Single-pass Streaming Lower Bounds for Multi-armed Bandits Exploration with Instance-sensitive Sample Complexity

Anonymous Author(s)
Affiliation
Address
email

Abstract

Motivated by applications to process massive datasets, we study streaming algorithms for pure exploration in Stochastic Multi Armed Bandits (MABs). This problem was first formulated by Assadi and Wang [STOC 2020] as follows: A collection of n arms with unknown rewards are arriving one by one in a stream, and the algorithm is only allowed to store a limited number of arms at any point. The goal is to find the arm with the largest reward while minimizing the number of arm pulls (sample complexity) and the maximum number of arms stored in the memory (space complexity). Assadi and Wang designed an algorithm for this problem that uses a memory of just one arm and still achieves the sample complexity of $O(n/\Delta_{[2]}^2)$ which is *worst-case* optimal even for non-streaming algorithms; here $\Delta_{[i]}$ is the gap between the rewards of the best and the i-th best arms.

In this paper, we extended this line of work to stochastic MABs in the streaming model with the *instance-sensitive* sample complexity, i.e. the sample complexity of $O(\sum_{i=2}^{n} \frac{1}{\Delta_{[i]}^2} \log\log(\frac{1}{\Delta_{[i]}}))$, similar in spirit to Karnin et.al. [ICML 2013] and Jamieson et.al. [COLT 2014] in the classical setting. We devise strong negative results under this setting: our results show that any streaming algorithm under a single pass has to use either asymptotically higher sample complexity than the instance-sensitive bound, or a memory of $\Omega(n)$ arms, even if the parameter $\Delta_{[2]}$ is known. In fact, the lower bound holds under much stronger assumptions, including the random order streams or a known quantity of the sample complexity. We complement our lower bounds by proposing a new algorithm that uses a memory of a single arm and achieves the instance-optimal sample complexity when all the strong assumptions hold simultaneously.

Our results are developed based on a novel *arm trapping lemma*. This generic complexity result shows that any algorithm to *trap* the index of the best arm among $o(n)$ indices (but not necessarily *find* it) has to use $\Theta(n/\Delta_{[2]}^2)$ sample complexity. This result is *not* restricted to the streaming setting, and to the best of our knowledge, this is the first result that captures the sample-space trade-off for 'trapping' arms in multi-armed bandits, and it can be of independent interest.

1 Introduction

The stochastic multi-armed bandits (MABs) is one of the most well-known models in machine learning and theoretical computer science. In this model, we are given n arms with unknown reward distributions, and we are required to find some arms toward certain goals. The *pure exploration* version of the stochastic MABs studies how to find the best arm(s), and it has been extensively investigated in the literature [19, 32, 5, 24, 25, 22, 12, 27, 9, 1, 14]. The model captures many

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

fundamental problems in finding the best items among many, has various applications in experiment design [34, 7], crowdsourcing [15, 41, 8], search ranking [18, 3, 33], networking [40, 38], and others.

We study the problem of finding the best (*one*) arm in stochastic MABs. Since the distribution of the rewards is unknown, the natural strategy is to pull each arm multiple times and use the empirical rewards to find the best arm. Accordingly, the goal of the algorithms in this problem is to identify the best arm with a high constant probability and a number of arm pulls (sample complexity) that is as small as possible. Following the classical upper bound of [19] and lower bound of [32], it is well known that $\Theta(n/\Delta_{[2]}^2)$ arm pulls are both necessary and sufficient – throughout, $\Delta_{[i]}$ is the gap between the rewards of the best and the i-th-best arms. Going beyond worst-case, the main focus of the literature has been on obtaining *instance-sensitive* algorithms for the problem, resulting in algorithms of [22] and [25] with $O(\sum_{i=2}^{n} \frac{1}{\Delta_{[i]}^2} \log\log\left(\frac{1}{\Delta_{[i]}}\right))$ arm pulls (see also [14] and references therein for various extensions). This upper bound has at most a doubly-logarithm gap from known instance-optimal lower bounds [32, 22, 14].

Almost the entirety of algorithms in the literature for best arm identification require the entire set of the arms to be repeatedly visited, and therefore, a memory of $\Omega(n)$ arms is inevitably needed. More recently, motivated by applications to processing massive number of arms (e.g., in online advertisements) or the situations when the arms are arriving online (e.g. crowd-sourcing), the *space* requirements of algorithms for stochastic MABs has also received considerable attention [4, 30, 31, 11, 29, 2]. In particular, [4] recently formulated and studied the problem of pure exploration in *streaming* stochastic multi-armed bandits. In their model, the arms arrive one by one in a stream, and the algorithm can only output or pull a past arm if it is stored. The largest number of arms stored at any point is named the *space complexity*; see also [30, 29, 23]. [4] designed an algorithm that given the knowledge of $\Delta_{[2]}$, finds the best arm with high constant probability using $O(n/\Delta_{[2]}^2)$ arm pulls and a memory of a *single* arm.

The algorithm of [4] achieves the worst-case optimal sample and space complexity simultaneously, thus asserting that no trade-off is needed between the sample and space for this problem. On the flip side, this upper bound is only worst-case optimal and *not* instance-sensitive. Very recently, [23] shows that if we are allowed to make $O(\log(1/\Delta_{[2]}))$ passes over the stream, it is possible to find the best arm with the instance-sensitive sample complexity in the same spirit of [22, 25], and only use a single-arm memory. However, allowing $O(\log(1/\Delta_{[2]}))$ passes seems to be a strong condition especially in streaming applications. Therefore, a natural open question is *whether we can achieve instance-sensitive sample complexity, while maintaining a small memory, preferably still a single arm, in a single pass of the streaming model*.

The main result of this paper is that finding the best arm with the instance-sensitive sample complexity in a single pass is indeed *hard*: if we are only given the parameter $\Delta_{[2]}$, any single-pass streaming algorithm must either use asymptotically higher sample complexity than the instance-sensitive bound, or use a memory of $\Omega(n)$ arms. Furthermore, the lower bound holds even if we are given additional strong assumption of either (i) the arms arrive in a random order or (ii) a known quantity of the sample complexity. En route to proving the above lower bounds, we also provide an *even stronger* lower bound for single-pass streaming algorithms without any parameter (not even $\Delta_{[2]}$): we show that in such a case, any single-pass streaming algorithm with a memory of $o(n)$ has to use an *unbounded* sample complexity. Finally, we complement our lower bounds with a single-pass streaming algorithm that uses a single-arm memory and achieves the instance-sensitive sample complexity, if given the parameter $\Delta_{[2]}$ and conditions of *both* (i) and (ii).

1.1 Instance-sensitive Sample Complexity

As stated earlier, the worst-case sample complexity of stochastic MABs is $\Theta(n/\Delta_{[2]}^2)$. However, in most instances of interest, one can hope for much better sample complexity that exploits the rewards of different arms in a more fine-grained way. This is the basic idea behind *instance-sensitive* algorithms that have received significant attention in the context of stochastic MABs. We follow the same approach in this paper. In particular, we aim to achieve the sample complexity of $O\big(\sum_{i=2}^{n} \frac{1}{\Delta_{[i]}^2} \log\log(\frac{1}{\Delta_{[i]}})\big)$ arm pulls – the same bound achieved by classical algorithms for non-streaming stochastic MABs [22, 25].

Following the standard notation used in the literature [22, 25, 12], we denote the instance-sensitive complexity as follows:

$$\mathbf{H}_1 := \sum_{i=2}^{n} \frac{1}{\Delta_{[i]}^2}, \qquad \mathbf{H}_2 := \sum_{i=2}^{n} \frac{1}{\Delta_{[i]}^2} \log\log\left(\frac{1}{\Delta_{[i]}}\right).$$

Thus, the algorithms of [22, 25] use $O(\mathbf{H}_2)$ arm pulls. The term '(almost) instance-sensitive' is derived from 'instance-optimal' in the literature [12, 1, 4]. We use such a term since there is no known instance-*optimal* upper bound for MAB; this term is shared by recent work like [26]. Moreover, since there is a constant difference between $\mathbf{H}_2$ and the instance-sensitive sample complexity (i.e., $\mathbf{H}_2$ vs. $O(\mathbf{H}_2)$), we denote the latter as *INST-complexity* in this work.

1.2 Our Contributions

Our main contribution is a set of lower bounds showing that it is almost impossible to achieve the instance-sensitive sample complexity with low memory for any single-pass streaming algorithm. Moreover, if no parameter is provided (especially $\Delta_{[2]}$ is not known), it is impossible to achieve any bounded sample complexity with low memory:

Result 1 (Informal). *There are families of stochastic multi-armed bandits instances such that:*

- (i) *If the knowledge of $\Delta_{[2]}$ is provided, but the knowledge of INST-complexity and the random arrival of arms are **not both** provided, the sample complexity of any single-pass streaming algorithm with $o(n)$-arm memory is $\omega(\mathbf{H}_2)$.*
- (ii) *If the knowledge of $\Delta_{[2]}$ is not provided and no other parameter is provided, the sample complexity of any single-pass streaming algorithm with $o(n)$-arm memory is **unbounded** as a function of n and $\Delta_{[2]}$;*

Note that (i) of Result 1 captures three cases: only $\Delta_{[2]}$ is known; $\Delta_{[2]}$ is known and the stream is in a random order; and $\Delta_{[2]}$ is known and the knowledge of INST-complexity is provided. Compared to the positive results in [4], our lower bounds rule out any instance-sensitive algorithms under the same parameter settings. Moreover, they provide a strong justification for the knowledge of $\Delta_{[2]}$ in the algorithm of [4] (and any streaming stochastic multi-armed bandit algorithms).

Another remark of Result 1 is the *separation* of the complexity for multi-armed bandits under the streaming and the classical models. The algorithm in [4] implies that with the worst-case sample complexity and a known $\Delta_{[2]}$, there is no separation of hardness for pure exploration MABs between the streaming and the classical settings, i.e. no sample-space trade-off. On the other hand, our results suggest that if we switch to the regime of instance-sensitive sample complexity, or if we remove the assumption of $\Delta_{[2]}$, a sample-space trade-off starts to emerge (note that there exist algorithms that achieve $\mathbf{H}_2$ sample complexity *without* the knowledge of $\Delta_{[2]}$ under the classical setting).

To complement the lower bounds, we design an algorithm that uses a single arm memory and achieves instance-sensitive sample complexity when $\Delta_{[2]}$ is not too small. The algorithm is conditioned on the knowledge of $\Delta_{[2]}$ and *both* the random order stream and the knowledge of INST-complexity.

Result 2 (Informal). *Assuming the knowledge of $\Delta_{[2]}$ and INST-complexity, and when the arms arrive in a random order, there is a streaming algorithm that finds the best arm in streaming stochastic multi-armed bandits with $O\left(\mathbf{H}_2(=\sum_{i=2}^{n}\frac{1}{\Delta_{[i]}^2}\log\log\left(\frac{1}{\Delta_{[i]}}\right)) + \text{poly}(\frac{\log n}{\Delta_{[2]}})\right)$ arm pulls and a memory of a single arm with high (constant) probability.*

We should note that the sample complexity of our algorithm in Result 2 has an additive term compared to the $O(\mathbf{H}_2)$ obtained in the non-streaming setting [22, 25]. This term however is typically negligible except for pathological cases where $\Delta_{[2]}$ is extremely small compared to both average values of $\Delta_{[i]}$'s and the number of arms.

A key ingredient of our proof towards Result 1 is an arm-trapping lemma that characterizes the hardness to *keep* the arm in the memory, shown as below:

Result 3 (Arm-trapping lemma, informal). *There are families of stochastic multi-armed*

bandits instances such that any algorithm that outputs $o(n)$ *arms that* contains the index *of the best arm (with high probability) has to use* $\Omega\left(\frac{n}{\Delta_{[2]}^2}\right)$ *arm pulls. As a result, any single-pass streaming algorithm that keeps the best arm by the end of the stream (with high probability) has to use either* $\Omega\left(\frac{n}{\Delta_{[2]}^2}\right)$ *arm pulls or an* $\Omega(n)$*-arm memory.*

To the best of our knowledge, this is the first Multi-armed Bandits lower bound that explicitly establishes trade-off between sample and space complexity for *trapping* an arm. Note that this statement is stronger than the usual lower bounds on *finding* arms: our lower bound holds even if the algorithm is only required to output a collection of indices that contains the index of the best arm (possibly without *knowing* which arm is the best).

1.3 Related Work

Originated in the work of Robbins [34] almost 70 years ago, the study of stochastic MABs basically follows two popular branches: the pure exploration problem, which aims to find the best arm [19, 32, 12, 27, 9, 1, 14]; and the exploration-exploitation problem, which focus on the regret minimization [39, 6, 28, 10, 17]. Covering the vast literature on regret minimization is beyond the scope of this paper but we should note that even regret minimization is studied in streaming setting although mostly for *multi-pass* algorithms [30, 11, 29, 2].

The worst-case optimal bounds on the sample complexity of exploration problem was established by [19] (upper bound) and [32] (lower bound). Besides the best arm identification, [19] also introduced the ε-best arm variation of the pure exploration problem. The problem is motivated by learning under the PAC(ε, δ)-framework, where the algorithm aims to find an arm with reward at most ε-far from the best arm. The ε-best arm problem is also studied extensively. It has recently been considered in the streaming model as well by [4, 30] who designed streaming algorithms for ε-best arm problem with $O(\log^*(n))$ memory and $O(n/\varepsilon^2)$ arm pulls (we note that [30] also designed a heuristic for this problem on random streams with $O(1)$ memory that works well empirically but does not have the same theoretical guarantee).

The instance-sensitive sample complexity of $O(\mathbf{H}_2)$ was first achieved by [25], followed up by [22] who achieved the same asymptotic complexity but a better performance in practice. [22] also discovered that for the problem with 2 arms, $\Omega(\frac{1}{\Delta_{[2]}^2}\log\log(\frac{1}{\Delta_{[2]}}))$ arm pulls are necessary – a lower bound that is due to a classical result of [20] from over half of a century ago. [22] conjectured that the lower bound holds for n arms (and the complexity will be 'instance-*optimal*'); nonetheless, a later work by [12] shows that for some instances, an algorithm with $O(\sum_{i=2}^{n}\frac{1}{\Delta_{[i]}^2})$ arm pulls is possible and formulated a 'gap-entropy' conjecture for capturing the instance-optimal bounds (see [14] for further progress on this conjecture).

2 Preliminaries

Problem Definition

The formal description of the stochastic multi-armed bandit model that we study is as follows.

Problem 1 (Pure exploration MABs)**.** *There is a set of* n *arms* $\{\mathsf{arm}_i\}_{i=1}^n$ *with unknown rewards, each following a sub-Gaussian distribution with means* $\{p_i\}_{i=1}^n$*. Our goal is to find the best arm with the highest expected reward, i.e.,* p_i*, by pulling them and observing the empirical rewards.*

We study the problem under the *streaming* model, which means the arms arrive one by one from the set; for each arriving arm, the algorithm can pull it multiple times to observe the empirical reward or store it to the memory (so it can be pulled later). However, if an arm is not stored, it is lost forever and the algorithm can neither pull nor restore it. We refer to the number of arm pulls by the algorithm as the *sample complexity*, and the maximum number of arms stored as the *space complexity*.

The sample complexity of stochastic MABs is characterized by the *gap parameters* $\{\Delta_{[i]}\}_{i=2}^n$, which denotes the difference between the reward of the best arm and the arm with the i-th highest reward. By definition, we have $\left(\Delta_{[2]} \leq \Delta_{[3]} \leq \cdots \leq \Delta_{[n]}\right)$. We assume that these parameters are greater than 0 – an assumption that is common in the literature [19, 24, 16, 36]. Note that the gap parameters are not necessarily *given* to the algorithm, although it is common to assume the knowledge of $\Delta_{[2]}$. In fact, we provide a comprehensive treatment to the necessity of the $\Delta_{[2]}$ in this work.

Due to space limitations, we postpone the technical preliminaries to Appendix A.

3 The Arm Trapping Lemma

Our lower bounds are built upon an arm trapping lemma, which is a information-theoretic sample-space trade-off for any algorithm to 'trap' the best arm (*not* restricted to the streaming setting). Compared to the classical lower bounds for finding arms in stochastic MABs, our lemma stands out by showing that even *trapping* the best arm (possibly without knowing which one is the best among the trapped arms) requires a lot of arm pulls. As a direct corollary, we can show that for any streaming algorithm with a memory of $o(n)$ arms, *storing* the best arm in the stream takes $\Omega(\frac{n}{\Delta_{[2]}^2})$ arm pulls.

Our hard distribution for the arm-trapping lemma is quite natural: consider a collection of n arms with one of them having a reward of $\frac{1}{2}$, and all other arms having a reward of $\frac{1}{2} - \beta$ for some β. The high level idea behind the proof is a non-standard reduction from this problem to the problem of distinguishing whether the reward of a single arm is $1/2$ or $1/2 - \beta$; a problem that provably requires $\Omega(1/\beta^2)$ samples. The reduction places this "special arm" in a collection of $n - 1$ arms with reward $1/2 - \beta$, uniformly permute them, and then gives them as input to any algorithm that can supposedly contradict Lemma 3.1. The proof then shows that if the algorithm can only place this special arm with non-trivial probability among the output 'trap' set, this arm actually has reward of $1/2$ and not $1/2 - \beta$. Combining this with a simple 'direct sum style' argument then proves the lower bound.

We now formally define the hard family of instances as follows.

`DIST`: **A hard distribution to trap the best arm**

1. An index $i^\star$ sampled uniform at random from $[n]$.
2. For $i \neq i^\star$, let the arms be with reward $p_i = \frac{1}{2} - \beta$.
3. For $i = i^\star$, let the arm be with reward $p_{i^\star} = \frac{1}{2}$.

Based on the definition of `DIST`, we can give the formal statement of our streaming arm trapping lemma.

Lemma 3.1 (Arm Trapping Lemma)**.** *Any algorithm that outputs (indices of) $\frac{n}{8}$ arms which contains the best arm on `DIST` with probability at least $\frac{2}{3}$ has to use at least $\frac{1}{1200} \cdot \frac{n}{\beta^2}$ arm pulls.*

The proof of Lemma 3.1 follows a (non-standard) reduction from the complexity to distinguish two arms with a β-reward gap, specified as follows:

Lemma 3.2. *Consider an arm with the reward from the following distribution*

- $p = \frac{1}{2}$*, with probability* $\frac{1}{2}$*;*
- $p = \frac{1}{2} - \beta$*, with probability* $\frac{1}{2}$*.*

Any algorithm to determine the reward of the arm with $\beta < \frac{1}{6}$ and a success probability of at least $\frac{7}{12}$ has to use $\frac{1}{144} \cdot \frac{1}{\beta^2}$ arm pulls.

Lemma 3.2 is an information-theoretic lower bound similar to the ones in [32]. The bound shows that for any algorithm to distinguish an arm with reward either $\frac{1}{2}$ or $(\frac{1}{2} - \beta)$, it has to pull the arm $\frac{n}{\beta^2}$ times (if β is sufficiently small). This argument is standard and is provided for completeness as similar results, such as classical Farrell's bound [20, 22, 12], have just subtle differences with this bound, which in fact prohibits us from using them. We defer the proof of Lemma 3.2 to Appendix C.

Proof of Lemma 3.1. We prove Lemma 3.1 by a reduction argument: we show that if there exists such an algorithm (we call it `ALG`) to trap the best arm in `DIST` with probability at least $\frac{2}{3}$ and $\frac{1}{1200} \cdot \frac{n}{\beta^2}$ arm pulls, it implies an algorithm that determines the reward of the arm in Lemma 3.2 with probability at least $\frac{7}{12}$ and less than $\frac{1}{144} \cdot \frac{1}{\beta^2}$ arm pulls, which is impossible by the statement of Lemma 3.2.

The reduction is given as follows: suppose such an `ALG` to trap the best arm exists, then we design an algorithm `ALG'` to determine the reward of the special arm (denote it as $\widetilde{\mathsf{arm}}$) as follows.

- `ALG'` samples $i^\star \in [n]$ uniformly at random, and puts $\widetilde{\mathsf{arm}}$ at position $i^\star$.

- ALG' creates $(n-1)$ 'dummy arms' with rewards $(\frac{1}{2}-\beta)$, and puts them into the positions $\mathsf{arm}_1, \cdots, \mathsf{arm}_{i^\star-1}, \mathsf{arm}_{i^\star+1}, \cdots, \mathsf{arm}_n$; note that any arm pulls on these arms does *not* count towards the sample complexity on $\widetilde{\mathsf{arm}}$.
- ALG' runs ALG with the instance (*without revealing* $i^\star$), and output with the following rules
  1. If ALG makes more than $\frac{1}{1200} \cdot \frac{n}{\beta^2}$ arm pulls at any time, terminate and output $\widetilde{p} = \frac{1}{2} - \beta$;
  2. Else, if ALG makes more than $\frac{1}{150} \cdot \frac{1}{\beta^2}$ arm pulls on $\widetilde{\mathsf{arm}}$, terminate and output $\widetilde{p} = \frac{1}{2}$;
  3. Otherwise (if ALG makes at most $\frac{1}{150} \cdot \frac{1}{\beta^2}$ arm pulls on $\widetilde{\mathsf{arm}}$), output with the following rules
     - $\widetilde{p} = \frac{1}{2}$ if $\widetilde{\mathsf{arm}}$ is among the arms returned by ALG.
     - $\widetilde{p} = \frac{1}{2} - \beta$ if $\widetilde{\mathsf{arm}}$ is *not* among the arms returned by ALG.

It is straightforward to see that the number of arm pulls on $\widetilde{\mathsf{arm}}$ is at most $\frac{1}{150} \cdot \frac{1}{\beta^2}$, as we explicitly terminate the algorithm once the sample complexity exceeds this limit.

We now verify that ALG' determines the reward of the arm with probability at least $\frac{7}{12}$. We first show that if $p = \frac{1}{2}$, the error probability is at most $\frac{1}{3}$. To see this, note that for ALG' to make mistakes when $p = \frac{1}{2}$, ALG needs to either makes more than $\frac{1}{1200} \cdot \frac{n}{\beta^2}$ arm pulls, or fail to include $\widetilde{\mathsf{arm}}$ among the arm it stores. By the assumption of the algorithm, the probability for either of the cases to happen is at most $\frac{1}{3}$.

On the other hand, when $p = \frac{1}{2} - \beta$, we show that the error probability is at most $\frac{1}{2}$. To see this, note that when $p = \frac{1}{2} - \beta$, ALG' could make mistakes on the following occasions.

- ALG fails (returns arbitrary results, but terminates with $\frac{1}{1200} \cdot \frac{n}{\beta^2}$ arm pulls).
- ALG returns correctly, but it also includes $\widetilde{\mathsf{arm}}$.
- ALG returns correctly, but uses more than $\frac{1}{40} \cdot \frac{1}{\beta^2}$ arm pulls on $\widetilde{\mathsf{arm}}$ such that ALG' falsely terminates and returns $p = \frac{1}{2}$.

The probability for the first scenario to happen is clearly at most $\frac{1}{3}$. For the second scenario, note that $\widetilde{\mathsf{arm}}$ has the same reward as other arms, and they become identical random variables. Therefore, we have

$$\begin{aligned}\Pr\left(\widetilde{\mathsf{arm}} \text{ included by } \texttt{ALG} \mid p = \frac{1}{2} - \beta\right) &= \Pr\left(\mathsf{arm}_j \text{ included by } \texttt{ALG} \mid p_j = \frac{1}{2} - \beta\right) \\ &= \Pr\left(\mathsf{arm}_j \text{ included by } \texttt{ALG}, \forall j \in [n]\right)\end{aligned}$$

Since ALG only returns $\frac{1}{8}$ fraction of the arms, the probability for $\widetilde{\mathsf{arm}}$ to be falsely included to the output is $\frac{1}{8}$.

For the third scenario, we show that conditioning on $p = \frac{1}{2} - \beta$, it is with lower probability that ALG excessively makes arm pulls on $\widetilde{\mathsf{arm}}$ while keeping the overall number of arm pulls low. By an argument that is similar to the previous scenario, the number of arm pulls on $\widetilde{\mathsf{arm}}$ is an identical random variable with the number of arm pulls on any of the other arms. Therefore, denote $T_{i^\star}$ as the number of arm pulls on $\widetilde{\mathsf{arm}}$ and T_j the number of arm pulls on any arm j, we have

$$\begin{aligned}\mathbb{E}\left[T_{i^\star} \,\middle|\, p = \frac{1}{2} - \beta\right] &= \mathbb{E}\left[T_j \,\middle|\, p_j = \frac{1}{2} - \beta, \forall j \in [n]\right] \\ &= \mathbb{E}\left[T_j\right], \forall j \in [n]\end{aligned}$$

Hence, the number of arm pulls on $\widetilde{\mathsf{arm}}$ under $p = \frac{1}{2} - \beta$ equals to the expectation on each arm (the *average*). More formally, we have

$$\mathbb{E}\left[\text{number of total arm pulls of } \texttt{ALG} \,\middle|\, p = \frac{1}{2} - \beta\right] = \mathbb{E}\left[\sum_{j=1}^{n} T_j \,\middle|\, p_j = \frac{1}{2} - \beta\right]$$

$$= \sum_{j=1}^{n} \mathbb{E}\left[T_j \,\middle|\, p_j = \frac{1}{2} - \beta\right]$$

(linearity of expectation)

$$= n \cdot \mathbb{E}\left[T_{i^\star} \,\middle|\, p_{i^\star} = \frac{1}{2} - \beta\right].$$

(identical random variables with equal expectations)

That is to say, conditioning on $p = \frac{1}{2} - \beta$, we have $\mathbb{E}\left[T_{i^\star} \,|\, p = \frac{1}{2} - \beta\right] = \frac{\mathbb{E}[\text{number of total arm pulls of } \texttt{ALG}]}{n} = \frac{1}{1200} \cdot \frac{1}{\Delta^2}$ arm pulls. By Markov bound, the probability for the number of arm pulls on $\widetilde{\text{arm}}$ to be more than $\frac{1}{150} \cdot \frac{1}{\Delta^2}$ is at most $\frac{1}{8}$. Therefore, by summarizing the above scenarios, we conclude that the probability for `ALG'` to make mistake under $p = \frac{1}{2} - \beta$ is bounded by $\frac{1}{3} + \frac{2}{3} \cdot (\frac{1}{8} + \frac{1}{8}) = \frac{1}{2}$.

Summarizing the above two cases on p gives us the failure probability of at most

$$\begin{aligned}\Pr(\texttt{ALG'} \text{ fails}) &= \frac{1}{2} \cdot \Pr\left(\texttt{ALG'} \text{ fails} \mid p = \frac{1}{2}\right) + \frac{1}{2} \cdot \Pr\left(\texttt{ALG'} \text{ fails} \mid p = \frac{1}{2} - \beta\right) \\ &= \frac{1}{2} \cdot \frac{1}{3} + \frac{1}{2} \cdot \frac{1}{2} = \frac{5}{12}.\end{aligned}$$

As such, `ALG'` can determine the reward of the arm in Lemma 3.2 with probability at least $\frac{7}{12}$ and $\frac{1}{150} \cdot \frac{1}{\Delta^2}$ arm pulls, which contradicts the lemma itself. Therefore, the lower bound is obtained. □

As a natural corollary of Lemma 3.1, any single-pass streaming algorithm to *store* the best arm with a memory of $o(n)$ arms has to use $\Omega\left(\frac{n}{\beta^2}\right)$ arm pulls. We formalize this result below as Corollary 3.3, and will use it as a standard tool in the proofs of the lower bounds.

Corollary 3.3. *Any streaming algorithm that stores the best arm on* `DIST` *with a memory of* $\frac{n}{8}$ *arms with probability at least* $\frac{2}{3}$ *has to use* $\Omega\left(\frac{n}{\beta^2}\right)$ *arm pulls.*

Remark 3.4. Note that it is possible to prove Lemma 3.1 without the reduction argument. For instance, one can use the approaches developed by [27, 13] to obtain a similar result. The merit of our proof is that it provides a *black-box* way to prove MAB lower bounds without resorting to previous techniques. Furthermore, the idea to prove lower bounds by 'embedding' a single instance into multiple copies and using the 'direct sum' argument is extensively used in the theoretical computer science literature. Our proof shows that such an idea works for machine learning lower bounds, which gives an interesting connection between the areas. We also note that very recently, [37] independently developed a proof for another problem with the technique similar to ours.

4 Streaming Lower Bounds for Multi-armed Bandits with Instance-sensitive Sample Complexity

In this section, we provide the main lower bounds for single-pass streaming MABs. On the high level, our lower bound states that for any streaming algorithm with a memory of $o(n)$ arms that finds the best arm, if some parameters are not given, the sample complexity of the algorithm can be far more than the instance-sensitive $O(\mathbf{H}_2)$ or even unbounded. We emphasize that such lower bounds are different from standard lower bounds in streaming setting, as they need to take into account sample complexity of the algorithm as well, i.e., establish a sample-space trade-off.

Remark 4.1. To be consistent with the definition of streaming MAB, in our lower bounds, we only measure the space of the algorithms in terms of the number of arms they store; the algorithms otherwise are allowed to maintain any arbitrary statistics on the input. What we show in all these lower bounds is that if the sample complexity is "small", then at some point the algorithm must have mistakenly discarded the best arm from its memory without the knowledge of even the index of this arm; thus at the end of the stream, the algorithm has no way of recovering the best arm (even given the arbitrary statistics it has collected).

4.1 A lower bound without the knowledge of $\Delta_{[2]}$ and other parameters

We begin with showing that if no additional parameter is provided other than the stream itself, the sample complexity can be *arbitrarily* high, and even $O(n/\Delta_{[2]}^2)$ arm pulls are not sufficient. In fact,

we show that in this case, the sample complexity cannot be bounded as a function of n and $\Delta_{[2]}$. In contrast, the algorithm in [4], which assumes the knowledge of $\Delta_{[2]}$, achieves $O(n/\Delta_{[2]}^2)$ arm pulls with the memory of a single arm. Therefore, our lower bound asserts the importance of $\Delta_{[2]}$ for exploring the best arm under the streaming model. This shows that the knowledge of $\Delta_{[2]}$ plays a very different role in the streaming model vs classical (RAM) model.

Theorem 1. *For any function $f : (0,1) \to (0,1)$, there always exists a family of streaming stochastic multi-armed bandit instances, such that any algorithm to find the best arm with a memory of $\frac{n-1}{8}$ arms and a success probability of at least $\frac{5}{6}$ has to use $\omega\left(\frac{n}{f(\Delta_{[2]})}\right)$ arm pulls, if no further knowledge other than the arms themselves are provided.*

We defer the proof of Theorem 1 to Appendix B.1. We remark that even with the random arrival of arms, the sample lower bound in Theorem 1 still holds. The proof is almost identical to the case with the known $\Delta_{[2]}$, and the techniques will be clear once we present the results in Theorem 3. To avoid redundant repetition of the details, we omit the proof for Theorem 1 with random arrival.

4.2 Instance-sensitive Lower bounds with the knowledge of $\Delta_{[2]}$

Going beyond the case where none of the conditions are given, we now show the lower bound with $\Delta_{[2]}$ provided. Recall that [4] already solved the problem with $O(\frac{n}{\Delta_{[2]}^2})$ arm pulls; therefore, it is pointless to expect any result as strong as Theorem 1. Instead, we prove lower bound of $\omega(\mathbf{H}_2)$ by exploiting the gap between $\Theta\left(n/\Delta_{[2]}^2\right)$ and $\Theta\left(\mathbf{H}_2\right)$. We first present the lower bound with *only* the knowledge of $\Delta_{[2]}$.

Theorem 2. *There exists a family of streaming stochastic multi-armed bandit instances, such that given the knowledge of $\Delta_{[2]}$, any algorithm to find the best arm with a memory of $\frac{n-2}{8}$ arms and a success probability of at least $\frac{5}{6}$ has to use $\omega\left(\mathbf{H}_2\right) = \omega\left(\sum_{i=2}^{n} \frac{1}{\Delta_{[i]}^2} \log\log\left(\frac{1}{\Delta_{[i]}}\right)\right)$ arm pulls.*

We have noted that it is pointless to expect any result as strong as Theorem 1. Indeed, the hard distribution in Section 4.1 does not work anymore – by the change of the reward of the final arm, the value of $\Delta_{[2]}$ is β in one case and $\frac{1}{6} >> \beta$ in the other. Therefore, by simply checking the parameter, an algorithm can determine whether to store the best arm among the first $(n-1)$ arms, which makes the distribution impossible to 'fool' the algorithm.

As an alternative idea, we modify the distribution of the instances to satisfy the following properties. Firstly, it forces the algorithm to make $\Omega(\frac{n}{\Delta_{[2]}^2})$ arm pulls to find the best arm. Secondly, it leaves a large gap between $\Delta_{[2]}$ and other $\Delta_{[i]}$, $\forall i \geq 2$ such that there is $\frac{n}{\Delta_{[2]}^2} = \omega(\mathbf{H}_2)$. And finally, $\Delta_{[2]}$ is static (the same version) for all the instances. We defer the formal proof of Theorem 2 to Appendix B.2.

We now provide lower bounds with either the Random Arrival or the quantity of $\mathbf{H}_2$ (but *not both*) in Theorem 3 and Theorem 4.

Theorem 3. *There exists a family of streaming stochastic multi-armed bandit instances, such that given the knowledge of $\Delta_{[2]}$ and the random arrival of arms, any algorithm to find the best arm with a memory of $(\frac{n}{24} - 1)$ arms and a success probability of at least $\frac{80}{81}$ has to use $\omega\left(\mathbf{H}_2\right) = \omega\left(\sum_{i=2}^{n} \frac{1}{\Delta_{[i]}^2} \log\log\left(\frac{1}{\Delta_{[i]}}\right)\right)$ arm pulls.*

Theorem 4. *There exists a family of streaming stochastic multi-armed bandit instances, such that given the knowledge of $\Delta_{[2]}$ and the INST-complexity $= O(\mathbf{H}_2)$, any algorithm to find the best arm with a memory of t arms for $s = o(n)$ and a success probability of at least $\frac{8}{9}$ has to use $\Omega\left(\frac{n}{s} \cdot \mathbf{H}_2\right) = \omega\left(\mathbf{H}_2\right)$ arm pulls.*

The idea to prove the lower bound with random arrival is quite simple: we show that with at least a constant probability, the distribution of an instance with random arrival will resemble the instance we used for Theorem 2. More specifically, suppose each arm arrives at any position with probability $\frac{1}{n}$. Then, with probability $\frac{1}{3}$, the arm with reward $\frac{1}{2}$ appears in the first $\frac{1}{3}$-fraction of the stream; Similarly, with probability $\frac{2}{3} \cdot \frac{2}{3}$, both of the last two arms in the distribution of Theorem 2 will not

pop up until the late $\frac{2}{3}$ fraction of the stream. The family of the instances with both of the above properties indeed resembles the distribution we used earlier, and we can show that the algorithm is forced to use a lot of arm pulls under this scenario. The hard instance for Theorem 4 is developed with a different idea. Due to space limits, we defer the proofs to Appendix B.

5 An Instance-sensitive Streaming Algorithm

We begin with introducing notations used in this section. We denote the best arm with arm^*. Furthermore, we denote the instance-sensitive complexity (*INST-complexity*) for solving the problem with probability of success at least $1-\delta$ as $\mathbf{H}_2^\delta := \sum_{i=2}^n \frac{1}{\Delta_{[i]}^2} \log\left(\frac{1}{\delta}\log\left(\frac{1}{\Delta_{[i]}}\right)\right)$. Note $\mathbf{H}_2$ and $\mathbf{H}_2^\delta$ only differ from the correct probability; if one sets δ to be a constant, $\mathbf{H}_2$ and $\mathbf{H}_2^\delta$ become the same asymptotically.

As mentioned before, our algorithm makes two assumptions on the stream of arms, namely:

(i) *Known INST-complexity*: The parameter $\mathbf{H}_2^\delta$ (or even some $O(1)$-approximation of it) is given;

(ii) *Random arrival stream*: the stream is a random permutation of the arms $\{\mathsf{arm}_i\}_{i=1}^n$.

Although the assumptions are strong, they are also common in the study of streaming models as well as stochastic MABs. The formal guarantees of our algorithm is as follows.

Theorem 5. *There is a streaming algorithm that for any confidence parameter $\delta > 0$, given n arms in a random order, gap parameter $\Delta_{[2]}$, and INST-complexity $\mathbf{H}_2^\delta = \sum_{i=2}^n \frac{1}{\Delta_{[i]}^2} \log\left(\frac{1}{\delta}\log\left(\frac{1}{\Delta_{[i]}}\right)\right)$, finds the best arm w.p. at least $1-\delta$ using $O\left(\mathbf{H}_2^\delta + \frac{1}{\Delta_{[2]}^2} \cdot \log(\frac{n}{\delta})^2 \cdot \log^2\left(\frac{1}{\delta}\log\left(\frac{1}{\Delta_{[2]}}\right)\right)\right)$ arm pulls and a memory of a single arm.*

The algorithm in Theorem 5 contains an additive term in sample complexity beside the standard $\mathbf{H}_2^\delta$ term. For most reasonable choices of parameters, this additive term is in fact negligible—in other words, our streaming algorithm in Theorem 5 achieves the same sample complexity as that of non-streaming algorithms of [25, 22]. Limited by the space, we defer the details and the analysis of the algorithm to Appendix D.

6 Discussion and Conclusion

In this paper, we studied the limitations for streaming MABs algorithm with the instance-sensitive sample complexity in a single pass. We proved a collection of lower bounds showing that achieving the instance-sensitive sample complexity with a small memory in a single-pass stream is *hard*: with the absence of various (subsets of) assumptions, including the knowledge of $\Delta_{[2]}$, the quantity of the sample complexity, and the random arrival of arms, no algorithm can achieve the instance-sensitive bound with $o(n)$-arm memory. Furthermore, as a byproduct of the lower bounds, we establish the crucial role of $\Delta_{[2]}$ in streaming MABs: without such a parameter, any algorithm with a memory even as large as a constant fraction of n cannot achieve any bounded sample complexity. We complemented our lower bounds by an algorithm that achieves the instance-sensitive bound under most realistic parameters and with a memory of a single arm when all assumptions hold. Our lower and upper bounds demonstrate a sharp dichotomy.

Our results open the door for various future work in this line of research. One open problem posed by our results is whether one can find a lower bound (or an instance-sensitive streaming algorithm) without the knowledge of $\Delta_{[2]}$, but with the knowledge of the quantity of the sample complexity and the random arrival of arms. A bizarre scenario as it may sound, if one could prove the lower bound with these conditions (or find an algorithm that finds the best arm with $O(\mathbf{H}_2)$ sample complexity), one can conclude and complete the problem of instance-sensitive exploration in single-pass streaming stochastic multi-armed bandits. Another interesting problem is whether one can refine the streaming algorithm (or design a new one) that drops the additive term on the sample complexity. Finally, a more challenging open problem is to understand the tight number of passes that are *necessary* and *sufficient* to achieve the instance-sensitive sample complexity (with no or minimum assumptions) under the streaming MABs model.

References

[1] A. Agarwal, S. Agarwal, S. Assadi, and S. Khanna. Learning with limited rounds of adaptivity: Coin tossing, multi-armed bandits, and ranking from pairwise comparisons. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017, Amsterdam, The Netherlands, 7-10 July 2017*, pages 39–75, 2017. 1, 1.1, 1.3, C

[2] A. Agarwal, S. Khanna, and P. Patil. A sharp memory-regret trade-off for multi-pass streaming bandits. *CoRR*, abs/2205.00984, 2022. 1, 1.3

[3] D. Agarwal, B. Chen, P. Elango, N. Motgi, S. Park, R. Ramakrishnan, S. Roy, and J. Zachariah. Online models for content optimization. In *Advances in Neural Information Processing Systems 21, Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 8-11, 2008*, pages 17–24, 2008. 1

[4] S. Assadi and C. Wang. Exploration with limited memory: streaming algorithms for coin tossing, noisy comparisons, and multi-armed bandits. In K. Makarychev, Y. Makarychev, M. Tulsiani, G. Kamath, and J. 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. 1, 1.1, 1.2, 1.3, 4.1, 4.2, A, A, A.4, D.1, (i), D.2

[5] J. Audibert, S. Bubeck, and R. Munos. Best arm identification in multi-armed bandits. In *COLT 2010 - The 23rd Conference on Learning Theory, Haifa, Israel, June 27-29, 2010*, pages 41–53, 2010. 1

[6] S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Found. Trends Mach. Learn.*, 5(1):1–122, 2012. 1.3

[7] G. Burtini, J. Loeppky, and R. Lawrence. A survey of online experiment design with the stochastic multi-armed bandit. *arXiv preprint arXiv:1510.00757*, 2015. 1

[8] W. Cao, J. Li, Y. Tao, and Z. Li. On top-k selection in multi-armed bandits and hidden bipartite graphs. In *NIPS*, volume 8, page 3, 2015. 1

[9] A. Carpentier and A. Locatelli. Tight (lower) bounds for the fixed budget best arm identification bandit problem. In *Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, USA, June 23-26, 2016*, pages 590–604, 2016. 1, 1.3

[10] A. R. Chaudhuri and S. Kalyanakrishnan. Regret minimisation in multi-armed bandits using bounded arm memory. *CoRR*, abs/1901.08387, 2019. 1.3

[11] A. R. Chaudhuri and S. 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, 2020. 1, 1.3

[12] L. Chen and J. Li. On the optimal sample complexity for best arm identification. *CoRR*, abs/1511.03774, 2015. 1, 1.1, 1.3, 3

[13] L. Chen, J. Li, and M. Qiao. Nearly instance optimal sample complexity bounds for top-k arm selection. In A. Singh and X. J. Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, volume 54 of *Proceedings of Machine Learning Research*, pages 101–110. PMLR, 2017. 3.4

[14] L. Chen, J. Li, and M. Qiao. Towards instance optimal bounds for best arm identification. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017, Amsterdam, The Netherlands, 7-10 July 2017*, pages 535–592, 2017. 1, 1.3

[15] X. Chen, P. N. Bennett, K. Collins-Thompson, and E. Horvitz. Pairwise ranking aggregation in a crowdsourced setting. In *Sixth ACM International Conference on Web Search and Data Mining, WSDM 2013, Rome, Italy, February 4-8, 2013*, pages 193–202, 2013. 1

[16] Y. Chen and C. Suh. Spectral MLE: top-k rank aggregation from pairwise comparisons. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 371–380, 2015. 2

[17] S. Dong, T. Ma, and B. V. Roy. On the performance of thompson sampling on logistic bandits. In *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, pages 1158–1160, 2019. 1.3

[18] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation methods for the web. In *Proceedings of the Tenth International World Wide Web Conference, WWW 10, Hong Kong, China, May 1-5, 2001*, pages 613–622, 2001. 1

[19] E. Even-Dar, S. Mannor, and Y. Mansour. PAC bounds for multi-armed bandit and markov decision processes. In *Computational Learning Theory, 15th Annual Conference on Computational Learning Theory, COLT 2002, Sydney, Australia, July 8-10, 2002, Proceedings*, pages 255–270, 2002. 1, 1.3, 2

[20] R. H. Farrell. Asymptotic behavior of expected sample size in certain one sided tests. *The Annals of Mathematical Statistics*, pages 36–72, 1964. 1.3, 3

[21] W. Hoeffding. Probability inequalities for sums of bounded random variables. In *The Collected Works of Wassily Hoeffding*, pages 409–426. Springer, 1994. A.3

[22] K. G. Jamieson, M. Malloy, R. D. Nowak, and S. Bubeck. lil' UCB : An optimal exploration algorithm for multi-armed bandits. In *Proceedings of The 27th Conference on Learning Theory, COLT 2014, Barcelona, Spain, June 13-15, 2014*, pages 423–439, 2014. 1, 1.1, 1.2, 1.3, 3, 5

[23] T. Jin, K. Huang, J. Tang, and X. Xiao. Optimal streaming algorithms for multi-armed bandits. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 5045–5054. PMLR, 2021. 1

[24] S. Kalyanakrishnan and P. Stone. Efficient selection of multiple bandit arms: Theory and practice. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, pages 511–518, 2010. 1, 2

[25] Z. S. Karnin, T. Koren, and O. Somekh. Almost optimal exploration in multi-armed bandits. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pages 1238–1246, 2013. 1, 1.1, 1.2, 1.3, 5, D.2

[26] N. Karpov and Q. Zhang. Instance-sensitive algorithms for pure exploration in multinomial logit bandit. *CoRR*, abs/2012.01499, 2020. 1.1

[27] E. Kaufmann, O. Cappé, and A. Garivier. On the complexity of best-arm identification in multi-armed bandit models. *J. Mach. Learn. Res.*, 17:1:1–1:42, 2016. 1, 1.3, 3.4

[28] J. Komiyama, J. Honda, and H. Nakagawa. Optimal regret analysis of thompson sampling in stochastic multi-armed bandit problem with multiple plays. In *International Conference on Machine Learning*, pages 1152–1161. PMLR, 2015. 1.3

[29] D. Liau, Z. Song, E. Price, and G. Yang. Stochastic multi-armed bandits in constant space. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*, pages 386–394, 2018. 1, 1.3

[30] A. Maiti, V. Patil, and A. Khan. Streaming algorithms for stochastic multi-armed bandits. *arXiv preprint arXiv:2012.05142*, 2020. 1, 1.3

[31] A. Maiti, V. Patil, and A. Khan. Multi-armed bandits with bounded arm-memory: Near-optimal guarantees for best-arm identification and regret minimization. In M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 19553–19565, 2021. 1

[32] S. Mannor and J. N. Tsitsiklis. Lower bounds on the sample complexity of exploration in the multi-armed bandit problem. In *Computational Learning Theory and Kernel Machines, 16th Annual Conference on Computational Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003, Proceedings*, pages 418–432, 2003. 1, 1.3, 3, C

[33] F. Radlinski, R. Kleinberg, and T. Joachims. Learning diverse rankings with multi-armed bandits. In *Proceedings of the 25th international conference on Machine learning*, pages 784–791, 2008. 1

[34] H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952. 1, 1.3

[35] R. J. Serfling. Probability inequalities for the sum in sampling without replacement. *The Annals of Statistics*, pages 39–48, 1974. A.3

[36] N. B. Shah and M. J. Wainwright. Simple, robust and optimal ranking from pairwise comparisons. *J. Mach. Learn. Res.*, 18:199:1–199:38, 2017. 2

[37] V. Srinivas, D. P. Woodruff, Z. Xu, and S. Zhou. Memory bounds for the experts problem. In *54th Annual ACM Symposium on Theory of Computing June 20-24, 2022 in Rome, Italy (to appear)*, 2022. 3.4

[38] M. S. Talebi, Z. Zou, R. Combes, A. Proutière, and M. Johansson. Stochastic online shortest path routing: The value of feedback. *IEEE Trans. Automat. Contr.*, 63(4):915–930, 2018. 1

[39] W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933. 1.3

[40] Y. Xue, P. Zhou, T. Jiang, S. Mao, and X. Huang. Distributed learning for multi-channel selection in wireless network monitoring. In *2016 13th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, pages 1–9. IEEE, 2016. 1

[41] Y. Zhou, X. Chen, and J. Li. Optimal PAC multiple arm identification with applications to crowdsourcing. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 217–225, 2014. 1

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We have discussed the problems left open by our work both in the corresponding sections and in Section 6. All the assumptions are explicitly stated in the paper.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] This is a theoretical work and we cannot identify any immediate positive or negative societal impacts.

(d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] Please see the appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]