$\mathrm{Regret}_\nu(T) \geq \mathbb{P}_\nu\big[N_{\mathcal{U}}(T) \leq m\big](T-m)\Delta = \Omega(T\Delta) = \widetilde{\Omega}\left(\frac{K}{B}\sqrt{T}\right),$$

where the last inequality holds due to $\Delta = \widetilde{\Theta}(1/\sqrt{m}) = \Theta(K/(B\sqrt{T}))$. This gives the desired rate and finishes the proof.

7 Conclusion

In this paper, we study regret minimization in K-armed bandits with A optimal arms. We first provide a sharper analysis of the previous sub-sampling algorithm and lower bounds for the problem. These results together establishes a $\widetilde{\Theta}\left(\frac{K-A}{\sqrt{KA}}\sqrt{T}\right)$ minimax regret. We also show that near-optimality imposes exploration constraints that prevent uniform adaptation across sufficiently many different models with distinct numbers of optimal arms. Directions for future work include closing the remaining logarithmic gaps, designing Pareto optimal algorithms and determining which aspects of the exploration behavior induced by sub-sampling are necessary for minimax optimality.

[2]See Appendix C.3 for the formal statement and proof.

AI USE STATEMENT

In this work, we used generative AI tools to formulate mathematical claims, identify key ideas for their proofs, and assist in drafting parts of the proofs. We also used these tools to summarize existing literature, identify relevant work, and improve the paper's readability. We did not use generative AI tools to develop theoretical models or conceptual frameworks, translate material, or interpret results. Because this is a purely theoretical work, generating synthetic data, designing or providing feedback on experiments or research methodology, implementing methods, cleaning or reformatting data, and conducting qualitative or thematic data analysis were not applicable. The authors reviewed all AI-assisted material. In particular, we checked and mathematical statements and proofs, reorganized proof arguments and rewrote proofs in our own words for rigorousness and human-readability. We also verified relevant literature suggested by the tools, and reviewed all AI-proposed language edits to make sure they accurately reflect the authors' intended meaning. We take responsibility for the final content of this work, including any text, claims, or other material produced with the assistance of generative AI.

ETHICS STATEMENT

This work is purely theoretical: it characterizes the minimax regret of multi-armed bandits with multiple optimal arms. The paper itself involves no human subjects, collects no data, and uses no datasets. While bandit algorithms, involving those studied in this paper, might be deployed in applications such as recommendation and model routing, where fairness and privacy considerations can arise, our results concern the statistical limits of learning and do not introduce risks beyond those of existing bandit methods.

REPRODUCIBILITY STATEMENT

This work is purely theoretical and contains no experiments. We provide the formal definition of the problem of interest, as well as all modeling assumptions in Section 3. Our major theoretical results, introduced in Section 4, Section 5, as well as intermediate and supplementary results in Section 6, are formally stated, if not in the main text, and proved in Appendix A, Appendix B and Appendix C. The high-probability lower bound mentioned in Remark 4.2 is formally stated and proved in Appendix D. All supplementary technical results required for the proofs throughout the appendix are stated in Appendix E with the corresponding references.

REFERENCES

Kumar Ashutosh, Jayakrishnan Nair, Anmol Kagrecha, and Krishna Jagannathan. Bandit algorithms: Letting go of logarithmic regret for statistical robustness. In *International Conference on Artificial Intelligence and Statistics*, pp. 622–630. PMLR, 2021.

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *Colt*, pp. 217–226, 2009.

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

Maryam Aziz, Jesse Anderton, Emilie Kaufmann, and Javed Aslam. Pure exploration in infinitely-armed bandit models with fixed-confidence. In *Algorithmic Learning Theory*, pp. 3–24. PMLR, 2018.

Mohsen Bayati, Nima Hamidi, Ramesh Johari, and Khashayar Khosravi. Unreasonable effectiveness of greedy algorithms in multi-armed bandit with many arms. *Advances in Neural Information Processing Systems*, 33:1713–1723, 2020.

Ilija Bogunovic, Arpan Losalka, Andreas Krause, and Jonathan Scarlett. Stochastic linear bandits robust to adversarial attacks. In *International Conference on Artificial Intelligence and Statistics*, pp. 991–999. PMLR, 2021.

Djallel Bouneffouf, Irina Rish, and Charu Aggarwal. Survey on applications of multi-armed and contextual bandits. In *2020 IEEE Congress on Evolutionary Computation (CEC)*, pp. 1–8, 2020.

Apostolos N Burnetas and Michael N Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142, 1996.

Arghya Roy Chaudhuri and Shivaram Kalyanakrishnan. Pac identification of a bandit arm relative to a reward quantile. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pp. 1777–1783, 2017.

Arghya Roy Chaudhuri and Shivaram Kalyanakrishnan. Quantile-regret minimisation in infinitely many-armed bandits. In *UAI*, pp. 425–434, 2018.

Rianne De Heide, James Cheshire, Pierre Ménard, and Alexandra Carpentier. Bandits with many optimal arms. *Advances in Neural Information Processing Systems*, 34:22457–22469, 2021.

Rémy Degenne and Wouter M Koolen. Pure exploration with multiple correct answers. *Advances in Neural Information Processing Systems*, 32, 2019.

Dylan J Foster, Claudio Gentile, Mehryar Mohri, and Julian Zimmert. Adapting to misspecification in contextual bandits. *Advances in Neural Information Processing Systems*, 33:11478–11489, 2020.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pp. 998–1027. PMLR, 2016.

Aurélien Garivier and Emilie Kaufmann. Nonasymptotic sequential tests for overlapping hypotheses applied to near-optimal arm identification in bandit models. *Sequential Analysis*, 40(1):61–96, 2021.

Aurélien Garivier, Pierre Ménard, and Gilles Stoltz. Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 44(2):377–399, 2019.

Gianmarco Genalti, Lupo Marsigli, Nicola Gatti, and Alberto Maria Metelli. (ε, u)-adaptive regret minimization in heavy-tailed bandits. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 1882–1915. PMLR, 2024.

Yonatan Gur, Ahmadreza Momeni, and Stefan Wager. Smoothness-adaptive contextual bandits. *Operations Research*, 70(6):3198–3216, 2022.

Hédi Hadiji. Polynomial cost of adaptation for x-armed bandits. *Advances in Neural Information Processing Systems*, 32, 2019.

Hédi Hadiji and Gilles Stoltz. Adaptation to the range in k-armed bandits. *Journal of Machine Learning Research*, 24(13):1–33, 2023.

Jiafan He, Dongruo Zhou, Tong Zhang, and Quanquan Gu. Nearly optimal algorithms for linear contextual bandits with adversarial corruptions. *Advances in neural information processing systems*, 35:34614–34625, 2022.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17(none):257 – 317, 2020. doi: 10.1214/18-PS321.

Naman Jain, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *International Conference on Learning Representations*, volume 2025, pp. 58791–58831, 2025.

Kaixuan Ji, Qiwei Di, Heyang Zhao, Qingyue Zhao, and Quanquan Gu. On the optimal sample complexity of offline multi-armed bandits with kl regularization. *arXiv preprint arXiv:2605.02141*, 2026.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. Swe-bench: Can language models resolve real-world github issues? In *International Conference on Learning Representations*, volume 2024, pp. 54107–54157, 2024.

Anand Kalvit and Assaf Zeevi. From finite to countable-armed bandits. *Advances in neural information processing systems*, 33:8259–8269, 2020.

Hideaki Kano, Junya Honda, Kentaro Sakamaki, Kentaro Matsuura, Atsuyoshi Nakamura, and Masashi Sugiyama. Good arm identification via bandit feedback. *Machine Learning*, 108(5): 721–745, 2019.

Julian Katz-Samuels and Kevin Jamieson. The true sample complexity of identifying good arms. In *International Conference on Artificial Intelligence and Statistics*, pp. 1781–1791. PMLR, 2020.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

Tor Lattimore. Refining the confidence level for optimistic bandit strategies. *Journal of Machine Learning Research*, 19(20):1–32, 2018.

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

Haolin Liu, Artin Tajdini, Andrew Wagenmaker, and Chen-Yu Wei. Corruption-robust linear bandits: Minimax optimality and gap-dependent misspecification. *Advances in Neural Information Processing Systems*, 37:24277–24325, 2024.

Andrea Locatelli and Alexandra Carpentier. Adaptivity to smoothness in x-armed bandits. In *Conference on Learning Theory*, pp. 1463–1492. PMLR, 2018.

Teodor Vanislavov Marinov and Julian Zimmert. The pareto frontier of model selection for general contextual bandits. *Advances in Neural Information Processing Systems*, 34:17956–17967, 2021.

Pranoy Panda, Raghav Magazine, Chaitanya Devaguptapu, Sho Takemori, and Vishal Sharma. Adaptive llm routing under budget constraints. In *EMNLP (Findings)*, pp. 23934–23949, 2025.

Reid Pryzant, Dan Iter, Jerry Li, Yin Lee, Chenguang Zhu, and Michael Zeng. Automatic prompt optimization with "gradient descent" and beam search. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 7957–7968, 2023.

Lan V Truong. Optimal best-arm identification under fixed confidence with multiple optima. *IEEE Transactions on Information Theory*, 2026.

Alexandre B Tsybakov and Vladimir Zaiats. *Introduction to nonparametric estimation*, volume 11. Springer, 2009.

Yizao Wang, Jean-Yves Audibert, and Rémi Munos. Algorithms for infinitely many-armed bandits. *Advances in Neural Information Processing Systems*, 21, 2008.

Chen-Yu Wei, Christoph Dann, and Julian Zimmert. A model selection approach for corruption robust reinforcement learning. In *International Conference on Algorithmic Learning Theory*, pp. 1043–1096. PMLR, 2022.

Bin Yu. Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam: research papers in probability and statistics*, pp. 423–435. Springer, 1997.

Heyang Zhao, Jiafan He, Dongruo Zhou, Tong Zhang, and Quanquan Gu. Variance-dependent regret bounds for linear bandits and reinforcement learning: Adaptivity and computational efficiency. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 4977–5020. PMLR, 2023.

Yinglun Zhu and Robert Nowak. On regret with multiple best arms. *Advances in Neural Information Processing Systems*, 33:9050–9060, 2020.

Yinglun Zhu and Robert Nowak. Pareto optimal model selection in linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 6793–6813. PMLR, 2022.

A Proofs of Theorems in Section 4

A.1 Proof of Theorem 4.1

Proof of Theorem 4.1. Let $\mathcal{E}_{\text{miss}} = \{\mathcal{B} \cap \mathcal{A}^{\text{opt}} = \varnothing\}$ be the event that the randomly sampled subset does not contain any optimal arms, and $\mathcal{E}_{\text{sub}} = \{\mathcal{B} \cap \mathcal{A}^{\text{sub}} \neq \varnothing\} \cap \{\mathcal{B} \cap \mathcal{A}^{\text{opt}} \neq \varnothing\}$ be the event that the randomly sampled subset contains at least one sub-optimal arm and at least one optimal arm. We first bound the probability of $\mathcal{E}_{\text{miss}}$. First, if $L = K - A + 1$ then every subset of L arms must contain at least one optimal arm, since there are only $K - A$ suboptimal arms. Therefore, $\mathbb{P}[\mathcal{E}_{\text{miss}}] = 0$. Otherwise, if $A \leq K/2$, we directly have

$$L \leq \left\lceil \frac{\log(2T)}{\log(K/(K-A))} \right\rceil.$$

In the complement case $A > K/2$, we also have

$$L = \lceil \log_2 T + 1 \rceil \leq \left\lceil \frac{\log(2T)}{\log 2} \right\rceil \leq \left\lceil \frac{\log(2T)}{\log(K/(K-A))} \right\rceil.$$

Since $\mathcal{B}$ is sampled uniformly without replacement, we have

$$\mathbb{P}[\mathcal{E}_{\text{miss}}] = \binom{K-A}{L} \binom{K}{L}^{-1} = \prod_{i=0}^{L-1} \frac{K-A-i}{K-i} \leq \left(\frac{K-A}{K}\right)^L,$$

where the inequality holds due to $(K-A-i)/(K-i) \leq (K-A)/K$ for every $i \in 0, \ldots, L-1$. Plugging in the value of L, we see that

$$\mathbb{P}[\mathcal{E}_{\text{miss}}] \leq \left(\frac{K-A}{K}\right)^L = \exp\left(- L \log\left(\frac{K}{K-A}\right)\right) \leq \exp\big(- \log(2T)\big) = \frac{1}{2T}.$$

Combining the two cases gives $\mathbb{P}[\mathcal{E}_{\text{miss}}] \leq (2T)^{-1}$.

We then bound the probability of event $\mathcal{E}_{\text{sub}} \cup \mathcal{E}_{\text{miss}}$. Trivially, we have $\mathbb{P}[\mathcal{E}_{\text{sub}} \cup \mathcal{E}_{\text{miss}}] \leq 1$. Moreover, by the fact that $\mathcal{E}_{\text{sub}} \cup \mathcal{E}_{\text{miss}} = \bigcup_{a \in \mathcal{A}^{\text{sub}}} \{a \in \mathcal{B}\}$

$$\mathbb{P}[\mathcal{E}_{\text{sub}} \cup \mathcal{E}_{\text{miss}}] \leq \sum_{a \in \mathcal{A}^{\text{sub}}} \mathbb{P}\big[\{a \in \mathcal{B}\}\big] \leq \frac{L(K-A)}{K},$$

which gives $\mathbb{P}[\mathcal{E}_{\text{sub}} \cup \mathcal{E}_{\text{miss}}] \leq 1 \wedge L(K-A)K^{-1}$

Now we are ready to bound the regret. On the event $\mathcal{E}_{\text{opt}} \coloneqq \big(\mathcal{E}_{\text{miss}} \cup \mathcal{E}_{\text{sub}}\big)^{\complement}$, $\mathcal{B} \subseteq \mathcal{A}^{\text{opt}}$, therefore no regret is incurred. On the event $\mathcal{E}_{\text{sub}}$, the restricted bandit instance over $\mathcal{B}$ contains at least one globally optimal arm. Thus, the best mean reward among the arms in $\mathcal{B}$ equals r^*. By the minimax regret guarantee of MOSS, there exists a universal constant $C > 0$ such that

$$\text{Regret}_{\mathcal{B}}(T) \leq C\sqrt{LT}.$$

for every realization of $\mathcal{B}$ satisfying $\mathcal{E}_{\text{sub}}$ and $T \geq K$ (see e.g., Lattimore & Szepesvári, 2020, Theorem 9.1 for a proof). Finally, On the event $\mathcal{E}_{\text{miss}}$, the per-round regret is at most 1, since all mean rewards belong to $[0, 1]$. Therefore, the total regret is at most T. Combining the analysis gives that

$$\begin{aligned}
&\text{Regret}(T) \\
&= \mathbb{E}\Big[\text{Regret}(T)\, \mathbb{1}\left(\mathcal{E}_{\text{opt}}\right)\Big] + \mathbb{E}\big[\text{Regret}(T)\, \mathbb{1}(\mathcal{E}_{\text{sub}})\big] + \mathbb{E}\big[\text{Regret}(T)\, \mathbb{1}(\mathcal{E}_{\text{miss}})\big] \\
&\leq C\sqrt{LT}\mathbb{P}[\mathcal{E}_{\text{sub}} \cup \mathcal{E}_{\text{miss}}] + T\mathbb{P}[\mathcal{E}_{\text{miss}}] \\
&\leq C\sqrt{LT}\left(\frac{L(K-A)}{K} \wedge 1\right) + \frac{1}{2}.
\end{aligned}$$

It remains to bound L. Using the elementary inequality $-\log(1-x) \geq x$ for all $x \in [0, 1)$, we obtain $\log(K/(K-A)) = -\log(1 - A/K) \geq A/K$, therefore

$$L \leq \left\lceil \frac{\log(2T)}{\log\big(K/(K-A)\big)} \right\rceil \leq 1 + \frac{\log(2T)}{\log\big(K/(K-A)\big)} \leq 1 + \frac{K}{A}\log(2T) = \widetilde{O}\left(\frac{K}{A}\right),$$

where the first inequality holds due to $\lceil x \rceil \le x + 1$. Substituting this inequality into the preceding regret bound yields

$$\mathrm{Regret}(T) = \widetilde{O}\left(\left(\frac{(K-A)}{A} \wedge 1\right)\sqrt{\frac{KT}{A}} + \frac{1}{2}\right) = \widetilde{O}\left(\frac{K-A}{\sqrt{KA}}\sqrt{T}\right).$$

Since this bound is achieved by a valid algorithm uniformly over all instances with exactly A optimal arms, the claimed minimax upper bound follows. □

B Proofs of Theorems in Section 5

B.1 Proof of Theorem 5.1

Proof of Theorem 5.1. The hard instance construction largely follows Theorem 15.2 of Lattimore & Szepesvári (2020) and a similar idea was also presented in De Heide et al. (2021). For simplicity, we define $N_{\mathcal{S}}(T) := \sum_{a\in\mathcal{S}} N_a(T)$ for subset $\mathcal{S} \subseteq \mathcal{A}$. For any algorithm Alg, we fix an arbitrary subset $\mathcal{S}_0 \subseteq \mathcal{A}$ such that $|\mathcal{S}_0| = A$. For a parameter $\Delta \in (0, 1/2]$ to be specified later, define the first bandit instance ν_0 by

$$r_0(a) := \begin{cases} \Delta, & a \in \mathcal{S}_0, \\ 0, & a \in \mathcal{S}_0^{\complement}. \end{cases}$$

Thus, ν_0 has exactly A optimal arms, namely the arms in $\mathcal{S}_0$. Let $\mathbb{P}_0$ and $\mathbb{E}_0$ denote probability and expectation under ν_0 and Alg. Since $\sum_{a\in\mathcal{S}_0^{\complement}} \mathbb{E}_0[N_a(T)] \le T$, there exists a subset $\mathcal{S}_1 \subseteq \mathcal{S}_0^{\complement}$ with $|\mathcal{S}_1| = A$ such that

$$\mathbb{E}_0[N_{\mathcal{S}_1}(T)] \le \frac{AT}{K-A}. \tag{B.1}$$

Indeed, one may choose $\mathcal{S}_1$ to consist of the A least-sampled arms in $\mathcal{S}_0^{\complement}$. Such a subset exists because $A \le K/4$, and hence $A < K - A$. We now define a second bandit instance ν_1 by

$$r_1(a) := \begin{cases} 2\Delta, & a \in \mathcal{S}_1, \\ \Delta, & a \in \mathcal{S}_0, \\ 0, & a \in (\mathcal{S}_0 \cup \mathcal{S}_1)^{\complement}. \end{cases}$$

The instance ν_1 also has exactly A optimal arms, namely the arms in $\mathcal{S}_1$. Similarly, we use $\mathbb{P}_1$ and $\mathbb{E}_1$ to denote probability and expectation under ν_1 and Alg, and let $\mathrm{Regret}_i(T)$ denote the expected regret under ν_i. Now we consider the event $\mathcal{E} := \{N_{\mathcal{S}_1}(T) \ge T/2\}$. Under ν_0, every arm in $\mathcal{S}_1$ is suboptimal with gap Δ. Consequently,

$$\mathrm{Regret}_0(T) \ge \Delta\mathbb{E}_0[N_{\mathcal{S}_1}(T)] \ge \frac{\Delta T}{2}\mathbb{P}_0[\mathcal{E}].$$

Under ν_1, every arm outside $\mathcal{S}_1$ has suboptimality gap at least Δ. Therefore,

$$\mathrm{Regret}_1(T) \ge \Delta\mathbb{E}_1[T - N_{\mathcal{S}_1}(T)] \ge \frac{\Delta T}{2}\mathbb{P}_1\big[\mathcal{E}^{\complement}\big].$$

Combining the preceding two inequalities gives

$$\mathrm{Regret}_0(T) + \mathrm{Regret}_1(T) \ge \frac{\Delta T}{2}\big(\mathbb{P}_0[\mathcal{E}] + \mathbb{P}_1[\mathcal{E}^{\complement}]\big). \tag{B.2}$$

It remains to bound the right-hand-side of (B.2). First, by Lemma E.4

$$\mathbb{P}_0[\mathcal{E}] + \mathbb{P}_1[\mathcal{E}^{\complement}] \ge \frac{1}{2}\exp\big(-\mathrm{KL}(\mathbb{P}_0\|\mathbb{P}_1)\big)$$

Since the two instances differ only on the arms in $\mathcal{S}_1$. For every $a \in \mathcal{S}_1$, the reward distribution is $\mathcal{N}(0,1)$ under ν_0 and $\mathcal{N}(2\Delta, 1)$ under ν_1. Using the chain rule for KL divergence under adaptive sampling (Lattimore & Szepesvári, 2020, Lemma 15.1),

$$\mathrm{KL}(\mathbb{P}_0\|\mathbb{P}_1) = \sum_{a\in\mathcal{S}_1} \mathbb{E}_0[N_a(T)]\,\mathrm{KL}\big(\mathcal{N}(0,1)\|\mathcal{N}(2\Delta,1)\big) = 2\Delta^2\mathbb{E}_0[N_{\mathcal{S}_1}(T)] \le \frac{2\Delta^2 AT}{K-A},$$

where the last inequality follows from (B.1). Now we choose $\Delta = \frac{1}{4}\sqrt{(K-A)(AT)^{-1}}$, which gives $\mathrm{KL}(\mathbb{P}_0\|\mathbb{P}_1) \le \frac{1}{8}$. As a result, $\mathbb{P}_0[\mathcal{E}] + \mathbb{P}_1[\mathcal{E}^{\complement}] \ge \exp(-1/8)/2$. Substituting this inequality into (B.2) and taking the max over ν_0 and ν_1 gives

$$\max_{\nu_0,\nu_1} \mathrm{Regret}(T) \ge \frac{e^{-1/8}}{8}\Delta T = \Omega\left(\sqrt{\frac{(K-A)T}{A}}\right) = \Omega\left(\sqrt{\frac{KT}{A}}\right),$$

which finishes the proof. □

B.2 Proof of Theorem 5.2

To prove Theorem 5.2, we need the following change-of-measure argument.

Proposition B.1. Let $(\Omega, \mathcal{G})$ be a measurable space, and let $\mathbb{P}$ and $\mathbb{Q}$ be probability measures on $(\Omega, \mathcal{G})$ satisfying $\mathbb{P} \ll \mathbb{Q}$. Define the log-likelihood ratio $L = \log(\mathrm{d}\mathbb{P}/\mathrm{d}\mathbb{Q})$. Then, for every $\mathcal{G}$-measurable random variable $X : \Omega \to [0, b]$ and every $\gamma \in \mathbb{R}$,

$$\mathbb{E}_{\mathbb{P}}[X] \le e^{\gamma}\mathbb{E}_{\mathbb{Q}}[X] + b\mathbb{P}[L > \gamma]. \tag{B.3}$$

Equivalently,

$$\mathbb{E}_{\mathbb{Q}}[X] \ge e^{-\gamma}\big[\mathbb{E}_{\mathbb{P}}[X] - b\mathbb{P}(L > \gamma)\big]. \tag{B.4}$$

Proof of Proposition B.1. Decompose the expectation according to whether the log-likelihood ratio exceeds γ:

$$\mathbb{E}_{\mathbb{P}}[X] = \mathbb{E}_{\mathbb{P}}\big[X\mathbf{1}\{L \le \gamma\}\big] + \mathbb{E}_{\mathbb{P}}\big[X\mathbf{1}\{L > \gamma\}\big].$$

For the first term, the change-of-measure identity gives

$$\mathbb{E}_{\mathbb{P}}\big[X\mathbf{1}\{L \le \gamma\}\big] = \mathbb{E}_{\mathbb{Q}}\big[Xe^{L}\mathbf{1}\{L \le \gamma\}\big] \le e^{\gamma}\mathbb{E}_{\mathbb{Q}}\big[X\mathbf{1}\{L \le \gamma\}\big] \le e^{\gamma}\mathbb{E}_{\mathbb{Q}}[X],$$

where the last inequality follows from $X \ge 0$. For the second term, since $X \le b$,

$$\mathbb{E}_{\mathbb{P}}\big[X\mathbf{1}\{L > \gamma\}\big] \le b\mathbb{P}(L > \gamma).$$

Combining the two bounds yields

$$\mathbb{E}_{\mathbb{P}}[X] \le e^{\gamma}\mathbb{E}_{\mathbb{Q}}[X] + b\mathbb{P}(L > \gamma),$$

which finishes the proof. □

Remark B.2. When specializing to $X = \mathbb{1}\{\mathcal{E}\}$, where $\mathcal{E} \in \mathcal{G}$ is any measurable event, Proposition B.1 reduces to $\mathbb{P}(\mathcal{E}) \le e^{\gamma}\mathbb{Q}(\mathcal{E}) + \mathbb{P}[L > \gamma]$. This change-of-measure inequality, in its event-based form, was first introduced for establishing instance-dependent lower bounds (Degenne & Koolen, 2019; Garivier & Kaufmann, 2021), as an modification to the previous KL-contraction-based technique (Garivier & Kaufmann, 2016; Garivier et al., 2019). Recently, Ji et al. (2026) adopted this change-of-measure technique, together with subsequent combinatorial arguments, to prove the worst case lower bound in offline MABs with multiple optimal arms.

Next we investigate the specific form of the likelihood ratio between two bandits μ and λ with Gaussian reward noise.

Proposition B.3 (Log-likelihood ratio for Gaussian bandits)**.** Fix a bandit algorithm and two bandit instance μ and λ with mean vectors $\boldsymbol{\lambda}, \boldsymbol{\mu} \in \mathbb{R}^K$. Define the filtration

$$\begin{aligned} &\mathcal{F}_0 := \{\emptyset, \Omega\}, \qquad \mathcal{F}_t := \sigma(a_1, r_1, \ldots, a_t, r_t), \qquad t \in [T], \\ &\mathcal{G}_{t+1} := \sigma(\mathcal{F}_t, a_{t+1}), \qquad t = 0, \ldots, T-1. \end{aligned}$$

Then we see that, under the instance λ, $r_t \mid \mathcal{G}_t \sim \mathcal{N}(\lambda_{a_t}, 1)$. Let $\mathbb{P}_\lambda$ and $\mathbb{P}_\mu$ denote the trajectory distributions induced by the same algorithm under λ and μ, respectively. For every stopping time $\tau \le T$, define

$$\mathbb{P}_\lambda^\tau := \mathbb{P}_\lambda|_{\mathcal{F}_\tau}, \qquad \mathbb{P}_\mu^\tau := \mathbb{P}_\mu|_{\mathcal{F}_\tau}.$$

Then the log-likelihood ratio between the two stopped trajectory distributions satisfies

$$L_{\lambda,\mu}(\tau) := \log\frac{\mathrm{d}\mathbb{P}_\lambda^\tau}{\mathrm{d}\mathbb{P}_\mu^\tau} = \sum_{t=1}^{\tau}\left[(\lambda_{a_t} - \mu_{a_t})(r_t - \lambda_{a_t}) + \frac{1}{2}(\lambda_{a_t} - \mu_{a_t})^2\right]. \tag{B.5}$$

Proof. For every $t \in [T]$, let $\pi_t(\cdot \mid h_{t-1})$ denote the conditional distribution used by the algorithm to select a_t given the history $h_{t-1} = (a_1, r_1, \ldots, a_{t-1}, r_{t-1})$. Since the same algorithm is used under both instances, the action-selection kernel π_t is identical under $\mathbb{P}_\lambda$ and $\mathbb{P}_\mu$. In particular, it does not depend on the underlying bandit instance.

Let $\phi(x) := (\sqrt{2\pi})^{-1} \exp(-x^2/2)$ denote the standard Gaussian density. Fix any deterministic $n \in [T]$. With respect to the product of the counting measure on actions and the Lebesgue measure on rewards, the density of the trajectory up to time n under λ and μ are

$$p_\lambda(h_n) = \prod_{t=1}^{n} \pi_t(a_t \mid h_{t-1}) \phi(r_t - \lambda_{a_t}),$$

$$p_\mu(h_n) = \prod_{t=1}^{n} \pi_t(a_t \mid h_{t-1}) \phi(r_t - \mu_{a_t}).$$

Therefore, the action-selection probabilities cancel, and we obtain

$$\frac{\mathrm{d}\,(\mathbb{P}_\lambda|_{\mathcal{F}_n})}{\mathrm{d}\,(\mathbb{P}_\mu|_{\mathcal{F}_n})} = \prod_{t=1}^{n} \frac{\phi(r_t - \lambda_{a_t})}{\phi(r_t - \mu_{a_t})}. \tag{B.6}$$

Define the likelihood-ratio process $Z_n = \mathrm{d}\mathbb{P}_\lambda|_{\mathcal{F}_n} / \mathrm{d}\mathbb{P}_\mu|_{\mathcal{F}_n}$ for all $n \in [T]$, and set $Z_0 := 1$. We next show that Z_τ is the Radon–Nikodym derivative between the two trajectory distributions restricted to $\mathcal{F}_\tau$. Indeed, for every event $E \in \mathcal{F}_\tau$, we have

$$\mathbb{E}_\mu\big[Z_\tau \, \mathbb{1}\{E\}\big] = \sum_{n=0}^{T} \mathbb{E}_\mu\big[Z_n \, \mathbb{1}\{E \cap \{\tau = n\}\}\big] = \sum_{n=0}^{T} \mathbb{P}_\lambda\big(E \cap \{\tau = n\}\big) = \mathbb{P}_\lambda(E),$$

where the second equality follows because $E \cap \{\tau = n\} \in \mathcal{F}_n$. Consequently, $Z_\tau = \mathrm{d}\mathbb{P}_\lambda^\tau / \mathrm{d}\mathbb{P}_\mu^\tau$.

It remains to compute the logarithm of Z_τ. For every $t \in [T]$,

$$\log \frac{\phi(r_t - \lambda_{a_t})}{\phi(r_t - \mu_{a_t})} = \frac{1}{2}\big[(r_t - \mu_{a_t})^2 - (r_t - \lambda_{a_t})^2\big] = (\lambda_{a_t} - \mu_{a_t})(r_t - \lambda_{a_t}) + \frac{1}{2}(\lambda_{a_t} - \mu_{a_t})^2.$$

Taking the logarithm of (B.6) and evaluating the resulting likelihood-ratio process at τ therefore gives

$$L_{\lambda,\mu}(\tau) = \log \frac{\mathrm{d}\mathbb{P}_\lambda^\tau}{\mathrm{d}\mathbb{P}_\mu^\tau} = \sum_{t=1}^{\tau} \left[(\lambda_{a_t} - \mu_{a_t})(r_t - \lambda_{a_t}) + \frac{1}{2}(\lambda_{a_t} - \mu_{a_t})^2\right],$$

which proves the proposition. □

Now we are ready to prove Theorem 5.2.

Proof of Theorem 5.2. The hard instance construction is inspired by Exercise 15.2 of Lattimore & Szepesvári (2020). We consider the multi-armed bandit class parameterized by mean vectors:

$$\mu \in \mathcal{V}_A := \left\{ \boldsymbol{\mu} \in \left\{\frac{1}{2}, \frac{1}{2} - \Delta\right\}^K \,\middle|\, \sum_{i=1}^{K} \mathbb{1}(\mu_i = 1/2) = A \right\},$$

where $\Delta > 0$ is a parameter to be determined later. We further consider an auxiliary instance θ such that $\boldsymbol{\theta} = (1/2, \cdots, 1/2)^\top \in \mathbb{R}^K$, and let $\mathcal{V} = \mathcal{V}_A \cup \{\theta\}$. Unless stated otherwise, we assume that noise of rewards is standard Gaussian. For any $\mu \in \mathcal{V}$, the corresponding instance is given by $([K], r_\mu, T)$ where the reward is given by $r_\mu(i) = \mu_i$. For any $\mu \in \mathcal{V}_A$, we use $\mathcal{A}^{\text{opt}}(\mu)$ to denote the set of optimal actions (those with reward $1/2$) and $\mathcal{A}^{\text{sub}}(\mu)$ to denote the set of sub-optimal arms. Our goal is to show that, for any algorithm `Alg`, there exists $\mu \in \mathcal{V}_A$ such that that

$$\mathrm{Regret}_{\mu,\texttt{Alg}}(T) = \widetilde{\Omega}\left(\frac{K - A}{K}\sqrt{T}\right).$$

In the subsequent proof, we will drop the subscript of `Alg` when there is no ambiguity.

Now we consider any $\mu \in \mathcal{V}_A$ and its difference between θ, in terms of the distribution of induced trajectories. Consider the random variable $X_\mu = \sum_{a \in \mathcal{A}^{\text{sub}}(\mu)} N_a(T)$, then we know that

$$\mathrm{Regret}_\mu(T) = \mathbb{E}_\mu[X_\mu]\Delta.$$

On the other side, since $X_\mu \le T$, Proposition B.1 tells us that for any $\gamma > 0$,

$$\mathbb{E}_\mu[X_\mu] \ge e^{-\gamma}\left[\mathbb{E}_\theta[X_\mu] - T\mathbb{P}_\theta\left(\log\frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\mathbb{P}_\mu} > \gamma\right)\right]. \tag{B.7}$$

To lower bound the right hand side, we need to properly choose the threshold γ, which requires the computation of $\log(\mathrm{d}\mathbb{P}_\theta/\mathrm{d}\mathbb{P}_\mu)$. Applying Proposition B.3 with $\tau = T$ gives that

$$\begin{aligned}\log\frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\mathbb{P}_\mu} &= \sum_{t=1}^{T}\left[(\theta_{a_t} - \mu_{a_t})(r_t - \theta_{a_t}) + \frac{1}{2}(\theta_{a_t} - \mu_{a_t})^2\right] \\ &= \underbrace{\frac{1}{2}\sum_{a\in\mathcal{A}^{\mathrm{sub}}(\mu)} N_a(T)(\theta_a - \mu_a)^2}_{I_1} + \underbrace{\sum_{t=1}^{T}\left(r_t - \frac{1}{2}\right)\Delta\,\mathbb{1}\{a_t \in \mathcal{A}^{\mathrm{sub}}(\mu)\}}_{I_2},\end{aligned}$$

Since $\sum_{a\in\mathcal{A}^{\mathrm{sub}}(\mu)} N_a(T) \le T$, we know that $I_1 \le T\Delta^2/2$. For the second term, we consider

$$M_t = \sum_{s=1}^{t}\left(r_s - \frac{1}{2}\right)\Delta\,\mathbb{1}\{\mathcal{E}_s(\mu)\}, \quad V_t = \sum_{s=1}^{t}\Delta^2\,\mathbb{1}\{\mathcal{E}_s(\mu)\}, \text{ where } \mathcal{E}_s(\mu) = \{a_s \in \mathcal{A}^{\mathrm{sub}}(\mu)\}$$

We further define

$$Z_t = \exp\left(\lambda M_t - \frac{\lambda^2 V_t}{2}\right), \quad t = 0, 1, \ldots, T.$$

Then under the probability distribution $\mathbb{P}_\theta$, we see that for any $\lambda \in \mathbb{R}$,

$$\mathbb{E}_\theta[Z_t|\mathcal{F}_{t-1}] = Z_{t-1}\mathbb{E}_\theta\Bigg[\underbrace{\exp\left(\lambda(r_t - 1/2)\Delta\,\mathbb{1}\{\mathcal{E}_t(\mu)\} - \frac{\lambda^2\Delta^2\,\mathbb{1}\{\mathcal{E}_t(\mu)\}}{2}\right)}_{(\star)}\Bigg|\mathcal{F}_{t-1}\Bigg]$$

For $(\star)$, we first take expectation on $\mathcal{G}_t$, which gives that

$$\begin{aligned}\mathbb{E}_\theta[(\star)|\mathcal{G}_t] &= 1 - \mathbb{1}\{\mathcal{E}_t(\mu)\} + \mathbb{1}\{\mathcal{E}_t(\mu)\}\mathbb{E}_\theta\left[\exp\left(\lambda(r_t - 1/2)\Delta - \frac{\lambda^2\Delta^2}{2}\right)\Big|\mathcal{G}_t\right] \\ &\le 1 - \mathbb{1}\{\mathcal{E}_t(\mu)\} + \mathbb{1}\{\mathcal{E}_t(\mu)\} \\ &= 1,\end{aligned}$$

where the inequality holds due to that $r_t - 1/2$ is 1-sub-gaussian conditioned on $\mathcal{G}_t$ on $\mathbb{P}_\theta$. This immediately gives $\mathbb{E}_\theta[(\star)|\mathcal{F}_{t-1}] \le 1$, verifying that Z_t is a super-martingale for any λ. Now invoking Lemma E.1 with $\tau = T$, since $I_2 = M_T$, we see that for any $\beta > 0$,

$$\mathbb{P}_\theta(I_2 \ge \beta) \le \exp\left(-\frac{\beta^2}{2T\Delta^2}\right).$$

Now combining the estimation of I_1 and I_2 and set $\beta = 2\Delta\sqrt{T\log K}$, and $\gamma = \beta + T\Delta^2/2$, we obtain that

$$\mathbb{P}_\theta\left(\log\frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\mathbb{P}_\mu} > \gamma\right) \le \exp\left(-\frac{\beta^2}{2T\Delta^2}\right) = \frac{1}{K^2}.$$

Plugging into (B.7), we see that

$$\begin{aligned}\mathbb{E}_\mu[X_\mu] &\ge \exp\left(-\beta - \frac{T\Delta^2}{2}\right)\left[\mathbb{E}_\theta[X_\mu] - T\mathbb{P}_\theta\left(\log\frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\mathbb{P}_\mu} > \gamma\right)\right] \\ &\ge \exp\left(-\beta - \frac{T\Delta^2}{2}\right)\left[\mathbb{E}_\theta[X_\mu] - \frac{T}{K^2}\right].\end{aligned}$$

Now we sum up over all $\mu \in \mathcal{V}_A$ and obtains that

$$\sum_{\mu\in\mathcal{V}_A}\mathbb{E}_\mu[X_\mu] \ge \exp\left(-\beta - \frac{T\Delta^2}{2}\right)\left(\sum_{\mu\in\mathcal{V}_A}\mathbb{E}_\theta[X_\mu] - \sum_{\mu\in\mathcal{V}_A}\frac{T}{K^2}\right)$$

$$= \exp\bigg(-\beta - \frac{T\Delta^2}{2}\bigg)\bigg(\sum_{\mu\in\mathcal{V}_A}\sum_{a\in\mathcal{A}^{\mathrm{sub}}(\mu)}\mathbb{E}_\theta[N_a(T)] - \binom{K}{K-A}\frac{T}{K^2}\bigg)$$
$$= \exp\bigg(-\beta - \frac{T\Delta^2}{2}\bigg)\bigg(\sum_{a\in\mathcal{A}}\sum_{\mu: a\in\mathcal{A}^{\mathrm{sub}}(\mu)}\mathbb{E}_\theta[N_a(T)] - \binom{K}{K-A}\frac{T}{K^2}\bigg)$$
$$= \exp\bigg(-\beta - \frac{T\Delta^2}{2}\bigg)\bigg(T\binom{K-1}{K-A-1} - \binom{K}{K-A}\frac{T}{K^2}\bigg),$$

where the second equation is given by $|\mathcal{V}_A| = \binom{K}{K-A}$ and the last equation holds due to $\sum_{a\in\mathcal{A}} N_a(T) = T$ and $|\{\mu : a \in \mathcal{A}^{\mathrm{sub}}(\mu)\}| = \binom{K-1}{K-A-1}$. Now take maximum over all $\mu \in \mathcal{V}_A$ and leverage the fact that maximum is larger that average, we see that

$$\sup_{\mu\in\mathcal{V}_A}\mathbb{E}_\mu[X_\mu] \geq T\exp\bigg(-\beta - \frac{T\Delta^2}{2}\bigg)\bigg[\binom{K}{K-A}^{-1}\binom{K-1}{K-A-1} - \frac{1}{K^2}\bigg]$$
$$\geq T\exp\bigg(-\beta - \frac{T\Delta^2}{2}\bigg)\bigg(\frac{K-A}{K} - \frac{1}{K^2}\bigg)$$
$$\geq \frac{(K-A)T}{2K}\exp\bigg(-\beta - \frac{T\Delta^2}{2}\bigg).$$

Recall that $\mathrm{Regret}_\mu(T) = \mathbb{E}_\mu[X_\mu]\Delta$, we see

$$\sup_{\mu\in\mathcal{V}_A}\mathrm{Regret}_\mu(T) \geq \frac{(K-A)T\Delta}{2K}\exp\bigg(-\beta - \frac{T\Delta^2}{2}\bigg).$$

Finally, recall that $\beta = 2\Delta\sqrt{T\log K}$. Now we select $\Delta = \sqrt{(T\log K)^{-1}}$, we see that $\beta = 2$ and $\gamma = 2 + (2\log K)^{-1} \leq 3$. This leads to

$$\sup_{\mu\in\mathcal{V}_A}\mathrm{Regret}_\mu(T) \gtrsim \frac{(K-A)T}{K\sqrt{T\log K}} = \widetilde{\Omega}\bigg(\frac{K-A}{K}\sqrt{T}\bigg),$$

for any algorithm `Alg`, which finishes the proof. □

C Proofs of Theorems in Section 6

C.1 Formal Version and Proof of Lemma 6.4

In this section, we provide the formal version of Lemma 6.4 and the proof of it.

Lemma C.1. For some $1 \leq A < K$, consider an algorithm `Alg` satisfying

$$\sup_{\nu:\,|\mathcal{A}^{\mathrm{opt}}(\nu)|=A}\mathrm{Regret}_{\nu,\mathtt{Alg}}(T) \leq \Gamma\frac{K-A}{K}\sqrt{T}$$

for some $\Gamma \geq 1$. Let θ denote the all-equal instance, then, for every $m \in [T]$ and every $\ell \geq 1$,

$$\max_{\mathcal{S}\subseteq\mathcal{A}:\,|\mathcal{S}|=K-A}\mathbb{P}_\theta\big[N_{\mathcal{S}}(T) \geq m\big] \geq \frac{K-A}{K} - \exp(-\ell) - 8\Gamma\frac{K-A}{K}\sqrt{\frac{m\ell}{T}}$$
$$\max_{\mathcal{S}\subseteq\mathcal{A}:\,|\mathcal{S}|=K-A}\mathbb{P}_\theta\big[N_{\mathcal{S}}(T) \geq m\big] \leq \exp(-\ell) + 8\Gamma\frac{K-A}{K}\sqrt{\frac{T\ell}{m}}.$$

Proof of Lemma C.1. Fix any $m \in [T]$ and $\ell \geq 1$, and for simplicity we write

$$q_{\mathtt{Alg}}(m) := \max_{\mathcal{S}\subseteq\mathcal{A}:\,|\mathcal{S}|=K-A}\mathbb{P}_\theta\big(N_{\mathcal{S}}(T) \geq m\big).$$

Recall that θ is the all-equal instance with mean vector $\boldsymbol{\theta} = \mathbf{1}/2$. For every subset $\mathcal{S} \subseteq \mathcal{A}$ with $|\mathcal{S}| = K - A$, construct an alternative Gaussian bandit instance $\nu_{\mathcal{S}}$ with mean vector $\boldsymbol{\nu}_{\mathcal{S}}$ given by

$$\nu_{\mathcal{S},a} := \begin{cases} 1/2 - \Delta, & a \in \mathcal{S}, \\ 1/2, & a \notin \mathcal{S}, \end{cases} \qquad \Delta := \frac{1}{4\sqrt{m\ell}}.$$

Since $|\mathcal{S}| = K - A$, the instance $\nu_{\mathcal{S}}$ has exactly A optimal arms. Moreover, $\Delta \le 1/4$, so all its reward means belong to $[0, 1]$. We show that, whenever the algorithm pulls only m or fewer than m times on the set $\mathcal{S}$, it cannot reliably distinguish θ and $\nu_{\mathcal{S}}$. To show this, for a fixed $\mathcal{S}$, we define the stopping time $\tau_{\mathcal{S}}$ at which the pull count on $\mathcal{S}$ is exactly m

$$H_{\mathcal{S}} := \{N_{\mathcal{S}}(T) \ge m\}, \quad \tau_{\mathcal{S}} := \inf\{t \in [T] : N_{\mathcal{S}}(t) = m\} \wedge T.$$

Then $\tau_{\mathcal{S}}$ is a bounded $(\mathcal{F}_t)_{t=0}^T$-stopping time, and

$$H_{\mathcal{S}} = \{N_{\mathcal{S}}(\tau_{\mathcal{S}}) = m\} \in \mathcal{F}_{\tau_{\mathcal{S}}}. \tag{C.1}$$

Now we show that, up to the stopping time $\tau_{\mathcal{S}}$, the distribution given by θ and $\nu_{\mathcal{S}}$ is close. We once again characterize this closeness via the log-likelihood ratio. In particular, consider the trajectory distributions restricted to $\mathcal{F}_{\tau_{\mathcal{S}}}$ by $\mathbb{P}_\theta^{\tau_{\mathcal{S}}} := \mathbb{P}_\theta|_{\mathcal{F}_{\tau_{\mathcal{S}}}}$ and $\mathbb{P}_{\nu_{\mathcal{S}}}^{\tau_{\mathcal{S}}} := \mathbb{P}_{\nu_{\mathcal{S}}}|_{\mathcal{F}_{\tau_{\mathcal{S}}}}$. Applying Proposition B.3 gives that

$$\begin{aligned}
\log \frac{\mathrm{d}\mathbb{P}_\theta^{\tau_{\mathcal{S}}}}{\mathrm{d}\mathbb{P}_{\nu_{\mathcal{S}}}^{\tau_{\mathcal{S}}}} &= \sum_{t=1}^{\tau_{\mathcal{S}}} \left[(\theta_{a_t} - \nu_{\mathcal{S},a_t})(r_t - \theta_{a_t}) + \frac{1}{2}(\theta_{a_t} - \nu_{\mathcal{S},a_t})^2 \right] \\
&= \Delta \sum_{t=1}^{\tau_{\mathcal{S}}} \mathbb{1}\{a_t \in \mathcal{S}\}(r_t - 1/2) + \frac{\Delta^2}{2} N_{\mathcal{S}}(\tau_{\mathcal{S}}).
\end{aligned}$$

Now we denote $M_t = \Delta \sum_{s=1}^t \mathbb{1}\{a_s \in \mathcal{S}\}(r_s - 1/2)$ and $V_t = \Delta^2 N_{\mathcal{S}}(t)$, then under the probability distribution $\mathbb{P}_\theta$, we see that for any $\lambda \in \mathbb{R}$,

$$\begin{aligned}
&\mathbb{E}_\theta\left[\exp\left(\lambda M_t - \frac{\lambda^2 V_t}{2} \right) \middle| \mathcal{F}_{t-1} \right] \\
&\quad = \exp\left(\lambda M_{t-1} - \frac{\lambda^2 V_{t-1}}{2} \right) \mathbb{E}_\theta\Bigg[\underbrace{\exp\left(\lambda(r_t - 1/2)\Delta\, \mathbb{1}\{a_t \in \mathcal{S}\} - \frac{\lambda^2 \Delta^2\, \mathbb{1}\{a_t \in \mathcal{S}\}}{2} \right)}_{(\star)} \Bigg| \mathcal{F}_{t-1} \Bigg]
\end{aligned}$$

For $(\star)$, we first take expectation on $\mathcal{G}_t$, which gives that

$$\begin{aligned}
\mathbb{E}_\theta[(\star)|\mathcal{G}_t] &= 1 - \mathbb{1}\{a_t \in \mathcal{S}\} + \mathbb{1}\{a_t \in \mathcal{S}\} \mathbb{E}_\theta\left[\exp\left(\lambda(r_t - 1/2)\Delta - \frac{\lambda^2 \Delta^2}{2} \right) \middle| \mathcal{G}_t \right] \\
&\le 1 - \mathbb{1}\{a_t \in \mathcal{S}\} + \mathbb{1}\{a_t \in \mathcal{S}\} \\
&= 1,
\end{aligned}$$

where the inequality holds due to that $r_t - 1/2$ is 1-sub-gaussian conditioned on $\mathcal{G}_t$ on $\mathbb{P}_\theta$. This immediately gives $\mathbb{E}_\theta[(\star)|\mathcal{F}_{t-1}] \le 1$, verifying that $\exp(\lambda M_t - \lambda^2 V_t/2)$ is a super-martingale for any λ. Now invoking Lemma E.1 with $x = \Delta\sqrt{2m\ell}$, since $V_\tau \le m\Delta^2$, we have

$$\mathbb{P}_\theta\big(M_\tau \ge \Delta\sqrt{2m\ell}\big) \le \exp\left(-\frac{2m\Delta^2\ell}{2m\Delta^2} \right) = \exp(-\ell).$$

Since $\Delta^2 N_{\mathcal{S}}(\tau_{\mathcal{S}})/2 \le m\Delta^2/2$, this immediately leads to

$$\mathbb{P}\left(\log \frac{\mathrm{d}\mathbb{P}_\theta^{\tau_{\mathcal{S}}}}{\mathrm{d}\mathbb{P}_{\nu_{\mathcal{S}}}^{\tau_{\mathcal{S}}}} \ge \Delta\sqrt{2m\ell} + m\Delta^2/2 \right) \le \exp(-\ell).$$

Recall that $\Delta = (4\sqrt{m\ell})^{-1}$, this means that $\Delta\sqrt{2m\ell} + m\Delta^2/2 = 1/2\sqrt{2} + 1/32\ell < \log 2$ since $\ell \ge 1$

$$\mathbb{P}\left(\log \frac{\mathrm{d}\mathbb{P}_\theta^{\tau_{\mathcal{S}}}}{\mathrm{d}\mathbb{P}_{\nu_{\mathcal{S}}}^{\tau_{\mathcal{S}}}} > \log 2 \right) \le \exp(-\ell). \tag{C.2}$$

We now prove the upper bound on $q_{\texttt{Alg}}(m)$. Fix any $\mathcal{S} \subseteq \mathcal{A}$ with $|\mathcal{S}| = K - A$. Since $H_{\mathcal{S}} \in \mathcal{F}_{\tau_{\mathcal{S}}}$, we may apply Proposition B.1 on the measurable space $(\Omega, \mathcal{F}_{\tau_{\mathcal{S}}})$ with $\mathbb{P} = \mathbb{P}_\theta^{\tau_{\mathcal{S}}}$, $\mathbb{Q} = \mathbb{P}_{\nu_{\mathcal{S}}}^{\tau_{\mathcal{S}}}$, $X = \mathbb{1}\{H_{\mathcal{S}}\}$, $b = 1$ and $\gamma = \log 2$. Using (C.2), we obtain

$$\mathbb{P}_{\nu_{\mathcal{S}}}(H_{\mathcal{S}}) \ge \frac{1}{2}\big[\mathbb{P}_\theta(H_{\mathcal{S}}) - \mathbb{P}_\theta(L_{\mathcal{S}} > \log 2)\big] \ge \frac{1}{2}\big[\mathbb{P}_\theta(H_{\mathcal{S}}) - \exp(-\ell)\big]. \tag{C.3}$$

Every arm in $\mathcal{S}$ is suboptimal under $\nu_{\mathcal{S}}$ with gap Δ. Hence, the assumed regret guarantee gives

$$\Gamma\frac{K-A}{K}\sqrt{T} \geq \mathrm{Regret}_{\nu_{\mathcal{S}},\mathtt{Alg}}(T) = \Delta\mathbb{E}_{\nu_{\mathcal{S}}}[N_{\mathcal{S}}(T)] \geq \Delta m\mathbb{P}_{\nu_{\mathcal{S}}}(H_{\mathcal{S}}) \geq \frac{\Delta m}{2}\big[\mathbb{P}_\theta(H_{\mathcal{S}}) - \exp(-\ell)\big].$$

Rearranging and substituting $\Delta = (4\sqrt{m\ell})^{-1}$ yields

$$\mathbb{P}_\theta\big(N_{\mathcal{S}}(T) \geq m\big) \leq \exp(-\ell) + \frac{2\Gamma(K-A)\sqrt{T}}{K\Delta m} = \exp(-\ell) + 8\Gamma\frac{K-A}{K}\sqrt{\frac{T\ell}{m}}.$$

Since this holds for every $|\mathcal{S}| = K - A$, taking maximum over all $|\mathcal{S}| = K - A$ gives the desired upper bound

$$q_{\mathtt{Alg}}(m) \leq \exp(-\ell) + 8\Gamma\frac{K-A}{K}\sqrt{\frac{T\ell}{m}}. \tag{C.4}$$

We next prove the lower bound on $q_{\mathtt{Alg}}(m)$. For every $\mathcal{S} \subseteq \mathcal{A}$ with $|\mathcal{S}| = K - A$, define $Y_{\mathcal{S}} = N_{\mathcal{S}}(T)\,\mathbb{1}\{N_{\mathcal{S}}(T) < m\} = N_{\mathcal{S}}(T)\,\mathbb{1}\{H_{\mathcal{S}}^{\complement}\}$. On $H_{\mathcal{S}}^{\complement}$, the stopping time satisfies $\tau_{\mathcal{S}} = T$, whereas on $H_{\mathcal{S}}$ both sides below vanish. Therefore, $Y_{\mathcal{S}} = N_{\mathcal{S}}(\tau_{\mathcal{S}})\,\mathbb{1}\{H_{\mathcal{S}}^{\complement}\}$. In particular, $Y_{\mathcal{S}}$ is $\mathcal{F}_{\tau_{\mathcal{S}}}$-measurable and takes values in $[0, T]$. Once again, applying Proposition B.1 once more on $(\Omega, \mathcal{F}_{\tau_{\mathcal{S}}})$, now with $\mathbb{P} = \mathbb{P}_\theta^{\tau_{\mathcal{S}}}$, $\mathbb{Q} = \mathbb{P}_{\nu_{\mathcal{S}}}^{\tau_{\mathcal{S}}}$, $X = Y_{\mathcal{S}}$, $b = T$ and $\gamma = \log 2$, and using (C.2), gives

$$\mathbb{E}_{\nu_{\mathcal{S}}}[Y_{\mathcal{S}}] \geq \frac{1}{2}\big[\mathbb{E}_\theta[Y_{\mathcal{S}}] - T\mathbb{P}_\theta(L_{\mathcal{S}} > \log 2)\big] \geq \frac{1}{2}\big[\mathbb{E}_\theta[Y_{\mathcal{S}}] - T\exp(-\ell)\big].$$

Since $Y_{\mathcal{S}} \leq N_{\mathcal{S}}(T)$, we have

$$\mathbb{E}_{\nu_{\mathcal{S}}}[N_{\mathcal{S}}(T)] \geq \mathbb{E}_{\nu_{\mathcal{S}}}[Y_{\mathcal{S}}] \geq \frac{1}{2}\big[\mathbb{E}_\theta\big[N_{\mathcal{S}}(T)\,\mathbb{1}\{H_{\mathcal{S}}^{\complement}\}\big] - T\exp(-\ell)\big].$$

Moreover, as $\mathbb{P}_\theta(H_{\mathcal{S}}) \leq q_{\mathtt{Alg}}(m)$, it follows that

$$\begin{aligned}
\mathbb{E}_{\nu_{\mathcal{S}}}[N_{\mathcal{S}}(T)] &\geq \frac{1}{2}\big[\mathbb{E}_\theta\big[N_{\mathcal{S}}(T)\,\mathbb{1}\{H_{\mathcal{S}}^{\complement}\}\big] - T\exp(-\ell)\big] \\
&\geq \frac{1}{2}\big[\mathbb{E}_\theta[N_{\mathcal{S}}(T)] - T\mathbb{P}_\theta(H_{\mathcal{S}}) - T\exp(-\ell)\big] \\
&\geq \frac{1}{2}\mathbb{E}_\theta[N_{\mathcal{S}}(T)] - \frac{T}{2}q_{\mathtt{Alg}}(m) - \frac{T}{2}\exp(-\ell).
\end{aligned}$$

Thus,

$$\mathbb{E}_{\nu_{\mathcal{S}}}[N_{\mathcal{S}}(T)] \geq \frac{1}{2}\mathbb{E}_\theta[N_{\mathcal{S}}(T)] - \frac{T}{2}q_{\mathtt{Alg}}(m) - \frac{T}{2}\exp(-\ell). \tag{C.5}$$

We now average (C.5) over all subsets $\mathcal{S} \subseteq \mathcal{A}$ with $|\mathcal{S}| = K - A$. Since every arm belongs to $\binom{K-1}{K-A-1}$ such subsets,

$$\frac{1}{\binom{K}{K-A}}\sum_{\mathcal{S}\subseteq\mathcal{A}:\,|\mathcal{S}|=K-A}\mathbb{E}_\theta[N_{\mathcal{S}}(T)] = \frac{\binom{K-1}{K-A-1}}{\binom{K}{K-A}}\sum_{a\in\mathcal{A}}\mathbb{E}_\theta[N_a(T)] = \frac{K-A}{K}T.$$

Therefore,

$$\frac{1}{\binom{K}{K-A}}\sum_{\mathcal{S}\subseteq\mathcal{A}:\,|\mathcal{S}|=K-A}\mathbb{E}_{\nu_{\mathcal{S}}}[N_{\mathcal{S}}(T)] \geq \frac{T}{2}\left[\frac{K-A}{K} - q_{\mathtt{Alg}}(m) - \exp(-\ell)\right]. \tag{C.6}$$

Since Every $\nu_{\mathcal{S}}$ has exactly A optimal arms, we have $\mathrm{Regret}_{\nu_{\mathcal{S}},\mathtt{Alg}}(T) = \Delta\mathbb{E}_{\nu_{\mathcal{S}}}[N_{\mathcal{S}}(T)]$. Together with (C.6), this implies that

$$\Gamma\frac{K-A}{K}\sqrt{T} \geq \frac{\Delta}{\binom{K}{K-A}}\sum_{\mathcal{S}\subseteq\mathcal{A}:\,|\mathcal{S}|=K-A}\mathbb{E}_{\nu_{\mathcal{S}}}[N_{\mathcal{S}}(T)] \geq \frac{\Delta T}{2}\left[\frac{K-A}{K} - q_{\mathtt{Alg}}(m) - \exp(-\ell)\right].$$

Rearranging and substituting $\Delta = (4\sqrt{m\ell})^{-1}$ gives

$$q_{\mathtt{Alg}}(m) \geq \frac{K-A}{K} - \exp(-\ell) - \frac{2\Gamma(K-A)\sqrt{T}}{K\Delta T} \tag{C.7}$$

$$= \frac{K-A}{K} - \exp(-\ell) - 8\Gamma\frac{K-A}{K}\sqrt{\frac{m\ell}{T}}. \tag{C.8}$$

Combining (C.4) and (C.8) proves

$$\frac{K-A}{K} - \exp(-\ell) - 8\Gamma\frac{K-A}{K}\sqrt{\frac{m\ell}{T}} \le q_{\texttt{Alg}}(m) \le \exp(-\ell) + 8\Gamma\frac{K-A}{K}\sqrt{\frac{T\ell}{m}},$$

as desired. □

C.2 Tightness of Lemma C.1

In this section, we show that both inequalities in Lemma C.1 cannot be improved in general, for all $m = \widetilde{O}(T)$ and $A \ge K/2$ and ℓ chosen to be the same order as the remaining factors in the bound. For simplicity, given any algorithm $\texttt{Alg}$, we write

$$q_{\texttt{Alg}}(m) := \max_{\mathcal{S}\subseteq\mathcal{A}:\,|\mathcal{S}|=K-A} \mathbb{P}_\theta\big(N_{\mathcal{S}}(T) \ge m\big).$$

Then a properly chosen ℓ, together with $m = \widetilde{O}(T)$ and $A \ge K/2$ reduces the lower envelop in Lemma C.1 to

$$q_{\texttt{Alg}}(m) = \widetilde{\Omega}\bigg(\frac{K-A}{K}\bigg),$$

and the upper envelop to

$$q_{\texttt{Alg}}(m) = \widetilde{O}\bigg(\frac{K-A}{K}\sqrt{\frac{T}{m}}\bigg).$$

We will show that there exist two algorithms that attain the reduced lower and upper envelop respectively.

Tightness of the Lower Envelope. We show that the lower bound is attained by the near-minimax optimal algorithm SMOSS. Recall that SMOSS first samples a subset $\mathcal{B} \subseteq \mathcal{A}$ of L arms uniformly without replacement and then runs a bandit algorithm only on the arms in $\mathcal{B}$ for all T rounds, where $L = \widetilde{\Theta}(1)$. We show that, under the all-equal instance θ, for every $m \le T/L$, $q_{\text{SMOSS}}(m) \le L(K-A)/K$.

Fix any $\mathcal{S} \subseteq \mathcal{A}$ with $|\mathcal{S}| = K - A$. Since SMOSS only pulls arms in the randomly sampled subset $\mathcal{B}$, the event $N_{\mathcal{S}}(T) \ge m$ for any $m \ge 1$ implies that $\mathcal{B} \cap \mathcal{S} \ne \emptyset$. Therefore,

$$\mathbb{P}_{\boldsymbol{\theta}}\big[N_{\mathcal{S}}(T) \ge m\big] \le \mathbb{P}\big[\mathcal{U} \cap \mathcal{S} \ne \emptyset\big] \le \sum_{a\in\mathcal{S}} \mathbb{P}[a \in \mathcal{U}] = \frac{L(K-A)}{K},$$

which gives that

$$q_{\text{SMOSS}}(m) \le \frac{L(K-A)}{K}. \tag{C.9}$$

Tightness of the upper envelope. For every prescribed scale $m \le T/2$, we construct a minimax-optimal algorithm satisfying

$$q_{\texttt{Alg}}(m) = \widetilde{\Omega}\bigg(\min\bigg\{1, \frac{K-A}{K}\sqrt{\frac{T}{m}}\bigg\}\bigg),$$

which matches the upper envelope in Lemma C.1. The constructed algorithm is described in Algorithm 2, which is built on the algorithm SMOSS.

We first show that Algorithm 2 is minimax-optimal. On the event $Z = 1$, we fix any fixed subset $|\mathcal{U}| = 2(K-A)$. Since the original instance contains only $K-A$ suboptimal arms in total, the set $\mathcal{U}$ of $2(K-A)$ arms contains at least $K-A$ optimal arms. Therefore, when SMOSS is run on $\mathcal{U}$ with parameters $K' = 2(K-A)$ and $A' = K-A$, the same subsampling analysis as in Section 4 gives

$$\sup_{\nu:\,|\mathcal{A}^{\text{opt}}(\nu)|=A} \text{Regret}^{\mathcal{U}}_{\nu,\text{SMOSS}}(2m) = \widetilde{O}(\sqrt{m}), \tag{C.10}$$

Algorithm 2 Tail-Saturating Algorithm for Bandit with Multiple Optimal Arms

Require: Action set $\mathcal{A}$, number of total rounds T, number of optimal arms A, and threshold $m \le T/2$.

1: Fix any subset $\mathcal{U} \subseteq \mathcal{A}$ with $|\mathcal{U}| = 2(K-A)$ together with a partition $\mathcal{U} = \mathcal{S}_0 \cup \mathcal{S}_1$, where $|\mathcal{S}_0| = |\mathcal{S}_1| = K - A$.
2: Set

$$p_m := \min\left\{1,\ \frac{K-A}{K}\sqrt{\frac{T}{m}}\right\},$$

and sample $Z \sim \mathrm{Bernoulli}(p_m)$.
3: **if** $Z = 0$ **then**
4: Run SMOSS on $\mathcal{A}$ for T rounds.
5: **else**
6: Sample a uniformly random permutation σ of the arms in $\mathcal{U}$.
7: Run SMOSS on $\sigma(\mathcal{U})$ for $2m$ rounds, using $K' = 2(K-A)$ and $A' = K - A$ as its input parameters, and map the selected actions back to $\mathcal{U}$ through σ^{-1}.
8: Restart SMOSS on the full action set $\mathcal{A}$ for the remaining $T - 2m$ rounds.
9: **end if**

where $\mathrm{Regret}^{\mathcal{U}}_{\nu,\mathrm{SMOSS}}(2m)$ denotes the regret accumulated during the exploration phase on $\mathcal{U}$. The remainder of Algorithm 2 simply runs SMOSS on the original problem, and hence contributes at most $\widetilde{O}((K-A)/K\sqrt{T})$ regret. Consequently,

$$\sup_{\nu:\,|\mathcal{A}^{\mathrm{opt}}(\nu)|=A} \mathrm{Regret}_{\nu,\mathtt{Alg}}(T) \le \widetilde{O}\left(\frac{K-A}{K}\sqrt{T}\right) + p_m\widetilde{O}(\sqrt{m}) = \widetilde{O}\left(\frac{K-A}{K}\sqrt{T}\right),$$

where the last equality follows from

$$p_m\sqrt{m} = \min\left\{\sqrt{m},\ \frac{K-A}{K}\sqrt{T}\right\} \le \frac{K-A}{K}\sqrt{T}.$$

Thus Algorithm 2 remains minimax-optimal up to logarithmic factors.

It remains to characterize its pull-count tail under the all-equal instance θ. Conditional on $Z = 1$, consider only the first $2m$ rounds. Let $Y := N_{\mathcal{S}_0}(2m)$. Since every arm has the same reward distribution under θ, and we apply a uniformly random permutation to $\mathcal{U}$ before running the exploration algorithm, the resulting trajectory distribution is invariant under permutations of the arms in $\mathcal{U}$. In particular, swapping $\mathcal{S}_0$ and $\mathcal{S}_1$ leaves the distribution unchanged. Moreover, every action during the first $2m$ rounds belongs to $\mathcal{U} = \mathcal{S}_0 \cup \mathcal{S}_1$. These leads to

$$Y = 2m - N_{\mathcal{S}_1}(2m) \overset{d}{=} 2m - Y,$$

which immediately gives

$$\mathbb{P}_\theta[Y \ge m \mid Z = 1] = \frac{1 + \mathbb{P}_\theta[Y = m \mid Z = 1]}{2} \ge \frac{1}{2}.$$

Since the number of pulls of $\mathcal{S}_0$ can only increase after the exploration phase, we obtain that

$$\begin{aligned} q_{\mathtt{Alg}}(m) &\ge \mathbb{P}_\theta\big[N_{\mathcal{S}_0}(T) \ge m\big] \\ &\ge \mathbb{P}[Z=1]\mathbb{P}_\theta\big[N_{\mathcal{S}_0}(2m) \ge m \,\big|\, Z = 1\big] \\ &\ge \frac{1}{2}\min\left\{1,\ \frac{K-A}{K}\sqrt{\frac{T}{m}}\right\}. \end{aligned}$$

Together with (C.9), it is indicated that both bounds in Lemma C.1 cannot be improved up to logarithmic factors.

C.3 Exploration profile of Near Minimax Optimal Algorithms

In this section, we provide the formal statement and proof of the exploration lower bound for algorithms that are near-minimax optimal on $\Xi_{K,B}$, referenced in Section 6.2. The argument follows the same strategy as Lemma C.1, but is applied in the opposite direction: instead of showing that an algorithm near-optimal on $\Xi_{K,A}$ rarely concentrates too many pulls on a fixed small set, we show that an algorithm near-optimal on $\Xi_{K,B}$ cannot avoid pulling a fixed set of size B sufficiently often.

Lemma C.2. Let $1 \le B < K$, and suppose that $\texttt{Alg}$ is minimax-optimal on $\Xi_{K,B}$ up to a factor $\Gamma \ge 1$, namely

$$\sup_{\nu:\,|\mathcal{A}^{\mathrm{opt}}(\nu)|=B} \mathrm{Regret}_{\nu,\texttt{Alg}}(T) \le \Gamma \frac{K-B}{\sqrt{KB}}\sqrt{T}.$$

Let θ denote the all-equal instance. Then, for every $\mathcal{U} \subseteq \mathcal{A}$ with $|\mathcal{U}| = B$, every $\ell \ge 1$, and every integer m with $1 \le m \le T/2$,

$$\mathbb{P}_\theta\big[N_{\mathcal{U}}(T) < m\big] \le \exp(-\ell) + 16\Gamma\sqrt{\frac{K\ell m}{BT}}. \tag{C.11}$$

In particular, for any $\delta = \Theta(1)$, taking $\ell = \log(2/\delta)$, sufficient large T and

$$m_B := \left\lfloor \frac{\delta^2 BT}{2^{10}\Gamma^2 K\ell} \right\rfloor = \widetilde{\Omega}\Big(\frac{TB}{K}\Big),$$

we obtain $\mathbb{P}_\theta\big[N_{\mathcal{U}}(T) \ge m_B\big] \ge 1-\delta$.

Proof of Lemma C.2. Fix $\mathcal{U} \subseteq \mathcal{A}$ with $|\mathcal{U}| = B$, $\ell \ge 1$, and an integer $m \le T/2$, and write $\Delta := (4\sqrt{m\ell})^{-1} \le 1/4$. Consider the instance μ obtained from the all-equal instance θ by increasing the mean rewards on $\mathcal{U}$ by Δ:

$$\mu_a := \begin{cases} 1/2+\Delta, & a \in \mathcal{U}, \\ 1/2, & a \in \mathcal{U}^{\complement}. \end{cases}$$

Since $|\mathcal{U}| = B$, every arm in $\mathcal{U}$ is strictly better than every arm outside $\mathcal{U}$, so μ has exactly B optimal arms, namely $\mathcal{A}^{\mathrm{opt}}(\mu) = \mathcal{U}$, and hence $\mu \in \Xi_{K,B}$. As in the proof of Lemma C.1, define

$$H := \{N_{\mathcal{U}}(T) \ge m\}, \qquad \tau := \inf\{t \in [T] : N_{\mathcal{U}}(t) = m\} \wedge T,$$

so that τ is a bounded $(\mathcal{F}_t)_{t=0}^T$-stopping time and, since the pull count $N_{\mathcal{U}}(t)$ increases by at most one per round, $H = \{N_{\mathcal{U}}(\tau) = m\} \in \mathcal{F}_\tau$. Consequently $H^{\complement} = \{N_{\mathcal{U}}(T) < m\}$ also belongs to $\mathcal{F}_\tau$, and on $H^{\complement}$ we have $\tau = T$.

We first show that, up to τ, the trajectory distributions induced by θ and μ remain close. Writing $\mathbb{P}_\theta^\tau := \mathbb{P}_\theta|_{\mathcal{F}_\tau}$ and $\mathbb{P}_\mu^\tau := \mathbb{P}_\mu|_{\mathcal{F}_\tau}$, and using $\theta_a - \mu_a = -\Delta\,\mathbb{1}\{a \in \mathcal{U}\}$, Proposition B.3 gives

$$\begin{aligned}\log\frac{\mathrm{d}\mathbb{P}_\theta^\tau}{\mathrm{d}\mathbb{P}_\mu^\tau} &= \sum_{t=1}^{\tau}\left[(\theta_{a_t}-\mu_{a_t})(r_t-\theta_{a_t}) + \frac{1}{2}(\theta_{a_t}-\mu_{a_t})^2\right] \\ &= -\Delta\sum_{t=1}^{\tau}\mathbb{1}\{a_t \in \mathcal{U}\}(r_t - 1/2) + \frac{\Delta^2}{2}N_{\mathcal{U}}(\tau).\end{aligned}$$

Denote $M_t := -\Delta\sum_{s=1}^t \mathbb{1}\{a_s \in \mathcal{U}\}(r_s - 1/2)$ and $V_t := \Delta^2 N_{\mathcal{U}}(t)$. Exactly as in the proof of Lemma C.1, since $r_t - 1/2$ is 1-sub-Gaussian conditioned on $\mathcal{G}_t$ under $\mathbb{P}_\theta$ (regardless of the sign of Δ), $\exp(\lambda M_t - \lambda^2 V_t/2)$ is a super-martingale under $\mathbb{P}_\theta$ for every $\lambda \in \mathbb{R}$. Invoking Lemma E.1 with $x = \Delta\sqrt{2m\ell}$, and using $V_\tau \le m\Delta^2$, we obtain

$$\mathbb{P}_\theta\big(M_\tau \ge \Delta\sqrt{2m\ell}\big) \le \exp\left(-\frac{2m\Delta^2\ell}{2m\Delta^2}\right) = \exp(-\ell).$$

Since $\Delta^2 N_{\mathcal{U}}(\tau)/2 \le m\Delta^2/2$, this gives

$$\mathbb{P}_\theta\left(\log\frac{\mathrm{d}\mathbb{P}_\theta^\tau}{\mathrm{d}\mathbb{P}_\mu^\tau} \ge \Delta\sqrt{2m\ell} + m\Delta^2/2\right) \le \exp(-\ell).$$

Recalling $\Delta = (4\sqrt{m\ell})^{-1}$, we have $\Delta\sqrt{2m\ell} + m\Delta^2/2 = 1/(2\sqrt{2}) + 1/(32\ell) < \log 2$ since $\ell \ge 1$, so

$$\mathbb{P}_\theta\left(\log\frac{\mathrm{d}\mathbb{P}_\theta^\tau}{\mathrm{d}\mathbb{P}_\mu^\tau} > \log 2\right) \le \exp(-\ell). \tag{C.12}$$

We now transfer the bound on $\mathbb{P}_\theta(H^{\mathtt{C}})$ to μ. Since $H^{\mathtt{C}} \in \mathcal{F}_\tau$, we apply Proposition B.1 on $(\Omega, \mathcal{F}_\tau)$ with $\mathbb{P} = \mathbb{P}_\theta^\tau$, $\mathbb{Q} = \mathbb{P}_\mu^\tau$, $X = \mathbb{1}\{H^{\mathtt{C}}\}$, $b = 1$ and $\gamma = \log 2$; using (C.12), this gives

$$\mathbb{P}_\mu\big(H^{\mathtt{C}}\big) \geq \frac{1}{2}\Big[\mathbb{P}_\theta\big(H^{\mathtt{C}}\big) - \exp(-\ell)\Big]. \tag{C.13}$$

It remains to bound the left-hand side of (C.13) using the regret guarantee of $\mathtt{Alg}$ on μ. Every arm in $\mathcal{U}^{\mathtt{C}}$ is suboptimal under μ with gap Δ, so

$$\mathrm{Regret}_{\mu,\mathtt{Alg}}(T) = \Delta\,\mathbb{E}_\mu\big[N_{\mathcal{U}^{\mathtt{C}}}(T)\big] = \Delta\big(T - \mathbb{E}_\mu[N_{\mathcal{U}}(T)]\big).$$

On the event $H^{\mathtt{C}} = \{N_{\mathcal{U}}(T) < m\}$ we have $N_{\mathcal{U}^{\mathtt{C}}}(T) > T - m$, so, using (C.13) and $m \leq T/2$,

$$\begin{aligned}
\mathrm{Regret}_{\mu,\mathtt{Alg}}(T) &\geq \Delta(T-m)\mathbb{P}_\mu\big(H^{\mathtt{C}}\big) \\
&\geq \frac{\Delta(T-m)}{2}\Big[\mathbb{P}_\theta\big(H^{\mathtt{C}}\big) - \exp(-\ell)\Big] \\
&\geq \frac{\Delta T}{4}\Big[\mathbb{P}_\theta\big(H^{\mathtt{C}}\big) - \exp(-\ell)\Big].
\end{aligned}$$

On the other hand, since $\mu \in \Xi_{K,B}$, the assumed minimax-optimality of $\mathtt{Alg}$ on $\Xi_{K,B}$ gives

$$\mathrm{Regret}_{\mu,\mathtt{Alg}}(T) \leq \Gamma\frac{K-B}{\sqrt{KB}}\sqrt{T} \leq \Gamma\sqrt{\frac{K}{B}}\sqrt{T},$$

where the second inequality uses $K - B \leq K$. Combining the last two displays and rearranging leads to

$$\mathbb{P}_\theta\big(H^{\mathtt{C}}\big) \leq \exp(-\ell) + \frac{4\Gamma}{\Delta\sqrt{T}}\sqrt{\frac{K}{B}} = \exp(-\ell) + 16\Gamma\sqrt{\frac{K\ell m}{BT}},$$

where the equality substitutes $\Delta^{-1} = 4\sqrt{m\ell}$. This proves the claimed result in (C.11). □

C.4 Proof of Theorem 6.1

Proof of Theorem 6.1. The proof follows a Le Cam two-point argument similar to the proof of Theorem 5.1. For simplicity, for any subset $\mathcal{S} \subseteq \mathcal{A}$, we define $N_{\mathcal{S}}(T) := \sum_{a\in\mathcal{S}} N_a(T)$. Fix an arbitrary subset $\mathcal{S}_0 \subseteq \mathcal{A}$ with $|\mathcal{S}_0| = A$. For a parameter $\Delta \in (0, 1/4]$ to be specified later, consider the reference instance ν_0 with reward function

$$r_0(a) := \begin{cases} 1/2, & a \in \mathcal{S}_0, \\ 1/2 - \Delta, & a \in \mathcal{S}_0^{\mathtt{C}}. \end{cases}$$

Thus, ν_0 has exactly A optimal arms, namely the arms in $\mathcal{S}_0$. Let $\mathbb{P}_0$ and $\mathbb{E}_0$ denote probability and expectation under ν_0 and Alg. By the assumed regret guarantee of Alg on every instance with A optimal arms,

$$\Delta\mathbb{E}_0[N_{\mathcal{S}_0^{\mathtt{C}}}(T)] = \mathrm{Regret}_{\nu_0,\mathsf{Alg}}(T) \leq R_A := \Gamma_{K,T}\frac{K-A}{\sqrt{KA}}\sqrt{T},$$

which gives that $\mathbb{E}_0[N_{\mathcal{S}_0^{\mathtt{C}}}(T)] \leq R_A/\Delta$.

Since $B \leq K - A$, we can choose a subset $\mathcal{S}_1 \subseteq \mathcal{S}_0^{\mathtt{C}}$ with $|\mathcal{S}_1| = B$ consisting of the B least-sampled arms in $\mathcal{S}_0^{\mathtt{C}}$. By averaging,

$$\mathbb{E}_0[N_{\mathcal{S}_1}(T)] \leq \frac{B}{K-A}\mathbb{E}_0[N_{\mathcal{S}_0^{\mathtt{C}}}(T)] \leq \frac{BR_A}{(K-A)\Delta}, \tag{C.14}$$

where the first inequality holds due to that $\mathcal{S}_1$ is the least sampled set under ν_0. We now construct the alternative instance ν_1 by promoting the arms in $\mathcal{S}_1$:

$$r_1(a) := \begin{cases} 1/2, & a \in \mathcal{S}_0, \\ 1/2 + \Delta, & a \in \mathcal{S}_1, \\ 1/2 - \Delta, & a \in (\mathcal{S}_0 \cup \mathcal{S}_1)^{\mathtt{C}}. \end{cases}$$

The instance ν_1 has exactly B optimal arms, namely the arms in $\mathcal{S}_1$. Let $\mathbb{P}_1$ and $\mathbb{E}_1$ denote probability and expectation under ν_1 and Alg.

We next apply a change-of-measure argument. Since $N_{\mathcal{S}_1}(T) \in [0, T]$, Lemma E.2 gives

$$\mathbb{E}_1[N_{\mathcal{S}_1}(T)] \le \mathbb{E}_0[N_{\mathcal{S}_1}(T)] + T\,\mathrm{TV}(\mathbb{P}_0, \mathbb{P}_1) \le \mathbb{E}_0[N_{\mathcal{S}_1}(T)] + T\sqrt{\frac{1}{2}\,\mathrm{KL}(\mathbb{P}_0\|\mathbb{P}_1)}, \tag{C.15}$$

where the last inequality holds due to Pinsker's inequality (Lemma E.3). To control the KL-divergence, we see that the instances ν_0 and ν_1 differ only on the arms in $\mathcal{S}_1$. For every $a \in \mathcal{S}_1$, the reward distribution is $\mathcal{N}(1/2-\Delta, 1)$ under ν_0 and $\mathcal{N}(1/2+\Delta, 1)$ under ν_1. Hence, by the KL decomposition for adaptive bandit experiments,

$$\begin{aligned}\mathrm{KL}(\mathbb{P}_0\|\mathbb{P}_1) &= \sum_{a\in\mathcal{S}_1} \mathbb{E}_0[N_a(T)]\,\mathrm{KL}\left(\mathcal{N}(1/2-\Delta,1)\|\mathcal{N}(1/2+\Delta,1)\right)\\ &= 2\Delta^2\mathbb{E}_0[N_{\mathcal{S}_1}(T)]\\ &\le \frac{2\Delta B R_A}{K-A},\end{aligned}$$

where the last inequality follows from (C.14).

Now choose $\Delta := \min\{1/4,\ (K-A)/(16BR_A)\}$. Then $\mathrm{KL}(\mathbb{P}_0\|\mathbb{P}_1) \le 1/8$, and consequently we know that $\mathrm{TV}(\mathbb{P}_0, \mathbb{P}_1) \le 1/4$. Moreover, by (C.14),

$$\mathbb{E}_0[N_{\mathcal{S}_1}(T)] \le \frac{BR_A}{K-A}\max\left\{4, \frac{16BR_A}{K-A}\right\} = \max\left\{\frac{4BR_A}{K-A}, \frac{16B^2R_A^2}{(K-A)^2}\right\} \le \frac{T}{4},$$

where the last inequality follows from the condition

$$T \ge \max\{16BR_A/(K-A),\ 64B^2R_A^2/(K-A)^2\},$$

which can be derived directly from $B \le A/(8\Gamma_{K,T})$ and $T \ge 256B^2\Gamma_{K,T}^2/(KA)$. Now we combine this inequality with (C.15) and $\mathrm{TV}(\mathbb{P}_0, \mathbb{P}_1) \le 1/4$, which leads to

$$\mathbb{E}_1[N_{\mathcal{S}_1}(T)] \le \frac{T}{4} + \frac{T}{4} = \frac{T}{2}.$$

Finally, under ν_1, every arm outside $\mathcal{S}_1$ has suboptimality gap at least Δ. Therefore,

$$\mathrm{Regret}_{\nu_1,\mathsf{Alg}}(T) \ge \Delta\mathbb{E}_1[T - N_{\mathcal{S}_1}(T)] \ge \frac{\Delta T}{2} = \min\left\{\frac{T}{8}, \frac{(K-A)T}{32BR_A}\right\}.$$

Recall the condition that $T \ge KA(16B^2\Gamma_{K,T}^2)^{-1}$, we have that $T/8 \ge (K-A)T/(32BR_A)$, leading to

$$\mathrm{Regret}_{\nu_1,\mathsf{Alg}}(T) \ge \frac{(K-A)T}{32BR_A} = \frac{\sqrt{KA}}{32B\Gamma_{K,T}}\sqrt{T}.$$

Since ν_1 has exactly B optimal arms, this proves the theorem. □

C.5 Proof of Theorem 6.2

In this section, we provide the proof of Theorem 6.2. We first need the following technical result.

Lemma C.3 (Aggregation of square-root tails)**.** Let $X_1, \ldots, X_q$ be nonnegative integer-valued random variables, not necessarily independent. Suppose that, for some $\delta, c \ge 0$, each X_i satisfies

$$\mathbb{P}[X_i \ge s] \le \delta + \frac{c}{\sqrt{s}}$$

for every $i \in [q]$ and every integer $s \ge 1$. Then, for any integer $m \ge 1$,

$$\mathbb{P}\left[\sum_{i=1}^{q} X_i > m\right] \le 2q\delta + \frac{3qc}{\sqrt{m}}.$$

Proof. For each $i \in [q]$, define the truncated random variable $\widetilde{X}_i := X_i \wedge m$. We decompose the event $\{\sum_{i=1}^q X_i > m\}$ according to whether any individual random variable exceeds the truncation level m. In particular,

$$\left\{ \sum_{i=1}^q X_i > m \right\} \subseteq \left\{ \max_{i\in[q]} X_i > m \right\} \cup \left\{ \sum_{i=1}^q \widetilde{X}_i > m \right\}.$$

For the first event, by a union bound and the assumed tail bound,

$$\mathbb{P}\left[\max_{i\in[q]} X_i > m \right] \le \sum_{i=1}^q \mathbb{P}[X_i > m] \le q\delta + \frac{qc}{\sqrt{m}}.$$

For the second event, since X_i is nonnegative and integer-valued, the tail-sum formula gives

$$\mathbb{E}[\widetilde{X}_i] = \sum_{s=1}^m \mathbb{P}[X_i \ge s] \le m\delta + c\sum_{s=1}^m \frac{1}{\sqrt{s}} \le m\delta + 2c\sqrt{m},$$

where the last inequality follows from $\sum_{s=1}^m 1/\sqrt{s} \le 2\sqrt{m}$. Therefore, by Markov's inequality,

$$\mathbb{P}\left[\sum_{i=1}^q \widetilde{X}_i > m \right] \le \frac{1}{m} \sum_{i=1}^q \mathbb{E}[\widetilde{X}_i] \le q\delta + \frac{2qc}{\sqrt{m}}.$$

Combining the two bounds yields

$$\mathbb{P}\left[\sum_{i=1}^q X_i > m \right] \le 2q\delta + \frac{3qc}{\sqrt{m}},$$

which finishes the proof. □

Now we are ready to prove Theorem 6.2.

Proof of Theorem 6.2. For convenience, define $R_A := \Gamma_{K,T}(K-A)\sqrt{T}/K$, which is the near-minimax optimal rate on $\Xi_{K,A}$. Fix an arbitrary subset $\mathcal{U} \subseteq \mathcal{A}$ with $|\mathcal{U}| = B$. For a parameter $\Delta > 0$ to be specified later, consider the bandit instance ν with reward function

$$r_\nu(a) := \begin{cases} 1/2 + \Delta, & a \in \mathcal{U}, \\ 1/2, & a \in \mathcal{U}^{\complement}. \end{cases}$$

Thus, ν has exactly B optimal arms, namely the arms in $\mathcal{U}$. Recall that θ denotes the all-equal reference instance with $r_\theta(a) = 1/2$ for every $a \in \mathcal{A}$. We use $\mathbb{P}_\nu$ and $\mathbb{P}_\theta$ to denote the trajectory distributions induced by Alg under ν and θ, respectively.

We first show that, under the reference instance θ, the algorithm samples $\mathcal{U}$ only $\widetilde{O}(q^2 R_A^2)$ times with constant probability, where $q := \lceil B/(K-A) \rceil$. Partition $\mathcal{U}$ into q disjoint subsets $\mathcal{U}_1, \dots, \mathcal{U}_q$ such that $|\mathcal{U}_i| \le K - A$ for every $i \in [q]$. For each $i \in [q]$, extend $\mathcal{U}_i$ to a subset $\mathcal{S}_i \subseteq \mathcal{A}$ satisfying $\mathcal{U}_i \subseteq \mathcal{S}_i$ and $|\mathcal{S}_i| = K - A$. Then

$$N_{\mathcal{U}}(T) = \sum_{i=1}^q N_{\mathcal{U}_i}(T) \le \sum_{i=1}^q N_{\mathcal{S}_i}(T).$$

Set $\ell := \log(16q)$. Then, for every $s \in [T]$ and every $i \in [q]$, applying Lemma C.1 gives

$$\mathbb{P}_\theta\big[N_{\mathcal{S}_i}(T) \ge s\big] \le \frac{1}{16q} + 8R_A\sqrt{\frac{\ell}{s}}.$$

Since $N_{\mathcal{S}_i}(T) \le T$, the same inequality holds trivially for $s > T$. Therefore, applying Lemma C.3 to $X_i := N_{\mathcal{S}_i}(T)$ with $\delta = 1/(16q)$ and $c = 8R_A\sqrt{\ell}$ gives, for every $m \ge 1$,

$$\mathbb{P}_\theta\left[\sum_{i=1}^q N_{\mathcal{S}_i}(T) > m \right] \le \frac{1}{8} + 24qR_A\sqrt{\frac{\ell}{m}}.$$

Now we choose $m = \lceil 4096q^2R_A^2\ell \rceil$. Then $24qR_A\sqrt{\ell/m} \le 3/8$, which leads to

$$\mathbb{P}_\theta\left[\sum_{i=1}^{q} N_{\mathcal{S}_i}(T) > m\right] \le \frac{1}{2}.$$

Since $N_{\mathcal{U}}(T) \le \sum_{i=1}^{q} N_{\mathcal{S}_i}(T)$, we conclude that

$$\mathbb{P}_\theta\big[N_{\mathcal{U}}(T) \le m\big] \ge \frac{1}{2}. \tag{C.16}$$

We now show that $T \ge 2m$. Since $B \ge K - A$, we have $q = \lceil B/(K-A) \rceil \le 2B/(K-A)$. Consequently, $\ell \le \log(32B/(K-A))$. Using the definition of R_A,

$$\begin{aligned} 4096q^2R_A^2\ell &\le 16384\frac{B^2}{(K-A)^2}\Gamma_{K,T}^2\frac{(K-A)^2T}{K^2}\log\left(\frac{32B}{K-A}\right) \\ &= 16384\frac{B^2\Gamma_{K,T}^2\log\big(32B/(K-A)\big)}{K^2}T. \end{aligned}$$

By the assumed condition

$$B\Gamma_{K,T}\sqrt{\log\left(\frac{32B}{K-A}\right)} \le \frac{K}{256},$$

the right-hand side is at most $T/4$. Therefore,

$$m = \lceil 4096q^2R_A^2\ell \rceil \le 4096q^2R_A^2\ell + 1 \le \frac{T}{4} + 1 \le \frac{T}{2},$$

where the last inequality follows from $T \ge 4$.

We now transfer (C.16) from the all-equal reference instance θ to the hard instance ν. Once again, we consider the stopping time

$$\mathcal{E} := \big\{N_{\mathcal{U}}(T) \le m\big\}, \qquad \tau := \inf\big\{t \in [T] : N_{\mathcal{U}}(t) = m+1\big\} \wedge T.$$

so that $\mathbb{P}_\theta[\mathcal{E}] \ge 1/2$. Notice that $\mathcal{E} \in \mathcal{F}_\tau$ and $N_{\mathcal{U}}(\tau) \le m+1$ almost surely. Since θ and ν differ only on the arms in $\mathcal{U}$, where the reward distribution is $\mathcal{N}(1/2,1)$ under θ and $\mathcal{N}(1/2+\Delta,1)$ under ν, the KL decomposition for the stopped trajectory gives

$$\mathrm{KL}\left(\mathbb{P}_\theta|_{\mathcal{F}_\tau} \| \mathbb{P}_\nu|_{\mathcal{F}_\tau}\right) = \frac{\Delta^2}{2}\mathbb{E}_\theta[N_{\mathcal{U}}(\tau)] \le \frac{\Delta^2(m+1)}{2}.$$

Now we choose $\Delta := 1/(4\sqrt{m+1})$. Then we see that $\mathsf{KL}\left(\mathbb{P}_\theta|_{\mathcal{F}_\tau} \| \mathbb{P}_\nu|_{\mathcal{F}_\tau}\right) \le 1/32$. By Pinsker's inequality (Lemma E.3), $\mathrm{TV}(\mathbb{P}_{\theta|}|_{\mathcal{F}_\tau}, \mathbb{P}_\nu|_{\mathcal{F}_\tau}) \le 1/8$. Since $\mathcal{E} \in \mathcal{F}_\tau$, it follows from (C.16) that

$$\mathbb{P}_\nu[\mathcal{E}] \ge \mathbb{P}_\theta[\mathcal{E}] - \mathrm{TV}\left(\mathbb{P}_\theta|_{\mathcal{F}_\tau}, \mathbb{P}_\nu|_{\mathcal{F}_\tau}\right) \ge \frac{3}{8}.$$

Recall that the arms in $\mathcal{U}$ are the unique optimal arms under ν, while every arm in $\mathcal{U}^{\mathsf{C}}$ has suboptimality gap Δ. Therefore,

$$\mathrm{Regret}_{\nu,\mathsf{Alg}}(T) = \Delta\mathbb{E}_\nu\big[T - N_{\mathcal{U}}(T)\big] \ge \Delta(T-m)\mathbb{P}_\nu[\mathcal{E}] \ge \frac{3\Delta T}{16} = \frac{3T}{64\sqrt{m+1}},$$

where the last inequality holds due to $m \le T/2$.

It remains to simplify this lower bound. Define $\bar{\ell} = \log(32B/(K-A))$. Since $B/(K-A) \ge 1$, we have $\ell = \log(16q) \ge \log(16B/(K-A)) \ge (4/5)\bar{\ell}$. Together with the assumed horizon condition $T \ge K^2/(B^2\Gamma_{K,T}^2\bar{\ell})$ gives

$$q^2R_A^2\ell \ge \frac{B^2}{(K-A)^2}\Gamma_{K,T}^2\frac{(K-A)^2T}{K^2}\ell \ge \frac{4}{5}\frac{B^2\Gamma_{K,T}^2T\bar{\ell}}{K^2} \ge \frac{4}{5}.$$

Consequently, $m+1 \le 4096q^2R_A^2\ell + 2 \le 8192q^2R_A^2\ell$. Therefore,

$$\mathrm{Regret}_{\nu,\mathsf{Alg}}(T) \ge \frac{3T}{64\sqrt{m+1}} \ge \frac{3}{8192\sqrt{2}}\frac{\sqrt{KA}}{B\Gamma_{K,T}\sqrt{\log(32B/(K-A))}}\sqrt{T},$$

where the second inequality uses $q \le 2B/(K-A)$, $K \ge A$ and $\ell \le \log(32B/(K-A))$. This finishes the proof. □

D HIGH-PROBABILITY LOWER BOUND

As discussed in Remark 4.2, one may ask whether the $\widetilde{O}((K-A)\sqrt{T}/K)$ rate in Theorem 4.1 can be extended to the high-probability setting. In this section, we will show that it is impossible and there is a necessary payoff to achieve the near-optimal expected rate. In particular, let $\{a_t\}_{t=1}^T$ be the actions selected by `Alg`. We write

$$R_T = \sum_{t=1}^{T} r^* - \sum_{t=1}^{T} r(a_t).$$

As a result, we have $\mathbb{E}[R_T] = \mathrm{Regret}(T)$. Our main hardness result is summarized in the following theorem:

Theorem D.1. Suppose $A > K/2$, so that $(K-A)/K < 1/2$, and fix $\Gamma \geq 1$. Define $\ell = \log(16K/(K-A))$ and $c_0 = 13/128$. Suppose that `Alg` is a strategy satisfying for any problem instance in $\Xi_{K,A}$,

$$\mathrm{Regret}(T) \leq \Gamma \frac{K-A}{K}\sqrt{T},$$

that is, a policy that achieves the minimax-optimal expected regret up to a factor of Γ. If $T \geq 8\Gamma^2\ell/c_0^2$, then there exists a problem instance $\mu \in \Xi_{K,A}$ such that

$$\mathbb{P}_{\mu,\texttt{Alg}}\left(R(T) \geq \frac{c_0}{8\sqrt{2}}\frac{\sqrt{T}}{\Gamma\ell}\right) \geq \frac{K-A}{32K}. \tag{D.1}$$

Theorem D.1 indicates that, for any near-minimax optimal algorithm `Alg`, there exists some instance, such that the pseudo-regret of `Alg` reaches $\widetilde{\Omega}(\sqrt{T})$, with probability $\Omega((K-A)/K)$.

Remark D.2. It is widely known that UCB-typed algorithm, while achieving a near-minimax-optimal rate, simultaneously achieves a $\widetilde{O}(\sqrt{KT}\mathrm{polylog}(\delta^{-1}))$ high-probability upper bound. Theorem D.1 shows that, the analogue rate, $\widetilde{O}\big((K-A)K^{-1}\sqrt{T}\cdot\mathrm{polylog}(\delta^{-1})\big)$ is impossible when $A = K - o(K)$. In particular, making $\widetilde{O}\big((K-A)K^{-1}\sqrt{T}\cdot\mathrm{polylog}(\delta^{-1})\big) = \widetilde{O}(\sqrt{T})$ requires $\delta \sim \exp(-K/(K-A)) \ll (K-A)/K$, which contradicts (D.1).

Proof of Theorem D.1. Recall that $\ell = \log(16K/(K-A))$ and $c_0 = 13/128$. Since $T \geq 8\Gamma^2\ell/c_0^2$, we can choose an integer m satisfying

$$\frac{c_0^2 T}{8\Gamma^2\ell} \leq m \leq \frac{c_0^2 T}{4\Gamma^2\ell}. \tag{D.2}$$

Let θ be the instance such that $\boldsymbol{\theta} = \mathbf{1}/2 \in \mathbb{R}^K$ denote the all-equal auxiliary instance. We first apply Lemma C.1 to characterize the exploration behavior of `Alg` under θ. By the lower bound in Lemma C.1,

$$\begin{aligned}
\max_{\mathcal{S}\subseteq[K]:\,|\mathcal{S}|=K-A} \mathbb{P}_\theta\big[N_{\mathcal{S}}(T) \geq m\big] &\geq \frac{K-A}{K} - \exp(-\ell) - 8\Gamma\frac{K-A}{K}\sqrt{\frac{m\ell}{T}} \\
&\geq \frac{K-A}{K} - \frac{K-A}{16K} - \frac{13(K-A)}{32K} \\
&\geq \frac{K-A}{2K},
\end{aligned}$$

where the second inequality follows from $\exp(-\ell) = (K-A)/(16K)$ and the upper bound on m in (D.2). Therefore, there exists a subset $\mathcal{S} \subseteq [K]$ with $|\mathcal{S}| = K - A$ such that

$$\mathbb{P}_\theta\big[N_{\mathcal{S}}(T) \geq m\big] \geq \frac{K-A}{2K}. \tag{D.3}$$

We now construct the hard instance μ by decreasing the mean rewards of the arms in $\mathcal{S}$ by Δ, with $\Delta = (4\sqrt{m\ell})^{-1}$, namely

$$\mu_a := \begin{cases} \frac{1}{2} - \Delta, & a \in \mathcal{S}, \\ \frac{1}{2}, & a \in \mathcal{S}^{\complement}. \end{cases}$$

Since $|\mathcal{S}| = K - A$, the instance μ has exactly A optimal arms and hence belongs to $\Xi_{K,A}$. Now we transfer (D.3) from the auxiliary instance θ to μ. As usual, we prove that at τ, the probability distributions induced by θ and μ are still close, therefore leading to that $\mathbb{P}_\mu[H] \gtrsim (K-A)/K$, where H and τ are defined by

$$H := \big\{N_{\mathcal{S}}(T) \geq m\big\}, \qquad \tau := \inf\big\{t \in [T] : N_{\mathcal{S}}(t) = m\big\} \wedge T.$$

In particular, consider the trajectory distributions restricted to $\mathcal{F}_\tau$ by $\mathbb{P}_\theta^\tau := \mathbb{P}_\theta|_{\mathcal{F}_\tau}$ and $\mathbb{P}_\mu^\tau := \mathbb{P}_\mu|_{\mathcal{F}_\tau}$. Applying Proposition B.3 gives that

$$\begin{aligned}\log \frac{\mathrm{d}\mathbb{P}_\theta|_{\mathcal{F}_\tau}}{\mathrm{d}\mathbb{P}_\mu|_{\mathcal{F}_\tau}} &= \sum_{t=1}^{\tau}\left[(\theta_{a_t} - \mu_{a_t})(r_t - \theta_{a_t}) + \frac{1}{2}(\theta_{a_t} - \mu_{a_t})^2\right] \\ &= \Delta\sum_{t=1}^{\tau}\mathbb{1}\{a_t \in \mathcal{S}\}(r_t - 1/2) + \frac{\Delta^2}{2}N_{\mathcal{S}}(\tau).\end{aligned}$$

Now we denote $M_t = \Delta\sum_{s=1}^{t}\mathbb{1}\{a_s \in \mathcal{S}\}(r_s - 1/2)$ and $V_t = \Delta^2 N_{\mathcal{S}}(t)$, then the probability distribution $\mathbb{P}_\theta$, we see that for any $\lambda \in \mathbb{R}$,

$$\begin{aligned}&\mathbb{E}_\theta\left[\exp\left(\lambda M_t - \frac{\lambda^2 V_t}{2}\right)\middle|\mathcal{F}_{t-1}\right] \\ &\quad= \exp\left(\lambda M_{t-1} - \frac{\lambda^2 V_{t-1}}{2}\right)\mathbb{E}_\theta\Bigg[\underbrace{\exp\left(\lambda(r_t - 1/2)\Delta\,\mathbb{1}\{a_t \in \mathcal{S}\} - \frac{\lambda^2\Delta^2\,\mathbb{1}\{a_t \in \mathcal{S}\}}{2}\right)}_{(\star)}\Bigg|\mathcal{F}_{t-1}\Bigg]\end{aligned}$$

For $(\star)$, we first take expectation on $\mathcal{G}_t$, which gives that

$$\begin{aligned}\mathbb{E}_\theta[(\star)|\mathcal{G}_t] &= 1 - \mathbb{1}\{a_t \in \mathcal{S}\} + \mathbb{1}\{a_t \in \mathcal{S}\}\mathbb{E}_\theta\left[\exp\left(\lambda(r_t - 1/2)\Delta - \frac{\lambda^2\Delta^2}{2}\right)\middle|\mathcal{G}_t\right] \\ &\leq 1 - \mathbb{1}\{a_t \in \mathcal{S}\} + \mathbb{1}\{a_t \in \mathcal{S}\} \\ &= 1,\end{aligned}$$

where the inequality holds due to that $r_t - 1/2$ is 1-sub-gaussian conditioned on $\mathcal{G}_t$ on $\mathbb{P}_\theta$. This immediately gives $\mathbb{E}_\theta[(\star)|\mathcal{F}_{t-1}] \leq 1$, verifying that $\exp(\lambda M_t - \lambda^2 V_t/2)$ is a super-martingale for any λ. Now invoking Lemma E.1 with $x = \Delta\sqrt{2m\ell}$, since $V_\tau \leq m\Delta^2$, we have

$$\mathbb{P}_\theta\big(M_\tau \geq \Delta\sqrt{2m\ell}\big) \leq \exp\left(-\frac{2m\Delta^2\ell}{2m\Delta^2}\right) = \exp(-\ell).$$

Since $\Delta^2 N_{\mathcal{S}}(\tau_{\mathcal{S}})/2 \leq m\Delta^2/2$, this immediately leads to

$$\mathbb{P}_\theta\left(\log\frac{\mathrm{d}\mathbb{P}_\theta^\tau}{\mathrm{d}\mathbb{P}_\mu^\tau} \geq \Delta\sqrt{2m\ell} + m\Delta^2/2\right) \leq \exp(-\ell).$$

Recall that $\Delta = (4\sqrt{m\ell})^{-1}$, this means that $\Delta\sqrt{2m\ell} + m\Delta^2/2 = 1/2\sqrt{2} + 1/32\ell < \log 2$ since $\ell \geq 1$. Therefore,

$$\mathbb{P}\left(\log\frac{\mathrm{d}\mathbb{P}_\theta^\tau}{\mathrm{d}\mathbb{P}_\mu^\tau} > \log 2\right) \leq \exp(-\ell).$$

Now we apply Proposition B.1 on the measurable space $(\Omega, \mathcal{F}_{\tau_{\mathcal{S}}})$ with $\mathbb{P} = \mathbb{P}_\theta^\tau$, $\mathbb{Q} = \mathbb{P}_\mu^\tau$, $X = \mathbb{1}\{H\}$, $b = 1$ and $\gamma = \log 2$ and obtain that

$$\mathbb{P}_\mu(H) = \mathbb{P}_\mu^\tau(H) \geq \frac{1}{2}\left[\mathbb{P}_\theta^\tau(H) - \mathbb{P}\left(\log\frac{\mathrm{d}\mathbb{P}_\theta^\tau}{\mathrm{d}\mathbb{P}_\mu^\tau} > \log 2\right)\right] \geq \frac{1}{2}\big[\mathbb{P}_\theta(H) - \exp(-\ell)\big],$$

where the equation holds due to that $H \in \mathcal{F}_\tau$.

Finally, every pull of an arm in $\mathcal{S}$ incurs regret Δ under μ. Hence, on the event H,

$$R(T) \geq \Delta m = \frac{\sqrt{m}}{4\sqrt{\ell}}.$$

Using the lower bound on m in (D.2), we obtain

$$\Delta m \geq \frac{1}{4\sqrt{\ell}} \sqrt{\frac{c_0^2 T}{8\Gamma^2 \ell}} = \frac{c_0}{8\sqrt{2}} \frac{\sqrt{T}}{\Gamma \ell}.$$

Therefore,

$$\mathbb{P}_\mu \left(R(T) \geq \frac{c_0}{8\sqrt{2}} \frac{\sqrt{T}}{\Gamma \ell} \right) \geq \frac{K - A}{32K},$$

which proves the desired result. □

E AUXILIARY LEMMAS

The following standard stopped martingale inequality is a direct consequence of the time-uniform Chernoff bound of Howard et al. (2020, Theorem 1).

Lemma E.1 (Stopped sub-Gaussian martingale inequality)**.** Let $(M_t, V_t)_{t=0}^T$ be an adapted process with $M_0 = V_0 = 0$ and $(V_t)_{t=0}^T$ nondecreasing. Suppose that, for every $s > 0$, $\exp(sM_t - s^2 V_t/2)$ is a nonnegative supermartingale. Then, for any stopping time $\tau \leq T$ and every $x, v > 0$,

$$\mathbb{P}\big(M_\tau \geq x,\, V_\tau \leq v\big) \leq \exp\left(-\frac{x^2}{2v}\right). \tag{E.1}$$

In particular, if $V_\tau \leq v$ almost surely, then

$$\mathbb{P}(M_\tau \geq x) \leq \exp\left(-\frac{x^2}{2v}\right).$$

The following lemma characterizes the change-of-measure under total variation, which is the core components in standard lower bound techniques like Le-Cam's method (Yu, 1997). We refer to Tsybakov & Zaiats (2009, Lemma 2.1) for the proof.

Lemma E.2. Let $(\Omega, \mathcal{F})$ be a measurable space, and let $\mathbb{P}$ and $\mathbb{Q}$ be probability measures on it. Then, for every $\mathcal{F}$-measurable function $f : \Omega \to [0, T]$,

$$\big|\mathbb{E}_\mathbb{P}[f] - \mathbb{E}_\mathbb{Q}[f]\big| \leq T \,\mathrm{TV}(\mathbb{P}, \mathbb{Q}),$$

where $\mathrm{TV}(\mathbb{P}, \mathbb{Q}) := \sup_{E \in \mathcal{F}} \big|\mathbb{P}(E) - \mathbb{Q}(E)\big|$.

Lemma E.3 (Pinsker's inequality)**.** Let $(\Omega, \mathcal{F})$ be a measurable space, and let $\mathbb{P}$ and $\mathbb{Q}$ be probability measures on it. Then

$$\mathrm{TV}(\mathbb{P}, \mathbb{Q}) \leq \sqrt{\frac{1}{2} \mathrm{KL}(\mathbb{P} \| \mathbb{Q})}.$$

The following lemma is standard for proof of minimax lower bound and we refer to Lattimore & Szepesvári (2020, Theorem 14.2) for the proof.

Lemma E.4 (Bretagnolle–Huber inequality)**.** Let $(\Omega, \mathcal{F})$ be a measurable space, and let $\mathbb{P}$ and $\mathbb{Q}$ be probability measures on it. Then for every event $E \in \mathcal{F}$,

$$\mathbb{P}(E) + \mathbb{Q}(E^\complement) \geq \frac{1}{2} \exp\big(-\mathrm{KL}(\mathbb{P} \| \mathbb{Q})\big).$$