Proceedings of Machine Learning Research vol 178:1–40, 2022 35th Annual Conference on Learning Theory

A Sharp Memory-Regret Trade-Off for Multi-Pass Streaming Bandits

Arpit Agarwal AA4931@COLUMBIA.EDU
Columbia University

Sanjeev Khanna SANJEEV@CIS.UPENN.EDU
Prathamesh Patil PPRATH@CIS.UPENN.EDU
University of Pennsylvania

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

The stochastic K-armed bandit problem has been studied extensively due to its applications in various domains ranging from online advertising to clinical trials. In practice however, the number of arms can be very large resulting in large memory requirements for simultaneously processing them. In this paper we consider a streaming setting where the arms are presented in a stream and the algorithm uses limited memory to process these arms. Here, the goal is not only to minimize regret, but also to do so in minimal memory. Previous algorithms for this problem operate in one of the two settings: they either use $\Omega(\log \log T)$ passes over the stream (Rathod, 2021; Chaudhuri and Kalyanakrishnan, 2020; Liau et al., 2018), or just a single pass (Maiti et al., 2021).

In this paper we study the trade-off between memory and regret when B passes over the stream are allowed, for any $B \geq 1$, and establish *tight* regret upper and lower bounds for any B-pass algorithm. Our results uncover a surprising *sharp transition phenomenon*: $O(1)$ memory is sufficient to achieve $\widetilde{\Theta}\left(T^{\frac{1}{2}+\frac{1}{2^{B+2}-2}}\right)$ regret in B passes, and increasing the memory to any quantity that is $o(K)$ has almost no impact on further reducing this regret, unless we use $\Omega(K)$ memory. Our main technical contribution is our lower bound which requires the use of *information-theoretic techniques* as well as ideas from *round elimination* to show that the *residual problem* remains challenging over subsequent passes.

1. Introduction

The stochastic multi-armed bandit problem is a widely studied problem with applications in many domains such as online advertising, recommendation systems, clinical trials, financial portfolio design etc. In this problem, there are K arms; in trial $t \in [T]$ the algorithm pulls an arm a_t and receives a reward drawn from the reward distribution of a_t with mean μ_{a_t}. The goal of the algorithm is to minimize the cumulative regret over T trials where the regret for trial t is defined as the gap between the largest mean reward $\max_{a \in [K]} \mu_a$ and μ_{a_t}.

In many practical applications such as online advertising and recommendation systems, the number of arms can be very large and the learner might not be able to store all the arms in memory. In these applications it can be more practical to process arms in a sequential manner with small memory that is sub-linear in the number of arms. Motivated by a long line of work on streaming algorithms in theoretical computer science (Alon et al., 1999), we consider a setting where the arms are presented in a (possibly adversarially chosen) stream and in each trial the algorithm decides whether to read the next arm from the stream into memory. The algorithm can only store the indices and statistics of M arms out of the K arms and can only play an arm if it is present in the memory. The goal in this setting is to minimize the regret given a bounded amount of memory.

Previously, Rathod (2021); Chaudhuri and Kalyanakrishnan (2020); Liau et al. (2018) developed algorithms for regret minimization in this limited-memory streaming bandits setting, but their algorithms require a relatively large number of passes over the stream, with the former requiring $O(\log\log T)$ passes, and the latter two requiring $O(\log T)$ passes. Moreover, it is not understood whether the trade-off between memory and regret obtained by these algorithms is tight as a corresponding multi-pass lower bound is not known. At the other extreme, Maiti et al. (2021) considered a 1-pass streaming setting and showed that any algorithm using M words (for $M < K$) of memory needs to incur $\Omega(T^{2/3}/M^{7/3})$ expected regret. Also, there is a simple 1-pass algorithm that uses $M = O(1)$ memory and achieves $O(K^{1/2}T^{2/3})$ regret. These results of Maiti et al. (2021) imply that the 1-pass setting exhibits a sharp trade-off between memory and regret as explained below.

The 1-pass regret as a function of memory M has a sharp transition: with $M = O(1)$ one can achieve $O(T^{2/3})$ regret, and increasing M beyond $O(1)$ has little impact on further reducing this regret, unless we increase M to K in which case one can achieve $O(\sqrt{T})$ regret.[1]

In this paper we study a streaming setting for multi-armed bandits where we are allowed B passes over the stream, for any $B \geq 1$. We seek to provide upper and lower bounds on the expected regret under a limited memory budget and B passes. We also seek to understand the trade-off between memory and regret as a function of the number of passes B. Does increasing memory beyond $O(1)$ help in this B-pass setting or is there again a sharp transition in regret similar to the 1-pass setting?

Our main result is to prove a lower bound on the regret of any B-pass algorithm that uses limited amount of memory. In particular, we show that any B-pass algorithm that uses $o(K/B^2)$ words of memory necessarily incurs $\Omega\left(4^{-B}T^{\frac{1}{2}+\frac{1}{2^{B+2}-2}}\right)$ regret in expectation. Note that for $B = 1$ our result implies a tighter lower bound of $\Omega\left(T^{2/3}\right)$ as compared to the $\Omega\left(T^{2/3}/M^{7/3}\right)$ bound in Maiti et al. (2021), for any 1-pass algorithm that uses $M < K/24$ words of memory.

Our lower bound exploits the main tension in the streaming setting: the algorithm has limited information about whether there are better arms further along in the stream, and hence, it is difficult to decide whether to keep exploring the current arms in memory or to read more arms into memory by throwing away some of the current arms from memory. We construct a distribution over *hard instances* such that, if in the first pass the algorithm performs *sufficient exploration* over (potentially 'bad') arms then it already incurs a large regret in expectation. If it performs *insufficient exploration* in the first pass then it will throw away many 'good' arms due to a limited memory budget and will be unable to isolate the underlying instance at the end of the first pass. One of the main technical difficulty is to show that the resulting *residual* distribution over instances is *challenging* in a way that leads to large regret in the remaining $B-1$ passes. We overcome this difficulty by using information-theoretic techniques to show that *insufficient exploration* leads to low *mutual information* which

1. Note that when $M = K$ one can simply read all the arms in memory at once and use any stochastic multi-armed bandit algorithm such as the UCB algorithm to achieve a regret of $O(\sqrt{TK})$.

further leads to *large entropy* in the residual distribution. We then inductively argue that any high entropy distribution over instances will lead to large regret in the remaining $B-1$ passes.

We complement our lower bound with a simple B-pass algorithm that uses $O(1)$ memory and achieves an expected regret upper bound of $\widetilde{O}\left(T^{\frac{1}{2}+\frac{1}{2^{B+2}-2}}\sqrt{KB}\right)$. This implies that $O(\log\log T)$ passes and $O(1)$ memory are sufficient to achieve an almost optimal regret of $\widetilde{O}(\sqrt{KT})$, and matches the recent $O(\log\log T)$ pass regret upper bound of Rathod (2021). When $B=1$, we also recover the $O(T^{2/3}\sqrt{K})$ upper bound of Maiti et al. (2021). In short, our algorithmic result interpolates the hitherto unknown space between the 1-pass $\widetilde{O}(T^{2/3}\sqrt{K})$ regret and the $(\log\log T)$-pass $\widetilde{O}(\sqrt{KT})$ regret upper bounds as function of the number of passes B.

Our algorithm is based on two key operations: (i) estimating the reward of the best arm, (ii) identifying sub-optimal arms based on this estimate. In each pass the algorithm sets a maximum budget for the number of pulls allowed for each arm, and this budget keeps increasing over successive passes. The algorithm reads an arm into memory and pulls this arm until it is identified as a sub-optimal arm or the maximum budget is exceeded. The estimate of the maximum reward is then updated and the next arm is read into memory. Since the budget keeps increasing over passes, the estimate for the maximum reward becomes more refined, and sub-optimal arms are identified more easily.

Our lower and upper bound together imply the following (perhaps surprising) sharp threshold phenomenon in our B-pass setting.

The B-pass regret as a function of memory M has a sharp transition: with $M=O(1)$ one can achieve $\widetilde{\Theta}\left(T^{\frac{1}{2}+\frac{1}{2^{B+2}-2}}\right)$ regret, and increasing M to any quantity that is $o(K/B^2)$ has almost no impact on further reducing this regret.

Related Work. The stochastic multi-armed bandit problem has been extensively studied in many fields including operations research, statistics and machine learning. We refer the reader to excellent surveys in Bubeck et al. (2013); Slivkins (2019), and only mention work that is directly relevant to our streaming setting. Liau et al. (2018) studied a limited memory setting for multi-armed bandits and showed that one can achieve (almost) instance-wise optimal regret in $O(\log T)$ passes and $O(1)$ memory. Chaudhuri and Kalyanakrishnan (2020) studied a similar setting and showed that with $O(\log T)$ passes and M memory one can achieve a regret upper bound of $\widetilde{O}(KM+\frac{K^{3/2}}{M}\sqrt{T})$. However, these works only considered a $O(\log T)$-pass setting and did not study the trade-off between memory and regret for any arbitrary number of passes $1\leq B<\log T$. A recent arXiv paper (Rathod, 2021) achieves a regret upper bound of $\widetilde{O}(\sqrt{KT})$ in $O(\log\log T)$ passes. However, their work does not address the question of the regret achievable (both upper and lower bounds) for any arbitrary number of passes $1\leq B<\log\log T$. As discussed earlier, Maiti et al. (2021) considered a 1-pass streaming setting, but their results do not apply more generally to $B>1$ passes, which is the main focus of our paper. There is also some work on best arm identification with limited memory in the streaming setting. Assadi and Wang (2020) show that one can identify the best arm with 1 pass over the stream and $O(1)$ memory using $O(K/\Delta^2)$ sample complexity where Δ is the minimum gap between the best arm and any other arm. Jin et al. (2021) further obtain instance-wise optimal sample complexity for this problem using $\log 1/\Delta$ passes and $O(1)$ memory.

The stochastic multi-armed bandits problem has also been studied under the setting of limited adaptivity (Gao et al., 2019; Perchet et al., 2015). Under this setting, an algorithm operates in rounds and in each round it plays arms according to a fixed distribution that can only depend on the outcomes from the previous rounds. Even though the tradeoff between rounds and regret in this setting is similar to the tradeoff between passes and regret given limited memory in our setting, the key difference between the two settings is that this setting necessarily requires at least 1 bit of information per arm for a total of $\Omega(K)$ memory, but cannot be adaptive within a batch, whereas in our setting, we can be fully adaptive within a pass but are given strictly less than K memory. Due to this difference the challenges in these two settings are quite different, which reflects in the techniques used in the respective lower bounds.

Very recently, independently of our work, Srinivas et al. (2022) studied the problem of online learning with expert advice in a streaming setting and established a trade-off between regret and memory in this setting. However, there are several fundamental differences between the multi-armed bandits problem studied here and the experts problem studied in Srinivas et al. (2022)– (1) in the experts problem one gets to see the loss of every expert at every trial, whereas in our problem one only gets to see the reward of the arm that is played, (2) in Srinivas et al. (2022) the losses on experts are generated adversarially whereas in our work the rewards of arms are generated stochastically, (3) in Srinivas et al. (2022) the stream consists of the prediction of experts for each trial, whereas in our work the stream consists of the arms. As a result, the two settings require very different techniques for proving lower and upper bounds, and neither result has any implications on the other.

Organization. In Section 2 we discuss the problem setting and set up relevant notation. We discuss our lower bound on regret in Section 3 which is the main result of our paper. We then provide an upper bound on regret in Section 4, and finally conclude in Section 5.

2. Problem Setting

We study the stochastic multi-armed bandit problem, where the instance consists of a finite set $\mathcal{K}$ of ($K = |\mathcal{K}|$) arms and a time horizon T of trials which is known ahead of time. When an arm $a \in \mathcal{K}$ is played in a trial, an i.i.d. reward is drawn from its corresponding reward distribution defined over $[0, 1]$ with mean μ_a of which the algorithm has no prior knowledge.[2] The objective here is to minimize the cumulative regret, which is defined as $R_T := \sum_{t=1}^{T}(\max_{a\in\mathcal{K}} \mu_a - \mu_{a_t})$ where a_t is the arm played in trial $t \in [T]$.

We assume a limited memory setting where the arms $\mathcal{K}$ are presented to the algorithm as an *arbitrarily (possibly adversarially) ordered read-only stream*, and the algorithm is restricted to store the *identities* and the corresponding *statistics* of at most $M < K$ arms simultaneously while being allowed at most $B \geq 1$ passes over the stream. The input parameters T, K, B and M are assumed to be stored for free ($O(1)$ space). Crucially, *the algorithm can only play an arm if it is in its memory*. Therefore, in each trial $t \in [T]$, the algorithm must decide to either play an arm currently present in its memory, which generates a reward (potentially incurring regret) and consumes a trial, or read the next arm from the stream into memory, which neither incurs regret nor consumes a trial. If the algorithm chooses to do the latter and the memory is full, then it must first discard some arm to accommodate the new arm, in which case both the statistics as well as the identity of the discarded arm are forgotten.

2. We assume that the support of the reward distributions is $[0, 1]$ for ease of analysis; our algorithmic results can be easily extended to sub-Gaussian distributions over arbitrary support.

Furthermore, the discarded arm cannot be read back into memory (and hence played) until it is encountered again in a future pass over the stream.

Remark 1 *In the above multi-pass streaming setting, the set of arms in any pass of the stream remains the same though their order may change arbitrarily between passes. One can also consider a modified setting where the algorithm is allowed to permanently discard arms from the stream so that they do not appear in future passes. For example, one might want to discard some arms if they are identified to be strictly suboptimal, in which case there is no need to process them any further. We note that both our lower bound and our algorithmic results also apply to this modified setting.*

Notation. In the rest of this paper, we use upper case letters to refer to instance dependent constants, such as the length of the time horizon T, number of arms K, number of passes B, and the memory size M. We use $\mathcal{B}$, $\mathcal{D}$, ψ and ϕ to refer to distributions, and $\mathcal{E}$ to refer to events. We use other upper case calligraphic letters to refer to sets, and other lower case English or Greek letters to refer to miscellaneous constants. Lastly, we use $\log$ base 2, and $\ln$ for natural logarithms.

We denote random variables in serif font, e.g., X. For a random variable X, $\text{supp}(\mathsf{X})$ denotes the support of X and $\text{dist}(\mathsf{X})$ denotes its distribution. We denote the *Shannon Entropy* of a random variable A by $\mathbb{H}(\mathsf{A})$ and the *mutual information* of two random variables A and B by $\mathbb{I}(\mathsf{A}\,;\mathsf{B}) = \mathbb{H}(\mathsf{A}) - \mathbb{H}(\mathsf{A} \mid \mathsf{B}) = \mathbb{H}(\mathsf{B}) - \mathbb{H}(\mathsf{B} \mid \mathsf{A})$. A summary of useful facts from information theory is given in Appendix A.

3. A Regret Lower Bound for Limited Memory Multi-Pass Algorithms

Our main result, which is an information-theoretic lower bound on the cumulative regret that can be achieved by any B-pass algorithm with limited memory, is presented in the following theorem.

Theorem 1 *Given a time horizon T, a stream of K arms, and passes $1 \le B < \log\log T$ over this stream, there exists a distribution over K-armed bandit instances such that any B-pass algorithm that uses at most $K \cdot (8B(B+1)\log e)^{-1}$ memory suffers $\Omega\left(4^{-B} T^{2^B/(2^{B+1}-1)}\right)$ regret in expectation.*

This lower bound paints a rather pessimistic picture for regret minimization in a limited memory streaming setting. Given any constant number of passes, we need $\Omega(K)$ memory to achieve $O(\sqrt{T})$ regret that is already achievable by a single pass algorithm with memory K. Furthermore, for any given memory M up to $o(K/\log^2\log T)$, a superconstant $\Omega(\log\log T)$ number of passes are required to achieve this optimal regret. In Section 4, we will show another surprising result on the threshold nature of memory: for a fixed number of passes B, the regret achieved by a constant memory algorithm is asymptotically no different from that achieved by any $o(K/B^2)$ memory algorithm. In other words, for any fixed number of passes B, the worst-case regret does not reduce with increasing memory unless we allow a relatively large $\Omega(K/B^2)$ memory.

To the best of our knowledge, this is the first regret lower bound for any $B > 1$ number of passes, and also improves upon the $\Omega(T^{2/3}/M^{7/4})$ lower bound of Maiti et al. (2021) for $B = 1$. We now present the key elements of our lower bound construction.

3.1. Overview of the Lower Bound

At a high level, our lower bound exploits the fact that any limited memory algorithm must operate conservatively due to the presence of arms for which it has absolutely no information until they are actually encountered in the stream. Since only a limited number of arms can be explored at any given time, any limited memory algorithm faces the following dilemma. (1) Spend enough time playing the arms it has in memory and gain some meaningful information about them, but then potentially run the risk of acquiring large regret in the event there is some high value arm yet to be seen, or (2) Try to quickly move ahead in the stream, discarding arms in memory after a few samples, but then potentially risk throwing away good arms due to lack of sufficient information. Since the decision to throw away arms is irrevocable, and the statistics and identities of the discarded arms are forgotten, the algorithm would then have one fewer pass to rectify its mistake in the event that no obviously high value arms are found ahead in the stream.

In the proof of our regret lower bound, without loss of generality, we will assume that the stream order does not change between passes[3]. Furthermore, we will restrict our attention to only deterministic algorithms, as a lower bound for deterministic algorithms on a suitable distribution over input instances also implies an identical lower bound for randomized algorithms (Yao, 1977). Formally, given any randomized algorithm with *low expected regret*, there exists a choice of random bits such that this algorithm achieves low expected regret given this choice of bits. Therefore, by conditioning on these random bits, we have a deterministic algorithm that achieves low expected regret.

Our lower bound is based on the general idea of 'round elimination' used for proving communication complexity lower bounds where one inductively argues that the residual instance at the end of the each round will remain 'hard' over subsequent rounds. Our B-pass lower bound constructs 'hard' instances over K arms by composing together $B+1$ layers of 'hard' instances over subsets of arms. We partition the stream of K arms into contiguous subsets of size $K/(B+1)$ and the j-th layer of hard instances is defined over the j-th $K/(B+1)$-sized subset. At a high level, we argue that after performing j passes, an algorithm will either incur 'large regret' or will only be able to 'peel-off' the last j layers. In other words, if the algorithm has not incurred 'high regret' at the end of j passes, then it still needs to solve a hard problem over at least $B+1-j$ layers with only $B-j$ passes left.

Within each layer $j \in [B+1]$, we generate a 'hard' instance by sampling a special arm i_j^* from a *near-uniform* distribution over the arms in that layer, whose mean reward is *nearly-equally-likely* to be either *low*, namely $\mu_{i_j^*} = 1/2$, or *high*, namely $\mu_{i_j^*} = 1/2 + \Delta_j$ where Δ_j is a parameter that we will specify shortly. All other arms in this layer have mean reward $1/2$. This potential 'high' reward of $1/2 + \Delta_j$ increases across layers, with Δ_1 being the smallest, and Δ_{B+1} being the largest. This intuitively forces any algorithm to rush through all of the initial B layers, because the regret would be massive if i_{B+1}^* realizes to have a high reward, the odds of which are nearly half. However, in doing so the algorithm will learn very little about the special arms in first B layers, and will have to solve a hard problem over these layers in the remaining $B-1$ rounds.

In order to formalize the above construction, we define a distribution over 'hard' instances for a single layer that is parameterized by the set of arms $\mathcal{A}$ in the layer, the mean reward parameter Δ for

3. The regret guarantees of our algorithm (Section 4) hold even when the arm order changes adversarially between passes.

the special arm in that layer, and a nearly-uniform joint distribution ψ over $\mathcal{A} \times \{0,1\}$ for sampling said special arm and its mean reward.

Distribution $\mathcal{D}_{\mathcal{A}}^{\Delta,\psi}$: Given a set of arms $\mathcal{A}$, a joint distribution ψ over support $\mathcal{A} \times \{0,1\}$, and parameter $\Delta \leq 1/4$

- Sample $(i^*, y) \sim \psi$ such that $i^* \in \mathcal{A}$ and $y \in \{0,1\}$. For all $i \in \mathcal{A}$, let

$$\mu_i = \begin{cases} \frac{1}{2} + y\Delta \,, & \text{if } i = i^* \\ \frac{1}{2} & \text{otherwise} \end{cases}$$

- Return the arms $\mathcal{A}$ with Bernoulli reward distributions with means $\{\mu_i\}_{i \in \mathcal{A}}$.

Note that the special arm i^* in layer is also a best arm within the layer. We will now define what is means for the distribution ψ of the special arm i^* to be γ-nearly uniform.

Definition 2 (γ-nearly uniform ψ) *Given a set of arms $\mathcal{A}$, a joint distribution ψ over support $\mathcal{A} \times \{0,1\}$, and $\gamma > 0$, we say that ψ is γ-nearly uniform if the random variables $(\mathsf{I}, \mathsf{Y}) \sim \psi$ are such that $\mathbb{H}(\mathsf{I}) \geq \log A - \gamma$ and $\mathbb{H}(\mathsf{Y}|\mathsf{I}) \geq \log 2 - \gamma$.*

This following key lemma quantifies how little any algorithm would actually learn about the special arm in a layer if this arm is γ-nearly uniformly distributed and the algorithm rushes through this layer, i.e. collects very few samples.

Lemma 3 *Given a time-horizon T, a set of arms $\mathcal{A}$ of size $A = |\mathcal{A}|$, with mean rewards generated according to a distribution $\mathcal{D}_{\mathcal{A}}^{\Delta,\psi}$ where ψ is γ-nearly uniform for some $\gamma \geq 0$. Let $(\mathsf{I}, \mathsf{Y}) \sim \psi$ and let* ALG *be any deterministic algorithm that* adaptively *pulls arms in $\mathcal{A}$. Let $\sigma \in [T]$ be the* randomly chosen *stopping time of the algorithm and $\mathcal{S}_\sigma = (j_t, r_t)_{t=1}^{\sigma}$ be execution history of* ALG *with j_t being the arm pulled and r_t being its observed reward in trial t, respectively. For a given input parameter $\beta < 1$, let $\mathcal{M} \subset \mathcal{A}$ be any set of size βA chosen to be retained in memory by* ALG *after observing the execution history $\mathcal{S}_\sigma$. If $\mathbb{E}[\sigma] \leq \frac{\epsilon^2}{6\Delta^2}$ for some $\epsilon > 0$, then the event*

$$\mathcal{E} = \Big(\mathsf{I} \notin \mathcal{M} \textit{ and } \mathbb{H}(\mathsf{I} \mid \mathcal{S}_\sigma, \mathsf{I} \notin \mathcal{M}) \geq \log((1-\beta)A) - \frac{\gamma + \epsilon}{1 - \log(1+\beta) - \gamma - \epsilon} \\ \textit{and } \mathbb{H}(\mathsf{Y} \mid \mathcal{S}_\sigma, \mathsf{I}, \mathsf{I} \notin \mathcal{M}) \geq \log 2 - \frac{\gamma + \epsilon}{1 - \log(1+\beta) - \gamma - \epsilon}\Big)\,,$$

occurs with probability at least $1 - \log(1+\beta) - \gamma - 3\epsilon$ over the samples seen by the algorithm.

A formal proof of this lemma is given in Appendix B, and its statement can be interpreted as follows – given a set of instances where the special arm and its mean reward are sampled from a γ-nearly uniform distribution, then no algorithm can hope to trap this special arm in a small subset (its memory) after a period of insufficient exploration with any considerable probability. Moreover, in the event that this arm is discarded, nothing meaningful is learned either as this special arm is nearly-equally likely to be any of the discarded arms, and its mean reward is nearly-equally likely to be either low or high. Thus, the entropy of the identity of the special arm as well as its mean reward remains large in the posterior distribution over the discarded arms induced by the samples observed by the algorithm. This observation will be important to show that in the event that the arm i^*_{B+1} with the largest potential reward realizes to have a low value, the algorithm still faces a hard distribution consisting of B layers while having depleted one of its passes. We will now "stitch" together these $(B+1)$ layer-wise hard distributions into a hard distribution over all K arms.

Distribution $\mathcal{D}_{\mathcal{K},B}^{\{\psi_j\}_{j=1}^{b+1}}$: Given a set of arms $\mathcal{K}$ of size $K = |\mathcal{K}|$, an integer $B \in \mathbb{N}_+$, and a set of $(b+1) \leq (B+1)$ joint distributions $\{\psi_j\}_{j=1}^{b+1}$ where each ψ_j is supported over $\mathcal{A}_j \times \{0,1\}$ with $\{\mathcal{A}_j\}_{j=1}^{b+1}$ being a contiguous and sequential partition of $\mathcal{K}$ into sets of equal size $K/(b+1)$.

- For $j \in [b+1]$, define

$$\Delta_j = \frac{T^{-\frac{2^B - 2^{j-1}}{2^{B+1}-1}}}{4}.$$

- For $j \in [b+1]$, sample mean reward parameters $\{\mu_i\}_{i \in \mathcal{A}_j}$ according to $\mathcal{D}_{\mathcal{A}_j}^{\Delta_j,\psi_j}$.
- Define

$$\mathcal{D}_{\mathcal{K},B}^{\{\psi^j\}_{j=1}^{b+1}} := \mathcal{D}_{\mathcal{A}_1}^{\Delta_1,\psi_1} \otimes \mathcal{D}_{\mathcal{A}_2}^{\Delta_2,\psi_2} \otimes \cdots \otimes \mathcal{D}_{\mathcal{A}_{b+1}}^{\Delta_{b+1},\psi_{b+1}}$$

- Return the arms $\mathcal{K}$ with the reward distribution of arm $i \in \mathcal{K}$ being Bernoulli $\mathcal{B}(\mu_i)$.

In the above distribution, one should think of B as an input parameter that corresponds to the number of passes allowed to the algorithm at the start, and b as the remaining number of passes at some intermediate step. We define our distribution for any $b \leq B$ as we need to show that the residual distribution over the instance remains 'hard' at every intermediate step in the algorithm. Hence, if there are b passes remaining, the algorithm still faces a $(b+1)$-layered 'hard' residual instance.

Armed with this hard distribution, we are now ready to prove the lower bound as follows. Let there be b passes remaining at an intermediate step in the algorithm, and let the distribution of rewards be according to $\mathcal{D}_{\mathcal{K},B}^{\{\psi_j\}_{j=1}^{b+1}}$ such that the special arm in each of the $b+1$ layers is nearly uniformly distributed. The algorithm is presented with each layer one by one in the stream. We divide the execution of the algorithm into $b+1$ epochs where the j-th epoch begins when the first arm in layer j is read into memory and ends right before the first arm from layer $j+1$ is read into memory.

Let $\alpha = 2^B/(2^{B+1}-1)$, and let the available memory be $\beta K/(b+1)$ for an appropriately chosen $\beta \in (0,1]$. Since the number of arms in each layer is $K/(b+1)$, the algorithm needs to discard at least $(1-\beta)$ fraction of the arms from each layer. Now suppose for any of the first $j \in [b]$ epochs, the algorithm actually collects at least ϵ^2/Δ_j^2 (for some small ϵ) samples in that epoch, then we are already done as the algorithm will suffer $\Omega(\epsilon^2\Delta_{b+1}/\Delta_j^2) = \Omega(T^\alpha)$ regret if the reward of i^*_{b+1} realizes to its high value, the odds of which are nearly half. On the other hand, if the algorithm does not explore enough in every epoch, then for sufficiently small β, ϵ, the bad event described in Lemma 3 will occur for all of the initial b epochs with constant probability. As a result, the posterior reward distributions over the $(1-\beta)$ fraction of arms discarded from every layer will provably remain hard (as per our definition of a hard instance for a layer). Therefore, if the reward of i^*_{b+1} realizes to its low value, the odds of which are nearly half, the algorithm now faces a hard distribution with b layers over $b-1$ passes, at which point we will appeal to induction to show that the regret of this algorithm in this case must also be large. This idea is formalized in the following lemma.

Lemma 4 *Let $K, T, B, b \in \mathbb{N}_+$ be any set of parameters such that $K \leq T$, and $1 \leq b \leq B < \log\log T$. Let $\{\mathcal{A}_j\}_{j=1}^{b+1}$ be a contiguous partition of arms $\mathcal{K}$ such that for each $j \in [b+1]$, $|\mathcal{A}_j| = A = K/(b+1)$. Furthermore, let $\{\psi_j\}_{j=1}^{b+1}$ be any set of distributions such that each ψ_j is γ-nearly uniform (see Definition 2) for $0 \leq \gamma \leq 1/(32b)$. Given a stream of arms $\mathcal{K}$ with mean rewards sampled according to $\mathcal{D}_{\mathcal{K},B}^{\{\psi_j\}_{j=1}^{b+1}}$, the expected regret R_T of any b-pass deterministic algorithm that uses at most $M = K(8b(b+1)\log e)^{-1}$ words of memory is bounded as*

$$\mathbb{E}[R_T] \geq \Omega\left(4^{-b} T^{\frac{2^B}{2^{B+1}-1}}\right).$$

The proof of our main result in Theorem 1 now follows easily from the above lemma by setting $b = B$. Note that even though the condition $B < \log\log T$ is not required in the proof of the above lemma, our lower bound becomes vacuous once $B \geq \log\log T$ as it becomes smaller than $\sqrt{T}$.

Proof (Sketch) We will prove this lemma using induction on the number of passes b. Let us consider a modified setting where the algorithm is allowed additional power: in every epoch, which begins when the first arm of that epoch is read into memory and ends right before the first arm of the next epoch is read into memory, the algorithm is allowed to store all arms of that epoch in memory. However, the algorithm may retain in memory at most (w.l.o.g, we can assume exactly) a β fraction of the arms from that epoch (in addition to the arms stored from previous epochs). This cannot hurt the regret as we are only allowing more memory, which can always be ignored.

Base Case ($b = 1$). Let ϵ be some small constant, and let σ be the number of trials in the first epoch.

Case 1. $[\mathbb{E}[\sigma] \geq \epsilon^2/(6\Delta_1^2)]$, with the expectation taken over the observations made by the algorithm.

In this case, observe that the entropy of the random variable $\mathbb{H}(\mathsf{Y}_2) \geq \mathbb{H}(\mathsf{Y}_2|\mathsf{I}_2) \geq \log 2 - \gamma$. Therefore, Y_2 is distributed as a Bernoulli $\mathcal{B}(p)$ with parameter p such that $|p - \frac{1}{2}| \leq \sqrt{5\ln(4)\gamma/16} = \sqrt{(5\ln 4)/2}/16$, which follows from Lemma 7 (See Appendix A) and the fact that $\gamma < 1/32$. Therefore, we have that the mean reward of i_2^* will realize to its high value with constant probability, which gives us that

$$\mathbb{E}[R_T] \geq \Omega(\Delta_1^{-2}\Delta_2) \geq \Omega(T^\alpha).$$

Case 2. $[\mathbb{E}[\sigma] < \epsilon^2/(6\Delta_1^2)]$.

Let $\mathcal{S}_\sigma$ be the outcomes observed by the algorithm over the arms sampled in epoch 1. Then by Lemma 3, we have that after observing the outcomes $\mathcal{S}_\sigma$, with constant probability, the best arm i_1^*

will be discarded by the algorithm (i.e. $\mathsf{I}_1 \notin \mathcal{M}$ where $\mathcal{M}$ is the set of arms from epoch 1 that are retained in memory by the algorithm), and the entropy of the posterior distribution

$$\mathbb{H}(\mathsf{Y}_1|\mathcal{S}_\sigma, \mathsf{I}_1, \mathsf{I}_1 \notin \mathcal{M}) \geq \log 2 - \frac{\gamma + \epsilon}{1 - (\log(1+\beta) + \gamma + \epsilon)} \geq \log 2 - c,$$

where c is a small constant. Furthermore, due to the fact that $\mathbb{E}(\sigma) < \epsilon^2/(6\Delta_1^2)$, a simple Markov's argument implies that the actual number of trials $\sigma < 1/\Delta_1^2 = o(T)$ in epoch 1 with probability at least $1 - \epsilon^2/6$, which is a large constant for a sufficiently small ϵ. Therefore, there are at least $T - 1/\Delta_1 = T - o(T) = \Omega(T)$ trials left in epoch 2. In this case, the algorithm will suffer a large regret when the best arm i_1^* in epoch 1 has a large mean reward, i.e. Y_1 realizes to have value 1, and the best arm i_2^* in epoch 2 realizes to have a low reward, i.e. Y_2 realizes to have value 0.

Observe that in the posterior distribution of the rewards of arms in epoch 1, the entropy in the reward of the best arm in the first epoch $\mathbb{H}(\mathsf{Y}_1|\mathcal{S}_\sigma, \mathsf{I}_1 \notin \mathcal{M}) \geq \mathbb{H}(\mathsf{Y}_1|\mathcal{S}_\sigma, \mathsf{I}_1, \mathsf{I}_1 \notin \mathcal{M}) \geq \log 2 - c$. Therefore, by Lemma 7, the posterior distribution of Y_1 is Bernoulli with parameter $p \geq 1/2 - \sqrt{(5c\ln 4)/16}$, which is a constant bounded away from 0. This implies that the best arm in the first epoch had realized a high mean reward with constant probability. Similarly, we have that in the prior distribution of the rewards of arms in epoch 2, the reward of the best arm in the second epoch $\mathbb{H}(\mathsf{Y}_2) \geq \mathbb{H}(\mathsf{Y}_2|\mathsf{I}_2) \geq \log 2 - \gamma$, and therefore, we have that the distribution of Y_2 is Bernoulli with parameter p such that $|p - \frac{1}{2}| \leq \sqrt{5\ln(4)\gamma/16} = \sqrt{(5\ln 4)/2}/16$. Therefore, the best arm in the second epoch realizes to have low reward with constant probability. This gives us that

$$\mathbb{E}[R_T] \geq \Omega((T - o(T))\Delta_1) = \Omega\left(T^\alpha\right).$$

Therefore, the expected regret is $\Omega(T^\alpha)$ in both cases, which proves the base case.

Induction Step: Let us assume our claim holds for any number of passes up to $b-1$, then we will show it also holds for b passes. Suppose for the sake of contradiction that the claim is not true for b, i.e. there exists a b-pass algorithm ALG with memory at most $K(8b(b+1)\log e)^{-1}$ whose expected regret over instances drawn from the distribution $\mathcal{D}_{\mathcal{K},B}^{\{\psi_j\}_{j=1}^{b+1}}$ is $o\left(T^\alpha 4^{-b}\right)$.

The general outline will be to show that if the algorithm ends any epoch after performing *sufficient exploration*, then it will incur a large regret in the case that arm i_{b+1}^* realizes to a large mean reward, whereas if the algorithm ends all epochs with *insufficient exploration*, then the algorithm will not just discard all special arms, but also the instance induced over the discarded arms will remain hard. Supposing the algorithm achieves low expected regret over this instance in $b-1$ passes, it would contradict our induction hypothesis.

Let $\epsilon = c_\epsilon/b^2$ for some sufficiently small constant c_ϵ. For any epoch $j \in [b+1]$, let t_j be the trial when epoch j begins, let $\mathcal{T}_j := \{t_j, t_j+1, \cdots, t_{j+1}-1\}$ be the trials that belong to epoch j, and let $\sigma_j = |\mathcal{T}_j|$ denote the number of trials in epoch j. Lastly, let $\mathcal{S}_{\sigma_j} = \cup_{r=1}^{j}\{(i_t, r_t)\}_{t\in\mathcal{T}_r}$ be the sequence of observations defining the execution history of the algorithm until the end of epoch j with i_t being the arm pulled and r_t being the reward realized in trial t, respectively.

Case 1. $\left[\mathbb{E}(\sigma_j) \geq \epsilon^2/(6\Delta_j^2) \text{ for some } j \in [b]\right]$, where the expectation is over the realization of rewards until epoch j.

In this case, observe that the entropy of the random variable $\mathbb{H}(\mathsf{Y}_{B+1}) \geq \mathbb{H}(\mathsf{Y}_{b+1}|\mathsf{I}_{b+1}) \geq \log 2 - \gamma$. Therefore, by Lemma 7, we have that Y_{b+1} is distributed as a Bernoulli with parameter p such that

$|p - \frac{1}{2}| \leq \sqrt{5\ln(4)\gamma/16} < 1/4 - 1/6 - c/b^3$. The latter inequality follows from the fact that $\gamma \leq 1/(32b)$ with c being some small constant for $b \geq 2$. Therefore, we have that the mean reward of i^*_{b+1} will realize to its high value with constant probability, which gives us that

$$\mathbb{E}(R_T(\text{ALG})) \geq \Omega(L_j\Delta_{b+1}) = \Omega(\epsilon^2\Delta_j^{-2} \cdot \Delta_{b+1}) = \Omega(T^\alpha b^{-4}),$$

contradicting our assumption that ALG had an expected regret of $o(4^{-b}T^\alpha)$.

Case 2. $\left[\mathbb{E}(\sigma_j) < \epsilon^2/(6\Delta_j^2) \text{ for all } j \in [b]\right]$.

In this case, we will leverage Lemma 3 to show that ALG would not just discard the best arms from all epochs, but also the conditional distributions (conditioned on the observations seen by the algorithm) of the identities of the best arms I_j, and their rewards $\mathsf{Y}_j|\mathsf{I}_j$ would remain essentially uniform over the arms discarded from memory in their corresponding epochs.

Consider any epoch $j \in [b]$. Given the execution history $\mathcal{S}_{\sigma_j}$, let $\mathcal{M}_j \subset \mathcal{A}_j$ be the set of arms from epoch j that were retained in memory, and let $\mathcal{R}_j = \mathcal{A}_j \setminus \mathcal{M}_j$ be the set of $|\mathcal{R}_j| = R$ arms that were rejected after epoch j. Then by Lemma 3, with high probability, the entropy of the posteriors

$$\mathbb{H}(\mathsf{I}_j|\mathcal{S}_{\sigma_j}, \mathsf{I}_j \in \mathcal{R}_j) = \mathbb{H}(\mathsf{I}_j|\mathcal{S}_{\sigma_j}, \mathsf{I}_j \notin \mathcal{M}_j) \geq \log R - \frac{\gamma + \epsilon}{1 - (\log{(1+\beta)} + \gamma + \epsilon)} \geq \log R - \frac{1}{32(b-1)},$$

where the final inequality follows by a sufficiently small choice of the constant c_ϵ. Similarly,

$$\mathbb{H}(\mathsf{Y}_j|\mathcal{S}_{\sigma_j}, \mathsf{I}_j, \mathsf{I}_j \in \mathcal{R}_j) \geq \log 2 - \frac{1}{32(b-1).}$$

At this point, we further argue that this supposed low regret algorithm ALG cannot spend too many trials on the first b epochs prior to processing the $(b+1)^{th}$ epoch with a large probability. Since we have that $\mathbb{E}(\sum_{j\in[b]}\sigma_j) \leq \sum_{j\in[b]}\mathbb{E}(\sigma_j) < \sum_{j\in[b]} L_j = \epsilon^2\sum_{j\in[b]}(6\Delta_j^2)^{-1}$, by Markov's inequality, it must be that $\Pr(\sum_{j\in[b]}\sigma_j \geq \sum_{j\in[b]}\Delta_j^{-2}) \leq \epsilon^2/6 < c_\epsilon^2/(6b^4)$. We define this event $\mathcal{E}_0 := (\sum_{j\in[b]}\sigma_j < \sum_{j\in[b]}\Delta_j^{-2})$ that the actual number of trials spent by the algorithm in the first b epochs is small, and the above argument gives us that $\Pr(\neg\mathcal{E}_0) < c_\epsilon^2/(6b^4)$. Now for every $j \in [b]$, let us define the event

$$\mathcal{E}_j := \Big(\mathsf{I}_j \in \mathcal{R}_j \text{ and } \mathbb{H}(\mathsf{I}_j \mid \mathcal{S}_{\sigma_j}, \mathsf{I}_j \in \mathcal{R}_j) \geq \log R - \frac{1}{32(b-1)} \\ \text{and } \mathbb{H}(\mathsf{Y}_j \mid \mathcal{S}_{\sigma_j}, \mathsf{I}_j, \mathsf{I}_j \in \mathcal{R}_j) \geq \log 2 - \frac{1}{32(b-1)}\Big).$$

Using Lemma 3, we have that $\Pr(\neg\mathcal{E}_j) \leq \log{(1+\beta)} + \gamma + 3\epsilon \leq 1/(6b)$, which follows by a sufficiently small choice of the constant c_ϵ. Let us also define the event $\mathcal{E}_{b+1} := \{\mu_{i^*_{b+1}} = 1/2\}$, where the mean reward of the best arm in epoch $b+1$ realizes to a low value. Since we have that $\mathbb{H}(\mathsf{Y}_{b+1}) \geq \mathbb{H}(\mathsf{Y}_{b+1}|\mathsf{I}_{b+1}) \geq \log 2 - \gamma$, following an identical calculation as Case 1, we have that Y_{b+1} is distributed as a Bernoulli $\mathcal{B}(p)$ with parameter p such that $|p - 1/2| \leq 1/4 - 1/6 - c/b^3$ where c is some absolute constant. Therefore, we have that $\Pr(\neg\mathcal{E}_{b+1}) \leq 1/2 + (1/4 - 1/6 - c/b^3)$.

Therefore, by a union bound over all these $(b+2)$ events, we have $\Pr(\mathcal{E}_0 \cap_{j\in[b+1]} \mathcal{E}_j) \geq 1 - \Pr(\neg\mathcal{E}_0) - \sum_{j\in[b+1]} \Pr(\neg\mathcal{E}_j) = 1 - c_\epsilon^2/(6b^4) - (1/2 + (1/4 - 1/6 - c/b^3)) - b/(6b) \geq 1/4$ by choosing a sufficiently small $c_\epsilon < \sqrt{6bc}$. We define this to be event $\mathcal{E} := (\mathcal{E}_0 \cap_{j\in[b+1]} \mathcal{E}_j)$.

Lastly, we argue that under event $\mathcal{E}$, the algorithm must necessarily spend $o(T)$ trials in the last epoch $b+1$. This is because under event $\mathcal{E}$, $\mu^*_{b+1} = 1/2$, and furthermore, in the posterior distribution of the reward of the best arm $\mu_{i_b^*}$ of the b^{th} epoch in the rejected set $\mathcal{R}_b$ is at least $\mathbb{H}(\mathsf{Y}_b|\mathcal{S}_{\sigma_b}, \mathsf{I}_b \notin \mathcal{M}_b) \geq \mathbb{H}(\mathsf{Y}_b|\mathcal{S}_{\sigma_b}, \mathsf{I}_b, \mathsf{I}_b \notin \mathcal{M}_b) \geq \log 2 - 1/(32(b-1))$. Therefore, by Lemma 7, the posterior distribution of Y_b is Bernoulli with parameter $p > 1/2 - 1/8$, which is the probability with which the rejected best arm in the b^{th} epoch actually had a large reward. Therefore, if the algorithm spends $\Omega(T)$ trials in the $(b+1)^{th}$ epoch, then the expected regret of the algorithm

$$\mathbb{E}[R_T(\text{ALG})] \geq \Pr(\mathcal{E})\,\mathbb{E}[R_T(\text{ALG})|\mathcal{E}] \geq \Omega(T)\cdot\Delta_b = \Omega\left(T^\alpha\right),$$

contradicting our assumption about the expected regret achieved by the algorithm. Therefore, under event $\mathcal{E}$, we have that the total number of trials spent by the algorithm in the first pass is $\sum_{j\in[b+1]} \sigma_j = o(T) + \sum_{j\in[b]} T^{1-\frac{2^j-1}{2^{B+1}-1}} = o(T)$, implying the number of trials T_- leftover is $T - o(T)$. Moreover, there are only $b-1$ passes left.

Now we shall use our assumption about the expected regret achievable by our algorithm to show that in order to achieve low expected regret overall, it must necessarily achieve low expected regret in the remaining passes too. Let $\mathbb{E}[R_{T_-}(\text{ALG})]$ denote the cumulative regret of the algorithm over the remaining $b-1$ passes. Therefore, we have

$$\mathbb{E}[R_T(\text{ALG})] \geq \mathbb{E}[R_{T_-}(\text{ALG}) \mid \mathcal{E}]\cdot\Pr(\mathcal{E}) = \frac{1}{4}\cdot\mathbb{E}[R_{T_-}(\text{ALG}) \mid \mathcal{E}]\,.$$

Therefore, $\mathbb{E}[R_{T_-}(\text{ALG}) \mid \mathcal{E}] \leq 4\,\mathbb{E}[R_T(\text{ALG})] = o\left(4\cdot T^\alpha\cdot 4^{-b}\right) = o\left(T^\alpha\cdot 4^{-(b-1)}\right)$. We shall use this fact to set up a contradiction to our induction hypothesis, which says that any $(b-1)$-pass algorithm with small memory must incur large regret. We begin by setting up the hard distribution.

In our new instance, we begin by discarding the arms $\mathcal{A}_{b+1}$ as under event $\mathcal{E}$, the mean reward of the best arm (and hence all arms) in this epoch $(b+1)$ has realized to a low value. Our new instance consists of all the rejected arms $\mathcal{R}_j$ for $j \in [b]$, the first b epochs. We refer to these set of arms as $\mathcal{K}' = \cup_{j\in[b]}\mathcal{R}_j$, whose size is exactly $|\mathcal{K}'| = K' = \sum_{j\in[b]}(1-\beta)K/(b+1) = (1-\beta)bK/(b+1)$.

We claim that the posterior distributions over $\mathsf{I}_j, \mathsf{Y}_j$ in $\mathcal{R}_j$ for $j \in [b]$ give us a hard distribution over arms $\mathcal{K}'$ for any $(b-1)$ pass algorithm. For $j \in [b]$, let ϕ_j be the joint distribution $\text{dist}((\mathsf{I}_j, \mathsf{Y}_j) \mid \mathsf{I}_j \in \mathcal{R}_j)$. Its easy to verify that $\mathcal{D}^{\{\phi_j\}_{j=1}^b}_{\mathcal{K}',B}$ is a hard distribution for any $(b-1)$ pass algorithm: the partitions $\{\mathcal{R}_j\}$ of $\mathcal{K}'$ are all of equal size $|\mathcal{K}'|/b$, and for any $j \in [b]$, random variables $(\mathsf{I}_j, \mathsf{Y}_j) \sim \phi_j$ satisfy the high entropy conditions $\mathbb{H}(\mathsf{I}_j) \geq \log R - \gamma_-$, and $\mathbb{H}(\mathsf{Y}_j|\mathsf{I}_j) \geq \log 2 - \gamma_-$, where $\gamma_- \leq 1/(32(b-1))$ (as indicated by event $\mathcal{E}$). Furthermore, the memory budget M' for a $(b-1)$ pass algorithm for this instance is

$$M' = \frac{K'}{8b(b-1)\log e} > \frac{\left(1-\frac{1}{b}\right)bK}{b+1}\cdot\frac{1}{8b(b-1)\log e} = \frac{K}{8b(b+1)\log e} = M,$$

which is in fact larger than the memory used by ALG. We shall show that we can use ALG in the subsequent $b-1$ passes under event $\mathcal{E}$ to construct a $(b-1)$-pass algorithm ALG_{b-1} that uses memory at most M' that over a time horizon of length $T - o(T)$, achieves regret $o(T^\alpha \cdot 4^{-(b-1)})$ in expectation over $\mathcal{D}_{\mathcal{K}',B}^{\{\phi_j\}_{j=1}^b}$, contradicting our induction hypothesis.

Let ALG_- denote the algorithm ALG for the remaining $b-1$ passes when event $\mathcal{E}$ occurs, then ALG_{b-1} is constructed as follows: if ALG_- pulls an arm in $\cup_{j\in[b]} \text{supp}(\phi_j)$, ALG_{b-1} pulls the corresponding arm in $\mathcal{K}'$ and returns the realized reward of the arm to ALG_-; otherwise, ALG_{b-1} simply samples from $\mathcal{B}(1/2)$ and returns the result to ALG_-. It is trivially true that $\mathbb{E}[R_{T_-}(\text{ALG}_{b-1})] \leq \mathbb{E}[R_{T_-}(\text{ALG}_-) \mid \mathcal{E}]$. This is because, given event $\mathcal{E}$, any other arm than the arms in $\cup_{j\in[b]}\text{supp}(\phi_j)$ is distributed as $\mathcal{B}(1/2)$. Hence, ALG_{b-1} is a $(b-1)$-pass algorithm with memory at most $K'(8b(b-1)\log e)^{-1}$ that achieves regret $o(T^\alpha \cdot 4^{-(b-1)})$ over a time horizon $T - o(T)$, which is a contradiction! This completes the proof of our lower bound.

■

In the following section, we present our algorithmic results for this problem. Specifically, we design an algorithm that achieves a regret of $\widetilde{O}\left(T^{\frac{2^B}{2^{B+1}-1}}\sqrt{KB}\right)$ in B passes given even just constant arm memory. Furthermore, our algorithm is able to achieve this regret, not just in expectation, but also with any polynomially high probability. This regret guarantee nearly matches the above lower bound, proving our above lower bound is nearly tight.

4. Limited Memory Multi-Pass Algorithms for Streaming Bandits

In this section, we present our worst-case and instance-dependent regret upper bounds for limited memory multi-pass streaming bandits. The following theorem characterizes our algorithmic results.

Theorem 5 (B-Pass Upper Bound) *Given a time horizon T, a stream of K arms, and number of passes $1 \leq B < \log\log T$, there exists a B-pass algorithm that uses $O(1)$ words of memory, and with probability $1 - 1/\text{poly}(T)$, achieves cumulative regret*

$$R_T \leq O\left(T^{\frac{2^B}{2^{B+1}-1}}\sqrt{KB\log T}\right).$$

Furthermore, supposing the arms $\mathcal{K}$ had mean rewards $\{\mu_j^*\}_{j\in\mathcal{K}}$, then given number of passes $1 \leq B < \log T$, there exists a B-pass algorithm that uses $O(1)$ words of memory, and with probability $1 - 1/\text{poly}(T)$, achieves a cumulative regret

$$R_T \leq O\left(\sum_{j\in\mathcal{S}} \frac{T^{1/(B+1)}\log T + B\log\left((\Delta_j^*)^2 T/\log T\right)}{\Delta_j^*}\right),$$

where $\mathcal{S} \subset \mathcal{K}$ is the set of strictly sub-optimal arms in $\mathcal{K}$, and for any sub-optimal arm $j \in \mathcal{S}$, $\Delta_j^* := \max_{\{i\in\mathcal{K}\}} \mu_i^* - \mu_j^*$ is the regret due to playing arm j.

Note that no assumptions are made about the stream order, and that these regret guarantees hold even when the order of arms is allowed to change (potentially adversarially) across rounds.

In the constant pass regime, the worst-case regret achievable matches our lower bound up to just a $\sqrt{K \log T}$ factor, implying our results are essentially tight for this regime. Our result further implies one can achieve a worst-case regret of $O(\sqrt{KT \log T \cdot \log \log T})$ in just $\log \log T$ passes over the stream, which matches the optimal regret achievable by even an unbounded memory algorithm up to a $\sqrt{\log T \log \log T}$ factor. With regards to instance-dependent regret, the picture is slightly different where we need $\log T$ passes (though still sublinear) over the stream to achieve regret $O(\sum_{j \in \mathcal{S}} (\Delta_j^*)^{-1} \log^2 T)$, which matches the instance-optimal regret achievable by an unbounded memory algorithm up to a $\log T$ factor.

Moreover, observe that our upper bound has no dependence on the available memory; the aforementioned regret guarantees can be achieved with even just constant memory. This effectively demonstrates a *sharp threshold* in the regret-memory tradeoff. In order to achieve near-optimal regret, one necessarily needs either linear memory, or a superconstant number of passes. Moreover, for any fixed number of passes B, there is no asymptotic difference between having $O(1)$ memory vs allowing a much larger $O(K/B^2)$ memory for minimizing regret. In the interest of space, we defer the algorithm description and its analysis to the Appendix, and can be found in section D.

5. Discussion and Conclusion

We studied the stochastic K-armed bandits problem in a limited memory, multi-pass streaming setting, where we study the interplay between the available memory M, the number of passes B, and the regret R_T over a time horizon T. We showed that any B-pass algorithm with memory $o(K/B^2)$ must necessarily incur $\Omega\left(4^{-B} T^{\frac{1}{2} + \frac{1}{2^{B+2}-2}}\right)$ regret in expectation. Moreover, we showed that it is possible to achieve $\widetilde{O}(T^{\frac{1}{2} + \frac{1}{2^{B+2}-2}} \sqrt{KB})$ regret with any polynomially large probability given B passes and just $O(1)$ memory. These results uncover a surprising phenomenon: increasing the memory beyond $O(1)$ memory to any quantity that is $o(K/B^2)$ has almost no effect on reducing the expected worst-case regret.

Our work highlights some interesting directions for future work. First, while our results are essentially tight for constant-pass algorithms, there is a gap of $1/2^B$ between our upper and lower bound on the regret when B is a superconstant. Second, it might also be worth exploring the regret landscape in the memory range of $\Omega(K/B^2)$ to $K-1$, for superconstant B. Finally, what is the best instance-dependent regret one can achieve in this limited-memory multi-pass streaming setting? Our work establishes an instance-dependent regret upper bound of $\widetilde{O}((T^{1/(B+1)} + B) \sum_{i \in \mathcal{S}} 1/\Delta_i^*)$, but leaves open the question of a matching lower bound.

Acknowledgments

This work was supported in part by NSF awards CCF-1763514, CCF-1934876, and CCF-2008304.

References

Arpit Agarwal, Sanjeev Khanna, and Prathamesh Patil. A sharp memory-regret trade-off for multi-pass streaming bandits, 2022. URL https://arxiv.org/abs/2205.00984.

Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *J. Comput. Syst. Sci.*, 58(1):137–147, 1999.

Sepehr Assadi and Chen Wang. Exploration with limited memory: streaming algorithms for coin tossing, noisy comparisons, and multi-armed bandits. In Konstantin Makarychev, Yury Makarychev, Madhur Tulsiani, Gautam Kamath, and Julia Chuzhoy, editors, *Proccedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, STOC 2020, Chicago, IL, USA, June 22-26, 2020*, pages 1237–1250. ACM, 2020.

Sébastien Bubeck, Tengyao Wang, and Nitin Viswanathan. Multiple identifications in multi-armed bandits. In *ICML*, 2013.

Arghya Roy Chaudhuri and Shivaram Kalyanakrishnan. Regret minimisation in multi-armed bandits using bounded arm memory. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 10085–10092. AAAI Press, 2020.

Thomas M. Cover and Joy A. Thomas. *Elements of information theory (2. ed.)*. Wiley, 2006. ISBN 978-0-471-24195-9.

Zijun Gao, Yanjun Han, Zhimei Ren, and Zhengqing Zhou. Batched multi-armed bandits problem. *Advances in Neural Information Processing Systems*, 32, 2019.

Alison L Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International statistical review*, 70(3):419–435, 2002.

Tianyuan Jin, Keke Huang, Jing Tang, and Xiaokui Xiao. Optimal streaming algorithms for multi-armed bandits. In *International Conference on Machine Learning*, pages 5045–5054. PMLR, 2021.

David Liau, Zhao Song, Eric Price, and Ger Yang. Stochastic multi-armed bandits in constant space. In Amos J. Storkey and Fernando Pérez-Cruz, editors, *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*, volume 84 of *Proceedings of Machine Learning Research*, pages 386–394. PMLR, 2018.

Arnab Maiti, Vishakha Patil, and Arindam Khan. Streaming algorithms for stochastic multi-armed bandits. In *NeurIPS*, 2021.

Vianney Perchet, Philippe Rigollet, Sylvain Chassang, Erik Snowberg, and Snowberg@caltech Edu. Batched Bandit Problems. In *COLT*, 2015.

Santanu Rathod. On reducing the order of arm-passes bandit streaming algorithms under memory bottleneck, 2021.

Aleksandrs Slivkins. Introduction to multi-armed bandits. *Found. Trends Mach. Learn.*, 12(1-2): 1–286, 2019.

Vaidehi Srinivas, David P Woodruff, Ziyu Xu, and Samson Zhou. Memory bounds for the experts problem. *arXiv preprint arXiv:2204.09837*, 2022.

Andrew Chi-Chin Yao. Probabilistic computations: Toward a unified measure of complexity. In *18th Annual Symposium on Foundations of Computer Science (sfcs 1977)*, pages 222–227. IEEE Computer Society, 1977.

Appendix A. Information-Theoretic Preliminaries

In this section, we record some basic facts about entropy and mutual information that are used in in this paper. The proofs can be found in Cover and Thomas (2006), Chapter 2. We also prove two crucial lemmas in this section, the first which highlights the difficulty of narrowing down the realization of a high entropy random variable to a small set of possibilities, and the second which bounds the parameter of a high entropy Bernoulli random variable.

Fact 1 *Let* A, B, *and* C *be three (possibly correlated) random variables.*

1. $0 \leq \mathbb{H}(\mathsf{A}) \leq \log |\mathsf{A}|$, *and* $\mathbb{H}(\mathsf{A}) = \log |\mathsf{A}|$ *iff* A *is uniformly distributed over its support.*
2. $\mathbb{I}(\mathsf{A}\,;\mathsf{B} \mid \mathsf{C}) \geq 0$. *The equality holds iff* A *and* B *are* independent *conditioned on* C.
3. *Conditioning can only drop the entropy:* $\mathbb{H}(\mathsf{A} \mid \mathsf{B}, \mathsf{C}) \leq \mathbb{H}(\mathsf{A} \mid \mathsf{B})$. *The equality holds iff* $\mathsf{A} \perp \mathsf{C} \mid \mathsf{B}$.
4. *Chain rule of mutual information:* $\mathbb{I}(\mathsf{A}, \mathsf{B}\,; \mathsf{C}) = \mathbb{I}(\mathsf{A}\,; \mathsf{C}) + \mathbb{I}(\mathsf{B}\,; \mathsf{C} \mid \mathsf{A})$.

For two distributions ϕ and ψ over the same probability space, the *Kullback-Leibler divergence* between ϕ and ψ is defined as $\mathbb{D}(\phi \mid\mid \psi) := \mathbb{E}_{\mathsf{A}\sim\phi}\left[\log \frac{\Pr_{\phi}(\mathsf{A})}{\Pr_{\psi}(\mathsf{A})}\right]$. For our proofs, we need the following relation between mutual information and KL-divergence.

Fact 2 *For random variables* $\mathsf{A}, \mathsf{B}, \mathsf{C}$,

$$\mathbb{I}(\mathsf{A}\,; \mathsf{B} \mid \mathsf{C}) = \mathop{\mathbb{E}}_{(b,c)\sim \text{dist}(\mathsf{B},\mathsf{C})} \Big[\mathbb{D}(\text{dist}(\mathsf{A} \mid \mathsf{C} = c) \mid\mid \text{dist}(\mathsf{A} \mid \mathsf{B} = b, \mathsf{C} = c))\Big].$$

The following fact can be proven by bounding the KL-divergence by χ^2-distance (see, e.g., Gibbs and Su (2002), Theorem 5).

Fact 3 *For any two parameters* $0 < p, q < 1$,

$$\mathbb{D}(\mathcal{B}(p) \mid\mid \mathcal{B}(q)) \leq \frac{(p-q)^2}{q \cdot (1-q)}$$

The following lemma outlines the difficulty in narrowing down the realization of a high-entropy random variable to a small number of possibilities.

Lemma 6 *Let* A *be a random variable supported over a set of size* A *with entropy* $\mathbb{H}(\mathsf{A}) \geq \log A - \gamma$ *for some* $\gamma \geq 0$. *Then for any set* $\mathcal{S}$ *of size* $|\mathcal{S}| = \beta A$ *for any* $\beta < 1$, *we have*

$$\Pr(\mathsf{A} \in \mathcal{S}) \leq \log(1+\beta) + \gamma \,.$$

Proof Suppose for the sake of contradiction, there exists a set $\mathcal{S}$ of size $|\mathcal{S}| = \beta A$ for some $\beta < 1$ such that $\Pr(\mathsf{A} \in \mathcal{S}) = \sum_{i \in \mathcal{S}} \Pr(\mathsf{A} = i) = \gamma' > \log(1+\beta) + \gamma$. Let $p_i = \Pr(\mathsf{A} = i)$. Then we have that:

$$\begin{aligned}
\mathbb{H}(\mathsf{A}) &= \sum_{i \in \mathcal{A}} p_i \log \frac{1}{p_i} \\
&= \sum_{i \in \mathcal{S}} p_i \log \frac{1}{p_i} + \sum_{i \notin \mathcal{S}} p_i \log \frac{1}{p_i} \\
&= \gamma' \sum_{i \in \mathcal{S}} \frac{p_i}{\gamma'} \log \frac{1}{p_i} + (1-\gamma') \sum_{i \notin \mathcal{S}} \frac{p_i}{(1-\gamma')} \log \frac{1}{p_i} \\
&\overset{(a)}{\leq} \gamma' \log \left(\sum_{i \in \mathcal{S}} \frac{p_i}{\gamma'} \cdot \frac{1}{p_i} \right) + (1-\gamma') \log \left(\sum_{i \notin \mathcal{S}} \frac{p_i}{(1-\gamma')} \cdot \frac{1}{p_i} \right) \\
&= \gamma' \log \frac{\beta A}{\gamma'} + (1-\gamma') \log \frac{(1-\beta) A}{(1-\gamma')} \\
&= \log A - \gamma' \underbrace{- \gamma' \log \frac{\gamma'}{2\beta} + (1-\gamma') \log \frac{1-\beta}{1-\gamma'}}_{f(\gamma')} \,,
\end{aligned}$$

where equation (a) follows by the Jensen's inequality, as the two summations are expectations over the concave $\log$ function over the set $\mathcal{S}$ and the set $\text{supp}(\mathsf{A}) \setminus \mathcal{S}$, respectively. One can verify that the function $f(\gamma')$ is concave, and is maximized at $\gamma' = 2\beta/(1+\beta)$ achieving a value of $\log(1+\beta)$. Therefore, we have that $-\gamma' + f(\gamma') \leq -\gamma' + \max_{\gamma'} f(\gamma') < -\gamma$ by choice of $\gamma' > \log(1+\beta) + \gamma$, giving us

$$\mathbb{H}(\mathsf{A}) < \log A - \gamma,$$

contradicting our initial assumption about the entropy of A. ■

Lastly, following lemma bounds the parameter of a high entropy Bernoulli random variable.

Lemma 7 *Given a Bernoulli random variable* $\mathsf{Y} \sim \mathcal{B}(p)$ *with entropy* $\mathbb{H}(\mathsf{Y}) \geq 1 - \gamma$ *for any* $\gamma \leq 1/4$, *then we have that*

$$\left| p - \frac{1}{2} \right| \leq \sqrt{\frac{5\gamma \ln 4}{16}}$$

Proof Suppose for the sake of contradiction, there exists a parameter $p := 1/2 + \Delta$ such that $\Delta > \sqrt{\frac{5\gamma \ln 4}{16}}$, and for $\mathsf{Y} \sim \mathcal{B}(p)$, the entropy $\mathbb{H}(\mathsf{Y}) \geq 1 - \gamma$. Then we have

$$\begin{aligned}
\mathbb{H}(\mathsf{Y}) &\leq (4p(1-p))^{1/\ln 4} \\
&= (4(1/2 - \Delta)(1/2 + \Delta))^{1/\ln 4} \\
&= (1 - 4\Delta^2)^{1/\ln 4} \\
&\leq \exp\left(-4\Delta^2/\ln 4\right) \\
&< \exp(-5\gamma/4) \\
&< 1 - \frac{5\gamma}{4} + \frac{25\gamma^2}{32} \\
&< 1 - \gamma,
\end{aligned}$$

where the final inequality follows by the fact that $\gamma \leq 1/4$, and thus $25\gamma^2/32 < \gamma/4$. This contradicts the assumption that $\mathbb{H}(\mathsf{Y}) \geq 1 - \gamma$. ■

Appendix B. Proof of Lemma 3

In this section, we shall prove Lemma 3, which is restated here for convenience.

Lemma 3 *Let T be the time horizon. Given any set of arms $\mathcal{A}$ of size $|\mathcal{A}| = A$, let $\mathcal{D}_{\mathcal{A}}^{\Delta,\psi}$ be the MAB distribution over instances defined in 3.1. Let $(\mathsf{I}, \mathsf{Y}) \sim \psi$, where I represents the random variable for the index of arm $i^* \in \mathcal{A}$, and Y represents the random variable that controls the mean reward μ_{i^*} of the best arm. Furthermore, let these be high-entropy random variables such that $\mathbb{H}(\mathsf{I}) \geq \log A - \gamma$ and $\mathbb{H}(\mathsf{Y}|\mathsf{I}) \geq \log 2 - \gamma$ for some parameter $\gamma > 0$. Let* ALG *be any deterministic multi-armed bandit algorithm that* adaptively *pulls arms in $\mathcal{A}$. Let $\sigma \in [T]$ be the* randomly chosen *stopping time of the algorithm and $\mathcal{S}_\sigma = (j_t, r_t)_{t=1}^{\sigma}$ be history of execution of* ALG *with j_t being the arm pulled and r_t being its observed reward in trial t, respectively. For a given input parameter $\beta < 1$, let $\mathcal{M} \subset \mathcal{A}$ be any set of size βA chosen to be retained in memory by* ALG *after observing the execution history $\mathcal{S}_\sigma$. If $\mathbb{E}[\sigma] \leq \frac{\epsilon^2}{6\Delta^2}$ for some $\epsilon > 0$, then the event*

$$\mathcal{E} = \Big(\mathsf{I} \notin \mathcal{M} \text{ and } \mathbb{H}(\mathsf{I} \mid \mathcal{S}_\sigma, \mathsf{I} \notin \mathcal{M}) \geq \log((1-\beta)A) - \frac{\gamma + \epsilon}{1 - \log(1+\beta) - \gamma - \epsilon}$$
$$\text{and } \mathbb{H}(\mathsf{Y} \mid \mathcal{S}_\sigma, \mathsf{I}, \mathsf{I} \notin \mathcal{M}) \geq \log 2 - \frac{\gamma + \epsilon}{1 - \log(1+\beta) - \gamma - \epsilon}\Big),$$

occurs with probability at least $1 - \log(1+\beta) - \gamma - 3\epsilon$ over the realizations of the arms sampled by the algorithm.

Proof Let $L = \frac{\epsilon^2}{6\Delta^2}$, J_s be the random variable for arm j_s pulled in trial s, and R_s be the random variable for the reward r_s observed in trial s, for $s \in [\sigma]$. For ease of calculation, we will expand the execution history beyond its stopping time σ and let $\mathsf{J}_s = 0$ and $\mathsf{R}_s = \frac{1}{2}$ for $s \in \{\sigma + 1, \cdots, T\}$. Finally for any trial $t \in [T]$, let $\mathsf{S}_t := \{(\mathsf{J}_s, \mathsf{R}_s)\}_{s \in [t]}$ be the sequence of random variable defining the execution history of the algorithm up until trial t, and let $\mathcal{S}_t$ be the realization of this sequence up until trial t.

We will begin by showing that in the event of insufficient exploration (i.e. when the algorithm stops quickly), little is learned about the identity of the best arm. In other words, the mutual information between the random variables I and S_T is small. Using the chain rule for mutual information, we have

$$\begin{aligned}
\mathbb{I}(\mathsf{I}\,;\mathsf{S}_T) &= \sum_{t=1}^{T} \mathbb{I}(\mathsf{I}\,;\mathsf{J}_t|\mathsf{S}_{t-1}) + \mathbb{I}(\mathsf{I}\,;\mathsf{R}_t|\mathsf{S}_{t-1},\mathsf{J}_t) \\
&= \sum_{t=1}^{T} 0 + \mathbb{I}(\mathsf{I}\,;\mathsf{R}_t|\mathsf{S}_{t-1},\mathsf{J}_t) && (\mathsf{J}_t \text{ is deterministic given } \mathsf{S}_{t-1}) \\
&= \sum_{t=1}^{T}\sum_{j\in\mathcal{A}}\sum_{\mathcal{S}_{t-1}} \Pr(\mathsf{S}_{t-1}=\mathcal{S}_{t-1},\mathsf{J}_t=j)\cdot \mathbb{I}(\mathsf{I}\,;\mathsf{R}_t|\mathsf{S}_{t-1}=\mathcal{S}_{t-1},\mathsf{J}_t=j). && (1)
\end{aligned}$$

Using Fact 2, we have

$$\mathbb{I}(\mathsf{I}\,;\mathsf{R}_t|\mathsf{S}_{t-1}=\mathcal{S}_{t-1},\mathsf{J}_t=j) = \mathop{\mathbb{E}}_{(i^*,y)\sim\psi}[\mathbb{D}(\mathrm{dist}(\mathsf{R}_t \mid \mathsf{S}_{t-1}=\mathcal{S}_{t-1},\mathsf{J}_t=j) \,||\, \mathrm{dist}(\mathsf{R}_t \mid \mathsf{I}=i^*,\mathsf{S}_{t-1}=\mathcal{S}_{t-1},\mathsf{J}_t=j))]$$

We will now prove that the average KL-divergence between the reward distributions for a single pull of an arm under different realizations of instances sampled from our hard distribution $\mathcal{D}_{\mathcal{A}}^{\Delta,\psi}$ is small. Therefore, a single pull of any arm can only provide limited information about the random variables of interest, and therefore, the total information that can be gathered from a small number of pulls is also small.

Claim 1 *For any arms $i^*, j \in \mathcal{A}$, trial $t \in [T]$, and any realization $\mathcal{S}_{t-1}$ of the execution history up until trial t, we have that*

$$\mathbb{D}(\mathrm{dist}(\mathsf{R}_t \mid \mathsf{S}_{t-1}=\mathcal{S}_{t-1},\mathsf{J}_t=j) \,||\, \mathrm{dist}(\mathsf{R}_t \mid \mathsf{I}=i^*,\mathsf{S}_{t-1}=\mathcal{S}_{t-1},\mathsf{J}_t=j)) \le 6\Delta^2\,.$$

Proof Let $p = \Pr(\mathsf{I}=j \mid \mathsf{S}_{t-1}=\mathcal{S}_{t-1},\mathsf{J}_t=j)$, and let $q=\Pr(\mathsf{Y}=1 \mid \mathsf{I}=j,\mathsf{S}_{t-1}=\mathcal{S}_{t-1},\mathsf{J}_t=j)$.

In the case where $i^*=j$, it is easy to observe that $\mathrm{dist}(\mathsf{R}_t \mid \mathsf{S}_{t-1}=\mathcal{S}_{t-1},\mathsf{J}_t=j)=\mathcal{B}(\frac{1}{2}+pq\Delta)$. Moreover, we also have that $\mathrm{dist}(\mathsf{R}_t \mid \mathsf{I}=i^*,\mathsf{S}_{t-1}=\mathcal{S}_{t-1},\mathsf{J}_t=j)=\mathcal{B}(\frac{1}{2}+q\Delta)$. We then have that

$$\begin{aligned}
\mathbb{D}(\mathrm{dist}(\mathsf{R}_t \mid \mathsf{S}_{t-1}=\mathcal{S}_{t-1},\mathsf{J}_t=j) \,||\, \mathrm{dist}(\mathsf{R}_t \mid \mathsf{I}=i^*,\mathsf{S}_{t-1}=\mathcal{S}_{t-1},\mathsf{J}_t=j)) &= \mathbb{D}\left(\mathcal{B}\left(\frac{1}{2}+pq\Delta\right) \,||\, \mathcal{B}\left(\frac{1}{2}+q\Delta\right)\right) \\
&\le \frac{(\frac{1}{2}+pq\Delta-\frac{1}{2}-q\Delta)^2}{(\frac{1}{2}+q\Delta)(1-\frac{1}{2}-q\Delta)} \\
&\le \frac{q^2(1-p)^2\Delta^2}{\frac{1}{4}-q^2\Delta^2} \\
&\le \frac{16q^2(1-p)^2\Delta^2}{4-q^2} \le 6\Delta^2\,,
\end{aligned}$$

where the first inequality above follows from Fact 3, and the final inequality follows due to $\Delta \leq 1/4$.

In the case where $i^* \neq j$, we have that $\text{dist}(\mathsf{R}_t \mid \mathsf{S}_{t-1} = \mathcal{S}_{t-1}, \mathsf{J}_t = j) = \mathcal{B}(\frac{1}{2} + pq\Delta)$. However, $\text{dist}(\mathsf{R}_t \mid \mathsf{I} = i^*, \mathsf{S}_{t-1} = \mathcal{S}_{t-1}, \mathsf{J}_t = j) = \mathcal{B}(\frac{1}{2})$. Using the same argument as above

$$\begin{aligned}\mathbb{D}(\text{dist}(\mathsf{R}_t \mid \mathsf{S}_{t-1} = S_{t-1}, \mathsf{J}_t = j) \mid\mid \text{dist}(\mathsf{R}_t \mid \mathsf{I} = i^*, \mathsf{S}_{t-1} = S_{t-1}, \mathsf{J}_t = j)) &= \mathbb{D}\left(\mathcal{B}\left(\frac{1}{2} + pq\Delta\right) \mid\mid \mathcal{B}\left(\frac{1}{2}\right)\right) \\ &\leq \frac{(\frac{1}{2} + pq\Delta - \frac{1}{2})^2}{(\frac{1}{2})(1 - \frac{1}{2})} \\ &\leq 4\Delta^2 .\end{aligned}$$

■

Using Eq. (1) and Claim 1 we have that

$$\begin{aligned}\mathbb{I}(\mathsf{I}\,;\mathsf{S}_T) &\leq \sum_{t=1}^{T} \sum_{j \in \mathcal{A}} \sum_{\mathcal{S}_{t-1}} \Pr(\mathsf{S}_{t-1} = \mathcal{S}_{t-1}, \mathsf{J}_t = j) \cdot 6\Delta^2 \\ &= \sum_{t=1}^{T} \sum_{j \in \mathcal{A}} \Pr(\mathsf{J}_t = j) \cdot 6\Delta^2 \\ &= \sum_{t=1}^{T} \Pr(\mathsf{J}_t \neq 0) \cdot 6\Delta^2 \\ &= \mathbb{E}[\sigma] \cdot 6\Delta^2 \\ &\leq L \cdot 6\Delta^2 = \epsilon^2 .\end{aligned}$$

This implies that that the conditional entropy of I given S_T is at least

$$\mathbb{H}(\mathsf{I} \mid \mathsf{S}_T) = \mathbb{H}(\mathsf{I}) - \mathbb{I}(\mathsf{I}\,;\mathsf{S}_T) \geq \mathbb{H}(\mathsf{I}) - \epsilon^2$$

We shall use an analogous argument to bound the mutual information between Y and S_T conditioned on I. Using the chain rule for mutual information, we have

$$\begin{aligned}\mathbb{I}(\mathsf{Y}\,;\mathsf{S}_T \mid \mathsf{I}) &= \sum_{t=1}^{T} \mathbb{I}(\mathsf{Y}\,;\mathsf{J}_t | \mathsf{S}_{t-1}, \mathsf{I}) + \mathbb{I}(\mathsf{Y}\,;\mathsf{R}_t | \mathsf{S}_{t-1}, \mathsf{J}_t, \mathsf{I}) \\ &= \sum_{t=1}^{T} 0 + \mathbb{I}(\mathsf{Y}\,;\mathsf{R}_t | \mathsf{S}_{t-1}, \mathsf{J}_t, \mathsf{I}) && (\mathsf{J}_t \text{ is deterministic given } \mathsf{S}_{t-1}) \\ &= \sum_{t=1}^{T} \sum_{i \in \mathcal{A}} \sum_{j \in \mathcal{A}} \sum_{\mathcal{S}_{t-1}} \Pr(\mathsf{S}_{t-1} = \mathcal{S}_{t-1}, \mathsf{J}_t = j, \mathsf{I} = i) \cdot \mathbb{I}(\mathsf{Y}\,;\mathsf{R}_t | \mathsf{S}_{t-1} = \mathcal{S}_{t-1}, \mathsf{J}_t = j, \mathsf{I} = i)\end{aligned} \tag{2}$$

We now calculate an upper bound on $\mathbb{I}(\mathsf{Y}\,;\mathsf{R}_t | \mathsf{S}_{t-1} = S_{t-1}, \mathsf{J}_t = j, \mathsf{I} = i)$. Using Fact 2, we have

$$\begin{aligned}&\mathbb{I}(\mathsf{Y}\,;\mathsf{R}_t | \mathsf{S}_{t-1} = \mathcal{S}_{t-1}, \mathsf{J}_t = j, \mathsf{I} = i) \\ &\quad= \mathop{\mathbb{E}}_{y \sim \psi | \mathsf{I} = i} [\mathbb{D}(\text{dist}(\mathsf{R}_t \mid \mathsf{S}_{t-1} = \mathcal{S}_{t-1}, \mathsf{J}_t = j, \mathsf{I} = i) \mid\mid \text{dist}(\mathsf{R}_t \mid \mathsf{Y} = y, \mathsf{S}_{t-1} = \mathcal{S}_{t-1}, \mathsf{J}_t = j, \mathsf{I} = i, \mathsf{Y} = y))]\end{aligned}$$

We now have an analogous claim, bounding the KL-divergence between the reward profiles of a single pull of an arm.

Claim 2 *For any arms $i, j \in \mathcal{A}$, $y \in \{0, 1\}$, trial $t \in [T]$, and any realization $\mathcal{S}_{t-1}$ of the execution history up until trial t, we have that*

$$\mathbb{D}(\text{dist}(\mathsf{R}_t \mid \mathsf{S}_{t-1} = \mathcal{S}_{t-1}, \mathsf{J}_t = j, \mathsf{I} = i) \,||\, \text{dist}(\mathsf{R}_t \mid \mathsf{Y} = y, \mathsf{S}_{t-1} = \mathcal{S}_{t-1}, \mathsf{J}_t = j, \mathsf{I} = i, \mathsf{Y} = y)) \leq 6\Delta^2\,.$$

Proof We begin with the simple case, when $i \neq j$. In this case, it is easy to observe that for any realization of Y, both the reward distributions will be $\mathcal{B}(1/2)$, due to which the KL Divergence will be 0. In the case that $i = j$, let $q = \Pr(\mathsf{Y} = 1 \mid S_{t-1} = \mathcal{S}_{t-1}, \mathsf{J}_t = j, \mathsf{I} = i)$.

We will first prove this bound in the case that $y = 1$. It easy to observe that $\text{dist}(\mathsf{R}_t \mid \mathsf{S}_{t-1} = \mathcal{S}_{t-1}, \mathsf{J}_t = j, \mathsf{I} = i) = \mathcal{B}(1/2 + q\Delta)$. Moreover, we also have that $\text{dist}(\mathsf{R}_t \mid \mathsf{S}_{t-1} = \mathcal{S}_{t-1}, \mathsf{J}_t = j, \mathsf{I} = i, \mathsf{Y} = y) = \mathcal{B}(1/2 + \Delta)$. We then have that

$$\begin{aligned}
\mathbb{D}(\text{dist}(\mathsf{R}_t \mid \mathsf{S}_{t-1} = \mathcal{S}_{t-1}, \mathsf{J}_t = j, \mathsf{I} = i) \,||\, &\text{dist}(\mathsf{R}_t \mid \mathsf{Y} = 1, \mathsf{S}_{t-1} = \mathcal{S}_{t-1}, \mathsf{J}_t = j, \mathsf{I} = i, \mathsf{Y} = y)) \\
&= \mathbb{D}(\mathcal{B}(1/2 + q\Delta) \,||\, \mathcal{B}(1/2 + \Delta)) \\
&\leq \frac{(\frac{1}{2} + q\Delta - \frac{1}{2} - \Delta)^2}{(\frac{1}{2} + \Delta)(\frac{1}{2} - \Delta)} \\
&\leq \frac{(1-q)^2\Delta^2}{\frac{1}{4} - \Delta^2} \\
&\leq \frac{16(1-q)^2\Delta^2}{3} \leq 6\Delta^2\,,
\end{aligned}$$

where the first inequality above follows from Fact 3, and the final inequality follows due to $\Delta < 1/4$.

In the case that $y = 0$, we again have $\text{dist}(\mathsf{R}_t \mid \mathsf{S}_{t-1} = \mathcal{S}_{t-1}, \mathsf{J}_t = j, \mathsf{I} = i) = \mathcal{B}(1/2 + q\Delta)$. However, $\text{dist}(\mathsf{R}_t \mid \mathsf{S}_{t-1} = \mathcal{S}_{t-1}, \mathsf{J}_t = j, \mathsf{I} = i, \mathsf{Y} = y) = \mathcal{B}(1/2)$. Using the same argument as above

$$\begin{aligned}
\mathbb{D}(\text{dist}(\mathsf{R}_t \mid \mathsf{S}_{t-1} = \mathcal{S}_{t-1}, \mathsf{J}_t = j, \mathsf{I} = i) \,||\, &\text{dist}(\mathsf{R}_t \mid \mathsf{Y} = 1, \mathsf{S}_{t-1} = \mathcal{S}_{t-1}, \mathsf{J}_t = j, \mathsf{I} = i, \mathsf{Y} = y)) \\
&= \mathbb{D}(\mathcal{B}(1/2 + q\Delta) \,||\, \mathcal{B}(1/2)) \\
&\leq \frac{(\frac{1}{2} + q\Delta - \frac{1}{2})^2}{(\frac{1}{2})(1 - \frac{1}{2})} \\
&\leq 4\Delta^2\,.
\end{aligned}$$

■

Using Eq. (2) and Claim 2 we have that

$$\begin{aligned}
\mathbb{I}(\mathsf{Y}\,;\mathsf{S}_T|\mathsf{I}) &\le \sum_{t=1}^{T}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{A}}\sum_{\mathcal{S}_{t-1}} \Pr(\mathsf{S}_{t-1}=\mathcal{S}_{t-1},\mathsf{J}_t=j,\mathsf{I}=i)\cdot 6\Delta^2\\
&= \sum_{t=1}^{T}\sum_{j\in\mathcal{A}} \Pr(\mathsf{J}_t=j)\cdot 6\Delta^2\\
&= \sum_{t=1}^{T} \Pr(\mathsf{J}_t\neq 0)\cdot 6\Delta^2\\
&= \mathbb{E}[\sigma]\cdot 6\Delta^2\\
&\le L\cdot 6\Delta^2 = \epsilon^2\,.
\end{aligned}$$

As before, we use this upper bound on the mutual information between Y and S_T given I to lower bound the conditional entropy of Y given S_T and I as

$$\mathbb{H}(\mathsf{Y}\mid\mathsf{S}_T,\mathsf{I}) = \mathbb{H}(\mathsf{Y}\mid\mathsf{I}) - \mathbb{I}(\mathsf{Y}\,;\mathsf{S}_T\mid\mathsf{I}) \ge \mathbb{H}(\mathsf{Y}\mid\mathsf{I}) - \epsilon^2\,.$$

These bounds demonstrate that in expectation, the entropies of the posterior distributions of I and $\mathsf{Y}|\mathsf{I}$ given the samples drawn by the algorithm will remain large if the algorithm does not draw sufficiently many samples. We shall further show that this must necessarily be the case, not just in expectation, but also with high probability.

Consider any realization $\mathcal{S}_T$ for S_T. We say that the realized outcome profile $\mathcal{S}_T$ is *ϵ-uninformative* iff both, $\mathbb{H}(\mathsf{I}\mid\mathsf{S}_T=\mathcal{S}_T)\ge\mathbb{H}(\mathsf{I})-\epsilon$, and $\mathbb{H}(\mathsf{Y}\mid\mathsf{I},\mathsf{S}_T=\mathcal{S}_T)\ge\mathbb{H}(\mathsf{Y}\mid\mathsf{I})-\epsilon$. Roughly speaking, whenever the outcome profile $\mathcal{S}_T$ is ϵ-uninformative, the algorithm is quite "uncertain" about both, the identity of i^*, as well as its reward μ_{i^*} (controlled through the variable y) and hence needs to estimate both among a large pool of possibilities in a later pass. To show that a realized outcome profile $\mathcal{S}_T$ will be ϵ-uninformative with high probability, let $C_I := \mathbb{H}(\mathsf{I})-\mathbb{H}(\mathsf{I}\mid\mathsf{S}_T)$. By Markov's inequality, we have that

$$\begin{aligned}
\Pr_{\mathcal{S}_T}\left(\mathbb{H}(\mathsf{I})-\mathbb{H}(\mathsf{I}\mid\mathsf{S}_T=\mathcal{S}_T)\ge C_I/\epsilon\right) &\le \frac{\mathbb{E}_{\mathcal{S}_T}\left[\mathbb{H}(\mathsf{I})-\mathbb{H}(\mathsf{I}\mid\mathsf{S}_T=S_T)\right]}{C_I/\epsilon}\\
&= \frac{(\mathbb{H}(\mathsf{I})-\mathbb{H}(\mathsf{I}\mid\mathsf{S}_T))}{C_I/\epsilon} = \epsilon\\
&\text{(by the choice of } C_I=\mathbb{H}(\mathsf{I})-\mathbb{H}(\mathsf{I}\mid\mathsf{S}_T))
\end{aligned}$$

Following an identical calculation with $C_Y := \mathbb{H}(\mathsf{Y}\mid\mathsf{I})-\mathbb{H}(\mathsf{Y}\mid\mathsf{I},\mathsf{S}_T)$, we have

$$\Pr_{\mathcal{S}_T}\left(\mathbb{H}(\mathsf{Y}\mid\mathsf{I})-\mathbb{H}(\mathsf{Y}\mid\mathsf{I},\mathsf{S}_T=\mathcal{S}_T)\ge C_Y/\epsilon\right)\le\epsilon$$

Since both, $C_I, C_Y\le\epsilon^2$, we have with probability at least $1-2\epsilon$ over realizations $\mathcal{S}_T$ of S_T,

$$\mathbb{H}(\mathsf{I})-\mathbb{H}(\mathsf{I}\mid\mathsf{S}_T=\mathcal{S}_T) < \frac{C_I}{\epsilon}\le\epsilon \implies \mathbb{H}(\mathsf{I}\mid\mathsf{S}_T=\mathcal{S}_T)\ge\log A-\gamma-\epsilon\,, \tag{3}$$

as well as

$$\mathbb{H}(\mathsf{Y} \mid \mathsf{I}) - \mathbb{H}(\mathsf{Y} \mid \mathsf{I}, \mathsf{S}_T = \mathcal{S}_T) < \frac{C_Y}{\epsilon} \leq \epsilon \implies \mathbb{H}(\mathsf{Y} \mid \mathsf{I}, \mathsf{S}_T = \mathcal{S}_T) \geq \log 2 - \gamma - \epsilon \,. \tag{4}$$

Henceforth, we shall use S_{ui} to refer to an ϵ-uninformative realization of S_T.

Now fix any ϵ-uninformative realization S_{ui}. Let $\mathcal{M} \subset \mathcal{A}$ be the set of arms of size $|\mathcal{M}| = \beta A$ chosen to be retained by the algorithm given its execution history S_{ui}. and let $\mathcal{R} = \mathcal{A} \setminus \mathcal{M}$ denote the remaining set of rejected arms. Using Lemma 6 we can argue that

$$\Pr(\mathsf{I} \in \mathcal{M} \mid \mathsf{S}_T = S_{\text{ui}}) \leq \log(1+\beta) + \gamma_I + \epsilon \implies \Pr(\mathsf{I} \in \mathcal{R} \mid \mathsf{S}_T = S_{\text{ui}}) > 1 - \log(1+\beta) - \gamma_I - \epsilon \,. \tag{5}$$

We will finally prove that in the event that the sequence of rewards observed by the algorithm was uninformative, and the algorithm actually did end up discarding the best arm from its memory, then the entropy of the identity of the best arm remains large amongst the arms the algorithm chose to reject at its stopping time.

Claim 3 *For any ϵ-uninformative realization S_{ui}, we have*

$$\mathbb{H}(\mathsf{I} \mid \mathsf{S}_T = S_{\text{ui}}, \mathsf{I} \notin \mathcal{M}) \geq \log((1-\beta)A) - \frac{\gamma + \epsilon}{1 - \log(1+\beta) - \gamma - \epsilon} \,.$$

Proof Suppose, for the sake of contradiction, that the above inequality is not true. Let X be an indicator random variable which is 1 when $\mathsf{I} \notin \mathcal{M}$, and 0 otherwise. Furthermore, let $p = \Pr(\mathsf{X} = 1 | \mathsf{S}_T = S_{\text{ui}}) = \Pr(\mathsf{I} \notin \mathcal{M} \mid \mathsf{S}_T = S_{\text{ui}})$. Then we have that

$$\begin{aligned}
\mathbb{H}(\mathsf{I} \mid \mathsf{S}_T = S_{\text{ui}}) &\overset{(a)}{\leq} \mathbb{H}(\mathsf{I}, \mathsf{X} \mid \mathsf{S}_T = S_{\text{ui}}) \\
&\overset{(b)}{=} \mathbb{H}(\mathsf{X} \mid \mathsf{S}_T = S_{\text{ui}}) + \mathbb{H}(\mathsf{I} \mid \mathsf{X}, \mathsf{S}_T = S_{\text{ui}}) \\
&= p\left(\mathbb{H}(\mathsf{I} \mid \mathsf{S}_T = S_{\text{ui}}, \mathsf{X} = 1) + \log\frac{1}{p}\right) + (1-p)\left(\mathbb{H}(\mathsf{I} \mid \mathsf{S}_T = S_{\text{ui}}, \mathsf{X} = 0) + \log\frac{1}{1-p}\right) \\
&\overset{(c)}{<} p\left(\log((1-\beta)A) - \frac{\gamma+\epsilon}{1-\log(1+\beta)-\gamma-\epsilon} + \log\frac{1}{p}\right) + (1-p)\left(\log(\beta A) + \log\frac{1}{1-p}\right) \\
&= \log A - \frac{p(\gamma+\epsilon)}{1-\log(1+\alpha)-\gamma-\epsilon} + p\log\frac{1-\beta}{p} + (1-p)\log\frac{\beta}{1-p} \\
&\overset{(d)}{\leq} \log A - \frac{p(\gamma+\epsilon)}{1-\log(1+\alpha)-\gamma-\epsilon} + \log\left(p \cdot \frac{(1-\beta)}{p} + (1-p)\cdot\frac{\beta}{1-p}\right) \\
&< \log A - \gamma - \epsilon,
\end{aligned}$$

where (a) follows due to the fact that the joint entropy in (I, X) is at least the entropy in I, (b) follows due to the chain rule for entropy, (c) follows by our assumption (for the sake of contradiction), (d) follows by Jensen's inequality, and the final inequality follows from bounding p through Equation 5. This contradicts the bound achieved in Equation 3. ■

We further argue a similar claim about the uncertainty in estimating the reward of the best arm in the event that the sequence of rewards observed by the algorithm is uninformative, and the algorithm did end up discarding the best arm from its memory.

Claim 4 *For any ϵ-uninformative realization S_{ui}, we have that*

$$\mathbb{H}(\mathsf{Y} \mid \mathsf{I}, \mathsf{S}_T = S_{\text{ui}}, \mathsf{I} \notin \mathcal{M}) \geq \log 2 - \frac{\gamma + \epsilon}{1 - \log(1+\beta) - \gamma - \epsilon} .$$

Proof Suppose, for the sake of contradiction, that the above inequality is not true. Let X be a random variable that takes value 1 when $\mathsf{I} \notin \mathcal{M}$, and 0 otherwise, and let $p = \Pr(\mathsf{X} = 1 \mid \mathsf{S}_T = S_{\text{ui}}) = \Pr(\mathsf{I} \notin \mathcal{M} \mid \mathsf{S}_T = S_{\text{ui}})$. We have that

$$\begin{aligned}
\mathbb{H}(\mathsf{Y} \mid \mathsf{I}, \mathsf{S}_T = S_{\text{ui}}) &\leq \mathbb{H}(\mathsf{Y}, \mathsf{X} \mid \mathsf{I}, \mathsf{S}_T = S_{\text{ui}}) \\
&= \mathbb{H}(\mathsf{Y} \mid \mathsf{I}, \mathsf{S}_T = S_{\text{ui}}, \mathsf{X}) + \mathbb{H}(\mathsf{X} \mid \mathsf{I}, \mathsf{S}_T = S_{\text{ui}}) \\
&\overset{(a)}{=} \mathbb{H}(\mathsf{Y} \mid \mathsf{I}, \mathsf{S}_T = S_{\text{ui}}, \mathsf{X}) \\
&= p\mathbb{H}(\mathsf{Y} \mid \mathsf{I}, \mathsf{S}_T = S_{\text{ui}}, \mathsf{X} = 1) + (1-p)\mathbb{H}(\mathsf{Y} \mid \mathsf{I}, \mathsf{S}_T = S_{\text{ui}}, \mathsf{X} = 0)
\end{aligned}$$

where (a) follows by observing that upon conditioning on the identity of the best arm I, as well as the observed outcome profile S_T, the value of the random variable X (i.e. whether the best arm was retained or discarded) is fixed, since the algorithm is deterministic. Therefore, $\mathbb{H}(\mathsf{X} \mid \mathsf{I}, \mathsf{S}_T = S_{\text{ui}}) = 0$. We now have

$$\begin{aligned}
\mathbb{H}(\mathsf{Y} \mid \mathsf{I}, \mathsf{S}_T = S_{\text{ui}}) &\leq p\mathbb{H}(\mathsf{Y} \mid \mathsf{I}, \mathsf{S}_T = S_{\text{ui}}, \mathsf{X} = 1) + (1-p)\mathbb{H}(\mathsf{Y} \mid \mathsf{I}, \mathsf{S}_T = S_{\text{ui}}, \mathsf{X} = 0) \\
&< p\left(\log 2 - \frac{\gamma + \epsilon}{1 - \log(1+\beta) - \gamma - \epsilon}\right) + (1-p)\log 2 \\
&< \log 2 - \gamma - \epsilon \,,
\end{aligned}$$

where the final inequality follows from bounding p through Equation 5, which contradicts the bound achieved in Equation 4. ■

We finally show that this outcome is not a rare event, but rather quite likely

$$\Pr(\mathsf{I} \notin \mathcal{M}) \geq 1 - \Pr(\mathcal{S}_T \text{ is informative}) - \Pr(\mathsf{I} \in \mathcal{M} \mid \mathcal{S}_T \text{ is uninformative}) \geq 1 - \log(1+\beta) - \gamma - 3\epsilon \,.$$

■

Appendix C. Proof of Lemma 4

In this section, we present the complete proof of Lemma 4 (restated here for convenience), which in turn implies Theorem 1.

Lemma 4 *Let $K, T, B, b \in \mathbb{N}_+$ be any set of parameters such that $T \geq K$, and $\log\log T \geq B \geq b \geq 1$. Given these parameters, let $\mathcal{K}$ be a set of $|\mathcal{K}| = K$ arms, and let $\{\mathcal{A}_j\}_{j=1}^{b+1}$ be a partition of $\mathcal{K}$ such that for each $j \in [b+1]$, $|\mathcal{A}_j| = A = K/(b+1)$. Furthermore, let $\{\psi_j\}_{j=1}^{b+1}$ be any set of joint distributions, where for each $j \in [b+1]$, distribution ψ_j supported over $\mathcal{A}_j \times \{0,1\}$ satisfies* *the condition that the random variables $\mathsf{I}_j, \mathsf{Y}_j \sim \psi_j$ drawn from the distribution have large entropy. Specifically, their entropy is such that $\mathbb{H}(\mathsf{I}_j) \geq \log A - \gamma$, and $\mathbb{H}(\mathsf{Y}_j \mid \mathsf{I}_j) \geq \log 2 - \gamma$ for some $0 \leq \gamma \leq (32b)^{-1}$. Consider an instance over arms $\mathcal{K}$ with mean rewards $\{\mu_i\}_{i\in\mathcal{K}}$ drawn according to $\mathcal{D}_{\mathcal{K},B}^{\{\psi_j\}_{j=1}^{b+1}}$, then over a time horizon of length T, the regret $R_T(\text{ALG})$ of any deterministic algorithm* ALG *that uses at most $M = K(8b(b+1)\log e)^{-1}$ memory and at most b passes over the arms $\mathcal{K}$ input to the algorithm as a stream is such that*

$$\mathbb{E}[R_T(\text{ALG})] \geq \Omega\left(T^{\frac{2^B}{2^{B+1}-1}} 4^{-b}\right)$$

Proof As described in our proof sketch, we will prove this lemma using induction on the number of passes b, and consider a modified setting where the algorithm is allowed additional power: in every epoch, which begins when the first arm of that epoch is read into memory and ends right before the first arm of the next epoch is read into memory, the algorithm is allowed to store all arms of that epoch in memory. However, the algorithm may retain in memory at most a β fraction of the arms from that epoch in addition to the arms stored from previous epochs (w.l.o.g, we can assume the algorithm retains exactly a β fraction of the arms as it can always choose to ignore the extra arms). This cannot hurt the regret as we are only allowing more memory, which can always be ignored. Formally, any algorithm that uses at most $\beta K/(B+1)$ memory (where $\beta = (8b\log e)^{-1}$) in the original setting can be used in this modified setting as it is allowed to use strictly more memory for each epoch in the modified setting. Also, note that the algorithm incurs the same regret in both settings. Hence, an optimal algorithm in this modified setting cannot incur more regret than an optimal algorithm in the original setting. Let $\alpha = 2^B/(2^{B+1}-1)$.

Base Case ($b = 1$): Let $\epsilon = 1/288$, $L = \epsilon^2/(6\Delta_1^2)$, and σ be the (random) length (number of trials) of the first epoch.

Case 1. $[\mathbb{E}[\sigma] \geq L]$, with the expectation taken over the observations made by the algorithm.

In this case, we claim that the algorithm will suffer an expected regret $\Omega(L\Delta_2)$. To see this, observe that $\mathbb{H}(\mathsf{Y}_2) \geq \mathbb{H}(\mathsf{Y}_2|\mathsf{I}_2) \geq \log 2 - \gamma$, which follows from Fact 1. Therefore, by Lemma 7, we have that the random variable Y_2 is distributed as a Bernoulli $\mathcal{B}(p)$ with parameter p such that $|p - \frac{1}{2}| \leq \sqrt{5\ln(4)\gamma/16} = \sqrt{(5\ln 4)/2}/16$, which follows from the fact that $\gamma < 1/32$. Therefore we have that the best arm i_2^* will realize to have a large reward $\mu_{i_2^*} = 1/2 + \Delta_2$ with probability at least $1/2 - \sqrt{(5\ln 4)/2}/16$, giving us that the expected regret of the algorithm

$$\begin{aligned}
\mathbb{E}[R_T(\text{ALG})] &\geq \left(\frac{1}{2} - \frac{\sqrt{(5\ln 4)/2}}{16}\right) L \cdot \Delta_2 \\
&= \left(\frac{1}{2} - \frac{\sqrt{(5\ln 4)/2}}{16}\right) \frac{\epsilon^2}{6} \cdot \frac{\Delta_2}{\Delta_1^2} \\
&= \Omega\left(T^{\frac{2^{B+1}-2}{2^{B+1}-1}} \cdot T^{-\frac{2^B-2}{2^{B+1}-1}}\right) = \Omega\left(T^\alpha\right)
\end{aligned}$$

Case 2. $[\mathbb{E}[\sigma] < L]$

Let $\mathcal{S}_\sigma$ be the outcomes observed by the algorithm over the arms sampled in epoch 1. Then by Lemma 3, we have that after observing the outcomes $\mathcal{S}_\sigma$, the best arm i_1^* will be discarded by the algorithm (i.e. $\mathsf{I}_1 \notin \mathcal{M}$ where $\mathcal{M}$ is the set of arms from epoch 1 that are retained in memory by the algorithm), and the entropy of the posterior distribution

$$\begin{aligned}
\mathbb{H}(\mathsf{Y}_1|\mathcal{S}_\sigma, \mathsf{I}_1, \mathsf{I}_1 \notin \mathcal{M}) &\geq \log 2 - \frac{\gamma+\epsilon}{1-(\log(1+\beta)+\gamma+\epsilon)} \\
&= \log 2 - \frac{\gamma+\epsilon}{1-\left(\log\left(1+\frac{1}{8\log e}\right)+\gamma+\epsilon\right)} \\
&\geq \log 2 - \frac{\gamma+\epsilon}{1-\left(\frac{\log e}{8\log e}+\gamma+\epsilon\right)} \\
&= \log 2 - \frac{10}{242}
\end{aligned}$$

with probability at least $1-(\log(1+\beta)-\gamma-3\epsilon) \geq \frac{5}{6}$. Furthermore, since we have that $\mathbb{E}(\sigma) < L = \epsilon^2/(6\Delta_1^2)$, by Markov's inequality, the actual number of trials σ spent in epoch 1 will be at most $1/\Delta_1^2 = o(T)$ with probability at least $1-\epsilon^2/6 \geq 1-10^{-6}$. There are least $T - 1/\Delta_1^2 = T - o(T) = \Omega(T)$ trials left in epoch 2 with a very high constant probability. In this case, the algorithm will suffer large regret when the best arm i_1^* in epoch 1 realizes to have a large reward $\mu_{i_1^*} = 1/2+\Delta_1$, i.e. Y_1 realizes to have value 1, and the best arm i_2^* in epoch 2 realizes to have a low reward of $\mu_{i_2^*} = 1/2$, i.e. Y_2 realizes to have value 0.

Observe that in the posterior distribution of the rewards of arms in epoch 1, the entropy in the reward of the best arm in the first epoch $\mathbb{H}(\mathsf{Y}_1|\mathcal{S}_\sigma, \mathsf{I}_1 \notin \mathcal{M}) \geq \mathbb{H}(\mathsf{Y}_1|\mathcal{S}_\sigma, \mathsf{I}_1, \mathsf{I}_1 \notin \mathcal{M}) \geq \log 2 - 10/242$, and therefore, by Lemma 7, the posterior distribution of Y_1 is Bernoulli with parameter $p \geq 1/2 - \sqrt{(5\cdot\ln 4\cdot 10)/(242\cdot 16)} = 1/2 - 5\sqrt{\ln 4}/44$ (a constant bounded away from 0) which is the probability with which the best arm in the first epoch actually had a large reward. Similarly, we have that in the prior distribution of the rewards of arms in epoch 2, the reward of the best arm in the second epoch $\mathbb{H}(\mathsf{Y}_2) \geq \mathbb{H}(\mathsf{Y}_2|\mathsf{I}_2) \geq \log 2 - \gamma$, and therefore, we have that the distribution of Y_2 is Bernoulli with parameter $p \leq 1/2 + \sqrt{(5\cdot\ln 4)/(32\cdot 16)} = 1/2 + \sqrt{(5\ln 4)/2}/16$. Therefore, the best arm in the second epoch realizes to have low reward with probability at least $1/2 - \sqrt{(5\ln 4)/2}/16$ (a constant bounded away from 0). Therefore, we have that the expected regret of the algorithm in this case

$$\begin{aligned}
\mathbb{E}[R_T(\text{ALG})] &\geq \frac{5}{6}\cdot\left(\frac{1}{2}-\frac{5\sqrt{\ln 4}}{44}\right)\cdot\left(\frac{1}{2}-\frac{\sqrt{(5\ln 4)/2}}{16}\right)(1-10^{-6})\cdot(T-o(T))\cdot\Delta_1 \\
&\geq \Omega\left((T-o(T))\cdot T^{-\frac{2^B-1}{2^{B+1}-1}}\right) = \Omega\left(T^\alpha\right)
\end{aligned}$$

Therefore, the expected regret is $\Omega(T^\alpha)$, which proves the base case.

Induction Step: Assuming the lemma is true for any number of passes up to $b-1$, will show that it also holds for b passes. Suppose for the sake of contradiction that the claim is not true for b, i.e. there exists a b-pass algorithm ALG with memory at most $K(8b(b+1)\log e)^{-1}$ whose expected regret over rewards drawn from the distribution $\mathcal{D}_{\mathcal{K},B}^{\{\psi_j\}_{j=1}^{b+1}}$ is $o\left(T^\alpha 4^{-b}\right)$.

The general outline will again be to show that if the algorithm ends any epoch after performing *sufficient exploration*, then it will incur large regret in the case that arm i^*_{b+1} realizes to a large mean reward, contradicting the assumption that ALG has small regret. On the other hand, if the algorithm ends all epochs with *insufficient exploration*, then the algorithm will not just discard all best arms, but also the instance induced over the discarded arms will remain hard. Supposing the algorithm achieves low expected regret over this instance in $b-1$ passes, it would contradict our induction hypothesis.

Consider any epoch $j \in [b]$. Let $\gamma_+ = 1/(32b)$, $\epsilon = 2\gamma_+/(9b)$, and $L_j := \epsilon^2/(6\Delta_j^2)$. For any epoch $j \in [b+1]$, let t_j be the trial when epoch j begins, let $\mathcal{T}_j := \{t_j, t_j+1, \cdots, t_{j+1}-1\}$ be the trials that belong to epoch j, and let $\sigma_j = |\mathcal{T}_j|$ denote the number of trials in epoch j. Lastly, let $\mathcal{S}_{\sigma_j} = \cup_{r=1}^{j}\{(i_t, r_t)\}_{t\in\mathcal{T}_r}$ be the sequence of observations defining the execution history of the algorithm until the end of epoch j with i_t being the arm pulled and r_t being the reward realized in trial t, respectively.

Case 1. $\left[\mathbb{E}(\sigma_j) \geq \epsilon^2/(6\Delta_j^2) \text{ for some } j \in [b]\right]$, where the expectation is over realizations of the rewards until epoch j.

In this case, observe that the expected regret of the algorithm is $\Omega(L_j \cdot \Delta_{b+1})$ in the event where the best arm in the final epoch (that has not been seen yet) realizes to have a large reward, i.e. $\mu_{i^*_{b+1}} = \frac{1}{2} + \Delta_{b+1}$. By definition of the input instance and Fact 1, we have that $\mathbb{H}(\mathsf{Y}_{b+1}) \geq \mathbb{H}(\mathsf{Y}_{b+1}|\mathsf{I}_{b+1}) \geq \log 2 - \gamma$. Therefore, by Lemma 7, we have that Y_{b+1} is distributed as a Bernoulli $\mathcal{B}(p)$ with parameter p such that $|p - 1/2| \leq \sqrt{(5\gamma \ln 4)/16}$. Since $\gamma \leq 1/(32b)$, and $b \geq 2$, we have $\sqrt{5\gamma \ln 4/16} \leq \sqrt{(10b^{-1}\ln 4)}/32 < 1/4 - 1/6 - 1/(200b^3)$ for $b \geq 2$. Therefore, we have that $\mu_{i^*_{b+1}} = 1/2 + \Delta_{b+1}$ with probability at least $1/2 - (1/4 - 1/6 - 1/(200b^3))$.

Therefore, the expected regret in the event that $\mathbb{E}[\sigma_j] > L_j$ for some $j \in [b]$

$$\mathbb{E}[R_T(\text{ALG})] \geq \Omega(L_j\Delta_{b+1}) = \Omega\left(T^{\frac{2^{B+1}-2^j}{2^{B+1}-1}} \cdot T^{-\frac{2^B-2^b}{2^{B+1}-1}} \cdot \epsilon^2\right) = \Omega\left(T^{\frac{2^B+2^b-2^j}{2^{B+1}-1}} \cdot \epsilon^2\right) = \Omega(T^{\alpha}b^{-4}),$$

which contradicts the assumption that the expected regret of the algorithm is $o\left(T^{\alpha}4^{-b}\right)$.

Case 2. $\left[\mathbb{E}(\sigma_j) < \epsilon^2/(6\Delta_j^2) \text{ for all } j \in [b]\right]$

In this case, we will leverage Lemma 3 to show that the algorithm will not be able to collect *sufficient information* about i^*_j and it will suffer large regret in the remaining number of passes. Using Lemma 3 we will show that the conditional distribution for I_j after epoch j will have high entropy. Let $\mathcal{T}_j \subseteq [T]$ be the trials that belong to epoch j. Let $\mathcal{S}_{\sigma_j} = \{(i_t, r_t)\}_{t\in\mathcal{T}_j}$ be the execution history of epoch j with i_t being the arm pulled and r_t being the reward realized in trial t, respectively. Given $\mathcal{S}_{\sigma_j}$, let $\mathcal{M}_j \subset \mathcal{A}_j$ be the set of $|\mathcal{M}_j| = \beta A = K/(8b(b+1)\log e)$ arms retained by the algorithm in memory, and let $\mathcal{R}_j = \mathcal{A}_j \setminus \mathcal{M}_j$ be the set of $|\mathcal{R}_j| = (1-\beta)A = R$ arms that were rejected after epoch j. Let $\gamma_- = 1/(32(b-1)) = b\gamma_+/(b-1)$. We first observe that by Lemma 3, the entropy of the posterior

$$\begin{aligned}
\mathbb{H}(\mathsf{I}_j|\mathcal{S}_{\sigma_j},\mathsf{I}_j\in\mathcal{R}_j)=\mathbb{H}(\mathsf{I}_j|\mathcal{S}_{\sigma_j},\mathsf{I}_j\notin\mathcal{M}_j) &\geq \log R-\frac{\gamma+\epsilon}{1-(\log(1+\beta)+\gamma+\epsilon)}\\
&\geq \log R-\frac{\gamma+\epsilon}{1-\left(\log\left(1+\frac{1}{8b\log e}\right)+\gamma+\epsilon\right)}\\
&\geq \log R+\frac{\left(1+\frac{2}{9b}\right)\gamma_+}{1-(\frac{\log e}{8b\log e}+\gamma_++\frac{2\gamma_+}{9b})}\\
&= \log R-\underbrace{\frac{\left(1+\frac{2}{9b}\right)}{1-(\frac{1}{8b}+\frac{1}{32b}+\frac{1}{144b^2})}}_{x}\gamma_+.
\end{aligned}$$

We claim that $x\leq b/(b-1)$, which would imply that the entropy of the posterior $\mathbb{H}(\mathsf{I}_j|S_{\sigma_j},\mathsf{I}_j\in E_j)\geq\log|E_j|-\gamma_-$. We have

$$\begin{aligned}
x&=\frac{\left(1+\frac{2}{9b}\right)}{1-(\frac{1}{8b}+\frac{1}{32b}+\frac{1}{144b^2})}\\
&<\frac{1+\frac{2}{9b}}{1-(5+\frac{2}{9})(\frac{1}{32b})}\\
&=\frac{288b+64}{288b-47}\\
&<\frac{b}{b-1},
\end{aligned}$$

where the final inequality follows by observing $\frac{288b+64}{288b-47}-\frac{b}{b-1}=\frac{-(177b+64)}{(b-1)(288b-47)}<0$. The proof of the fact that $\mathbb{H}(\mathsf{Y}_j|\mathcal{S}_{\sigma_j},\mathsf{I}_j,\mathsf{I}_j\in\mathcal{R}_j)\geq\log 2-\gamma_-$ follows by the exact same calculation.

At this point, we further argue that this supposed low regret algorithm ALG cannot spend too many trials on the first b epochs prior to processing the $(b+1)^{th}$ epoch with a large probability. Since we have that $\mathbb{E}(\sum_{j\in[b]}\sigma_j)<\sum_{j\in[b]}L_j=\epsilon^2\sum_{j\in[b]}(6\Delta_j^2)^{-1}$, by Markov's inequality, it must be that $\Pr(\sum_{j\in[b]}\sigma_j\geq\sum_{j\in[b]}\Delta_j^{-2})\leq\epsilon^2/6<(10b)^{-4}$. We define the event $\mathcal{E}_0:=(\sum_{j\in[b]}\sigma_j<\sum_{j\in[b]}\Delta_j^{-2})$ where the actual number of trials spent by the algorithm in the first b epochs is small, and the above calculation gives us that $\Pr(\neg\mathcal{E}_0)<(10b)^{-4}$. Now for every $j\in[b]$, let us define the event

$$\mathcal{E}_j:=\left(\mathsf{I}_j\in\mathcal{R}_j \text{ and } \mathbb{H}(\mathsf{I}_j\mid\mathcal{S}_{\sigma_j},\mathsf{I}_j\in\mathcal{R}_j)\geq\log R-\gamma_- \text{ and } \mathbb{H}(\mathsf{Y}_j\mid\mathcal{S}_{\sigma_j},\mathsf{I}_j,\mathsf{I}_j\in\mathcal{R}_j)\geq\log 2-\gamma_-\right).$$

Using Lemma 3, we have that

$$\begin{aligned}
\Pr(\neg\mathcal{E}_j)&\leq\log\left(1+\frac{1}{8b\log e}\right)+\gamma+3\epsilon\\
&\leq\frac{\log e}{8b\log e}+\gamma+\frac{6\gamma}{9b}\\
&=\frac{1}{8b}+\frac{1}{32b}+\frac{1}{96b}=\frac{1}{6b}.
\end{aligned}$$

where the final inequality follows by observing $b \geq 2$.

Let us also define the event $\mathcal{E}_{b+1} := \{\mu_{i^*_{b+1}} = 1/2\}$. By definition of the input instance and Fact 1, we have that $\mathbb{H}(\mathsf{Y}_{b+1}) \geq \mathbb{H}(\mathsf{Y}_{b+1}|\mathsf{I}_{b+1}) \geq \log 2 - \gamma$. Therefore, by Lemma 7, we have that Y_{b+1} is distributed as a Bernoulli $\mathcal{B}(p)$ with parameter p such that $|p - 1/2| \leq \sqrt{(5\gamma \ln 4)/16}$. Since $\gamma \leq 1/(32b)$, and $b \geq 2$, we have $\sqrt{5\gamma \ln 4/16} \leq \sqrt{5 \ln 4}/32 < 1/4 - 1/6 - 1/(200b^3)$. Therefore, we have that $\Pr(\neg\mathcal{E}_{b+1}) \leq 1/2 + (1/4 - 1/6 - 1/(200b^3))$. Therefore, by a union bound over all these $(b+2)$ events, we have

$$\Pr(\mathcal{E}_0 \cap_{j \in [b+1]} \mathcal{E}_j) \geq 1 - \frac{1}{(10b)^4} - \frac{b}{6b} - \left(\frac{1}{2} + \frac{1}{4} - \frac{1}{6} - \frac{1}{200b^3}\right) \geq \frac{1}{4}$$

Lastly, we argue that under event $\mathcal{E}$, the algorithm must necessarily spend $o(T)$ trials in the last epoch $b+1$. This is because under event $\mathcal{E}$, $\mu^*_{b+1} = 1/2$, and furthermore, in the posterior distribution of the reward of the best arm $\mu_{i^*_b}$ of the b^{th} epoch in the rejected set $\mathcal{R}_b$ is at least $\mathbb{H}(\mathsf{Y}_b|\mathcal{S}_{\sigma_b}, \mathsf{I}_b \notin \mathcal{M}_b) \geq \mathbb{H}(\mathsf{Y}_b|\mathcal{S}_{\sigma_b}, \mathsf{I}_b, \mathsf{I}_b \notin \mathcal{M}_b) \geq \log 2 - \gamma_-$, where $\gamma_- \leq 1/(32(b-1))$. Therefore, by Lemma 7, the posterior distribution of Y_b is Bernoulli with parameter $p \geq 1/2 - \sqrt{(5 \cdot \ln 4 \cdot \gamma_-)/16} = 1/2 - \sqrt{(5 \ln 4)/2}/16 > 1/2 - 1/8$, which is the probability with which the rejected best arm in the b^{th} epoch actually had a large reward. Therefore, if the algorithm spends $\Omega(T)$ trials in the $(b+1)^{th}$ epoch, then the expected regret of the algorithm

$$\begin{aligned}
\mathbb{E}[R_T(\textsc{alg})] \geq \Pr(\mathcal{E})\, \mathbb{E}[R_T(\textsc{alg})|\mathcal{E}] &\geq \frac{1}{4} \cdot \frac{3}{8} \cdot \Omega(T) \cdot \Delta_b \\
&= \Omega\left(T \cdot T^{-\frac{2^B - 2^{b-1}}{2^{B+1}-1}}\right) \\
&= \Omega\left(T^{\frac{2^B + 2^{b-1} - 1}{2^{B+1}-1}}\right) = \Omega\left(T^{\alpha}\right),
\end{aligned}$$

contradicting our assumption about the expected regret achieved by the algorithm. Therefore, under event $\mathcal{E}$, we have that the total number of trials spent by the algorithm in the first pass is necessarily $\sum_{j \in [b+1]} \sigma_j = o(T) + \sum_{j \in [b]} \Delta_j^{-2} = o(T) + \sum_{j \in [b]} T^{1 - \frac{2^j - 1}{2^{B+1}-1}} = o(T)$. Therefore, the number of trials T_- left is necessarily $T_- = T - o(T)$, and only $b-1$ passes left.

Now we shall use our assumption about the expected regret achievable by our algorithm to prove that in order to achieve low expected regret overall, it must necessarily achieve low expected regret in the remaining passes too. Let $\mathbb{E}[R_{T_-}(\textsc{alg})]$ denote the cumulative regret of the algorithm over the remaining $b-1$ passes. Therefore, we have

$$\mathbb{E}[R_T(\textsc{alg})] \geq \mathbb{E}[R_{T_-}(\textsc{alg}) \mid \mathcal{E}] \cdot \Pr(\mathcal{E}) = \frac{1}{4} \cdot \mathbb{E}[R_{T_-}(\textsc{alg}) \mid \mathcal{E}]\,.$$

Therefore, $\mathbb{E}[R_{T_-}(\textsc{alg}) \mid \mathcal{E}] \leq 4\, \mathbb{E}[R_T(\textsc{alg})] = o\left(4 \cdot T^{\alpha} \cdot 4^{-b}\right) = o\left(T^{\alpha} \cdot 4^{-(b-1)}\right)$. We shall use this fact to set up a contradiction to our induction hypothesis, which at a high level says that any $(b-1)$-pass algorithm with small memory must incur large regret. We begin by setting up the hard distribution.

In our new instance, we begin by discarding the arms $\mathcal{A}_{b+1}$ as under event $\mathcal{E}$, the reward of the best arm (and hence all arms) in this epoch $(b+1)$ has realized to a low value. Our new instance consists of all the rejected arms $\mathcal{R}_j$ for $j \in [b]$, the first b epochs. We refer to these set of arms as $\mathcal{K}' = \cup_{j\in[b]}\mathcal{R}_j$, whose size is exactly $|\mathcal{K}'| = K = \sum_{j\in[b]}(1-\beta)K/(b+1) = (1-\beta)bK/(b+1)$.

Next, we claim that the posterior distributions over $\mathsf{I}_j, \mathsf{Y}_j$ in $\mathcal{R}_j$ for $j \in [b]$ to give us a hard distribution over arms $\mathcal{K}'$ for the $(b-1)$ pass algorithm. For $j \in [b]$, let ϕ_j be the joint distribution $\text{dist}((\mathsf{I}_j, \mathsf{Y}_j) \mid \mathsf{I}_j \in \mathcal{R}_j)$, and let $\mathcal{D}_{\mathcal{K}',B}^{\{\phi_j\}_{j=1}^b}$. Its easy to verify that $\mathcal{D}_{\mathcal{K}',B}^{\{\phi_j\}_{j=1}^b}$ satisfies the requirements for a hard distribution for a $(b-1)$ pass algorithm, as the partitions $\{\mathcal{R}_j\}$ of $\mathcal{K}'$ are all of equal size $|\mathcal{K}'|/b$, and for random variables $(\mathsf{I}_j, \mathsf{Y}_j) \sim \phi_j$ for any $j \in [b]$ satisfy the high entropy condition $\mathbb{H}(\mathsf{I}_j) \geq \log|\mathcal{R}_j| - \gamma_-$, and $\mathbb{H}(\mathsf{Y}_j|\mathsf{I}_j) \geq \log 2 - \gamma_-$, where $\gamma_- \leq (32(b-1))^{-1}$ (as indicated by event $\mathcal{E}$). Furthermore, the memory budget for a $(b-1)$ pass algorithm for this instance is

$$\frac{K'}{8b(b-1)\log e} = \frac{(1-\beta)bK}{b+1}\cdot\frac{1}{8b(b-1)\log e} > \frac{\left(1-\frac{1}{b}\right)bK}{b+1}\cdot\frac{1}{8b(b-1)\log e} = \frac{K}{8b(b+1)\log e},$$

which is in fact larger than the memory used by $\textsc{Alg}$. We shall show that we can use the behavior of $\textsc{Alg}$ in the subsequent $b-1$ passes under event $\mathcal{E}$ to construct a $(b-1)$-pass algorithm with low memory that achieves $o(T^\alpha \cdot 4^{-(b-1)})$ expected regret on the above hard instance, which contradicts our induction hypothesis.

Let $\textsc{Alg}_-$ denote the algorithm $\textsc{Alg}$ for the remaining $b-1$ passes when event $\mathcal{E}$ occurs. Under the assumption that the regret $R_{T_-}(\textsc{Alg})$ is small, we will construct a $(b-1)$-pass algorithm $\textsc{Alg}_{b-1}$ with small memory that achieves small regret over time horizon $T_- = T - o(T)$ on the instance $\mathcal{D}_{\mathcal{K}',B}^{\{\phi^j\}_{j=1}^b}$. $\textsc{Alg}_{b-1}$ is constructed as follows: if $\textsc{Alg}_-$ pulls an arm in $\cup_{j\in[b]}\,\text{supp}(\phi_j)$, $\textsc{Alg}_{b-1}$ also pulls the corresponding arm in $\mathcal{K}'$ and returns the realized reward of the arm to $\textsc{Alg}_-$; otherwise, $\textsc{Alg}_{b-1}$ simply pulls an arm with distribution $\mathcal{B}(1/2)$ and returns the result to $\textsc{Alg}_-$.

It is trivially true that $\mathbb{E}[R_{T_-}(\textsc{Alg}_{b-1})] \leq \mathbb{E}[R_{T_-}(\textsc{Alg}_-) \mid \mathcal{E}]$. This is because, given event $\mathcal{E}$, any other arm than the arm in $\cup_{j\in[b]}\text{supp}(\phi_j)$ is distributed as $\mathcal{B}(1/2)$. Hence, $\textsc{Alg}_{b-1}$ is a $(b-1)$-pass algorithm with memory at most $K'(8b(b-1)\log e)^{-1}$ that achieves regret $o(T^\alpha \cdot 4^{-(b-1)})$ over a time horizon $T - o(T)$, which is a contradiction! This completes the proof of our lower bound.

■

Appendix D. Algorithm and Proof of Theorem 5

In this section, we describe and formally present our algorithm, and analyze its regret guarantees which prove Theorem 5.

D.1. Algorithm Description

Our proposed algorithm builds upon the classical Sequential-Elimination algorithm, where one maintains an "active set" of arms which are played in a round-robin manner until sufficient evidence is gathered indicating the sub-optimality of some arm, at which point it is permanently discarded from the active set.

In our limited memory setting, it is not possible to have all arms in the active set as the number of statistics we can save at any given time is bounded by M. Therefore, our active set is of size roughly equal to our memory, and we play the least played arm $i_{\min} \in \mathcal{M}$ in our active set until we gather sufficient evidence to discard a sub-optimal arm, after which the next arm from the stream is read into our active set. This requires storing 2 statistics per arm $i \in \mathcal{M}$ in memory – the cumulative reward observed r_i, as well as the number of times the arm was played n_i. For ease of exposition, we shall assume that both of these can be stored in a single word of memory.

In addition to these arms in our active set, we reserve an additional word of memory to store the arm $\tilde{i}$ (and its statistics $\tilde{\ell}$) we have estimated to be the best. This arm $\tilde{i}$ serves two important purposes. Firstly, it is exploited until the end of the time horizon after we have exhausted our budget on the number of passes. Secondly, its stored statistic $\tilde{\ell}$ which is a lower bound on its estimated mean reward, is used to quickly identify and discard sub-optimal arms from memory. This is necessary in this limited memory streaming setting as unlike the full memory setting, it is not possible to permanently discard bad arms. Even after establishing the sub-optimality of bad arms, their identity is forgotten when they are discarded from memory, and will be repeatedly encountered in subsequent passes at which point it eliminating them without incurring too much regret becomes crucial.

However, as established in our lower bound construction, the limited memory setting has an inherent risk associated: there can be some high value arm somewhere ahead in the stream that has not been read into memory yet, due to which overplaying the arms currently in memory can lead to large regret. As a result, we need to maintain a careful balance between playing arms in memory and exploring further into the stream. To address this problem, we borrow an idea from the limited-adaptivity framework for multi-armed bandits (Perchet et al., 2015; Gao et al., 2019), where we additionally impose a cap on the maximum number of times any arm can be played in a single pass, effectively limiting the length of exploration to (roughly) N^b in any single pass $b \in [B]$. If all arms in the active set have been played equal to the cap for that pass without any arm being discarded, then an arbitrary arm is ejected to make room for the next arm in the stream. This cap grows across passes, and in pass b is roughly $T^{\frac{2^{B+1}(1-1/2^b)}{2^{B+1}-1}}$ if the objective is to minimize worst-case regret $(w = 1)$, and $T^{\frac{b}{B+1}}$ if the objective is to minimize instance-dependent regret $(w = 0)$. Intuitively, one can think of this as approximating the mean rewards of arms with an increasingly finer precision across passes. If a crude estimate of the reward suffices to discard a suboptimal arm, then it does so. Otherwise, this specific choice of the cap ensures that this arm has not been explored enough to incur significant regret. The following is a formal description of this algorithm.

It is clear that Algorithm 1 uses memory at most M, and performs B passes over the stream given any input parameters M, B. Furthermore, observe that while we allow for larger memory, our algorithm just needs $M = 2$ words of memory. We shall now analyze the regret guarantees of Algorithm 1, which are restated here for convenience.

Theorem 5 *Given a time horizon T, a stream of K arms, memory $2 \leq M < K$, and number of passes $1 \leq B < \log\log T$, Algorithm 1 set for worst-case regret minimization $(w = 1)$ achieves cumulative regret*

$$R_T \leq O\left(T^{\frac{2^B}{2^{B+1}-1}}\sqrt{KB\log T}\right),$$

Algorithm 1 Memory Bounded Successive Elimination

Input. Memory M; number of passes B; time horizon T;

variable $w = \begin{cases} 1, & \text{if minimizing worst-case regret} \\ 0, & \text{if minimizing instance-dependent regret} \end{cases}$

Let arms in memory $\mathcal{M} \leftarrow \emptyset$, $N^0 \leftarrow 1$.

Set aside a single word of memory: set (estimated) best arm $\tilde{i} \leftarrow \emptyset$, lower confidence bound $\tilde{\ell} \leftarrow 0$

for pass $b = 1, \ldots, B$ **do**

  Set the maximum number of pulls across all arms in pass b,

$$N^b \leftarrow \begin{cases} T^{\frac{2^B}{2^{B+1}-1}} \sqrt{N^{b-1}}, & \text{if } w = 1 \text{ (minimize worst-case regret)} \\ T^{\frac{1}{B+1}} N^{b-1}, & \text{if } w = 0 \text{ (minimize instance-dependent regret)} \end{cases}$$

  For all arms $i \in \mathcal{M}$, set $n_i^b \leftarrow 0$, and $r_i^b \leftarrow 0$

  while pass is not finished **do**

    while $|\mathcal{M}| < M - 1$ **do**

      $\mathcal{M} \leftarrow \mathcal{M} \cup \{i\}$, where i is the next arm in the stream that is not already in memory.

      Set number of pulls $n_i^b \leftarrow 0$, cumulative reward $r_i^b \leftarrow 0$

    end while

    Let $i_{\min} \leftarrow \text{argmin}_{i \in \mathcal{M}} n_i^b$ be the least played arm in memory (ties broken arbitrarily)

    if $\left(n_{i_{\min}}^b \geq N^b/(KB) \text{ and } w = 1\right)$ or $\left(n_{i_{\min}}^b \geq N^b \text{ and } w = 0\right)$ **then**

      Discard an arbitrary arm $i \in \mathcal{M}$ from memory; $\mathcal{M} \leftarrow \mathcal{M} \setminus \{i\}$

    else

      Play arm $i_{\min}$ once, and observe reward r

      Update $r_{i_{\min}}^b \leftarrow r_{i_{\min}}^b + r$, and $n_{i_{\min}}^b \leftarrow n_{i_{\min}}^b + 1$

    end if

    Update $\tilde{\ell} \leftarrow \max_{i \in \mathcal{M}} r_i^b/n_i^b - \sqrt{(5 \log T)/n_i^b}$; and $\tilde{i} \leftarrow \text{argmax}_{i \in \mathcal{M}} r_i^b/n_i^b - \sqrt{(5 \log T)/n_i^b}$

    if there exists an arm $j \in \mathcal{M}$ such that $r_j^b/n_j^b + \sqrt{(5 \log T)/n_j^b} < \tilde{\ell}$ **then**

      Discard arm j from memory; $\mathcal{M} \leftarrow \mathcal{M} \setminus \{j\}$

    end if

  end while

end for

Play the estimated best arm $\tilde{i}$ until the end of the time horizon

with probability $1 - 2/T$ *using at most* B *passes and at most* M *words of memory. Furthermore, supposing the arms* $\mathcal{K}$ *had mean rewards* $\{\mu_j^*\}_{j \in \mathcal{K}}$*, then given number of passes* $1 \leq B < \log T$*, Algorithm 1 set for instance-dependent regret minimization* $(w = 0)$ *achieves cumulative regret*

$$R_T \leq O\left(\sum_{j \in \mathcal{S}} \frac{T^{1/(B+1)} \log T + B \log\left((\Delta_j^*)^2 T / \log T\right)}{\Delta_j^*}\right),$$

with probability $1 - 2/T$ *using at most* B *passes and at most* M *words of memory, where* $\mathcal{S} \subset \mathcal{K}$ *is the set of strictly sub-optimal arms in* $\mathcal{K}$*, and for any sub-optimal arm* $j \in \mathcal{S}$*,* $\Delta_j^* := \max_{\{i \in \mathcal{K}\}} \mu_i^* - \mu_j^*$ *is the regret due to playing arm* j*.*

D.2. Analysis of Algorithm 1 for Worst-Case Regret ($w = 1$)

As mentioned earlier, our algorithm cleverly balances playing the arms currently in memory, thereby gathering valuable information about them, and quickly exploring ahead into the stream to find potential high value arms by setting a cap ($\approx N^b$) on the maximum number of times any arm i can be played in any pass b. This cap is raised across passes in a systematic way, with the choice of growth rate guaranteeing that the total regret incurred due to playing suboptimal arms will be small.

At a high level, our proof shows that if across all passes $b \in [B]$, the observed mean rewards r_j^b/n_j^b for all arms $j \in \mathcal{K}$ are not too far from their true mean rewards μ_j^*, then the estimated best arm $\tilde{i}$ at the end of any pass b is a good proxy for the true best arm i^* for that pass. Specifically, the mean reward of $\tilde{i}$ closer to the mean reward of i^* than the precision ($\approx \sqrt{1/N^b}$) with which we estimate the means in that pass. This can be used to eliminate any bad arms that are distinguishable from the best arm in pass b (based on the precision set for that pass), but only starting the following pass $b+1$. Due the delayed nature of this information, i.e. the estimated best arm becomes "good enough" for elimination purposes in a pass b only after the true best arm, (or a good proxy for it) are encountered in the stream in that pass b. Therefore, suboptimal arms that transitioned from being indistinguishable in pass $b-1$ to distinguishable in pass b, and appeared early on in the stream can potentially be overplayed because the estimated best arm has not been updated yet. This would incur more regret than is desirable, but only up to a multiplicative $T^{2^B/(2^{B+1}-1)}$ factor (matching our B-pass lower bound) due to our choice (N^b) of the cap on the number of times an arm can be played in any pass. This gives our final guarantee on the upper bound on the regret of our algorithm.

We shall now formally prove this bound on the worst-case regret achieved by our algorithm. We begin with the following simple lemma, which bounds the deviation in the observed mean rewards of any arm from its true mean reward. Specifically, this lemma says that whenever an arm is stored in memory, its true mean reward will lie within a confidence ball of radius $O(\sqrt{\log T/n})$ if it has been played n times since being read into memory. Furthermore, this property would hold for all arms across all rounds with a polynomially large probability.

Lemma 8 *Let $\mathcal{K}$ be the set of arms in the stream, where arm $i \in \mathcal{K}$ has mean reward μ_i^*, and let B be the total number of passes. Whenever arm an $i \in \mathcal{K}$ is present in memory in pass $b \in [B]$ of the stream, we define the event*

$$\mathcal{E}_{i,b} := \left| \mu_i^* - \frac{r_i^b}{n_i^b} \right| < \sqrt{\frac{c \log T}{n_i^b}},$$

where r_i^b represents the observed cumulative reward of arm i in the b^{th} pass, n_i^b represents the number of times arm i was played in the b^{th} pass, and $c \geq 5$ is any constant. Then we have that the event $\mathcal{E} := \cap_{i \in \mathcal{K}, b \in [B]} \mathcal{E}_{i,b}$ occurs with probability at least $1 - 2/T^{c-4}$.

Proof Consider any fixed arm $i \in \mathcal{K}$. We shall assume that the rewards for this arm are sampled from the corresponding reward distribution, and written on a tape (of length at most T). Whenever the algorithm chooses to play this arm i in some round, it simply reads the realized reward from the next cell on the tape. By definition, as long as an arm is in memory, the algorithm maintains a running average of the observed rewards of that arm, resetting the running average every time it starts a new pass or loads the arm into memory, treating the last cell on the tape as a new starting point for counting rewards for this arm. For a fixed starting point on the tape after which it started keeping count of the rewards for the arm, the probability that $|\mu_i^* - r_i/n_i| \geq \sqrt{c \log T/n_i}$ for a fixed value of n_i is at most $2/T^c$ by Hoeffding's inequality. Therefore, the probability that this event occurs for some starting point on the tape and some value of n_i, by a union bound, is at most $2/T^{c-2}$. Observe that this event is exactly $\neg\mathcal{E}_{i,b}$ for some fixed pass b, as the running average is reset (either when the arm was in memory at the start of the pass, or was first loaded into memory at some point during that pass) at most once during that pass. Therefore, for any fixed arm $i \in \mathcal{K}$, the probability that event $\neg\mathcal{E}_{i,b}$ occurs for some pass $b \in [B]$, by a union bound over the passes, is at most $2B/T^{c-2}$. Finally, taking a union bound over all arms, the probability that event $\neg\mathcal{E}_{i,b}$ occurs for some pass arm $i \in \mathcal{K}$, and some pass $b \in [B]$ is at most $2KB/T^{c-2}$. This gives us our claimed bound, since $K, B \leq T$. ■

For simplicity, we assume $c = 5$ in Algorithm 1 and in the subsequent proof, which gives us that the above defined "good event" of interest occurs with probability at least $1 - 2/T$. Henceforth, we shall assume that this event occurs, following which our regret guarantees hold with probability 1.

Proof [of Theorem 5 (worst-case upper bound)] Let $\mathcal{E} := \cap_{i\in\mathcal{K}, b\in[B]} \mathcal{E}_{i,b}$ be the good event of interest defined in Lemma 8. Then we shall prove that conditioned on event $\mathcal{E}$, the cumulative regret R_T of Algorithm 1 set for worst-case regret minimization $(w = 1)$ is

$$R_T \leq O\left(T^{\frac{2^B}{2^{B+1}-1}} \sqrt{KB \log T}\right),$$

with probability 1. Prior to formally proving this, observe that this also automatically implies a $\mathbb{E}[R_T] \leq O\left(T^{\frac{2^B}{2^{B+1}-1}} \sqrt{KB \log T}\right)$ result for the expected regret of our algorithm as

$$\begin{aligned}
\mathbb{E}[R_T] &= \Pr(\mathcal{E})\, \mathbb{E}[R_T \mid \mathcal{E}] + (1 - \Pr(\mathcal{E}))\, \mathbb{E}[R_T \mid \neg\mathcal{E}] \\
&\leq \left(1 - \frac{1}{\text{poly}(T)}\right) \mathbb{E}(R_T|\mathcal{E}) + \frac{1}{\text{poly}(T)} \cdot T \\
&\leq O\left(T^{\frac{2^B}{2^{B+1}-1}} \sqrt{KB \log T}\right)
\end{aligned}$$

Let $\mu^*_{\max} := \max_{i\in\mathcal{K}} \mu_i^*$ be the largest expected reward of any arm in the stream, and let $\mathcal{S} := \{j \in \mathcal{K} : \ \mu_j^* < \mu^*_{\max}\}$ be the set of all suboptimal arms, with $\Delta_j := \mu^*_{\max} - \mu_j^*$ being the regret due to playing any suboptimal arm $j \in \mathcal{S}$. Furthermore, let $i^* \in \mathcal{K} \setminus \mathcal{S}$ be any arbitrary optimal arm, which we shall henceforth refer to as the best arm.

For any pass $b \in [B]$ and any arm $j \in \mathcal{K}$, let m_j^b be the maximum number of times arm j was played in pass b. Furthermore, let $R_j^b = m_j^b \Delta_j$ be the regret incurred by the algorithm by playing arm j in the b^{th} pass, and subsequently, let $R^b = \sum_{j\in\mathcal{S}} R_j^b$ be the total regret incurred in the b^{th} pass. Finally, let $R_{\tilde{i}}$ be the regret incurred by playing the estimated best arm $\tilde{i}$ at the end of the B^{th} pass until the end of the time horizon. Therefore, we have that

$$R_T = \sum_{b\in[B]} R^b + R_{\tilde{i}} = \sum_{j\in\mathcal{S}} R_j^1 + \sum_{b=2}^{B} \sum_{j\in\mathcal{S}} R_j^b + R_{\tilde{i}} \leq \frac{T^{\frac{2^B}{2^{B+1}-1}}}{B} + \sum_{b=2}^{B} \sum_{j\in\mathcal{S}} R_j^b + R_{\tilde{i}},$$

where the final inequality follows by observing that the maximum number of times any arm is played in the first pass is at most $T^{\frac{2^B}{2^{B+1}-1}}/(KB)$.

We shall now present the key implication of event $\mathcal{E}$, which basically guarantees that the best arm i^* will necessarily be played the maximum allowable number of times in every pass, i.e. i^* cannot be prematurely discarded from memory after it has been read in the stream. This gives us certain desirable guarantees about the true mean reward of the estimated best arm saved in $\tilde{i}$, as well as the maximum number of times m_j^b any suboptimal arm $j \in \mathcal{S}$ can be played in any pass $b \geq 2$, which will be crucial in bounding the cumulative regret of the algorithm.

Claim 5 *Given that event $\mathcal{E}$ occurs, arm i^* will necessarily be played $N^b/(KB)$ times in every pass $b \in [B]$. Consequently, for any pass $b \geq 2$, we have*

$$\tilde{\ell} \geq \mu_{\max}^* - 2\sqrt{(5KB\log T)/N^{b-1}},$$

at all times in pass b.

Proof This follows by observing that $\tilde{\ell}$ can never exceed $\mu_{\max}^*$ at any point during the execution of the algorithm. This is because for any arm $j \in \mathcal{K}$, event $\mathcal{E}$ guarantees that its observed mean reward $r_j^b/n_j^b < \mu_j^* + \sqrt{(5\log T)/n_j^b}$ in every pass $b \in [B]$, which guarantees $r_j^b/n_j^b - \sqrt{(5\log T)/n_j^b} < \mu_j^* \leq \mu_{\max}^*$. Furthermore, for the best arm i^*, we have that $r_{i^*}^b/n_{i^*}^b > \mu_{\max}^* - \sqrt{(5\log T)/n_{i^*}^b}$ in every pass $b \in [B]$, which guarantees $r_{i^*}^b/n_{i^*}^b + \sqrt{(5\log T)/n_{i^*}^b} > \mu_{\max}^*$. Therefore, the only way the best arm can be discarded from memory in any pass is if the memory was full with all arms in memory having been played $N^b/(KB)$ times without being eliminated.

To see why this implies a lower bound on the value of $\tilde{\ell}$ in any pass $b \geq 2$, observe that the best arm i^* was played $m_{i^*}^{b-1} = N^{b-1}/(KB)$ times in the previous pass, implying that the observed mean reward of the best arm i^* at the end of pass $b-1$ was $r_{i^*}^{b-1}/m_{i^*}^{b-1} \geq \mu_{\max}^* - \sqrt{(5KB\log T)/N^{b-1}}$. Therefore, the value of $\tilde{\ell}$ at the end of pass $b-1$ is at least $\mu_{\max}^* - 2\sqrt{(5KB\log T)/N^{b-1}}$, with the claim following from the fact that $\tilde{\ell}$ is a strictly increasing value. ■

We are now ready to bound the cumulative regret R_j^b due to playing any suboptimal arm $j \in \mathcal{S}$ in any pass $b \geq 2$. Let m_j^b be the final time arm j was played in round b. Since suboptimal arm j was played that one last time, it must be the case that

$$r_j^b/(m_j^b - 1) + \sqrt{(5\log T)/(m_j^b - 1)} \geq \tilde{\ell}, \tag{6}$$

where $\tilde{\ell}$ was the value of the largest lower confidence bound at that moment. Furthermore, by definition of event $\mathcal{E}$ and Claim 5, it must be that

$$r_j^b/(m_j^b - 1) < \mu_j^* + \sqrt{(5\log T)/(m_j^b - 1)}, \text{ and } \tilde{\ell} \geq \mu_{\max}^* - 2\sqrt{(5KB\log T)/N^{b-1}}. \tag{7}$$

Substituting these bounds into Equation 6, we get

$$\Delta_j < 2\sqrt{(5\log T)/(m_j^b - 1)} + 2\sqrt{(5KB\log T)/N^{b-1}},$$

and therefore, the cumulative regret due to playing arm $j \in \mathcal{S}$ in pass $b \geq 2$ is at most

$$\begin{aligned} R_j^b &= m_j^b \Delta_j \\ &< 2m_j^b \sqrt{(5 \log T)/(m_j^b - 1)} + 2m_j^b \sqrt{(5KB \log T)/N^{b-1}} \\ &< 2\sqrt{6 m_j^b \log T} + 2m_j^b \sqrt{(5KB \log T)/N^{b-1}}, \end{aligned}$$

Therefore, the total regret in pass $b \geq 2$ is given by

$$\begin{aligned} R^b &= \sum_{j \in \mathcal{S}} \mathbb{E}(R_j^b | \mathcal{E}) \\ &\leq 2 \sum_{j \in \mathcal{S}} \sqrt{6 m_j^b \log T} + \sum_{j \in \mathcal{S}} 2 m_j^b \sqrt{\frac{5KB \log T}{N^{b-1}}} \\ &\leq 2 \sum_{j \in \mathcal{S}} \sqrt{6 m_j^b \log T} + \frac{2N^b}{B} \sqrt{\frac{5KB \log T}{N^{b-1}}} \\ &\leq 2 \sum_{j \in \mathcal{S}} \sqrt{6 m_j^b \log T} + 2T^{\frac{2^B}{2^{B+1}-1}} \sqrt{\frac{5K \log T}{B}}, \end{aligned}$$

where the penultimate inequality follows due to the fact that in any pass $b \in [B]$, $\sum_{j \in \mathcal{S}} m_j^b \leq |\mathcal{S}| N^b/(KB) \leq N^b/B$, and the final inequality follows due to the fact that $N^b = T^{2^B/(2^{B+1}-1)} \sqrt{N^{b-1}}$. Therefore, the cumulative regret in the first B passes is

$$\begin{aligned} \sum_{b=2}^{B} R^b &\leq 2 \sum_{b=2}^{B} \sum_{j \in \mathcal{S}} \sqrt{6 m_j^b \log T} + \sum_{b=2}^{B} 2T^{\frac{2^B}{2^{B+1}-1}} \sqrt{\frac{5K \log T}{B}} \\ &\leq 4\sqrt{6 \log T} \sum_{b=2}^{B} \sum_{j \in \mathcal{S}} \sqrt{m_j^b} + 2T^{\frac{2^B}{2^{B+1}-1}} \sqrt{5KB \log T}. \end{aligned}$$

Due to the nature of the total number of pulls N^b in any pass $b \in [B]$, we have $\sum_{b=2}^{B} \sum_{j \in \mathcal{S}} m_j^b \leq 2N^B/B$. Applying Jensen's inequality to the concave function $f(x) = \sqrt{x}$, we have

$$\frac{1}{|\mathcal{S}|B} \sum_{b=2}^{B} \sum_{j \in \mathcal{S}} \sqrt{m_j^b} \leq \sqrt{\frac{1}{|\mathcal{S}|B} \sum_{b=2}^{B} \sum_{j \in \mathcal{S}} m_j^b} \leq \frac{1}{B} \sqrt{\frac{2N^B}{|\mathcal{S}|}},$$

giving us

$$\sum_{b=2}^{B} \sum_{j \in \mathcal{S}} \sqrt{m_j^b} \leq \sqrt{2|\mathcal{S}|N^B} \leq \sqrt{2KN^B}.$$

Now observe that $N^B = T^{1 - \frac{1}{2^{B+1}-1}}$, giving us our final bound on the cumulative regret of the algorithm in the first B passes as

$$\sum_{b=2}^{B} R^b \leq O\left(T^{\frac{2^B}{2^{B+1}-1}} \sqrt{KB \log T} \right).$$

We shall finally bound the cumulative regret $R_{\tilde{i}}$ due to playing the estimated best arm $\tilde{i}$ until the end of the time horizon. Since this estimated best arm $\tilde{i}$ was responsible for setting the final value of $\tilde{\ell}$ at the end of the B^{th} pass, it must be the case that $\mu^*_{\tilde{i}} > \tilde{\ell} > \mu^*_{\max} - 2\sqrt{(5KB\log T)/N^B}$, with the final inequality following due to Claim 5, which guarantees that

$$\Delta_{\tilde{i}} < 2\sqrt{(5KB\log T)/N^B}.$$

Furthermore, arm $\tilde{i}$ can be played at most T times after pass B until the end of the time horizon, giving us that the regret due to playing arm j

$$R_{\tilde{i}} < T\Delta_{\tilde{i}} < 2T\sqrt{(5KB\log T)/N^B} = T^{\frac{2^B}{2^{B+1}-1}}\sqrt{5KB\log T},$$

where the final inequality follows by observing $N^B = T^{1-\frac{1}{2^{B+1}-1}}$, and therefore, $\frac{T}{\sqrt{N^B}} = T^{\frac{1}{2}+\frac{1}{2(2^{B+1}-1)}} = T^{\frac{2^B}{2^{B+1}-1}}$. Combining these bounds on R^1, $\sum_{b=2}^{B} R^b$, and $R_{\tilde{i}}$, we get our claimed bound on the cumulative regret achieved by our algorithm. This concludes the proof of Theorem 5 for worst-case regret.

■

We shall now analyze the regret of Algorithm 1 when it is set to minimizing instance-dependent regret ($w = 0$).

D.3. Analysis of Algorithm 1 for Instance-Dependent Regret ($w = 0$)

The analysis of the instance-dependent regret of Algorithm 1 is conceptually identical to the previous analysis. It is based on this same intuition that the estimated best arm is a good enough proxy to eliminate distinguishable bad arms from memory, but there is this delay in information due to the best arm (or a proxy for the best arm) can be encountered very late in the stream in some pass, causing some bad arms that just transitioned from being indistinguishable in the previous pass to being distinguishable in the current pass, to be potentially overplayed but only up to a multiplicative $T^{1/(B+1)}$ factor. The additional additional B factor comes from the fact that the identities and statistics of discarded arms is forgotten, due to which the suboptimality of even distinguishable arms has to be repeatedly established in every subsequent pass. Even though the suboptimality of any distinguishable arm can be established in a future pass while incurring the optimal regret in that pass, this process has to be repeated in every pass until B. We now formally prove the claimed instance-dependent regret guarantee.

Proof [of Theorem 5 (instance-dependent upper bound)] Let $\mathcal{E} := \cap_{i\in\mathcal{K}, b\in[B]}\mathcal{E}_{i,b}$ be the good event of interest defined in Lemma 8. We shall prove that conditioned on event $\mathcal{E}$, the cumulative regret of Algorithm 1 set for instance-dependent regret minimization ($w = 0$) is

$$R_T \leq O\left(\sum_{j\in\mathcal{S}} \frac{T^{1/(B+1)}\log T + B\log\left((\Delta^*_j)^2 T/\log T\right)}{\Delta^*_j}\right),$$

with probability 1. We note that since event $\mathcal{E}$ occurs with a polynomially large probability as proved in Lemma 8, this also implies the same bound on the expected regret of the aforementioned algorithm.

Let $\mu^*_{\max} := \max_{i\in\mathcal{K}} \mu^*_i$ be the largest expected reward of any arm in the stream, and let $\mathcal{S} := \{j \in \mathcal{K} : \mu^*_j < \mu^*_{\max}\}$ be the set of all suboptimal arms, with $\Delta_j := \mu^*_{\max} - \mu^*_j$ being the regret due to playing a suboptimal arm $j \in \mathcal{S}$. Furthermore, let $i^* \in \mathcal{K} \setminus \mathcal{S}$ be any arbitrary optimal arm, which we shall henceforth refer to as the best arm.

For any pass $b \in [B]$ and any arm $j \in \mathcal{K}$, let m^b_j be the maximum number of times arm j was played in pass b. Furthermore, let $R^b_j = m^b_j \Delta_j$ be the regret incurred by the algorithm by playing arm j in the b^{th} pass, and subsequently, let $R_j = \sum_{b\in[B]} R^b_j$ be the total regret due to playing a suboptimal arm $j \in \mathcal{S}$. Finally, let $R_{\tilde{i}}$ be the regret incurred by playing the estimated best arm $\tilde{i}$ at the end of the B^{th} pass until the end of the time horizon. Therefore, we have that

$$R_T = \sum_{j\in\mathcal{S}} R_j + R_{\tilde{i}} = \sum_{j\in\mathcal{S}} \sum_{b\in[B]} R^b_j + R_{\tilde{i}}.$$

As before, we shall use a crucial implication of event $\mathcal{E}$ to bound the instance-dependent regret of our algorithm. The following claim is the analog of Claim 5 in this setting, with a minor difference due to the fact that an arm is played a maximum of N^b times in epoch b when minimizing instance-dependent regret $(w = 0)$ as compared to $N^b/(KB)$ when minimizing worst-case regret $(w = 1)$.

Claim 6 *Given that event $\mathcal{E}$ occurs, arm i^* will necessarily be played $N^b/(KB)$ times in every pass $b \in [B]$. Consequently, for any pass $b \geq 2$, we have*

$$\tilde{\ell} \geq \mu^*_{\max} - 2\sqrt{(5 \log T)/N^{b-1}},$$

at all times in pass b.

The proof of this claim follows identically to that of Claim 5. For any suboptimal arm $j \in \mathcal{S}$, we define the *distinguishing pass* b_j to be the smallest value of $b \in [B]$ such that $\Delta^*_j > 4\sqrt{(5\log T)/T^{b/(B+1)}}$. Intuitively, this represents the pass in which the precision to which we estimate the gap parameters exceeds the value of Δ^*_j, due to which it becomes possible to efficiently infer the sub-optimality of arm j. We now claim that in any pass $b > b_j$, arm j will be played $m^b_j \leq 80 \log T/(\Delta^*_j)^2$ times in that pass. To see this, observe that in any pass $b > b_j$ we have that

$$\begin{aligned}\tilde{\ell} &\geq \mu^*_{\max} - 2\sqrt{(5 \log T)/N^{b-1}} \\ &\geq \mu^*_{\max} - 2\sqrt{(5 \log T)/N^{b_j}} \\ &> \mu^*_{\max} - \Delta^*_j/2 \\ &= (\mu^*_{\max} + \mu^*_j)/2.\end{aligned}$$

Now in pass any pass $b > b_j$, event $\mathcal{E}$ further guarantees that after any n_j^b pulls of arm j, we will have

$$\begin{aligned} r_j^b/n_j^b + \sqrt{(5\log T)/n_j^b} &< \mu_j^* + 2\sqrt{(5\log T)/n_j^b} \\ &\overset{(a)}{<} \mu_j^* + \Delta_j^*/2 \\ &= (\mu_{\max}^* + \mu_j^*)/2, \end{aligned}$$

where equation (a) follows by supposing arm j was actually played $80\log T/(\Delta_j^*)^2$ times in that pass. This would guarantee that arm j will be discarded from memory after $80\log T/(\Delta_j^*)^2$ pulls. Therefore, we have that the number of times arm j is played in pass $b \in [B]$ is bounded as $m_j^b \leq N^b$ for $b \leq b_j$ and $m_j^b \leq 80\log T/(\Delta_j^*)^2$ for $b > b_j$, giving us the total regret due to playing arm j as

$$\begin{aligned} R_j &= \sum_{b\in[B]} R_j^b \\ &= \sum_{b\in[B]} m_j^b \Delta_j^* \\ &\leq \sum_{b\leq b_j} N^b \Delta_j^* + \sum_{b>b_j} \frac{80\log T}{(\Delta_j^*)^2}\Delta_j^* \\ &\leq \Delta_j^* \sum_{b=1}^{b_j} T^{b/(B+1)} + (B - b_j - 1)\frac{80\log T}{\Delta_j^*} \\ &\leq \Delta_j^* \left(\frac{T^{1/(B+1)}(T^{b_j/(B+1)} - 1)}{T^{1/(B+1)} - 1} \right) + (B - b_j - 1)\frac{80\log T}{\Delta_j^*} \\ &= \Delta_j^* \left(\left(1 + \frac{1}{T^{1/(B+1)} - 1}\right) \left(T^{b_j/(B+1)} - 1\right) \right) + (B - b_j - 1)\frac{80\log T}{\Delta_j^*} \end{aligned}$$

Observe that b_j is the smallest value of $b \in [B]$ such that $\Delta_j^* > 4\sqrt{(5\log T)/T^{b/(B+1)}}$, implying $\Delta_j^* \leq 4\sqrt{(5\log T)/T^{(b_j-1)/(B+1)}}$, giving us

$$T^{b_j/(B+1)} \leq \frac{80}{(\Delta_j^*)^2} T^{1/(B+1)} \log T, \text{ and } b_j > \frac{(B+1)}{\log T} \log\left(\frac{80\log T}{(\Delta_j^*)^2}\right).$$

Therefore, substituting these values into the above equation, and using the fact that $(B+1) \leq \log T$, we get

$$R_j \leq O\left(\frac{T^{1/(B+1)}\log T + B\log\left((\Delta_j^*)^2 T/\log T\right)}{\Delta_j^*} \right),$$

and therefore,

$$\sum_{j\in\mathcal{S}} R_j \leq O\left(\sum_{j\in\mathcal{S}} \frac{T^{1/(B+1)}\log T + B\log\left((\Delta_j^*)^2 T/\log T\right)}{\Delta_j^*} \right)$$

To bound $R_{\tilde{i}}$, observe that for arm $\tilde{i}$, it must have been the case that $\Delta^*_{\tilde{i}} < 4\sqrt{(5\log T)/N^B}$ due to Claim 6. Therefore, the regret due to playing any arm $\tilde{i}$ until the end of the time horizon can be bounded by

$$R_{\tilde{i}} \leq T\Delta^*_{\tilde{i}} \leq \frac{T}{\Delta^*_{\tilde{i}}}(\Delta^*_{\tilde{i}})^2 \leq \frac{80T^{1/(B+1)}\log T}{\Delta^*_{\tilde{i}}},$$

where the final inequality follows from the fact that $N^B = T^{\frac{B}{B+1}}$. Combining these two bounds gives us our claimed upper bound on the cumulative regret as

$$R_T \leq O\left(\sum_{j\in\mathcal{S}} \frac{T^{1/(B+1)}\log T + B\log\left((\Delta^*_j)^2 T/\log T\right)}{\Delta^*_j}\right)$$

■