Collaborative Learning with Limited Interaction: Tight Bounds for Distributed Exploration in Multi-Armed Bandits

Chao Tao*, Qin Zhang†, Yuan Zhou‡
*Computer Science Department, Indiana University
Email: taochao@iu.edu
†Computer Science Department, Indiana University
Email: qzhangcs@indiana.edu
‡Computer Science Department, Indiana University
and
Department of ISE, University of Illinois at Urbana-Champaign
Email: yuanz@illinois.edu

***Abstract*—Best arm identification (or, pure exploration) in multi-armed bandits is a fundamental problem in machine learning. In this paper we study the distributed version of this problem where we have multiple agents, and they want to learn the best arm collaboratively. We want to quantify the power of collaboration under limited interaction (or, communication steps), as interaction is expensive in many settings. We measure the running time of a distributed algorithm as the *speedup* over the best centralized algorithm where there is only one agent. We give almost tight round-speedup tradeoffs for this problem, along which we develop several new techniques for proving lower bounds on the number of communication steps under time or confidence constraints.**

***Index Terms*—communication complexity, foundations of machine learning, lower bounds, parallel computation**

I. Introduction

One of the biggest challenges in machine learning is to make learning scalable. A natural way to speed up the learning process is to introduce multiple learners/agents, and let them learn the target function collaboratively. A fundamental question in this direction is to *quantify the power of collaboration under limited interaction*, as interaction is expensive in many settings. In this paper we approach this general question via the study of a central problem in online learning – *best arm identification* (or, *pure exploration*) in *multi-armed bandits*. We present efficient collaborative learning algorithms and complement them with almost tight lower bounds.

Best Arm Identification. In multi-armed bandits (MAB) we have n alternative arms, where the i-th arm is associated with an unknown reward distribution $\mathcal{D}_i$ with mean θ_i. Without loss of generality we assume that each $\mathcal{D}_i$ has support on $[0,1]$; this can always be satisfied with proper rescaling. We also assume that $\theta_i \in [\iota, 1-\iota]$ for any positive constant $\iota > 0$.[1] We are interested in the best arm identification problem in MAB, in which we want to identify the arm with the largest mean. In the standard setting we only have one agent, who tries to identify the best arm by a sequence of arm pulls. Upon each pull of the i-th arm the agent observes an *i.i.d.* sample/reward from $\mathcal{D}_i$. At any time step, the index of the next pull (or, the final output at the end of the game) is decided by the indices and outcomes of all previous pulls and the randomness of the algorithm (if any). Our goal is to identify the best arm using the minimum amount of arm pulls, which is equivalent to minimizing the *running time* of the algorithm; we can just assume that each arm pull takes a unit time.

MAB has been studied for more than half a century [2], [3], due to its wide practical applications in clinical trials [4], adaptive routings [5], financial portfolio design [6], model selection [7], computer game play [8], stories/ads display on website [9], just to name a few. In many of these scenarios we are interested in finding out the best arm (strategy, choice, etc.) as soon as possible and committing to it. For example, in the Monte Carlo Tree Search used by computer game play engines, we want to find out the best move among a huge number of possible moves. In the task of high-quality website design, we hope to find out the best design among a set of alternatives for display. In almost all such applications the arm pull is the most expensive component: in the real-time decision making of computer game play, it is time-expensive to perform a single Monte Carlo simulation; in website design tasks, having a user to test each alternative is both time and capital expensive (often a fixed monetary reward is paid for each trial a tester carries out).

Chao Tao is supported in part by NSF IIS-1633215. Qin Zhang is supported in part by NSF IIS-1633215 and CCF-1844234.

[1] This assumption is due to minor technical reasons, and is also made in many existing bandit lower bounds (e.g. [1]). It does not affect our claims by much, since the most interesting and the hardest instances remain covered by the assumption.

In the literature of best arm identification in MAB, two variants have been considered:

1) *Fixed-time best arm:* Given a time budget T, identify the best arm with the smallest error probability.[2]
2) *Fixed-confidence best arm:* Given an error probability δ, identify the best arm with error probability at most δ using the smallest amount of time.

We will study both variants in this paper.

Collaborative Best Arm Identification. In this paper we study best arm identification in the collaborative learning model, where we have K agents who try to learn the best arm together. The learning proceeds in rounds. In each round each agent pull a (multi)set of arms *without* communication. For each agent at any time step, based on the indices and outcomes of all previous pulls, all the messages received, and the randomness of the algorithm (if any), the agent, if not in the *wait* mode, takes one of the following actions: (1) makes the next pull; (2) requests for a communication step and enters the wait mode; (3) terminates and outputs the answer. A communication step starts if *all non-terminated* agents are in the wait mode. After a communication step all non-terminated agents exit the wait mode and start a new round. During each communication step each agent can broadcast a message to every other agent. While we do not restrict the size of the message, in practice it will not be too large – the information of all pull outcomes of an agent can be described by an array of size at most n, with each coordinate storing a pair (c_i, sum_i), where c_i is the number of pulls on the i-th arm, and sum_i is sum of the rewards of the c_i pulls. Once terminated, the agent will not make any further actions. The algorithm terminates if all agents terminate. When the algorithm terminates, each agent should agree on the *same* best arm; otherwise we say the algorithm fails. The *number of rounds* of computation, denoted by R, is the number of communication steps *plus one*.

Our goal in the collaborative learning model is to minimize the number of rounds R, and the *running time* $T = \sum_{r\in[R]} t_r$, where t_r is the *maximum* number of pulls made among the K agents in round r. The motivation for minimizing R is that initiating a communication step always comes with a big time overhead (due to network bandwidth, latency, protocol handshaking), and energy consumption (e.g., think about robots exploring in the deep sea and on Mars). Round-efficiency is one of the major concerns in all parallel/distributed computational models such as the BSP model [10] and MapReduce [11]. The *total cost* of the algorithm is a *weighted sum* of R and T, where the coefficients depend on the concrete applications. We are thus interested in the best *round-time tradeoffs* for collaborative best arm identification.

Speedup in Collaborative Learning. As the time complexity of the best arm identification in the centralized setting is already well-understood (see, e.g. [1], [12]–[18]), we would like to interpret the running time of a collaborative learning algorithm as the *speedup* over that of the best centralized algorithm, which also expresses *the power of collaboration*. Intuitively speaking, if the running time of the best centralized algorithm is $T_{\mathcal{O}}$, and that of a proposed collaborative learning algorithm $\mathcal{A}$ is $T_{\mathcal{A}}$, then we say the speedup of $\mathcal{A}$ is $\beta_{\mathcal{A}} = T_{\mathcal{O}}/T_{\mathcal{A}}$. However, due to the parameters in the definition of the best arm identification *and* the instance-dependent bounds for the best centralized algorithms, the definition of the speedup of a collaborative learning algorithm needs to be a bit more involved.

Recall that an MAB instance is a set of random variables $\{X_1, \ldots, X_n\}$ each of which has support on $[0, 1]$. Since we are interested in the instance-dependent bounds, we assume that a random permutation is "built-in" to the input, that is, the $X_1, \ldots, X_n$ are randomly permuted before being fed to the algorithm. This is a standard assumption in the literature of MAB, since otherwise no conceivable algorithm can achieve instance-optimality – the foolish algorithm that always outputs the first arm will work perfectly in the instance in which the first arm has the largest mean.

For any fixed-time algorithm $\mathcal{A}$ and an input instance I, we let $\delta_{\mathcal{A}}(I, T)$ be the error probability of $\mathcal{A}$ on I given time budget T. For any fixed-confidence algorithm $\mathcal{A}$ and an input instance I, we let $T_{\mathcal{A}}(I, \delta)$ be the expected time used by $\mathcal{A}$ on I given the confidence parameter $(1 - \delta)$. In both definitions, the randomness is taken over both $\mathcal{A}$ and I. We also extend the definition $T_{\mathcal{A}}(I, \delta)$ to any fixed-time algorithm $\mathcal{A}$ by letting it be the smallest T such that $\delta_{\mathcal{A}}(I, T) \le \delta$.

We now define the key notion of *speedup* for a collaborative algorithm. We say that an instance I is T-solvable by an algorithm $\mathcal{O}$ (for both fixed-budget and fixed-time and fixed-confidence settings), if $T_{\mathcal{O}}(I, 1/3) \le T$. For any T, the speedup of a collaborative learning algorithm $\mathcal{A}$ (which can be either fixed-budget or fixed-time) for instances T-solvable by a centralized algorithm is defined as follows.

$$\beta_{\mathcal{A}}(T) = \inf_{\text{centralized } \mathcal{O}} \inf_{\text{instance } I} \inf_{\delta\in(0,1/3]:T_{\mathcal{O}}(I,\delta)\le T} \frac{T_{\mathcal{O}}(I, \delta)}{T_{\mathcal{A}}(I, \delta)}. \quad (1)$$

Here the most inner $\inf$ returns $+\infty$ if the set of candidate δ is empty. Note that the most natural definition for speedup would be for *all* instances. However, since our upper bound result logarithmically degrades as T grows, we have to introduce the T parameter in the definition, that is, we only consider those instances I for which the centralized algorithm can finish within time T under error δ.

Finally, we let $\beta_{K,R}(T) = \sup_{\mathcal{A}} \beta_{\mathcal{A}}(T)$ where the $\sup$ is taken over all R-round algorithms $\mathcal{A}$ for the collaborative learning model with K agents.[3]

Clearly there is a tradeoff between R and $\beta_{K,R}$: When $R = 1$ (i.e., there is *no* communication step), each agent needs to solve the problem by itself, and thus $\beta_{K,1} \le 1$. When R

[2]In the literature this is often called *fixed-budget best arm*. Here we use *time* instead of *budget* in order to be consistent with the collaborative learning setting, where it is easier to measure the performance of the algorithm by its running time.

[3]A similar concept of *speedup* was introduce in the previous work [19]. However, no formal definition was given in [19].

increases, $\beta_{K,R}$ may increase. On the other hand we always have $\beta_{K,R} \leq K$. Our goal is to find the best *round-speedup* tradeoffs, which is essentially equivalent to the *round-time* tradeoffs that we mentioned earlier.

As one of our goals is to understand the scalability of the learning process, we are particularly interested in one end of the tradeoff curve: What is the *smallest* R such that $\beta_{K,R} = \Omega(K)$? In other words, how many rounds are needed to make best arm identification fully scalable in the collaborative learning model? In this paper we will address this question by giving almost tight round-speedup tradeoffs.

Our Contributions. Our results are shown in Table I. For convenience we use the ' $\tilde{}$ ' notation on O, Ω, Θ to hide logarithmic factors, which will be made explicit in the actual theorems. Our contributions include:

1) *Almost tight round-speedup tradeoffs for fixed-time.* In particular, we show that any algorithm for the fixed-time best arm identification problem in the collaborative learning model with K agents that achieves $(K/\ln^{O(1)} K)$-speedup needs at least $\Omega(\ln K/\ln\ln K)$ rounds (for $T \geq K^{\Omega(1)}$). We complement this lower bound with an algorithm that runs in $\ln K$ rounds and achieves $\tilde{\Omega}(K)$-speedup.
2) *Almost tight round-speedup tradeoffs for fixed-confidence.* In particular, we show that any algorithm for the fixed confidence best arm identification problem in the collaborative learning model with K agents that achieves $(K/\ln^{O(1)} K)$-speedup needs at least $\Omega\left(\ln\frac{1}{\Delta_{\min}}/(\ln\ln K + \ln\ln\frac{1}{\Delta_{\min}})\right)$ rounds (for $T \geq \Delta_{\min}^{-\Omega(1)}$), which almost matches an algorithm in [19] that runs in $\ln\frac{1}{\Delta_{\min}}$ rounds and achieves $\tilde{\Omega}(K)$-speedup. Here $\Delta_{\min}$ is the difference between the mean of the best arm and that of the second best arm in the input.
3) *A separation for two problems.* The two results above give a separation on the round complexity of fully scalable algorithms between the fixed-time case and the fixed-confidence case. In particular, the fixed-time case has smaller round complexity for input instances with $\Delta_{\min} < 1/K$ (and when $T \geq \Delta_{\min}^{-\Omega(1)}$), which indicates that knowing the "right" time budget is useful to reduce the number of rounds of the computation.
4) *A generalization of the round-elimination technique.* In the lower bound proof for the fixed-time case, we develop a new technique which can be seen as a generalization of the standard round-elimination technique: we perform the round reduction on *classes* of input distributions. We believe that this new technique will be useful for proving round-speedup tradeoffs for other problems in collaborative learning.
5) *A new technique for instance-dependent round complexity.* In the lower bound proof for the fixed-confidence case, we develop a new technique for proving instance-dependent lower bound for round complexity. The *distribution exchange lemma* we introduce for handling different input distributions at different rounds may be of independent interest.

[4]We note again that the number of rounds equals to the number of communication steps *plus one*.

Related Works. There are two main research directions in literature for MAB in the centralized setting, *regret minimization* and *pure exploration*. In the regret minimization setting (see e.g. [20]–[22]), the player aims at maximizing the total reward gained within the time horizon, which is equivalent to minimizing the *regret* which is defined to be the difference between the total reward achieved by the offline optimal strategy (where all information about the input instance is known beforehand) and the total reward by the player. In the pure exploration setting (see, e.g. [1], [12], [14], [15], [18], [23]), the goal is to maximize the probability to successfully identify the best arm, while minimizing the number of sequential samples used by the player. Motivated by various applications, other exploration goals were also studied, e.g., to identify the top-k best arms [24]–[26], and to identify the set of arms with means above a given threshold [27].

The collaborative learning model for MAB studied in this paper was first proposed by [19], and has proved to be practically useful – authors of [28] and [29] applied the model to distributed wireless network monitoring and collective sensemaking.

Agarwal et al. [30] studied the problem of minimum adaptivity needed in pure exploration. Their model can be viewed as a restricted collaborative learning model, where the agents are *not* fully adaptive and have to determine their strategy at the beginning of each round. Some solid bounds on the round complexity are proved in [30], including a lower bound using the round elimination technique. As we shall discuss shortly, we develop a generalized round elimination framework and prove a much better round complexity lower bound for a more sophisticated hard instance.

There are other works studying the regret minimization problem under various distributed computing settings. For example, motivated by the applications in cognitive radio network, a line of research (e.g., [31]–[33]) studied the regret minimization problem where the radio channels are modeled by the arms and the rewards represent the utilization rates of radio channels which could be deeply discounted if an arm is simultaneously played by multiple agents and a collision occurs. Regret minimization algorithms were also designed for the distributed settings with an underlying communication network for the peer-to-peer environments (e.g., [34]–[36]). In [37], [38], the authors studied distributed regret minimization in the adversarial case. Authors of [39] studied the regret minimization problem in the batched setting.

Blum et al. [40] studied PAC learning of a general function in the collaborative setting, and their results were further strengthened by [41], [42]. However, in the collaborative learning model they studied, each agent can only sample from

problem	number of rounds[4]	$\beta_{K,R}(T)$	UB/LB	ref.
fixed-time	1	1	–	trivial
	2	$\tilde{\Omega}(\sqrt{K})$	UB	[19]
	2	$\tilde{O}(\sqrt{K})$	LB	[19]
	R	$\tilde{\Omega}(K^{\frac{R-1}{R}})$	UB	**new**
	$\Omega\left(\frac{\ln \tilde{K}}{\ln\ln \tilde{K}+\ln \frac{K}{\beta}}\right)$ when $\beta \in [K/\tilde{K}^{0.1}, K]$	β	LB	**new**
fixed-confidence	R	$\tilde{\Omega}\left((\Delta_{\min})^{\frac{2}{R-1}} K\right)$	UB	[19]
	$\Omega\left(\min\left\{\frac{\ln \frac{1}{\widetilde{\Delta_{\min}}}}{\ln\left(1+\frac{K(\ln K)^2}{\beta}\right)+\ln\ln \frac{1}{\widetilde{\Delta_{\min}}}}, \sqrt{\frac{\beta}{(\ln K)^3}}\right\}\right)$	β	LB	**new**

TABLE I: Our results for collaborative best arm identification in multi-armed bandits. K is the number of agents. $\Delta_{\min}$ is the difference between the mean of the best arm and that of the second best arm in the input. In the lower bound for the fixed-time setting, we set $\tilde{K} = \min\{K, \sqrt{T}\}$; in the lower bound for the fixed-confidence setting, we set $\widetilde{\Delta_{\min}}^{-1} = \min\{\Delta_{\min}^{-1}, T\}$.

one particular distribution, and is thus different from the model this paper focuses on.

II. Techniques Overview

In this section we summarize the high level ideas of our algorithms and lower bounds. For convenience, the parameters used in this overview are only for illustration purposes.

Lower bound for fixed-time algorithms. A standard technique for proving round lower bounds in communication/sample complexity is the *round elimination* [43]. Roughly speaking, we show that if there exists an r-round algorithm with error probability δ_r and sample complexity $f(n_r)$ on an input distribution σ_r, then there also exists an $(r-1)$-round algorithm with error probability δ_{r-1} and sample complexity $f(n_{r-1})$ on an input distribution σ_{r-1}. Finally, we show that there is *no* 0-round algorithm with error probability $\delta_0 \ll 1$ on a nontrivial input distribution σ_0.

In [30] the authors used the round elimination technique to prove an $\Omega(\ln^* n)$ round lower bound for the best arm identification problem under the total pull budget $\tilde{O}(n/\Delta_{\min}^2)$.[5] In their hard input there is a single best arm with mean $\frac{1}{2}$, and $(n-1)$ arms with means $(\frac{1}{2} - \Delta_{\min})$. This "one-spike" structure makes it relatively easy to perform the standard round elimination. The basic arguments in [30] go as follows: Suppose the best arm is chosen from the $n_r = n$ arms uniformly at random. If the agents do not make enough pulls in the first round, then conditioned on the pull outcomes of the first round, the posterior distribution of the index of the best arm can be written as a convex combination of a set of distributions, each of which has support size at least $n_{r-1} \approx \log n$ and is close (in terms of the total variation distance) to the *uniform* distribution on its support, and is thus again hard for a $(r-1)$-round algorithm.

However, since our goal is to prove a much higher *logarithmic* round lower bound, we have to restrict the total pull budget within the *instance dependent* parameter $\tilde{O}(H) = \tilde{O}\left(\sum_{i=2}^n 1/\Delta_i^2\right)$ (Δ_i is the difference between the mean of the best arm and that of the i-th best arm in the input), and create a hard input distribution with logarithmic levels of arms in terms of their means.[6] Roughly speaking, we take $\frac{n}{2}$ random arms and assign them with mean $(\frac{1}{2} - \frac{1}{4})$, $\frac{n}{4}$ random arms with mean $(\frac{1}{2} - \frac{1}{8})$, and so on. With such a "pyramid-like" structure, it seems difficult to take the same path of arguments as that for the one-spike structure in [30]. In particular, it is not clear how to decompose the posterior distribution of the means of arms into a convex combination of a set of distributions, each of which is close to the same pyramid-like distribution. We note that such a decomposition is non-trivial even for the one-spike structure. Now with a pyramid-like structure we have to guarantee that arms of the $(\ell+1)$-th level are chosen randomly from the arms in the union of the $(\ell+1)$-th level and the ℓ-th level for each level ℓ, which looks to be technically challenging.

We take a different approach. We perform the round elimination on *classes* of input distributions. More precisely, we show that if there is *no* $(r-1)$-round algorithm with error probability δ_{r-1} and pull complexity $f(n_{r-1})$ on *any* distribution in distribution class Σ_{r-1}, then there is *no* r-round algorithm with error probability δ_r and pull complexity $f(n_r)$ on *any* distribution in distribution class Σ_r. When working with a class of distributions, we do *not* need to show that the posterior distribution ν' of some input distribution $\nu \in \Sigma_r$ is close to a particular distribution, but only that $\nu' \in \Sigma_{r-1}$.

Although we now have more flexibility on selecting hard input distribution, we still want to find classes of distributions that are easy to work with. To this end we introduce two more ideas. First, at the beginning we sample the mean of

[5] $\ln^* n$ is the number of times the logarithm function must be iteratively applied before the result is less than or equal to 1.

[6] $H = O(\sum_{i=2}^n 1/\Delta_i^2)$ is a standard parameter for describing the pull complexity of algorithms in the multi-armed bandits literature (see, e.g., [21]).

each arm independently from the *same* distribution, in which the pyramid-like structure is encoded. We found that making the means of arms independent of each other at any time (conditioned on the observations obtained so far) can dramatically simplify the analysis. Second, we choose to *publish* some arms after each round r to make the posterior distribution of the set of unpublished arms stay within the distribution class Σ_{r-1}. By publishing an arm we mean to exploit the arm and learn its mean exactly. With the ability of publishing arms we can keep the classes of distributions $\Sigma_r, \Sigma_{r-1}, \ldots$ relatively simple for the round elimination process.

Further different from [30] in which the set of arms pulled by each agent in each round is pre-determined at the beginning (i.e., the pulls are *oblivious* in each round), we allow the agents to act *adaptively* in each round. Allowing adaptivity inside each round adds another layer of technical challenge to our lower bound proof. Using a *coupling*-like argument, we manage to show that when the number of arms n is *smaller* than the number of agents K, adaptive pulls do not have much advantage against oblivious pulls in each round. We note that such an argument does not hold when $n \gg K$, and this is why we can only prove a round lower bound of $\Omega(\ln K/\ln\ln K)$ in the adaptive case compared with a round lower bound of $\Omega(\ln n/\ln\ln n)$ in the oblivious case when the speedup $\beta = \tilde{\Omega}(K)$. Surprisingly, this is almost the best that we can achieve – our next result shows that there is an $\tilde{\Omega}(K)$-speedup adaptive algorithm using $\ln K$ rounds of computation.

Upper bound for fixed-time algorithms. Our algorithm is conceptually simple, and goes by two phases. The goal of the first phase is to eliminate most of the suboptimal arms and make sure that the number of the remaining arms is at most K, which is the number of agents. This is achieved by assigning each arm to a random agent, and each agent uses $T/2$ time budget to identify the best arm among its assigned arms using the start-of-the-art centralized algorithm. Note that no communication is needed in this phase, and there are still R rounds left for the second phase. We allow each of the R rounds to use $T/(2R)$ time budget. The goal of the r-th round in the second phase is to reduce the number of arms to at most $K^{\frac{R-r}{R}}$, so that after the R-th round, only the optimal arm survives. To achieve this, we uniformly spend the time budget on each remaining arm. We are able to prove that this simple strategy works, and our analysis crucially relies on the the guarantee that there are at most $K^{\frac{R-r+1}{R}}$ arms at the beginning of the r-th round.

We note that when $R = 2$, the speedup of our algorithm is $\tilde{\Omega}(\sqrt{K})$, matching that of the 2-round algorithm presented in [19]. Our algorithm also provides the optimal speedup guarantee for $R > 2$, matching our lower bound result mentioned above.

The algorithm mentioned above only guarantees to identify the best arm with constant error probability. When the input time horizon T is larger, one would expect an algorithm with an error probability that diminishes exponentially in T. To this end, we strengthen our basic algorithm to a meta-algorithm that invokes the basic algorithm several times in parallel and returns the plurality vote. One technical difficulty here is that the optimal error probability depends on the input instance and is not known beforehand. One has to guess the right problem complexity and make sure that the basic algorithm does not consistently return the same suboptimal arm when the given time horizon is less than the problem complexity (otherwise the meta algorithm would recognize the suboptimal arm as the best arm with high confidence).

We manage to resolve this issue via novel algorithmic ideas that may be applied to strengthen fixed-time bandit algorithms in general. In particular, in the first phase of our basic algorithm, we assign a *random* time budget (instead of the fixed $T/2$ as described above) to the centralized algorithm invoked by each agent, and this proves to be useful to prevent the algorithm from identifying a suboptimal arm with overwhelmingly high probability. We note that in [19], the authors got around this problem by allowing the algorithm to have access to *both* the time horizon and the confidence parameters, which does not fall into the standard fixed-time category.

Lower bound for fixed-confidence algorithms. We first reduce the lower bound for best arm identification algorithms to the task of showing round lower bound for a closely related problem, SIGNID, which has proved to be a useful proxy in studying the lower bounds for bandit exploration in the centralized setting [15], [18], [44]. The goal of SIGNID is to identify (with fixed confidence) whether the mean reward of the only input arm is greater or less than $1/2$. The difference between $1/2$ and the mean of the arm, denoted by Δ, corresponds to $\Delta_{\min}$ in the best arm identification problem, and our new task becomes to show a round lower bound for the SIGNID problem that increases as Δ approaches 0.

While our lower bound proof for fixed-time setting can be viewed as a generalization of the round elimination technique, our lower bound for the SIGNID problem in the fixed-confidence setting uses a completely different approach due to the following reasons. First, the online learning algorithm that our lower bound is against aims at achieving an instance dependent optimal time complexity as it gradually learns the underlying distribution. In other words, the hardness stems from the fact that the algorithm does *not* know the underlying distribution beforehand, while traditional round elimination proofs do *not* utilize this property. Second, our lower bound proof introduces a sequence of arm distributions and inductively shows that any algorithm needs at least r rounds on the r-th input distribution. While traditional round elimination manages to conduct this induction via embedding the $(r-1)$-st input distribution into the r-th input distribution, it is not clear how to perform such an embedding in our proof, as our distributions are very different.

Intuitively, in our inductive proof we set the r-th input distribution to be the Bernoulli arm with $\Delta = \Delta_r = 1/\zeta^r$ and $\zeta > 1$ depends on K (the number of agents) and β (the speedup of the algorithm). We hope to show that any algorithm

needs r rounds on the r-th input distribution. Suppose we have shown the lower bound for the r-th input distribution. Since the algorithm has β-speedup, it performs at most $O(\Delta_r^{-2}K/\beta)$ pulls for the r-th instance. We will show via a *distribution exchange lemma* (which will be explained in details shortly) that this amount of pulls is not sufficient to tell $\Delta = \Delta_r$ from $\Delta = \Delta_{r+1}$. Hence the algorithm also uses at most $O(\Delta_r^{-2}K/\beta)$ pulls during the first r rounds on the $(r+1)$-st instance, which is not sufficient to decide the sign of the $(r+1)$-st instance. Therefore the algorithm needs at least $(r+1)$ rounds on the $(r+1)$-st instance, completing the induction for the $(r+1)$-st instance.

To make the intuition rigorous, we need to strengthen our inductive hypothesis as follows. The goal of the r-th inductive step is to show that for $\Delta = \Delta_r$, any algorithm needs at least r rounds and makes at most $o(\Delta_r^{-2})$ pulls across the K agents during the first r rounds. While the 0-th inductive step holds straightforwardly as the induction basis, we go from the r-th inductive step to the $(r+1)$-st inductive step via a *progress lemma* and the distribution exchange lemma mentioned above.

Given the hypothesis for the r-th inductive step, the progress lemma guarantees that the algorithm has to proceed to the $(r+1)$-st round and perform more pulls. Thanks to the strengthened hypothesis, the total number of pulls performed in the first r rounds is $o(\Delta_r^{-2})$. Hence the statistical difference between the pulls drawn from the r-th input distribution and its negated distribution (where the outcomes 0 and 1 are flipped) is at most $o(1)$ due to Pinsker's inequality, and this is not enough for the algorithm to correctly decide the sign of the arm.

The distribution exchange lemma guarantees that the algorithm performs no more than $O(\Delta_r^{-2}K/\beta)$ pulls across the agents during the first $(r+1)$ rounds on the $(r+1)$-st input distribution. By setting $\zeta = \omega(K/\beta)$, one can verify that $O(\Delta_r^{-2}K/\beta) = o(\Delta_{r+1}^{-2})$, and the hypothesis for the $(r+1)$-st inductive step is proved. The intuition behind the distribution exchange lemma is as follows. While the algorithm needs $(r+1)$ rounds on the r-th input distribution (by the progress lemma), we know that the algorithm cannot use more than $\Omega(\Delta_r^{-2}K/\beta)$ pulls by the β-speedup constraint. These many pulls are not enough to tell the difference between the r-th and the $(r+1)$-st distribution, and hence we can change the underlying distribution and show that the same happens for the $(r+1)$-st input distribution.

However, this intuition is not easy to be formalized. If we simply use the statistical difference between the distributions induced by Δ_r and Δ_{r+1} to upper bound the probability difference between each agent's behavior for the two input arms, we will face a probability error of $\Theta(\sqrt{1/\beta})$ for each agent. In total, this becomes a probability error of $\Theta(K\sqrt{1/\beta}) \gg 1$ throughout all K agents, which is too much. To overcome this difficulty, we need to prove a more refined probabilistic upper bound on the behavior discrepancy of each agent for different arms. This is achieved via a technical lemma that provides a much better upper bound on the difference between the probabilities that two product distributions assign to the same event, given that the event does not happen very often. This technical lemma may be of independent interest.

III. Lower Bounds for Fixed-Time Distributed Algorithms

In this section we prove a lower bound for the fixed-time collaborative learning algorithms. We start by considering the non-adaptive case, where in each round each agent fixes the (multi-)set of arms to pull as well as the order of the pulls at the very beginning. We will then extend the proof to the adaptive case.

When we write $c = a \pm b$ we mean c is in the range of $[a-b, a+b]$.

A. Lower Bound for Non-Adaptive Algorithms

We prove the following theorem in this section.

Theorem 1. *For any time budget $T > 0$, any $\alpha \in [1, n^{0.2}]$, any (K/α)-speedup randomized non-adaptive algorithm for the fixed-time best arm identification problem in the collaborative learning model with K agents and $n \le \sqrt{T}$ arms needs $\Omega(\ln n/(\ln\ln n + \ln\alpha))$ rounds in expectation.*

Parameters. We list a few parameters to be used in the proof. Let $\alpha \in [1, n^{0.2}]$ be the parameter in the statement of Theorem 1. Set $B = \alpha(\ln n)^{100}$ (thus $(\ln n)^{100} \le B \le (\ln n)^{100} n^{0.2}$), $\gamma = \alpha(\ln n)^{100}$, $\rho = (\ln n)^3$, and $\kappa = (\ln n)^2$.

1) The Class of Hard Distributions

We first define a class of distributions which is hard for the best arm identification problem.

Let L be a parameter to be chosen later (in (8)). Define $\mathcal{D}_j(\eta)$ to be the class of distributions π with support

$$\{B^{-1}, \dots, B^{-(j-1)}, B^{-j}, \dots, B^{-L}\},$$

such that if $X \sim \pi$, then

1) $\Pr\left[(X = B^{-1}) \vee \cdots \vee (X = B^{-(j-1)})\right] \le n^{-9}$, (only defined for $j \ge 2$)
2) For any $\ell = j, \dots, L$, $\Pr[X = B^{-\ell}] = \lambda_j \cdot B^{-2\ell} \cdot \left(1 \pm \rho^{-\ell}\eta\right)$, where λ_j is a normalization factor (to make $\sum_{\ell=1}^{L} \Pr[X = B^{-\ell}] = 1$).

Note that when $\eta = 0$, $\mathcal{D}_1(0)$ only contains a single distribution; slightly abusing the notation, define $\mathcal{D}_1 \triangleq \mathcal{D}_1(0)$ to denote that particular distribution. For $j \ge 2$, define $\mathcal{D}_j \triangleq \mathcal{D}_j(\rho^{j-1})$. That is, we set $\eta = \rho^{j-1}$ by default, and consequently $\lambda_j = \left(1 \pm \frac{2}{\rho}\right) B^{2j}$.

We introduce a few threshold parameters: $\zeta_1 = \left(\frac{1}{2} - B^{-(j+1)}\right)\gamma B^{2j} - \sqrt{10\gamma \ln n} B^j$, $\zeta_2 = \frac{\gamma B^{2j}}{2} - B^{j+0.6}$, $\zeta_3 = \frac{\gamma B^{2j}}{2} + B^{j+0.6}$. It is easy to see that $\zeta_2 < \zeta_1 < \zeta_3$.

The following lemma gives some basic properties of pulling from an arm with mean $\left(\frac{1}{2} - B^{-\ell}\right)$. We leave the proof to Appendix A.

Lemma 2. *Consider an arm with mean $\left(\frac{1}{2} - X\right)$. We pull the arm γB^{2j} times. Let $\Theta = (\Theta_1, \Theta_2, \dots, \Theta_{\gamma B^{2j}})$ be the pull outcomes, and let $|\Theta| = \sum_{i \in [\gamma B^{2j}]} \Theta_i$. We have the followings.*

1) *If* $X = B^{-\ell}$ *for* $\ell > j$*, then* $|\Theta| \in [\zeta_2, \zeta_3]$ *with probability at least* $1 - n^{-10}$.
2) *If* $X = B^{-\ell}$ *for* $\ell \le j$*, then* $|\Theta| < \zeta_1$ *with probability at least* $1 - n^{-10}$.
3) *If* $X = B^{-\ell}$ *for* $\ell > j$*, then* $|\Theta| \ge \zeta_1$ *with probability at least* $1 - n^{-10}$.

The next lemma states important properties of distributions in classes $\mathcal{D}_j$. Intuitively, if the mean of an arm is distributed according to some distribution in class $\mathcal{D}_j$, then after pulling it γB^{2j} times, we can learn by Lemma 2 that at least one of the followings hold: (1) the sequence of pull outcomes is very rare; (2) very likely the mean of the arm is at most $(\frac{1}{2} - B^{-j})$; (3) very likely the mean of the arm is more than $(\frac{1}{2} - B^{-j})$. In the first two cases we *publish* the arm, that is, we fully exploit the arm and learn its mean exactly. We will show that if the arm is not published, then the posterior distribution of the mean of the arm (given the outcomes of the γB^{2j} pulls) belongs to class $\mathcal{D}_{j+1}$.

Lemma 3. *Consider an arm with mean* $\left(\frac{1}{2} - X\right)$ *where* $X \sim \mu \in \mathcal{D}_j$ *for some* $j \in [L-1]$*. We pull the arm* γB^{2j} *times. Let* $\Theta = (\Theta_1, \Theta_2, \ldots, \Theta_{\gamma B^{2j}})$ *be the pull outcomes, and let* $|\Theta| = \sum_{i \in [\gamma B^{2j}]} \Theta_i$*. If* $|\Theta| \notin [\zeta_1, \zeta_3]$*, then we publish the arm. Let* ν *be the posterior distribution of* X *after observing* Θ*. If the arm is not published, then we must have* $\nu \in \mathcal{D}_{j+1}$.

Proof. We analyze the posterior distribution of X after observing $\Theta = \theta$ for any θ with $|\theta| \in [\zeta_1, \zeta_3]$.

Let $\chi_{\le j}$ denote the event that $(X = B^{-1}) \vee \cdots \vee (X = B^{-j})$, and let $\chi_{>j}$ denote the event that $(X = B^{-(j+1)}) \vee \cdots \vee (X = B^{-L})$. Since $X \sim \mu \in \mathcal{D}_j$, we have

$$\begin{aligned} & \Pr[\chi_{>j}] \\ \ge\ & \Pr[X = B^{-(j+1)}] \\ =\ & \left(1 \pm \frac{2}{\rho}\right) B^{2j} \cdot B^{-2(j+1)} \cdot \left(1 \pm \rho^{-(j+1)} \rho^{j-1}\right) \\ \ge\ & 1/(2B^2). \end{aligned} \tag{2}$$

For the convenience of writing, let $m = \gamma B^{2j}$. Thus $\zeta_1 = m \cdot (\frac{1}{2} - z)$ where $z = B^{-j}\left(B^{-1} + \sqrt{\frac{10 \ln n}{\gamma}}\right)$. Let $\epsilon = B^{-j}$, and $\epsilon' = B^{-(j+1)}$.

For any θ with $|\theta| \ge \zeta_1$, we have

$$\begin{aligned} & \Pr[\chi_{\le j} \mid \Theta = \theta] \\ =\ & \frac{\Pr[\Theta = \theta \mid \chi_{\le j}] \cdot \Pr[\chi_{\le j}]}{\Pr[\Theta = \theta]} \\ =\ & \frac{\Pr[\Theta = \theta \mid \chi_{\le j}] \cdot \Pr[\chi_{\le j}]}{\Pr[\Theta = \theta \mid \chi_{\le j}] \cdot \Pr[\chi_{\le j}] + \Pr[\Theta = \theta \mid \chi_{>j}] \cdot \Pr[\chi_{>j}]} \\ \le\ & \frac{\Pr[\Theta = \theta \mid X = \epsilon] \cdot 1}{0 + \Pr[\Theta = \theta \mid X = \epsilon'] \cdot 1/(2B^2)} \\ & \quad \text{(by (2) and monotonicity)} \\ =\ & 2B^2 \cdot \frac{(1/2 - \epsilon)^{|\theta|}(1/2 + \epsilon)^{m - |\theta|}}{(1/2 - \epsilon')^{|\theta|}(1/2 + \epsilon')^{m - |\theta|}} \\ \le\ & 2B^2 \cdot \frac{(1/2 - \epsilon)^{\zeta_1}(1/2 + \epsilon)^{m - \zeta_1}}{(1/2 - \epsilon')^{\zeta_1}(1/2 + \epsilon')^{m - \zeta_1}} \quad \text{(by monotonicity)} \\ =\ & 2B^2 \cdot A^m, \end{aligned} \tag{3}$$

where

$$A = \frac{(1 - 2\epsilon)^{1/2 - z}(1 + 2\epsilon)^{1/2 + z}}{(1 - 2\epsilon')^{1/2 - z}(1 + 2\epsilon')^{1/2 + z}}. \tag{4}$$

We next analyze A. For small enough $\epsilon > 0$, we have $\epsilon - \frac{\epsilon^2}{2} \le \ln(1+\epsilon) \le \epsilon - \frac{\epsilon^2}{2} + \epsilon^3$, and $-\epsilon - \frac{\epsilon^2}{2} - \epsilon^3 \le \ln(1-\epsilon) \le -\epsilon - \frac{\epsilon^2}{2}$. Taking the natural logarithm on both sides of (4) and using two inequalities for $\ln(1+\epsilon)$ and $\ln(1-\epsilon)$ above, we have

$$\begin{aligned} \ln A \ \le\ & (1/2 - z)\left(-2\epsilon - 2\epsilon^2 + 2(\epsilon') + 2(\epsilon')^2 + 8(\epsilon')^3\right) \\ & + (1/2 + z)\left(2\epsilon - 2\epsilon^2 + 8\epsilon^3 - 2(\epsilon') + 2(\epsilon')^2\right) \\ =\ & 1/2 \cdot \left(-4\epsilon^2 + 8\epsilon^3 + 4(\epsilon')^2 + 8(\epsilon')^3\right) \\ & + z(4\epsilon + 8\epsilon^3 - 4(\epsilon') - 8(\epsilon')^3) \\ \le\ & -2B^{-2j} + B^{-j}\left(B^{-1} + \sqrt{\frac{10 \ln n}{\gamma}}\right) 4B^{-j} \\ & + O(B^{-2j-1}) \\ \le\ & -B^{-2j}. \end{aligned} \tag{5}$$

Plugging (5) back to (3), we have

$$\Pr[\chi_{\le j} \mid \Theta = \theta] \le 2B^2 \cdot e^{-B^{-2j} \cdot \gamma B^{2j}} \le n^{-9}. \tag{6}$$

where the last inequality holds since $B \le (\ln n)^{100} n^{0.2}$ and $\gamma \ge (\ln n)^{100}$. Therefore ν satisfies the first condition of the distribution class $\mathcal{D}_{j+1}$.

For any θ with $|\theta| \in [\zeta_1, \zeta_3]$ and $\ell = j+1, \ldots, L$, we have

$$\begin{aligned} & \Pr[X = B^{-\ell} \mid \Theta = \theta] \\ =\ & \frac{\Pr[\Theta = \theta \mid X = B^{-\ell}] \cdot \Pr[X = B^{-\ell}]}{\Pr[\Theta = \theta]} \\ =\ & \frac{1}{\Pr[\Theta = \theta]} \cdot \Big(\Pr\left[\Theta = \mathbb{E}[\Theta] \mid X = B^{-\ell}\right] \\ & \cdot (1 \pm B^{-\ell})^{B^{j+0.61}}\Big) \cdot \lambda_j B^{-2\ell}\left(1 \pm \rho^{-\ell}\eta\right) \\ =\ & \frac{1}{\Pr[\Theta = \theta]} \cdot \left(\frac{1}{2\sqrt{2\pi\gamma B^{2j}}} \cdot \frac{1}{\sqrt{1 - 4B^{-2\ell}}}\right. \\ & \left. \cdot (1 \pm B^{-\ell})^{B^{j+0.7}}\right) \cdot \lambda_j B^{-2\ell}\left(1 \pm \rho^{-\ell}\eta\right) \\ =\ & \left(\frac{1}{\Pr[\Theta = \theta]} \cdot \frac{1}{2\sqrt{2\pi\gamma B^{2j}}} \cdot \lambda_j\right) \cdot \frac{1}{\sqrt{1 - 4B^{-2\ell}}} \\ & \cdot (1 \pm B^{-\ell})^{B^{j+0.7}} \cdot B^{-2\ell}\left(1 \pm \rho^{-\ell}\eta\right) \\ =\ & \lambda_j' \cdot (1 \pm 3B^{-2\ell}) \cdot (1 \pm B^{-\ell+j+0.8}) \cdot B^{-2\ell}\left(1 \pm \rho^{-\ell}\eta\right) \\ =\ & \lambda_j' \cdot B^{-2\ell}\left(1 \pm \rho^{-\ell}\eta'\right), \end{aligned} \tag{7}$$

where

- λ_j' is a normalization factor.
- The second equality holds since we have $|\theta| \in [\zeta_1, \zeta_3]$, and thus $\left|\theta - \mathbb{E}[\Theta \mid X = B^{-\ell}]\right| \le B^{j+0.61}$.
- In the third equality, we have used the Stirling's approximation for factorials (i.e., $n! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n\left(1 + \Theta(\frac{1}{n})\right)$) when calculating $\Pr\left[\Theta = \mathbb{E}[\Theta] \mid X = B^{-\ell}\right]$.

- The fifth inequality holds since $\frac{1}{\sqrt{1-4B^{-2\ell}}} = 1 \pm 3B^{-2\ell}$.
- In the last equality, since $B \geq (\ln n)^{100}$, $\rho = (\ln n)^3$, $\eta = \rho^{j-1}$ and $\ell \geq j+1$, we can set $\eta' = \rho^j$.

Therefore ν satisfies the second condition of the distribution class $\mathcal{D}_{j+1}$.

By (6) and (7), we have $\nu \in \mathcal{D}_{j+1}$. □

2) The Hard Input Distribution

Input Distribution σ: We pick the hard input distribution for the best arm identification problem as follows: the mean of each of the n arms is $\left(\frac{1}{2} - X\right)$, where $X \sim \mathcal{D}_1$.

Set $n = B^{2L}/\lambda_1$, where $\lambda_1 = \Theta(B^2)$ is the normalization factor of the distribution $\mathcal{D}_1$. This implies

$$L = \ln(n\lambda_1)/(2\ln B) = \Theta(\ln n/(\ln\ln n + \ln\alpha)). \quad (8)$$

We will use the running time of a good deterministic sequential algorithm as an upper bound for that of any collaborative learning algorithm that has a good speedup.

Let $\mathcal{E}_0$ be the event that there is one and only one best arm with mean $(\frac{1}{2} - B^{-L})$ when $I \sim \sigma$.

Lemma 4. *Given budget $W = n\ln^3 n \cdot B^2$, the deterministic sequential algorithm in [1] has expected error $o(1)$ on input distribution σ conditioned on $\mathcal{E}_0$.*

Proof. Given budget W, the error of the algorithm in [1] (denoted by $\mathcal{A}_{\text{ABM}}$) on an input instance I is bounded by

$$\text{err}(I) \leq n^2 \cdot \exp\left(-\frac{W}{2\ln n \cdot H(I)}\right), \quad (9)$$

where

$$H(I) = \sum_{i=2}^{n} \frac{1}{\Delta_i^2}, \quad (10)$$

where Δ_i is the difference between the mean of the best arm and that of the i-th best arm in I. We try to upper bound $H(I)$ when $I \sim \sigma = (\mathcal{D}_1)^n$ conditioned on $\mathcal{E}_0$.

Recall that in the distribution $\mathcal{D}_1$, $\Pr[X = B^{-\ell}] = \lambda_1 B^{-2\ell}$ for $\ell = 1, \ldots, L$ where $\lambda_1 = \Theta(B^2)$ is a normalization factor. Let k_ℓ be the number of arms with mean $(\frac{1}{2} - B^{-\ell})$. By Chernoff-Hoeffding bound and union bound, we have that with probability $(1 - e^{-B})$, for all $\ell = 1, \ldots, L-1$,

$$k_\ell = \Theta(\lambda_1 B^{-2\ell} n) = \Theta(B^{2L-2\ell}).$$

Thus for a large enough universal constant c_H, with probability $(1 - e^{-B})$,

$$H(I) = \sum_{\ell=1}^{L-1} k_\ell \cdot \frac{1}{\left(B^{-\ell} - B^{-L}\right)^2} \leq c_H L B^{2L}. \quad (11)$$

Plugging-in (11) to (9), we get

$$\text{err}(I) \leq n^2 \cdot \exp\left(-\frac{n\ln^3 n \cdot B^2}{2\ln n \cdot c_H L B^{2L}}\right) = o(1), \quad (12)$$

where the equality holds since $n = \Theta(B^{2L}/B^2)$ and $L = O(\ln n/\ln\ln n)$. Therefore, conditioned on $\mathcal{E}_0$ and under time budget W, the expected error of $\mathcal{A}_{\text{ABM}}$ on input distribution σ is at most $o(1) + e^{-B} = o(1)$.

□

3) Proof of Theorem 1

We say a collaborative learning algorithm is z-cost if the *total* number of pulls made by K agents is z. Since $n \leq \sqrt{T}$, we have $W = n\ln^3 n \cdot B^2 \leq T$. By Lemma 4 and the definition of speedup (Eq. (1)), if there is a (K/α)-speedup collaborative learning algorithm, then there must be a $\left(\frac{W}{K/\alpha} \cdot K\right) = (\alpha W)$-cost collaborative learning algorithm that has expected error $o(1)$ on input distribution σ conditioned on $\mathcal{E}_0$. By this observation, Theorem 1 follows immediately from the following lemma and Yao's Minimax Lemma [45].

Lemma 5. *Any deterministic (αW)-cost non-adaptive algorithm that solves the best arm identification problem in the collaborative learning model with K agents and n arms with error probability 0.99 on input distribution σ conditioned on $\mathcal{E}_0$ needs $\Omega(\ln n/(\ln\ln n + \ln\alpha))$ rounds.*

Let $I_j = \left(\left(1 \pm \frac{1}{L}\right) B^{-2}\right)^{j-1} n$. In the rest of this section we prove Lemma 5 by induction.

The Induction Step. The following lemma intuitively states that if there is no good $(r-1)$-round (αW)-cost non-adaptive algorithm, then there is no good r-round (αW)-cost non-adaptive algorithm.

Lemma 6. *For any $j \leq \frac{L}{2} - 1$, if there is no $(r-1)$-round (αW)-cost deterministic non-adaptive algorithm with error probability $\delta + O\left(\frac{1}{\kappa}\right)$ on any input distribution in $(\mathcal{D}_{j+1})^{n_{j+1}}$ for any $n_{j+1} \in I_{j+1}$, then there is no r-round (αW)-cost deterministic non-adaptive algorithm with error probability δ on any input distribution in $(\mathcal{D}_j)^{n_j}$ for any $n_j \in I_j$.*

Proof. Consider any r-round (αW)-cost deterministic non-adaptive algorithm $\mathcal{A}$ that succeeds with probability δ' on any input distribution in $\mu \in (\mathcal{D}_j)^{n_j}$ for any $n_j \in I_j$. Since we are considering a non-adaptive algorithm, at the beginning of the first round, the total number of pulls by the K agents on each of the n_j arms in the first round are fixed. Let $(t_1, \ldots, t_{n_j})$ be such a pull configuration, where t_z denotes the number of pulls on the z-th arm. For an (αW)-cost algorithm, by a simple counting argument, at least $(1 - \frac{1}{\kappa})$ fraction of t_z satisfies $t_z \leq \alpha\kappa\frac{W}{n_j}$. Let S be the set of arms z with $t_z > \gamma B^{2j}$. Since

$$\alpha\kappa\frac{W}{n_j} \leq \alpha\kappa\frac{n\ln^3 n B^2}{\left(\left(1 - \frac{1}{L}\right) B^{-2}\right)^{j-1} n} \leq \gamma B^{2j},$$

we have $|S| \leq \frac{1}{\kappa} \cdot n_j$.

We augment the first round of Algorithm $\mathcal{A}$ as follows.

Algorithm Augmentation.

1) We publish all arms in S.
2) For the rest of the arms $z \in [n_j] \backslash S$, we keep pulling them until the total number of pulls reaches γB^{2j}. Let $\Theta_z = (\Theta_{z,1}, \ldots, \Theta_{z,\gamma B^{2j}})$ be the γB^{2j} pull outcomes. If $|\Theta_z| \notin [\zeta_1, \zeta_3]$, we publish the arm.

3) If the number of unpublished arms is not in the range of I_{j+1}, or there is a published arm with mean $\left(\frac{1}{2}-B^{-L}\right)$, then we return "error".

We note that the first two steps will only help the algorithm, and thus will only lead to a stronger lower bound. We will show that the extra error introduced by the last step is small, which will be counted in the error probability increase in the induction.

The following claim bounds the number of arms that are not published after the first round.

Claim 7. *For any $j \le \frac{L}{2}-1$, with probability at least $1-O\left(\frac{1}{\kappa}\right)$, the number of unpublished arms after the first round is in the range I_{j+1}.*

Proof. For each arm $z \in [n_j]\backslash S$, let $\left(\frac{1}{2}-X\right)$ be its mean where $X \sim \pi \in \mathcal{D}_j$. Let Y_z be the indicator variable of the event that arm z is not published. By Lemma 2,

$$\begin{aligned}\Pr[Y_z=1] &= \sum_{\ell>j}\Pr[X=B^{-\ell}]\pm n^{-9}\\ &= \left(1\pm\frac{1}{B}\right)\cdot\left(1\pm\frac{2}{\rho}\right)B^{2j}\\ &\quad\cdot B^{-2(j+1)}\left(1\pm\rho^{-(j+1)}\cdot\rho^{j-1}\right)\pm n^{-9}\\ &= \left(1\pm\frac{1}{L^2}\right)\cdot B^{-2},\end{aligned}$$

where the second inequality holds since $\Pr[X = B^{-\ell}]$ decreases at a rate of approximately B^{-2} when ℓ increments, and the last inequality holds since $\rho=(\ln n)^3$ and $L<\ln n$.

By Chernoff-Hoeffding bound, and the fact that we publish all arms in S, we have

$$\sum_{z\in[n_j]} Y_z = \left(1\pm\frac{2}{L^2}\right)B^{-2}(n_j-|S|)$$

with probability $1-e^{-\Omega(n_j(BL)^{-4})}\ge 1-O\left(\frac{1}{\kappa}\right)$. Plugging the fact that $|S|\le\frac{1}{\kappa}\cdot n_j$, we have that with probability $1-O\left(\frac{1}{\kappa}\right)$ over distribution μ,

$$\sum_{z\in[n_j]} Y_z = \left(1\pm\frac{2}{L^2}\right)\left(1\pm\frac{1}{\kappa}\right)B^{-2}n_j = \left(1\pm\frac{1}{L}\right)B^{-2}n_j.$$

Therefore, if $n_j \in I_j$, then with probability $1-O\left(\frac{1}{\kappa}\right)$, $\sum_{z\in[n_j]} Y_z \in I_{j+1}$. □

The following claim shows that the best arm is not likely to be published in the first round.

Claim 8. *For any $j \le \frac{L}{2}-1$, the probability that there is a published arm with mean $\left(\frac{1}{2}-B^{-L}\right)$ is at most $O\left(\frac{1}{\kappa}\right)$.*

Proof. Since the input distribution to $\mathcal{A}$ belongs to the class $(\mathcal{D}_j)^{n_j}$, the probability that S contains an arm with mean $\left(\frac{1}{2}-B^{-L}\right)$, conditioned on $|S|\le\frac{1}{\kappa}\cdot n_j$, can be upper bounded by

$$\begin{aligned}&1-\left(1-\lambda_j B^{-2L}\cdot(1+\rho^{-L+j})\right)^{\frac{n_j}{\kappa}}\\ \le\ &1-\left(1-\lambda_j B^{-2L}\cdot(1+\rho^{-L+j})\right)^{\left(\left(1+\frac{1}{L}\right)B^{-2}\right)^{j-1}\cdot\frac{n}{\kappa}}\\ =\ &1-\left(1-\frac{\lambda_j}{B^{2L}}\cdot(1+\rho^{-L+j})\right)^{\left(\left(1+\frac{1}{L}\right)B^{-2}\right)^{j-1}\cdot\frac{B^{2L}}{\lambda_1}\frac{1}{\kappa}}\\ =\ &O\left(\frac{1}{\kappa}\right).\end{aligned}$$

For each arm $z\in[n]\backslash S$ arms, by Lemma 2 we have that if arm z has mean $\left(\frac{1}{2}-B^{-L}\right)$, then with probability at least $(1-n^{-9})$ we have $|\Theta_z|\in[\zeta_1,\zeta_3]$. The lemma follows by a union bound. □

By Claim 7, Claim 8 and Lemma 3 (which states that if an arm is not published, then its posterior distribution belongs to $\mathcal{D}_{j+1}$), for $j\le\frac{L}{2}-1$, if there is no $(r-1)$-round (αW)-cost algorithm with error probability δ' on any input distribution in $(\mathcal{D}_{j+1})^{n_{j+1}}$ for any $n_{j+1}\in I_{j+1}$, then there is no r-round (αW)-cost algorithm with error probability $\left(\delta'-O\left(\frac{1}{\kappa}\right)\right)$ on any input distribution in $(\mathcal{D}_j)^{n_j}$ for any $n_j\in I_j$, which proves Lemma 6. □

The Base Case. Recall that in our collaborative learning model, if an algorithm uses 0 round then it needs to output the answer immediately (without any further arm pull). We have the following lemma.

Lemma 9. *Any 0-round deterministic algorithm must have error probability at least $(1-o(1))$ on any distribution in $(\mathcal{D}_{\frac{L}{2}})^{n_{\frac{L}{2}}}$ (for any $n_{\frac{L}{2}}\in I_{\frac{L}{2}}$) conditioned on $\mathcal{E}_0$.*

Proof. First we have

$$\begin{aligned}n_{\frac{L}{2}} &= \left(\left(1\pm\frac{1}{L}\right)B^{-2}\right)^{\frac{L}{2}-1}n\\ &= \left(\left(1\pm\frac{1}{L}\right)B^{-2}\right)^{\frac{L}{2}-1}\frac{B^{2L}}{B^2}\\ &= \Theta(B^L).\end{aligned}\tag{13}$$

Thus the probability that there exists at least one arm with mean $\left(\frac{1}{2}-B^{-L}\right)$ is

$$1-\left(1-\left(1\pm\frac{1}{B}\right)B^{-L}\cdot\left(1\pm\rho^{-L}\cdot\rho^{\frac{L}{2}}\right)\right)^{n_{\frac{L}{2}}}=\Theta(1).$$

For each arm i in the $n_{\frac{L}{2}}$ arms, the probability that i and only i has mean $\left(\frac{1}{2}-B^{-L}\right)$ is

$$\begin{aligned}&\lambda_{\frac{L}{2}}B^{-2L}(1\pm\rho^{-\frac{L}{2}})\left(1-\lambda_{\frac{L}{2}}B^{-2L}(1\pm\rho^{-\frac{L}{2}})\right)^{n_{\frac{L}{2}}-1}\\ =\ &\Theta\left(1/n_{\frac{L}{2}}\right).\end{aligned}$$

Therefore any 0-round deterministic algorithm computes the best arm on any distribution in $(\mathcal{D}_{\frac{L}{2}})^{n_{\frac{L}{2}}}$ conditioned on $\mathcal{E}_0$ with probability at most $O\left(1/n_{\frac{L}{2}}\right)=o(1)$. □

Lemma 5 follows from Lemma 6 and Lemma 9. Note that the extra error accumulated during the induction process is bounded by $L \cdot O\left(\frac{1}{\kappa}\right) = o(1)$ since $L = \Theta(\ln n/(\ln\ln n + \ln\alpha))$.

B. Lower Bound for Adaptive Algorithms

In this section we consider general adaptive algorithms. We prove the following theorem.

Theorem 10. *Let $\tilde{K} = \min\{K, \sqrt{T}\}$. For any $\alpha \in [1, \tilde{K}^{0.1}]$, any (K/α)-speedup randomized algorithm for the fixed-time best arm identification problem in the collaborative learning model with K agents needs $\Omega(\ln\tilde{K}/(\ln\ln\tilde{K} + \ln\alpha))$ rounds in expectation.*

The high level idea for proving Theorem 10 is the following: We show that adaptivity cannot give much advantage to the algorithm under the input distribution σ (defined in Section III-A2) when the number of arms n is smaller than the number of agents K. For this purpose we choose n such that

$$nB^2 = \tilde{K}, \tag{14}$$

where $\tilde{K} = \min\{K, \sqrt{T}\}$, and $B = \alpha(\ln n)^{100}$ is the parameter defined at the beginning of Section III-A. We thus have $n \le \sqrt{T}$, and if $\alpha \le \tilde{K}^{0.1}$ then we have $\alpha \le n^{0.2}$; both conditions are needed if we are going to "call" Theorem 1 (for the non-adaptive case) later in the proof, that is, we will use the proof for the non-adaptive case as a subroutine in the proof for the adaptive case.

We will focus on the case when $\sqrt{T} \ge K$; the proof for the other case is essentially the same.

We make use of the same induction (including notations and the algorithm augmentation) as that for the non-adaptive case in Section III-A. Clearly, the base case (i.e., Lemma 9) still holds in the adaptive case since no pull is allowed.

Lemma 11. *Any 0-round deterministic algorithm must have error probability $1 - o(1)$ on any distribution in $(\mathcal{D}_{\frac{L}{2}})^{n_{\frac{L}{2}}}$ (for any $n_{\frac{L}{2}} \in I_{\frac{L}{2}}$) conditioned on $\mathcal{E}_0$.*

Our task is to show the following induction step.

Lemma 12. *For any $j \le \frac{L}{2} - 1$, if there is no $(r-1)$-round (K/α)-speedup deterministic adaptive algorithm with error probability $\delta + O\left(\frac{1}{\kappa}\right)$ on any input distribution in $(\mathcal{D}_{j+1})^{n_{j+1}}$ for any $n_{j+1} \in I_{j+1}$, then there is no r-round (K/α)-speedup deterministic adaptive algorithm with error probability δ on any input distribution in $(\mathcal{D}_j)^{n_j}$ for any $n_j \in I_j$.*

We comment that Lemma 12 does not hold when $n \gg K$ (e.g., $n \ge K^2$), and this is why we can only prove a lower bound of $\Omega(\ln K/(\ln\ln K + \ln\alpha))$ (Theorem 10) instead of $\Omega(\ln n/(\ln\ln n + \ln\alpha))$ (Theorem 1). In the rest of this section we prove Lemma 12.

Proof. Let $\mathcal{E}_1$ denote the event that all the n_j arms have means $(\frac{1}{2} - B^{-\ell})$ for $\ell \ge j$. Since the input is sampled from a distribution in $(\mathcal{D}_j)^{n_j}$, we have

$$\Pr[\mathcal{E}_1] \ge (1 - n^{-9})^{n_j} \ge 1 - n^{-7}. \tag{15}$$

Let $(\Theta_1, \ldots, \Theta_t)$ be the outcomes of t pulls when running the adaptive algorithm $\mathcal{A}$ on an input distributed according to $\mu \in (\mathcal{D}_j)^{n_j}$. We have the following simple fact.

Fact 13. *For any $t \ge 1$, for any possible set of outcomes $(\theta_1, \ldots, \theta_t) \in \{0,1\}^t$, we have*

$$\Pr[(\Theta_1, \ldots, \Theta_t) = (\theta_1, \ldots, \theta_t) \mid \mathcal{E}_1] = \left(\frac{1}{2} \pm B^{-j}\right)^t.$$

Let us conduct a thought experiment. During the run of the adaptive algorithm $\mathcal{A}$, whenever $\mathcal{A}$ pulls an arm, we sample instead an *unbiased* coin and let the result be the pull outcome. Let $(\Theta'_1, \ldots, \Theta'_t)$ be the outcomes of t pulls. It is easy to see that for any $(\theta_1, \ldots, \theta_t) \in \{0,1\}^t$, we have

$$\begin{aligned} q(\theta_1, \ldots, \theta_t) &= \Pr[(\Theta'_1, \ldots, \Theta'_t) = (\theta_1, \ldots, \theta_t) \mid \mathcal{E}_1] \\ &= \left(\frac{1}{2}\right)^t. \end{aligned} \tag{16}$$

In a (K/α)-speedup deterministic algorithm $\mathcal{A}$, each agent can make at most $t = \alpha W/K$ pulls. By Claim 13, (16), and the fact that we have set $n = K/B^2$, for any possible pull outcomes $(\theta_1, \ldots, \theta_t) \in \{0,1\}^t$, conditioned on $\mathcal{E}_1$, it holds that

$$\begin{aligned} \frac{p(\theta_1, \ldots, \theta_t)}{q(\theta_1, \ldots, \theta_t)} &= \frac{\left(\frac{1}{2} \pm B^{-j}\right)^t}{\left(\frac{1}{2}\right)^t} = (1 \pm 2B^{-j})^{\frac{\alpha W}{K}} \\ &= (1 \pm 2B^{-j})^{\alpha \ln^3 n} = \left[\frac{1}{2}, 2\right]. \end{aligned} \tag{17}$$

Let $X_{i,z}$ be the expected number of pulls to arm z by agent i when running $\mathcal{A}$ on input distribution μ. Let $Y_{i,z}$ be the expected number of pulls to arm z by agent i when we we simply feed random $0/1$ outcome to $\mathcal{A}$ at each pull step. By (17) we have that conditioned on $\mathcal{E}_1$.

$$\forall i \in [K], \forall z \in [n_j], \quad \frac{Y_{i,z}}{2} \le X_{i,z} \le 2Y_{i,z}. \tag{18}$$

Since $\sum_{i\in[K]}\sum_{z\in[n_j]} X_{i,z} \le \alpha W$, conditioned on $\mathcal{E}_1$ we have

$$\sum_{i\in[K]}\sum_{z\in[n_j]} Y_{i,z} \le 2\alpha W. \tag{19}$$

The key observation is that running $\mathcal{A}$ with random $0/1$ pull outcomes is more like running a non-adaptive algorithm. Indeed, we can sample a random bit string of length equal to the number of pulls at the beginning of the algorithm, and then the sequence of indices of arms that will be pulled are fully determined by the random bit string and the decision tree of the deterministic algorithm $\mathcal{A}$. In other words, all $Y_{i,z}$'s can be computed before the run of the algorithm $\mathcal{A}$.

By (19) and a simple counting argument, conditioned on $\mathcal{E}_1$, we have that for at most $1/\kappa$ fraction of arms $z \in [n_j]$, it holds that

$$\sum_{i\in[K]} Y_{i,z} \ge \frac{2\alpha\kappa W}{n_j}. \tag{20}$$

Denote the set of such z's by Q; we thus have $|Q| \le 1/\kappa \cdot n_j$. Note that Q can again be computed before the run of the

algorithm $\mathcal{A}$. By (18) and (20), we have that conditioned on $\mathcal{E}_1$, for any $z \in [n_j] \backslash Q$,

$$\sum_{i \in [K]} X_{i,z} \leq \frac{4\alpha\kappa W}{n_j} \leq \gamma B^{2j}. \tag{21}$$

Inequality (21) tells that for any arms $z \in [n_j] \backslash Q$, the total number of pulls on z over the K agents is at most γB^{2j}, which is the same as that in the proof for the non-adaptive case in Lemma 6 (Q corresponds to S in the proof of Lemma 6). We also have $\Pr[\neg \mathcal{E}_1] \leq n^{-7} \leq 1/\kappa$ which will contribute to the extra error in the induction. The rest of the proof simply follows from that for Lemma 6. □

IV. Fixed-Time Distributed Algorithms

In this section we present our fixed-time collaborative learning algorithm for the best arm identification problem. The algorithm takes a set $S = [n]$ of n arms, a time horizon T, and a round parameter R as input, and is guaranteed to terminate by the T-th time step and uses at most R rounds. We assume without loss of generality that $1 \in S$ is the best arm. We state the following theorem as our main algorithmic result.

Theorem 14. *Let $H = H(I)$ be the complexity parameter of the input instance I defined in (10). There exists a collaborative learning algorithm with time budget T and round budget R that returns the best arm with probability at least*

$$1 - n \cdot \exp\left(-\Omega\left(\frac{TK^{\frac{R-1}{R}}}{H \ln(HK)(\ln(TK^{\frac{R-1}{R}}/H))^2}\right)\right).$$

We now show that the algorithm in Theorem 14 has $\tilde{\Omega}(K^{\frac{R-1}{R}})$ speedup.

Theorem 15. *For any $R \geq 1$, there exists a fixed-time algorithm $\mathcal{A}$ such that $\beta_{\mathcal{A}}(T) = \Omega(K^{\frac{R-1}{R}} \ln(nTK)^{-4})$ for sufficiently large T. When $R = \Theta(\ln K)$, the speedup of the algorithm is $\tilde{\Omega}(K)$.*

Proof. It is know [1] that for every instance I, it holds that

$$\inf_{\text{centralized } \mathcal{O}} \delta_{\mathcal{O}}(I, T) \geq \frac{1}{2} \cdot \exp(-O(T/H)).$$

Therefore, for every $\delta \leq 1/3$, we have that

$$\inf_{\text{centralized } \mathcal{O}} T_{\mathcal{O}}(I, \delta) \geq \Omega(H \ln(1/\delta)). \tag{22}$$

On the other hand, let $\mathcal{A}$ be the algorithm in Theorem 14, for $\delta \leq 1/3$, we have that

$$T_{\mathcal{A}}(I, \delta) \leq O\left(HK^{-\frac{R-1}{R}} \ln(nHK/\delta)^4\right). \tag{23}$$

Combining (22) and (23), we have

$$\inf_{\text{centralized } \mathcal{O}} \inf_{\delta \in (0,1/3]: T_{\mathcal{O}}(I,\delta) \leq T} \frac{T_{\mathcal{O}}(I, \delta)}{T_{\mathcal{A}}(I, \delta)} \geq \Omega\left(\frac{K^{\frac{R-1}{R}}}{\ln(nTK)^4}\right),$$

which implies that $\beta_{\mathcal{A}}(T) = \Omega(K^{\frac{R-1}{R}} \ln(nTK)^{-4})$. □

The rest of this section is devoted to the proof of Theorem 14. In Section IV-A, we first prove a special case of Theorem 14 when $T = \Theta(HK^{-\frac{R-1}{R}} \ln(HK))$, for which the algorithm is guaranteed to output the best arm with constant probability. Then, in Section IV-B, we prove Theorem 14 by performing a technical modification to Algorithm 1 and a reduction from general parameter settings to several independent runs of modified Algorithm 1 with different parameters.

A. Special Case when $T = \Theta(HK^{-\frac{R-1}{R}} \ln(HK))$

Our algorithm for the special case when $T = \Theta(HK^{-\frac{R-1}{R}} \ln(HK))$ is presented in Algorithm 1. We have the following guarantees.

Theorem 16. *Let H be the instance dependent complexity parameter defined in (10). There exists a universal constant $c_{\mathrm{ALG}} > 0$ such that if $T \geq c_{\mathrm{ALG}} HK^{-\frac{R-1}{R}} \ln(HK)$, then Algorithm 1 returns the best arm with probability at least* 0.97.

Algorithm 1 uses a fixed-confidence centralized procedure $\mathcal{A}_{\mathrm{C}}$ as a building block, with the following guarantees.

Lemma 17. *(See, e.g. [14], [15], [18], [23]) There exists a centralized algorithm $\mathcal{A}_{\mathrm{C}}(I, \delta)$ where I is the input and δ is the error probability parameter, such that the algorithm returns the best arm and uses at most $O(H(I)(\ln H(I) + \ln \delta^{-1}))$ pulls with probability at least $(1 - \delta)$.*

We describe Algorithm 1 briefly in words. At a high level, the algorithm goes by R iterations. We keep a set of active arms, denoted by S_{r-1}, at the beginning of each iteration r with $S_0 = [n]$. During each iteration r, the agents collectively learn more information about the active arms in S_{r-1} and eliminate a subset of arms to form S_r. This is done in four steps. In the *preparation* step, each agent ℓ is assigned with exactly one arm $i_\ell^{(r)}$, which is the one it will learn in the later steps. If there are more agents than active arms, we simply assign each arm to $K/|S_{r-1}|$ agents. Otherwise, we first assign each arm to a random agent (which can be done by shared randomness without communication), and then each agent uses the centralized procedure $\mathcal{A}_{\mathrm{C}}$ to identify $i_\ell^{(r)}$ as the best arm among the set of assigned arms. We note that the latter case will only happen during iteration $r = 1$ (if it ever happens). Then each agent ℓ plays $i_\ell^{(r)}$ in the *learning* step and shares his own observation in the *communication and aggregation* step. In the *elimination* step, we calculate the confidence interval (CI) for each active arm using a carefully designed dependence on T, K, and R, and eliminate the arms whose CI does not overlap with the best arm. We note that this algorithm uses R communication steps, and therefore needs $(R + 1)$ rounds. In Section IV-A1, we describe a trick to shave 1 communication step and make the algorithm runs in R rounds.

For convenience, we assume without loss of generality that arm 1 is the best arm in the input set S. We first establish the following lemma which concerns about Lines 3–6 in Algorithm 1.

Lemma 18. *For large enough constant $c_{\mathrm{ALG}} > 0$ and $T \geq c_{\mathrm{ALG}} HK^{-\frac{R-1}{R}} \ln(HK)$, suppose Lines 3–6 are executed*

Algorithm 1: Fixed-Time Collaborative Learning Best Arm Identification with Constant Error Probability

Input: a set of arms $S = [n]$, time horizon T and communication steps R ($R \le O(\ln K)$)

initialize $S_0 \leftarrow S$

for *iteration* $r = 1$ *to* R **do**

`/* Step 1: preparation */`

if $|S_{r-1}| > K$ **then**

randomly assign each arm in S_{r-1} to one of the K agents, and let A_ℓ be the set of arms assigned to agent ℓ

for *agent* $\ell = 1$ *to* K **do**

$i_\ell^{(r)} \leftarrow \mathcal{A}_{\mathrm{C}}(A_\ell, 0.01)$, if $\mathcal{A}_{\mathrm{C}}$ does not terminate within $T/2$ pulls, stop the procedure anyways and set $i_\ell^{(r)} \leftarrow \perp$

else

assign each arm in S_{r-1} to $K/|S_{r-1}|$ agents (so that each agent is assigned with exactly one arm), and let $i_\ell^{(r)}$ be the arm assigned to agent ℓ

`/* Step 2: learning */`

for *agent* $\ell = 1$ *to* K **do**

play arm $i_\ell^{(r)}$ for $\frac{1}{2} \cdot T/R$ times and let $\hat{p}_\ell^{(r)}$ be the average of the observed rewards (if $i_\ell^{(r)} \neq \perp$)

`/* Step 3: communication and aggregation */`

for *agent* $\ell = 1$ *to* K **do**

broadcast $i_\ell^{(r)}$ and $\hat{p}_\ell^{(r)}$

$\tilde{S}_r \leftarrow \{i_\ell^{(r)} : \ell \in [K]\}$

Let $\hat{q}_i^{(r)} = \frac{1}{|\{\ell \in [K] : i_\ell = i\}|} \sum_{\ell \in [K] : i_\ell = i} \hat{p}_l^{(r)}$ for each $i \in \tilde{S}_r$

`/* Step 4: elimination */`

$$S_r \leftarrow \tilde{S}_r \setminus \left\{ i \in \tilde{S}_r : \text{ there exists an arm } j \text{ with } \hat{q}_j^{(r)} \ge \hat{q}_i^{(r)} + 2 \cdot \sqrt{\frac{R \ln(200KR)}{\max\{1, K/|S_{r-1}|\} \cdot T}} \right\}$$

return the only arm in S_R if $|S_R| = 1$, and $\perp$ otherwise

during iteration $r = 1$, then after the preparation step, with probability at least 0.98, there exists an agent $\ell \in [K]$ such that $i_\ell^{(r)} = 1$.

Proof. Let ℓ^* be the agent such that $1 \in A_{\ell^*}$. Since $H(A_{\ell^*}) = \sum_{i \in A_{\ell^*} \setminus \{1\}} \Delta_i^{-2}$. By linearity of expectation, we have that $\mathbb{E}[H(A_{\ell^*})] = \sum_{i \in S \setminus \{1\}} \Delta_i^{-2}/K = H/K$. By Markov's Inequality, and for large enough c_{ALG} and $T \ge c_{\mathrm{ALG}} H K^{-\frac{R-1}{R}} \ln H \ge c_{\mathrm{ALG}} H \ln(HK)/K$, we have that with probability at least 0.99, $T/2$ is greater than or equal to the sample complexity bound in Lemma 17 for $S = A_{\ell^*}$ and $\delta = 0.01$. Taking a union bound with the event that the run of $\mathcal{A}_{\mathrm{C}}(A_{\ell^*}, 0.01)$ is as described in Lemma 17, we have that $\Pr[i_{\ell^*}^{(r)} = 1] \ge 0.98$.

□

The following lemma concerns about the learning and elimination steps of Algorithm 1.

Lemma 19. *During each iteration r, assuming that $1 \in \tilde{S}_r$, with probability at least $(1 - 0.01/R)$,*

1) *we have that $1 \in S_r$;*
2) *if we further assume 1) $T \ge c_{\mathrm{ALG}} H K^{-\frac{R-1}{R}} \ln(HK)$ for sufficiently large $c_{\mathrm{ALG}} > 0$ and 2) either $r = 1$ or $|S_{r-1}| \le K^{\frac{R-r+1}{R}}$, we have that $|S_r| \le K^{\frac{R-r}{R}}$.*

Proof. Note that for each $i \in \tilde{S}_r$, we have that $|\{\ell \in [K] : i_\ell^{(r)} = i\}| \ge \max\{1, K/|S_{r-1}|\}$. Therefore, $\hat{q}_i^{(r)}$ is the average of at least $\max\{1, K/|S_{r-1}|\} \cdot \frac{1}{2} T/R$ pulls of arm i. By Chernoff-Hoeffding bound, we have

$$\Pr\left[\left|\hat{q}_i^{(r)} - \theta_i\right| > \sqrt{\frac{R \ln(200KR)}{\max\{1, K/|S_{r-1}|\} \cdot T}}\right] \le \frac{1}{20KR}. \tag{24}$$

We now condition on the event that

$$\forall i \in \tilde{S}_r : \left|\hat{q}_i^{(r)} - \theta_i\right| \le \sqrt{\frac{R \ln(200KR)}{\max\{1, K/|S_{r-1}|\} \cdot T}},$$

which holds with probability at least $(1 - 0.01/R)$ by (24), the fact that $|\tilde{S}_r| \le K$, and a union bound. Let $\mathcal{E}_3$ denote this event.

For the first item in the lemma, it is straightforward to verify that $1 \in S_r$ since for any suboptimal arm $i \in \tilde{S}_r \setminus \{1\}$, it holds that

$$\begin{aligned}\hat{q}_i^{(r)} - \hat{q}_1^{(r)} &\le \theta_i - \theta_1 + 2\sqrt{\frac{R \ln(200KR)}{\max\{1, K/|S_{r-1}|\} \cdot T}} \\ &< 2\sqrt{\frac{R \ln(200KR)}{\max\{1, K/|S_{r-1}|\} \cdot T}}.\end{aligned}$$

We now show the second item in the lemma. With the additional assumptions (in the second item), we have that $\max\{1, K/|S_{r-1}|\} \ge K^{\frac{r-1}{R}}$. Thus conditioned on $\mathcal{E}_3$, for all arms $i \in \tilde{S}_r$ it holds that

$$\left|\hat{q}_i^{(r)} - \theta_i\right| \le \sqrt{\frac{R \ln(200KR)}{K^{\frac{r-1}{R}} T}}.$$

For any suboptimal arm $i \in \tilde{S}_r$, the corresponding gap Δ_i has to be less or equal to $4\sqrt{\frac{R \ln(200KR)}{K^{\frac{r-1}{R}} T}}$ so that it may stay in S_r. This is because otherwise we have

$$\begin{aligned}\hat{q}_i^{(r)} + 2\sqrt{\frac{R \ln(200KR)}{K^{\frac{r-1}{R}} T}} &\le \theta_i + 3\sqrt{\frac{R \ln(200KR)}{K^{\frac{r-1}{R}} T}} \\ &\le \theta_1 - \sqrt{\frac{R \ln(200KR)}{K^{\frac{r-1}{R}} T}} \le \hat{q}_1^{(r)},\end{aligned}$$

and the arm will be eliminated at Line 15. Since $H \le \frac{TK^{\frac{R-1}{R}}}{c_{\mathrm{ALG}} H \ln(HK)}$ and c_{ALG} is a large enough constant, the number of suboptimal arms i such that $\Delta_i \le 4\sqrt{\frac{R\ln(200KR)}{K^{\frac{r-1}{R}}T}}$ can be upper bounded by

$$\frac{TK^{\frac{R-1}{R}}}{c_{\mathrm{ALG}} H \ln(HK)} \cdot 16 \cdot \frac{R\ln(200KR)}{K^{\frac{r-1}{R}}T} < K^{\frac{R-r}{R}},$$

and therefore $|S_r| \le K^{\frac{R-r}{R}}$. □

Analysis of Algorithm 1. By Lemma 18, we have $1 \in \tilde{S}_1$ with probability 0.98, conditioned on which and applying Lemma 19, we have both $1 \in S_R$ and $|S_R| \le 1$ with probability 0.99 (by a union bound over all R iterations). Therefore, Algorithm 1 outputs arm 1 (the best arm) with probability 0.97. □

1) Further Improvement on the Round Complexity

We have proved that Algorithm 1 satisfies the requirement in Theorem 16 using R communication steps, and therefore $(R+1)$ rounds. Now we sketch a trick to further reduce the number of communication steps of Algorithm 1 by one, and therefore the algorithm only uses R rounds, fully proving Theorem 16.

The main modification is made to the first iteration ($r = 1$) of Algorithm 1. In the preparation step, if $|S_0| > K^{\frac{R-1}{R}}$, then we randomly assign each arm in S_0 to $100K^{\frac{1}{R}}$ agents, and each agent uses the same procedure to identify $i_\ell^{(1)}$. Otherwise, the routine of the algorithm remains the same.

If $|S_0| > K^{\frac{R-1}{R}}$, in the elimination step, we first set $\tilde{S}_1$ to be the set of arms that are identified by at least $K^{\frac{1}{R}}$ agents in the preparation step. Then the elimination rule in Line 15 remains the same.

The rest iterations $r = 2, 3, \ldots$ remains the same. However, we only need to proceed to the $(R-1)$-st iteration and therefore the algorithm uses $(R-1)$ communication steps and R rounds.

To analyze the modified algorithm, the main difference is that we can strengthen Lemma 18 by showing that $1 \in \tilde{S}_1$ with probability at least 0.9. This is because by Markov's Inequality, for each agent ℓ such that $1 \in A_\ell$, with probability at least 0.99, $T/2$ is greater than or equal to the sample complexity of the instance A_ℓ (with error probability $\delta = 0.01$)), and therefore $\Pr[i_\ell^{(1)} = 1] \ge 0.98$. Therefore, the expected number of agents that identify arm 1 is at least $0.98 \cdot 100K^{\frac{1}{R}} \ge 50K^{\frac{1}{R}}$. Applying Markov's Inequality, we show that $\Pr[1 \in \tilde{S}_1] \ge 0.98$.

We also have that $|\tilde{S}_1| \le K^{\frac{R-1}{R}}$. Therefore, we iteratively apply a similar argument of Lemma 19 to the rest of the $(R-1)$ iterations, we have that with probability at least 0.97, for each $r = 2, 3, \ldots, R-1$, it holds that $1 \in S_r$ and $|S_r| \le R^{\frac{R-r-1}{R}}$. Therefore, the algorithm returns arm 1 after $(R-1)$ iterations with probability at least 0.97.

B. Algorithm for General Parameter Settings

For conciseness of the presentation, we only extend Algorithm 1 (that uses $(R+1)$ rounds) to general parameter settings. It is easy to verify that the same technique works for the algorithm described in Section IV-A1, which will fully prove Theorem 14. In the following of this subsection, we prove Theorem 14 with an algorithm with round complexity $(R+1)$.

We first make a small modification to Algorithm 1 and strengthen its theoretical guarantee. To do this, we need to introduce the following stronger property on the fixed-confidence centralized procedure $\mathcal{A}_{\mathrm{C}}$.

Lemma 20. *There exists a centralized algorithm $\mathcal{A}_{\mathrm{C}}(S, \delta)$ where the input is a set S of arms, such that there exists a cost function f_{C} such that*

$$f_{\mathrm{C}}(S, \delta) \le O(H(S)(\ln H(S) + \ln \delta^{-1})),$$

and the function is monotone in inversed gaps $\Delta_2^{-1}, \Delta_3^{-1}, \ldots, \Delta_{|S|}^{-1}$ where Δ_i is the difference between the mean of the best arm and that of the i-th best arm, and

$$\Pr[\textit{algorithm returns the best arm and uses at least } f_{\mathrm{C}}(S,\delta) \textit{ and at most } 100 f_{\mathrm{C}}(S,\delta) \textit{ pulls}] \ge 1 - \delta.$$

It can be easily verified that the Successive Elimination algorithm in [23] is a valid candidate algorithm for Lemma 20.

We now describe our technical change to Algorithm 1.

> **Algorithm 1′**: In Line 6 of Algorithm 1, instead of choosing $T/2$ as the time threshold, each agent ℓ independently chooses $\tau_\ell \in \{T/200, T/2\}$ uniformly at random and uses τ_ℓ as the time threshold.

It is straightforward to see that for a large enough constant c_{ALG}, Theorem 16 still holds for the Algorithm 1′. We now state the additional guarantee for the Algorithm 1′.

Lemma 21. *For any T and any suboptimal arm $i \in S$, the probability that Algorithm 1′ returns i is at most 0.86.*

Proof. For any fixed suboptimal arm $i \in S$, let p be the probability that Algorithm 1′ returns i.

If Lines 3–6 are not executed during iteration $r = 1$ or there exists an agent ℓ such that the corresponding $i_\ell^{(1)}$ at Line 6 equals to the best arm (arm 1), by Lemma 19 we know that $\Pr[1 \in S_R] \ge 0.99$, and thus the probability that i is returned is at most 0.01. For now on, we focus on the case that Lines 3–6 are executed during iteration $r = 1$ and none of $i_\ell^{(1)}$ equals to 1.

By Lemma 19, we know that $\Pr[\exists \ell : i_\ell^{(1)} = i] \ge p - 0.01$. We further have

$$\Pr[\exists \ell : i_\ell^{(1)} = i \text{ and } \tau_\ell = T/200] \ge p - 0.51$$

since $\Pr[\tau_\ell = T/200] = 0.5$. By Lemma 20, we have that

$$\Pr[\exists \ell : \text{best arm of } A_\ell \text{ is } i \text{ and } f_{\mathrm{C}}(A_\ell, 0.01) \le T/200] \ge p - 0.52. \quad (25)$$

Now consider a new partition of arms $\{A'_\ell\}_{\ell \in [K]}$ which is almost identical to $\{A_\ell\}$ except for that the assignments for arms 1 and i are exchanged. We note that first, the

marginal distribution of $\{A'_\ell\}$ is still the uniform distribution; and second, when i is the best arm of A_ℓ, we have that $f_{\mathrm{C}}(A_\ell, 0.01) \geq f_{\mathrm{C}}(A'_\ell, 0.01)$ due to the monotonicity of f_{C} and the gaps of $H(A_\ell)$ are point-wisely less than or equal to that of $H(A'_\ell)$. By (25),

$$\begin{aligned} & \Pr[\exists \ell : \text{best arm of } A_\ell \text{ is 1 and } f_{\mathrm{C}}(A_\ell, 0.01) \leq T/200] \\ = \; & \Pr[\exists \ell : \text{best arm of } A'_\ell \text{ is 1 and } f_{\mathrm{C}}(A'_\ell, 0.01) \leq T/200] \\ \geq \; & p - 0.52. \end{aligned}$$

By Lemma 19 and Lemma 20, we have that

$$\begin{aligned} \Pr[1 \in S_R] \; \geq \; & \Pr[\exists \ell : i_\ell^{(1)} = 1] - 0.01 \\ \geq \; & \Pr[\exists \ell : \text{best arm of } A_\ell \text{ is 1 and} \\ & f_{\mathrm{C}}(A_\ell, 0.01) \leq T/200 \text{ and } \tau_\ell = T/2] - 0.02 \\ \geq \; & \frac{p - 0.52}{2} - 0.02 \\ = \; & \frac{p}{2} - 0.28. \end{aligned}$$

Since $1 \in S_R$ is a disjoint event from the event that i is returned by the algorithm, we have $p + p/2 - 0.28 \leq 1$, leading to that $p \leq 1.28/1.5 < 0.86$. □

We are now ready to prove the main algorithmic result (Theorem 14).

Proof of Theorem 14. We build a meta algorithm that independently runs the Algorithm 1′ for several times with different parameters.

> **Meta Algorithm**: For each $s = 1, 2, 3, \ldots$, we run Algorithm 1′ with time horizon $\frac{T}{s^2 10^s} \cdot \frac{6}{\pi^2}$ and communication step parameter R for 10^s times, and let the returned values be $i_{s,1}, i_{s,2}, \ldots, i_{s,10^s}$. Finally, the algorithm will find the largest s such that the most frequent element in $\{i_{s,\cdot}\}$ has frequency greater than 0.9 and output the corresponding element, or output $\perp$ if no such s exists.

We note that we can still do this in R communication steps and the total run time will be at most

$$\sum_s 10^s \cdot \frac{T}{s^2 10^s} \cdot \frac{6}{\pi^2} \leq T.$$

Let s^* be the largest $s \geq 1$ such that $\frac{T}{s^2 10^s} \cdot \frac{6}{\pi^2} \geq c_{\mathrm{ALG}} H K^{-\frac{R-1}{R}} \ln(HK)$, where c_{ALG} is the constant in Theorem 16 for Algorithm 1′. If no such s exists, it is easy to verify that the theorem holds trivially. Otherwise, we have that $2^{s^*} = \Omega(TK^{\frac{R-1}{R}} / (H \ln(HK)(\ln(TK^{\frac{R-1}{R}}/H))^2))$.

By Theorem 16 and Chernoff-Hoeffding bound, we have that

$$\begin{aligned} & \Pr[\text{frequency of 1 in } \{i_{s^*,\cdot}\} > 0.9] \\ & \geq 1 - \exp\left(-\Omega\left(\frac{TK^{\frac{R-1}{R}}}{H \ln(HK)(\ln(TK^{\frac{R-1}{R}}/H))^2}\right)\right). \end{aligned}$$

On the other hand, for each $s = s^* + j$ (where $j \geq 1$), by Lemma 21, Chernoff-Hoeffding bound, and a union bound, we have that

$$\begin{aligned} & \Pr[\exists \text{suboptimal arm } i : \text{frequency of } i \text{ in } \{i_{s,\cdot}\} > 0.9] \\ & \leq n \cdot \exp\left(-2^j \cdot \Omega\left(\frac{TK^{\frac{R-1}{R}}}{H \ln(HK)(\ln(TK^{\frac{R-1}{R}}/H))^2}\right)\right). \end{aligned}$$

Finally, we have

$$\begin{aligned} & \Pr[\text{Meta Algorithm returns 1}] \\ \geq \; & \Pr[\text{frequency of 1 in } \{i_{s^*,\cdot}\} > 0.9] \\ & - \sum_{j=1}^{+\infty} \Pr[\exists \text{suboptimal arm } i : \text{freq. of } i \text{ in } \{i_{s^*+j,\cdot}\} > 0.9] \\ \geq \; & 1 - \sum_{j=0}^{+\infty} n \cdot \exp\left(-2^j \cdot \Omega\left(\frac{TK^{\frac{R-1}{R}}}{H \ln(HK)(\ln(TK^{\frac{R-1}{R}}/H))^2}\right)\right) \\ \geq \; & 1 - n \cdot \exp\left(-\Omega\left(\frac{TK^{\frac{R-1}{R}}}{H \ln(HK)(\ln(TK^{\frac{R-1}{R}}/H))^2}\right)\right). \end{aligned}$$

□

V. Lower Bounds for Fixed-Confidence Distributed Algorithms

In this section, we prove the following lower bound theorem for fixed-confidence collaborative learning algorithms.

Theorem 22. *For any large enough T, suppose that a randomize algorithm $\mathcal{A}$ for the fixed-confidence best arm identification problem in the collaborative learning model with K agents satisfies that $\beta_{\mathcal{A}}(T) \geq \beta$, then we have that $\mathcal{A}$ uses*

$$\Omega\left(\min\left\{\frac{\min\{\ln \Delta_{\min}^{-1}, \ln T\}}{\ln\left(1 + \frac{K(\ln K)^2}{\beta}\right) + \min\{\ln\ln \Delta_{\min}^{-1}, \ln\ln T\}}, \sqrt{\beta/(\ln K)^3}\right\}\right)$$

rounds in expectation.

To prove the theorem, we work with the following simpler problem.

The SignID problem. In the SignID problem, there is only one Bernoulli arm with mean reward denoted by $(\frac{1}{2} + \Delta)$ (where $\Delta \in [-\frac{1}{2}, \frac{1}{2}] \setminus \{0\}$). The goal for the agent is to make a few pulls on the arm and decide whether $\Delta > 0$ or $\Delta < 0$. Let $I(\Delta)$ denote the input instance. Throughout this section, we use the notations $\Pr_{I(\Delta)}[\cdot]$ and $\mathbb{E}_{I(\Delta)}[\cdot]$ to denote the probability and expectation when the underlying input instance is $I(\Delta)$. We say a collaborative learning algorithm $\mathcal{A}$ is δ-error and β-fast for the instance $I(\Delta)$, if we have that

$$\Pr_{I(\Delta)}\left[\mathcal{A} \text{ returns the correct decision within } \Delta^{-2}/\beta \text{ running time}\right] \geq 1 - \delta.$$

We first provide the following theorem on the round complexity lower bound for the SIGNID problem (which will be formally proved in Section V-A). Then we will show how these statements imply the round complexity lower bound for the best arm identification problem in the fixed confidence setting.

Theorem 23. *Let $\Delta^* \in (0, 1/8)$. If $\mathcal{A}$ is a $(1/K^5)$-error and β-fast algorithm for every* SIGNID *problem instance $I(\Delta)$ where $|\Delta| \in [\Delta^*, 1/8)$, then there exists $\Delta^\flat \geq \Delta^*$ such that*

$$\Pr_{I(\Delta^\flat)}\left[\mathcal{A} \text{ uses } \Omega\left(\min\left\{\frac{\ln(1/\Delta^*)}{\ln(1+K/\beta)+\ln\ln(1/\Delta^*)}, \sqrt{\beta/(\ln K)}\right\}\right) \text{ rounds}\right] \geq \frac{1}{2}.$$

Since we can easily convert a $(1/3)$-error and β-fast algorithm to a δ-error and $\beta/O(\ln \delta^{-1})$-fast algorithm for any $\delta < 0$, we have the following corollary.

Corollary 24. *Let $\Delta^* \in (0, 1/8)$. If $\mathcal{A}$ is a $(1/3)$-error and β-fast algorithm for every* SIGNID *problem instance $I(\Delta)$ where $|\Delta| \in [\Delta^*, 1/8)$, then there exists $\Delta^\flat \geq \Delta^*$ such that*

$$\Pr_{I(\Delta^\flat)}\left[\mathcal{A} \text{ uses } \Omega\left(\min\left\{\frac{\ln(1/\Delta^*)}{\ln(1+(K\ln K)/\beta)+\ln\ln(1/\Delta^*)}, \sqrt{\beta/(\ln K)^2}\right\}\right) \text{ rounds}\right] \geq \frac{1}{2}.$$

We now show how Theorem 23 implies the round complexity lower bound for the best arm identification problem. The proof of our main Theorem 22 will come after the following theorem.

Theorem 25. *Let $\Delta^* \in (0, 1/8)$. Given any randomized algorithm $\mathcal{A}_{\mathrm{BAI}}$ for the fixed-confidence best arm identification problem in the collaborative learning model with K agents, if for any 2-arm instance J where $\Delta_{\min}(J) \in [\Delta^*, 1/8)$,*

$$\Pr\left[\mathcal{A}_{\mathrm{BAI}} \text{ returns the best arm of } J \text{ within } \Delta_{\min}^{-2}/\beta \text{ running time}\right] \geq \frac{2}{3},$$

then there exists a 2-arm instance J^* where $\Delta_{\min}(J^*) \in [\Delta^*, 1/8)$, such that

$$\Pr\left[\mathcal{A}_{\mathrm{BAI}} \text{ uses } \Omega\left(\min\left\{\frac{\ln(1/\Delta^*)}{\ln(1+(K\ln K)/\beta)+\ln\ln(1/\Delta^*)}, \sqrt{\beta/(\ln K)^2}\right\}\right) \text{ rounds on } J^*\right] \geq \frac{1}{2}. \quad (26)$$

Proof. We first show that given such algorithm $\mathcal{A}_{\mathrm{BAI}}$ that uses no more than $R = R(\Delta_{\min})$ rounds of communication in expectation, there exists an algorithm $\mathcal{A}$ for SIGNID such that $\mathcal{A}$ is $(1/3)$-error and $\Omega(\beta)$-fast for all instances $I(\Delta)$ where $\Delta \in [\Delta^*, 1/8)$, and $\mathcal{A}$ uses at most $R(\Delta)$ rounds of communication in expectation.

To construct the algorithm $\mathcal{A}$, we set up a best arm identification instance J where one of the two arms (namely the *reference arm*) is set to be a Bernoulli arm with mean reward $1/2$, and the other arm (namely the *unknown arm*) is the one in the SIGNID instance. $\mathcal{A}$ simulates $\mathcal{A}_{\mathrm{BAI}}$ and plays the arm in the SIGNID instance once whenever $\mathcal{A}_{\mathrm{BAI}}$ wishes to play the unknown arm. $\mathcal{A}$ returns '< 0' if and only if $\mathcal{A}_{\mathrm{BAI}}$ returns the reference arm, and $\mathcal{A}$ returns '> 0' if and only if $\mathcal{A}_{\mathrm{BAI}}$ returns the unknown arm.

Suppose $I(\Delta)$ is the given SIGNID instance, we have that $\Delta_{\min}(J) = \Delta$, and therefore $\mathcal{A}$ uses $R(\Delta)$ rounds of communication in expectation. Also one can verify that $\mathcal{A}$ is a $(1/3)$-error and β-fast algorithm for $I(\Delta)$ whenever $\Delta \in [\Delta^*, 1/8)$. By Corollary 24, there exists $\Delta^\flat \geq \Delta^*$ such that

$$\Pr_{I(\Delta^\flat)}\left[\mathcal{A} \text{ uses } \Omega\left(\min\left\{\frac{\ln(1/\Delta^*)}{\ln(1+(K\ln K)/\beta)+\ln\ln(1/\Delta^*)}, \sqrt{\beta/(\ln K)^2}\right\}\right) \text{ rounds}\right] \geq \frac{1}{2}.$$

This implies that for the 2-arm instance J^* where $\Delta_{\min}(J^*) = \Delta^\flat$, we have that (26) holds. □

Proof of Theorem 22. Let $J(\Delta)$ be the 2-arm instance where one of the two arms is a Bernoulli arm with mean reward $1/2$ and the other arm is a Bernoulli arm with mean reward $1/2-\Delta$. By the lil'UCB algorithm in [15], we know that there exists a centralized algorithm $\mathcal{O}$ such that $T_{\mathcal{O}}(J(\Delta), 1/3) \leq O(\Delta^{-2}\ln\ln\Delta^{-1})$ for all $\Delta \in (0, 1/4)$. Therefore, there exists a universal constant $c > 0$ such that for any large enough T, we have $T_{\mathcal{O}}(J(cT/\ln T), 1/3) \leq T$ for all $\Delta \in (0, 1/4)$.

For any $\Delta_{\min}$, we set $\Delta^* = \max\{\Delta_{\min}, cT/\ln T\}$. By the definition of $\beta_{\mathcal{A}}(T)$ (in (1)) and the assumption that $\beta_{\mathcal{A}}(T) \geq \beta$, we have that for all instance $J(\Delta)$ where $\Delta \in [\Delta^*, 1/4)$, it holds that

$$\frac{T_{\mathcal{O}}(J(\Delta), 1/3)}{T_{\mathcal{A}}(J(\Delta), 1/3)} \geq \beta,$$

which implies that

$$T_{\mathcal{A}}(J(\Delta), 1/3) \leq \frac{T_{\mathcal{O}}(J(\Delta), 1/3)}{\beta} = O(\Delta^{-2}\ln\ln\Delta^{-1}/\beta) = O(\Delta^{-2}\ln\ln T/\beta).$$

We now invoke Theorem 25, and get that there exists J^* and $\mathcal{O}$ such that $T_{\mathcal{O}}(J^*, 1/3) \leq T$ and

$$\Pr_{J^*}\left[\mathcal{A} \text{ uses } \Omega\left(\min\left\{\frac{\ln(1/\Delta^*)}{\ln(1+(K\ln K\ln\ln T)/\beta)+\ln\ln(1/\Delta^*)}, \sqrt{\frac{\beta}{(\ln K)^2\ln\ln T}}\right\}\right) \text{ rounds}\right] \geq \frac{1}{2}.$$

Note that $\ln(1/\Delta^*) \leq O(\ln T)$. When $\ln T = \Omega(K)$, the second term in the $\min\{.,.\}$ function becomes smaller.

Therefore, in the first term, we can assume that $\ln T = O(K)$ and get the following simplified statement.

$$\Pr_{J^*}\left[\mathcal{A} \text{ uses } \Omega\left(\min\left\{\frac{\ln(1/\Delta^*)}{\ln(1+(K(\ln K)^2)/\beta)+\ln\ln(1/\Delta^*)}, \sqrt{\frac{\beta}{(\ln K)^3}}\right\}\right) \text{ rounds}\right] \geq \frac{1}{2}.$$

□

A. Proof of Theorem 23

Suppose $\mathcal{A}$ is a δ-error β-fast algorithm. We define the following events. For any integer $\alpha \geq 0$, let $\mathcal{E}(\alpha, T)$ to denote the event that $\mathcal{A}$ uses at least α rounds and at most T time steps before the end of the α-th round, and let $\mathcal{E}^*(\alpha, T)$ to denote the event that $\mathcal{A}$ uses at least $(\alpha+1)$ rounds and at most T time steps before the end of the α-th round.

We will make use of two lemmas: the *progress* lemma and the *distribution exchange* lemma. The progress lemma basically says that if the algorithm $\mathcal{A}$ only performs $o(\Delta^2)$ pulls by the end of the α-th round, then it must move forward to the $(\alpha+1)$-st round and perform more pulls.

Lemma 26 (Progress Lemma)**.** *Recall that $\mathcal{A}$ is a δ-error β-fast algorithm, and $\mathcal{E}$ and $\mathcal{E}^*$ are defined at the beginning of this section. For any $\Delta \in [\Delta^*, 1/8)$, any $\alpha \geq 0$, and any $q \geq 1$, so long as*

$$\Pr_{I(\Delta)}[\mathcal{E}(\alpha, \Delta^{-2}/(Kq))] \geq 1/2,$$

we have that

$$\Pr_{I(\Delta)}[\mathcal{E}^*(\alpha, \Delta^{-2}/(Kq))] \geq \Pr_{I(\Delta)}[\mathcal{E}(\alpha, \Delta^{-2}/(Kq))] - 2\delta - \frac{4}{\sqrt{3q}}, \quad (27)$$

where K is the number of agents $\mathcal{A}$ uses in parallel.

We defer the proof of Lemma 26 to Section V-B. Intuitively, Lemma 26 holds because of the following reason: If $\mathcal{A}$ uses at most $\Delta^{-2}/(Kq)$ time steps, it may perform at most Δ^{-2}/q pulls throughout all K agents. When q is large, this is not enough information to tell $I(\Delta)$ from $I(-\Delta)$, and therefore $\mathcal{A}$ cannot make a decision on the sign of the arm, and has to proceed to the next round.

The distribution exchange lemma basically says that if the algorithm $\mathcal{A}$ uses $(\alpha+1)$ rounds for instance $I(\Delta)$, then its $(\alpha+1)$-st round must conclude before time Δ^{-2}/β for instance $I(\Delta')$ where $\Delta' \leq \Delta$.

Lemma 27 (Distribution Exchange Lemma)**.** *Recall that $\mathcal{A}$ is a δ-error β-fast algorithm, and $\mathcal{E}$ and $\mathcal{E}^*$ are defined at the beginning of this section. For any $\Delta \in [\Delta^*, 1/8)$, any $\alpha \geq 0$, any $q \geq 100$, and any $\zeta \geq 1$, we have that*

$$\begin{aligned}&\Pr_{I(\Delta/\zeta)}[\mathcal{E}(\alpha+1, \Delta^{-2}/(Kq)+\Delta^{-2}/\beta)] \\ \geq\ &\Pr_{I(\Delta)}[\mathcal{E}^*(\alpha, \Delta^{-2}/(Kq))] - \delta \\ &- \left(\exp\left(5\sqrt{(3\ln K)/\beta}\right) - 1\right) - 1/K^5 - \frac{8}{\sqrt{3q}}. \quad (28)\end{aligned}$$

We defer the proof of Lemma 27 to Section V-D. At a higher level, we prove Lemma 27 using the following intuition. For instance $I(\Delta)$, since $\mathcal{A}$ is a δ-error β-fast algorithm, each agent is very likely to use at most Δ^{-2}/β pulls during the $(\alpha+1)$-st round, and only sees at most $(\Delta^{-2}/(Kq)+\Delta^{-2}/\beta)$ pull outcomes before the next communication (given the event $\mathcal{E}^*(\alpha, \Delta^{-2}/(Kq))$), which is insufficient to tell between $I(\Delta)$ and $I(\Delta/\zeta)$. Therefore, if the instance is $I(\Delta/\zeta)$, each agent is also very likely to use at most Δ^{-2}/β pulls during the $(\alpha+1)$-st round, and hence the whole algorithm finishes the $(\alpha+1)$-st round before $(\Delta^{-2}/(Kq)+\Delta^{-2}/\beta)$ time with high probability.

However, it is not technically easy to formalize this intuition. If we simply use the statistical difference between the two distributions (under $I(\Delta)$ and $I(\Delta/\zeta)$) for the Δ^{-2}/β pulls during the $(\alpha+1)$-st round to upper bound the probability difference between each agent's behavior for the two instances, we will face a probability error of $\Theta(\sqrt{1/\beta})$ for each agent. In total, this becomes a probability error of $\Theta(K\sqrt{1/\beta}) \gg 1$ throughout all K agents, which is too much. To overcome this difficulty, in Section V-C, we establish a technical lemma to derive a much better upper bound on the difference between the probabilities that two product distributions assign to the same event, given that the event does not happen very often.

We are now ready to prove Theorem 23.

Proof of Theorem 23. Combining Lemma 26 and Lemma 27, when $\Delta \in [\Delta^*, 1/8)$, $\alpha \geq 0$, $q \geq 100$, $\zeta \geq 1$ and $\Pr_{I(\Delta)}[\mathcal{E}(\alpha, \Delta^{-2}/(Kq))] \geq 1/2$, we have

$$\begin{aligned}&\Pr_{I(\Delta/\zeta)}[\mathcal{E}(\alpha+1, \Delta^{-2}/(Kq)+\Delta^{-2}/\beta)] \\ \geq\ &\Pr_{I(\Delta)}[\mathcal{E}(\alpha, \Delta^{-2}/(Kq))] - 3\delta \\ &- \left(\exp\left(5\sqrt{(3\ln K)/\beta}\right) - 1\right) - 1/K^5 - \frac{12}{\sqrt{3q}}. \quad (29)\end{aligned}$$

Set $\zeta = \sqrt{1+(Kq)/\beta}$, and (29) becomes

$$\begin{aligned}&\Pr_{I(\Delta/\zeta)}[\mathcal{E}(\alpha+1, (\Delta/\zeta)^{-2}/(Kq))] \\ \geq\ &\Pr_{I(\Delta)}[\mathcal{E}(\alpha, \Delta^{-2}/(Kq))] - 3\delta \\ &- \left(\exp\left(5\sqrt{(3\ln K)/\beta}\right) - 1\right) - 1/K^5 - \frac{12}{\sqrt{3q}}. \quad (30)\end{aligned}$$

Let t_0 be the largest integer such that

$$0.1 \cdot (1+(K \cdot 1000 t_0^2)/\beta)^{-t_0/2} \geq \Delta^*,$$

and we have $t_0 = \Omega\left(\frac{\ln(1/\Delta^*)}{\ln(1+K/\beta)+\ln\ln(1/\Delta^*)}\right)$. Let $t = \min\{t_0, \lfloor c_R\sqrt{\beta/(\ln K)}\rfloor\}$ for some small enough universal

constant $c_R > 0$. We also set $q = 1000t_0^2$. By the definition of event $\mathcal{E}(\cdot,\cdot)$ and the numbering of the steps of the communications, we have that $\mathcal{E}(0, 100/(Kq))$ always holds, and therefore

$$1 = \Pr_{I(1/10)}[\mathcal{E}(0, 100/(Kq))]. \tag{31}$$

Starting from (31), we iteratively apply (30) for t times. Let $\Delta^\flat = 0.1 \cdot (1 + (Kq)/\beta)^{-t/2} \geq \Delta^*$, we have that

$$\Pr_{I(\Delta^\flat)}[\mathcal{E}(t, \Delta^\flat/(Kq))] \geq 1 - \left(3\delta\left(\exp\left(5\sqrt{(3\ln K)/\beta}\right) - 1\right) + 1/K^5 + \frac{12}{\sqrt{3000t_0^2}}\right)t, \tag{32}$$

so long as

$$\left(3\delta + \left(\exp\left(5\sqrt{(3\ln K)/\beta}\right) - 1\right) + 1/K^5 + \frac{12}{\sqrt{3000t_0^2}}\right)t \leq \frac{1}{2}. \tag{33}$$

We see that (33) holds as long as $\delta \leq 1/K^5$ and c_R is small enough (note that when $\beta < \ln K/c_R^2$ then $t = 0$). Therefore, we conclude that

$$\Pr_{I(\Delta^\flat)}\left[\mathcal{A} \text{ uses } \Omega\left(\min\left\{\frac{\ln(1/\Delta^*)}{\ln(1 + K/\beta) + \ln\ln(1/\Delta^*)}, \sqrt{\beta/(\ln K)}\right\}\right) \text{ rounds}\right] \geq \frac{1}{2}.$$

□

B. Proof of the Progress Lemma (Lemma 26)

Proof of Lemma 26. Let F denote the event that $\mathcal{A}$ uses exactly α rounds, and uses at most $\Delta^{-2}/(Kq)$ time steps. It is clear that

$$\Pr_{I(\Delta)}[\mathcal{E}^*(\alpha, \Delta^{-2}/(Kq))] \geq \Pr_{I(\Delta)}[\mathcal{E}(\alpha, \Delta^{-2}/(Kq))] - \Pr_{I(\Delta)}[F].$$

Therefore it suffices to show that

$$\Pr_{I(\Delta)}[F] \leq 2\delta + \frac{4}{\sqrt{3q}}. \tag{34}$$

Note that

$$\Pr_{I(\Delta)}[F] = \Pr_{I(\Delta)}[F \wedge \mathcal{A} \text{ returns '}> 1/2\text{'}] + \Pr_{I(\Delta)}[F \wedge \mathcal{A} \text{ returns '}< 1/2\text{'}]. \tag{35}$$

We first focus on the first term of the Right-Hand Side (RHS) of (35). Let $\mathcal{D}_\Delta$ denote the product distribution $\mathcal{B}(1/2 + \Delta)^{\otimes \Delta^{-2}/q}$, and let $\mathcal{D}_{-\Delta}$ denote $\mathcal{B}(1/2 - \Delta)^{\otimes \Delta^{-2}/q}$, where $\mathcal{B}(\theta)$ is the Bernoulli distribution with the expectation θ. By Pinsker's inequality (Lemma 31) and simple KL-divergence calculation we have that when $\Delta \in (0, 1/8)$, it holds that

$$\|\mathcal{D}_\Delta - \mathcal{D}_{-\Delta}\|_{\mathrm{TV}} \leq \sqrt{\frac{1}{2}\mathrm{KL}(\mathcal{D}_\Delta \| \mathcal{D}_{-\Delta})} \leq \frac{4}{\sqrt{3q}}.$$

On the other hand, since when event F happens, $\mathcal{A}$ uses at most $\Delta^{-2}/(Kq) \cdot K = \Delta^{-2}/q$ pulls (over all agents), we have

$$\begin{aligned}
&\Pr_{I(\Delta)}[F \wedge \mathcal{A} \text{ returns '}> 1/2\text{'}] \\
\leq\ &\Pr_{I(-\Delta)}[F \wedge \mathcal{A} \text{ returns '}> 1/2\text{'}] + \|\mathcal{D}_\Delta - \mathcal{D}_{-\Delta}\|_{\mathrm{TV}} \\
\leq\ &\Pr_{I(-\Delta)}[\mathcal{A} \text{ returns '}> 1/2\text{'}] + \frac{4}{\sqrt{3q}} \leq \delta + \frac{4}{\sqrt{3q}}.
\end{aligned} \tag{36}$$

For the second term of the RHS of (35), we have

$$\Pr_{I(\Delta)}[F \wedge \mathcal{A} \text{ returns '}< 1/2\text{'}] \leq \Pr_{I(\Delta)}[\mathcal{A} \text{ returns '}< 1/2\text{'}] \leq \delta. \tag{37}$$

Combining (35), (36), and (37), we prove (34). □

C. Probability Discrepancy under Product Distributions for Infrequent Events

In this section, we prove the following lemma to upper bound the difference between the probabilities that two product distributions assign to the same event. Given that the event does not happen very often, our upper bound is significantly better than the total variation distance between the two product distributions.

Lemma 28. *Suppose $0 \leq \Delta' \leq \Delta \leq 1/8$. For any positive integer $m = \Delta^{-2}/\xi$ where $\xi \geq 100$, let $\mathcal{D}$ denote the product distribution $\mathcal{B}(1/2 + \Delta)^{\otimes m}$ and let $\mathcal{D}'$ denote the product distribution $\mathcal{B}(1/2 + \Delta')^{\otimes m}$, where $\mathcal{B}(\mu)$ is the Bernoulli distribution with the expectation μ. Let $\mathcal{X}$ be any probability distribution with sample space X. For any event $A \subseteq \{0,1\}^m \times X$ such that $\Pr_{\mathcal{D}\otimes\mathcal{X}}[A] \leq \gamma$, we have that*

$$\Pr_{\mathcal{D}'\otimes\mathcal{X}}[A] \leq \gamma \cdot \exp\left(5\sqrt{(3\ln Q)/\xi}\right) + 1/Q^6,$$

holds for all $Q \geq \xi$.

Proof. Let $L = \{\ell \in \{0,1\}^m : |\ell| \geq m/2 - z/\Delta\}$ where $|\ell|$ denotes the number of 1's in the vector ℓ and $z \geq 0$ is a parameter to be decided later. We have that

$$\begin{aligned}
&\Pr_{(\ell,x)\sim\mathcal{D}'\otimes\mathcal{X}}[(\ell, x) \in A] \\
\leq\ &\Pr_{(\ell,x)\sim\mathcal{D}'\otimes\mathcal{X}}[(\ell, x) \in A \wedge \ell \in L] + \Pr_{\ell\sim\mathcal{D}'}[\ell \notin L].
\end{aligned} \tag{38}$$

We first focus on the first term of the RHS of (38). Note that

$$\begin{aligned}
&\Pr_{(\ell,x)\sim\mathcal{D}'\otimes\mathcal{X}}[(\ell, x) \in A \wedge \ell \in L] \\
=\ &\sum_{\ell\in L} \Pr_{x\sim\mathcal{X}}[(\ell, x) \in A \mid \ell \in L] \cdot (1/2+\Delta')^{|\ell|}(1/2-\Delta')^{m-|\ell|}
\end{aligned} \tag{39}$$

When $\ell \in L$, by monotonicity, we have

$$\begin{aligned}
&\frac{(1/2+\Delta')^{|\ell|}(1/2-\Delta')^{m-|\ell|}}{(1/2+\Delta)^{|\ell|}(1/2-\Delta)^{m-|\ell|}} \\
\leq\ & \frac{(1/2+\Delta')^{m/2-z/\Delta}(1/2-\Delta')^{m/2+z/\Delta}}{(1/2+\Delta)^{m/2-z/\Delta}(1/2-\Delta)^{m/2+z/\Delta}} \\
=\ & \left(\frac{1/4-(\Delta')^2}{1/4-\Delta^2}\right)^{m/2}\left(\frac{(1/2-\Delta')(1/2+\Delta)}{(1/2+\Delta')(1/2-\Delta)}\right)^{z/\Delta} \\
\leq\ & \left(\frac{1}{1-4\Delta^2}\right)^{m/2}\left(\frac{1+2\Delta}{1-2\Delta}\right)^{z/\Delta}. \qquad (40)
\end{aligned}$$

Since $(1-\epsilon)^{-1/\epsilon} \leq e^{1.2}$ for all $\epsilon \in (0,1/4)$ and $(1+\epsilon)^{1/\epsilon} \leq e$ for all $\epsilon \in (0,1)$, for $\Delta \in (0,1/8)$, we have

$$\begin{aligned}
&\left(\frac{1}{1-4\Delta^2}\right)^{m/2}\left(\frac{1+2\Delta}{1-2\Delta}\right)^{z/\Delta} \\
\leq\ & \exp\left(1.2\cdot 4\Delta^2\cdot m/2 + 1.2\cdot 2\Delta\cdot z/\Delta + 2\Delta\cdot z/\Delta\right) \\
=\ & \exp(2.4/\xi + 4.4z). \qquad (41)
\end{aligned}$$

Combining (39), (40), (41), we have

$$\begin{aligned}
&\Pr_{(\ell,x)\sim\mathcal{D}'\otimes\mathcal{X}}[(\ell,x)\in A\wedge \ell\in L] \\
\leq\ & \exp(2.4/\xi+4.4z)\cdot \Pr_{(\ell,x)\sim\mathcal{D}\otimes\mathcal{X}}[(\ell,x)\in A\wedge \ell\in L] \\
\leq\ & \exp(2.4/\xi+4.4z)\cdot \Pr_{(\ell,x)\sim\mathcal{D}\otimes\mathcal{X}}[(\ell,x)\in A] \\
\leq\ & \gamma\cdot\exp(2.4/\xi+4.4z). \qquad (42)
\end{aligned}$$

For the second term of the RHS of (38), by Chernoff-Hoeffding bound, we have

$$\Pr_{\ell\sim\mathcal{D}'}[\ell\notin L] \leq \exp\left(-2m(z/(\Delta m))^2\right) = \exp\left(-2z^2\xi\right). \qquad (43)$$

□

Combining (38), (42), and (43), we have

$$\Pr_{(\ell,x)\sim\mathcal{D}'\otimes\mathcal{X}}[(\ell,x)\in A] \leq \gamma\cdot\exp(2.4/\xi+4.4z)+\exp\left(-2z^2\xi\right).$$

Setting $z = \sqrt{(3\ln Q)/\xi}$ and for $\xi \geq 100$ and $Q \geq \xi$, we have

$$\begin{aligned}
&\Pr_{(\ell,x)\sim\mathcal{D}'\otimes\mathcal{X}}[(\ell,x)\in A] \\
\leq\ & \gamma\cdot\exp\left(2.4/\xi+4.4\sqrt{(3\ln Q)/\xi}\right)+1/Q^6 \\
\leq\ & \gamma\cdot\exp\left(5\sqrt{(3\ln Q)/\xi}\right)+1/Q^6.
\end{aligned}$$

D. Proof of the Distribution Exchange Lemma (Lemma 27)

We first introduce a simple mathematical lemma, whose proof can be found in Appendix A.

Lemma 29. *For any $\gamma_1,\ldots,\gamma_K \in [0,1]$ and $x \geq 0$, it holds that*

$$\prod_{i=1}^K \max\{1-\gamma_i-\gamma_i x, 0\} \geq \prod_{i=1}^K(1-\gamma_i) - x.$$

Proof of Lemma 27. We will only prove (28) for $\mathcal{A}$ as a deterministic algorithm, i.e. when there is no randomness in $\mathcal{A}$ except for the observed rewards drawn from the arm. Once this is established, we can easily deduce that the same inequality holds for randomized $\mathcal{A}$ by taking expectation on both sides of (28) over the (possibly shared) random bits used by each agent of the collaborative learning algorithm $\mathcal{A}$.

Let $\ell \in \{0,1\}^{\Delta^{-2}/q}$ be the rewards from the first Δ^{-2}/q plays of the arm. Once conditioned on ℓ, $\mathcal{E}^*(\alpha,\Delta^{-2}/(Kq))$ becomes a deterministic event, since $\mathcal{A}$ is deterministic and the event only depends on the first Δ^{-2}/q rewards. In light of this, we let $\mathfrak{S}$ denote the set of ℓ conditioned on which $\mathcal{E}^*(\alpha,\Delta^{-2}/(Kq))$ holds. We have

$$\sum_{s\in\mathfrak{S}} \Pr_{I(\Delta)}[\ell = s] = \Pr_{I(\Delta)}[\mathcal{E}^*(\alpha,\Delta^{-2}/(Kq))]. \qquad (44)$$

For each agent $i \in [K]$, let G_i be the event that the agent uses more than Δ^{-2}/β pulls during the $(\alpha+1)$-st round. Since $\mathcal{A}$ is deterministic, conditioned on $\ell \in \mathfrak{S}$, G_i only depends on the random rewards observed by the i-th agent during the $(\alpha+1)$-st round, and is independent from G_j for any $j \neq i$. Since $\mathcal{A}$ is a δ-error β-fast algorithm, we have

$$\begin{aligned}
\delta\ \geq\ & \Pr_{I(\Delta)}[\mathcal{A}\text{ uses } > \Delta^{-2}/\beta \text{ time}] \\
\geq\ & \sum_{s\in\mathfrak{S}} \Pr_{I(\Delta)}[\ell=s]\cdot \Pr_{I(\Delta)}[G_1\vee G_2\vee\cdots\vee G_K \mid \ell = s] \\
=\ & \sum_{s\in\mathfrak{S}} \Pr_{I(\Delta)}[\ell=s]\cdot\left(1-\prod_{i=1}^K\left(1-\Pr_{I(\Delta)}[G_i\mid \ell=s]\right)\right) \\
=\ & \Pr_{I(\Delta)}[\mathcal{E}^*(\alpha,\Delta^{-2}/(Kq))] \\
& -\sum_{s\in\mathfrak{S}} \Pr_{I(\Delta)}[\ell=s]\cdot\prod_{i=1}^K\left(1-\Pr_{I(\Delta)}[G_i\mid \ell=s]\right),
\end{aligned}$$

where the last equality is because of (44). We thus have

$$\begin{aligned}
&\sum_{s\in\mathfrak{S}} \Pr_{I(\Delta)}[\ell=s]\cdot\prod_{i=1}^K\left(1-\Pr_{I(\Delta)}[G_i\mid \ell=s]\right) \\
&\qquad \geq \Pr_{I(\Delta)}[\mathcal{E}^*(\alpha,\Delta^{-2}/(Kq))] - \delta. \qquad (45)
\end{aligned}$$

We also have

$$\begin{aligned}
&\Pr_{I(\Delta/\zeta)}[\mathcal{E}(\alpha+1,\Delta^{-2}/(Kq)+\Delta^{-2}/\beta)] \\
\geq\ & \sum_{s\in\mathfrak{S}} \Pr_{I(\Delta/\zeta)}[\ell=s]\cdot \Pr_{I(\Delta/\zeta)}[\neg G_1\wedge\neg G_2\wedge\cdots\wedge\neg G_K\mid \ell=s] \\
=\ & \sum_{s\in\mathfrak{S}} \Pr_{I(\Delta/\zeta)}[\ell=s]\cdot\prod_{i=1}^K\left(1-\Pr_{I(\Delta/\zeta)}[G_i\mid \ell=s]\right). \qquad (46)
\end{aligned}$$

We next to fuse (45) and (46). Invoking Lemma 28 with $Q = K$ and $\xi = \beta$, we have

$$\begin{aligned}
&\sum_{s\in\mathfrak{S}} \Pr_{I(\Delta/\zeta)}[\ell = s] \cdot \prod_{i=1}^{K} \left(1 - \Pr_{I(\Delta/\zeta)}[G_i \mid \ell = s]\right) \\
\geq\ & \sum_{s\in\mathfrak{S}} \Pr_{I(\Delta/\zeta)}[\ell = s] \cdot \prod_{i=1}^{K} \max\Bigg\{1 \\
& - \Pr_{I(\Delta)}[G_i \mid \ell = s] \cdot \exp\left(5\sqrt{(3\ln K)/\beta}\right) - 1/K^6, 0\Bigg\} \\
\geq\ & \sum_{s\in\mathfrak{S}} \Pr_{I(\Delta/\zeta)}[\ell = s] \cdot \Bigg(\prod_{i=1}^{K} \max\Bigg\{1 - \Pr_{I(\Delta)}[G_i \mid \ell = s] \\
& \cdot \exp\left(5\sqrt{(3\ln K)/\beta}\right), 0\Bigg\} - 1/K^5\Bigg) \\
\geq\ & \sum_{s\in\mathfrak{S}} \Pr_{I(\Delta/\zeta)}[\ell = s] \cdot \Bigg(\prod_{i=1}^{K} \left(1 - \Pr_{I(\Delta)}[G_i \mid \ell = s]\right) \\
& - \left(\exp\left(5\sqrt{(3\ln K)/\beta}\right) - 1\right) - 1/K^5\Bigg) \\
\geq\ & \sum_{s\in\mathfrak{S}} \Pr_{I(\Delta/\zeta)}[\ell = s] \cdot \prod_{i=1}^{K} \left(1 - \Pr_{I(\Delta)}[G_i \mid \ell = s]\right) \\
& - \left(\exp\left(5\sqrt{(3\ln K)/\beta}\right) - 1\right) - 1/K^5, \qquad (47)
\end{aligned}$$

where the second to the last inequality is due to Lemma 29. Finally, we have

$$\begin{aligned}
&\sum_{s\in\mathfrak{S}} \Pr_{I(\Delta/\zeta)}[\ell = s] \cdot \prod_{i=1}^{K} \left(1 - \Pr_{I(\Delta)}[G_i \mid \ell = s]\right) \\
\geq\ & \sum_{s\in\mathfrak{S}} \Pr_{I(\Delta)}[\ell = s] \cdot \prod_{i=1}^{K} \left(1 - \Pr_{I(\Delta)}[G_i \mid \ell = s]\right) \\
& - \sum_{s\in\mathfrak{S}} \left|\Pr_{I(\Delta/\zeta)}[\ell = s] - \Pr_{I(\Delta)}[\ell = s]\right|, \qquad (48)
\end{aligned}$$

where by Pinsker's inequality (Lemma 31) and simple KL-divergence calculation for $\Delta \in (0, 1/8)$, we have

$$\sum_{s\in\mathfrak{S}} \left|\Pr_{I(\Delta/\zeta)}[\ell = s] - \Pr_{I(\Delta)}[\ell = s]\right| \leq \frac{8}{\sqrt{3q}}. \qquad (49)$$

Combining (46), (47), (48), and (49), we have

$$\begin{aligned}
&\Pr_{I(\Delta/\zeta)}[\mathcal{E}(\alpha + 1, \Delta^{-2}/(Kq) + \Delta^{-2}/\beta)] \\
\geq\ & \sum_{s\in\mathfrak{S}} \Pr_{I(\Delta)}[\ell = s] \cdot \prod_{i=1}^{K} \left(1 - \Pr_{I(\Delta)}[G_i \mid \ell = s]\right) \\
& - \left(\exp\left(5\sqrt{(3\ln K)/\beta}\right) - 1\right) - 1/K^5 - \frac{8}{\sqrt{3q}} \\
\geq\ & \Pr_{I(\Delta)}[\mathcal{E}^*(\alpha, \Delta^{-2}/(Kq))] - \delta \\
& - \left(\exp\left(5\sqrt{(3\ln K)/\beta}\right) - 1\right) - 1/K^5 - \frac{8}{\sqrt{3q}}, \qquad (50)
\end{aligned}$$

where the last inequality is due to (45). □

References

[1] J. Audibert, S. Bubeck, and R. Munos, "Best arm identification in multi-armed bandits," in *COLT*, 2010, pp. 41–53.

[2] H. Robbins, "Some aspects of the sequential design of experiments," *Bulletin of the American Mathematical Society*, vol. 58, no. 5, pp. 527–535, 1952.

[3] J. C. Gittins, "Bandit processes and dynamic allocation indices," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 41, no. 2, pp. 148–164, 1979.

[4] W. H. Press, "Bandit solutions provide unified ethical models for randomized clinical trials and comparative effectiveness research," *Proceedings of the National Academy of Sciences*, vol. 106, no. 52, pp. 22 387–22 392, 2009.

[5] B. Awerbuch and R. Kleinberg, "Online linear optimization and adaptive routing," *Journal of Computer and System Sciences*, vol. 74, no. 1, pp. 97–114, 2008.

[6] W. Shen, J. Wang, Y. Jiang, and H. Zha, "Portfolio choices with orthogonal bandit learning," in *IJCAI*, 2015, p. 974.

[7] O. Maron and A. W. Moore, "Hoeffding races: Accelerating model selection search for classification and function approximation," in *NIPS*, 1993, pp. 59–66.

[8] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot *et al.*, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, p. 484, 2016.

[9] D. Agarwal, B. Chen, P. Elango, N. Motgi, S. Park, R. Ramakrishnan, S. Roy, and J. Zachariah, "Online models for content optimization," in *NIPS*, 2008, pp. 17–24.

[10] L. G. Valiant, "A bridging model for parallel computation," *Communications of the ACM*, vol. 33, no. 8, pp. 103–111, 1990.

[11] J. Dean and S. Ghemawat, "Mapreduce: Simplified data processing on large clusters," *Communications of the ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[12] E. Even-Dar, S. Mannor, and Y. Mansour, "PAC bounds for multi-armed bandit and markov decision processes," in *COLT*, 2002, pp. 255–270.

[13] S. Mannor and J. N. Tsitsiklis, "The sample complexity of exploration in the multi-armed bandit problem," *Journal of Machine Learning Research*, vol. 5, no. Jun, pp. 623–648, 2004.

[14] Z. Karnin, T. Koren, and O. Somekh, "Almost optimal exploration in multi-armed bandits," in *ICML*, 2013, pp. 1238–1246.

[15] K. Jamieson, M. Malloy, R. Nowak, and S. Bubeck, "lil' UCB : An optimal exploration algorithm for multi-armed bandits," in *COLT*, 2014, pp. 423–439.

[16] E. Kaufmann, O. Cappé, and A. Garivier, "On the complexity of best-arm identification in multi-armed bandit models," *Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1–42, 2016.

[17] A. Carpentier and A. Locatelli, "Tight (lower) bounds for the fixed budget best arm identification bandit problem," in *COLT*, 2016, pp. 590–604.

[18] L. Chen, J. Li, and M. Qiao, "Towards instance optimal bounds for best arm identification," in *COLT*, 2017, pp. 535–592.

[19] E. Hillel, Z. S. Karnin, T. Koren, R. Lempel, and O. Somekh, "Distributed exploration in multi-armed bandits," in *NIPS*, 2013, pp. 854–862.

[20] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, vol. 47, no. 2-3, pp. 235–256, 2002.

[21] S. Bubeck, N. Cesa-Bianchi *et al.*, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Foundations and Trends® in Machine Learning*, vol. 5, no. 1, pp. 1–122, 2012.

[22] T. Lattimore and C. Szepesvári, "Bandit algorithms," *preprint*, 2018.

[23] E. Even-Dar, S. Mannor, and Y. Mansour, "Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems," *Journal of Machine Learning Research*, vol. 7, no. Jun, pp. 1079–1105, 2006.

[24] S. Bubeck, T. Wang, and N. Viswanathan, "Multiple identifications in multi-armed bandits," in *ICML*, 2013, pp. 258–265.

[25] Y. Zhou, X. Chen, and J. Li, "Optimal pac multiple arm identification with applications to crowdsourcing," in *ICML*, 2014, pp. 217–225.

[26] J. Chen, X. Chen, Q. Zhang, and Y. Zhou, "Adaptive multiple-arm identification," in *ICML*, 2017, pp. 722–730.

[27] A. Locatelli, M. Gutzeit, and A. Carpentier, "An optimal algorithm for the thresholding bandit problem," in *ICML*, 2016, pp. 1690–1698.

[28] Y. Xue, P. Zhou, T. Jiang, S. Mao, and X. Huang, "Distributed learning for multi-channel selection in wireless network monitoring," in *IEEE SECON*, 2016, pp. 1–9.
[29] P. Krafft, K. Zhou, I. Edwards, K. Starbird, and E. S. Spiro, "Centralized, parallel, and distributed information processing during collective sensemaking," in *CHI*, 2017, pp. 2976–2987.
[30] 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 *COLT*, 2017, pp. 39–75.
[31] K. Liu and Q. Zhao, "Distributed learning in multi-armed bandit with multiple players," *IEEE Transactions on Signal Processing*, vol. 58, no. 11, pp. 5667–5681, 2010.
[32] J. Rosenski, O. Shamir, and L. Szlak, "Multi-player bandits - a musical chairs approach," in *ICML*, 2016, pp. 155–163.
[33] I. Bistritz and A. Leshem, "Distributed multi-player bandits - a game of thrones approach," in *NeurIPS*, 2018, pp. 7222–7232.
[34] B. Szörényi, R. Busa-Fekete, I. Hegedűs, R. Ormándi, M. Jelasity, and B. Kégl, "Gossip-based distributed stochastic bandit algorithms," in *ICML*, 2013, pp. 19–27.
[35] P. Landgren, V. Srivastava, and N. E. Leonard, "On distributed cooperative decision-making in multiarmed bandits," in *ECC*, 2016, pp. 243–248.
[36] J. Xu, C. Tekin, S. Zhang, and M. Van Der Schaar, "Distributed multi-agent online learning based on global feedback," *IEEE Transactions on Signal Processing*, vol. 63, no. 9, pp. 2225–2238, 2015.
[37] B. Awerbuch and R. D. Kleinberg, "Competitive collaborative learning," in *COLT*, 2005, pp. 233–248.
[38] N. Cesa-Bianchi, C. Gentile, Y. Mansour, and A. Minora, "Delay and cooperation in nonstochastic bandits," in *COLT*, 2016, pp. 605–622.
[39] V. Perchet, P. Rigollet, S. Chassang, and E. Snowberg, "Batched bandit problems," in *COLT*, 2015, p. 1456.
[40] A. Blum, N. Haghtalab, A. D. Procaccia, and M. Qiao, "Collaborative PAC learning," in *NIPS*, 2017, pp. 2389–2398.
[41] J. Chen, Q. Zhang, and Y. Zhou, "Tight bounds for collaborative PAC learning via multiplicative weights," in *NeurIPS*, 2018, pp. 3602–3611.
[42] H. L. Nguyen and L. Zakynthinou, "Improved algorithms for collaborative PAC learning," in *NeurIPS*, 2018, pp. 7642–7650.
[43] P. B. Miltersen, N. Nisan, S. Safra, and A. Wigderson, "On data structures and asymmetric communication complexity," *Journal of Computer and System Sciences*, vol. 57, no. 1, pp. 37–49, 1998.
[44] R. H. Farrell, "Asymptotic behavior of expected sample size in certain one sided tests," *The Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 36–72, 1964.
[45] A. C. Yao, "Probabilistic computations: Toward a unified measure of complexity (extended abstract)," in *FOCS*, 1977, pp. 222–227.
[46] M. S. Pinsker, *Information and Information Stability of Random Variables and Processes*. Holden-Day, 1964.

APPENDIX

The following lemma states Chernoff-Hoeffding bound.

Lemma 30. *Let $X_1, X_2, \ldots, X_n$ be independent random variables bounded by $[0,1]$. Let $X = \sum_{i=1}^n X_i$. For additive error, for every $t \geq 0$, it holds that*

$$\Pr\left[|X - \mathbb{E}[X]| \geq t\right] \leq 2\exp\left(-\frac{2t^2}{n}\right).$$

For multiplicative error, for every $\delta \in [0,1]$, it holds that

$$\Pr\left[|X - \mathbb{E}[X]| \geq \delta\,\mathbb{E}[X]\right] \leq 2\exp\left(-\frac{\delta^2\,\mathbb{E}[X]}{3}\right).$$

The following lemma states Pinsker's inequality [46].

Lemma 31. *If P and Q are two discrete probability distributions on a measurable space (X, Σ), then for any measurable event $A \in \Sigma$, it holds that*

$$|P(A) - Q(A)| \leq \sqrt{\frac{1}{2}\mathrm{KL}(P\|Q)}$$

where

$$\mathrm{KL}(P\|Q) = \sum_{x\in X} P(x)\ln\left(\frac{P(x)}{Q(x)}\right)$$

is the Kullback–Leibler divergence.

A. Proof of Lemma 3

Proof. Let $S_\ell = |\Theta|_{|X=B^{-\ell}}$. We have $\mathbb{E}[S_\ell] = \gamma B^{2j}\cdot\left(\frac{1}{2} - B^{-\ell}\right)$.

For the first item, we have for any $\ell > j$,

$$\mathbb{E}[S_\ell] = \gamma B^{2j}\cdot\left(\frac{1}{2} - B^{-\ell}\right) = \frac{\gamma B^{2j}}{2} - \gamma B^{2j-\ell} = \frac{\gamma B^{2j}}{2} \pm \gamma B^{j-1}.$$

Since $B = \gamma \geq (\ln n)^{100}$, by Chernoff-Hoeffding bound we have that for any $\ell > j$, with probability at least $1 - n^{-10}$,

$$S_\ell = \frac{\gamma B^{2j}}{2} \pm B^{j+0.6}.$$

Now consider the second and third items. If $\ell > j$, then by Chernoff-Hoeffding bound,

$$\Pr\left[S_\ell \leq \left(\frac{1}{2} - B^{-(j+1)}\right)\gamma B^{2j} - \sqrt{10\gamma\ln n}B^j\right] \leq \Pr\left[S_\ell \leq \mathbb{E}[S_\ell] - \sqrt{10\gamma B^{2j}\ln n}\right] \leq 1/n^{10}. \quad (51)$$

If $\ell \leq j$, then

$$\Pr\left[S_\ell \geq \left(\frac{1}{2} - B^{-j}\right)\gamma B^{2j} + \sqrt{10\gamma\ln n}B^j\right] \leq \Pr\left[S_\ell \geq \mathbb{E}[S_\ell] + \sqrt{10\gamma B^{2j}\ln n}\right] \leq 1/n^{10}. \quad (52)$$

Since $B \geq (\ln n)^{100}$, we have

$$\left(\frac{1}{2} - B^{-j}\right)\gamma B^{2j} + \sqrt{10\gamma\ln n}B^j < \zeta_1 = \left(\frac{1}{2} - B^{-(j+1)}\right)\gamma B^{2j} - \sqrt{10\gamma\ln n}B^j. \quad (53)$$

The last two items follows from (51), (52) and (53).

□

B. Proof of Lemma 29

Proof. Note that when $x \geq \min_{i\in[K]}\left\{\frac{1-\gamma_i}{\gamma_i}\right\}$, the Left-Hand Side (LHS) of the desired inequality becomes 0 and the RHS is less than or equal to 0. Therefore, we only need to prove the inequality assuming $x < \min_{i\in[K]}\left\{\frac{1-\gamma_i}{\gamma_i}\right\}$.

Now the LHS becomes $\prod_{i=1}^K(1-\gamma_i-\gamma_i x)$. Let $f(t) = \prod_{i=1}^K(1-\gamma_i-\gamma_i t)$ for $t \in [0,x]$. Note that $f'(t) =$

$-\sum_{i=1}^{K}\gamma_i\prod_{j\neq i}(1-\gamma_j-\gamma_j t) \geq f'(0)$ for $t \in [0,x]$. We have

$$\begin{aligned}
&\prod_{i=1}^{K}(1-\gamma_i-\gamma_i x) = f(x) \geq f(0)+f'(0)x \\
= &\prod_{i=1}^{K}(1-\gamma_i) - \left(\sum_{i=1}^{K}\gamma_i\prod_{j\neq i}(1-\gamma_j)\right)x \\
\geq &\prod_{i=1}^{K}(1-\gamma_i) - \left(\prod_{i=1}^{K}(\gamma_i+(1-\gamma_i))\right)x \\
= &\prod_{i=1}^{K}(1-\gamma_i) - x.
\end{aligned}$$

□