ICQ: A Quantization Scheme for Best-Arm Identification Over Bit-Constrained Channels

Fathima Zarin Faizal[1], Adway Girish[2], Manjesh Kumar Hanawal[3], and Nikhil Karamchandani[1]

[1]Department of Electrical Engineering, IIT Bombay
[2]School of Computer and Communication Sciences, EPFL
[3]Industrial Engineering and Operations Research, IIT Bombay

April 25, 2023

Abstract

We study the problem of best-arm identification in a distributed variant of the multi-armed bandit setting, with a central learner and multiple agents. Each agent is associated with an arm of the bandit, generating stochastic rewards following an unknown distribution. Further, each agent can communicate the observed rewards with the learner over a bit-constrained channel. We propose a novel quantization scheme called Inflating Confidence for Quantization (ICQ) that can be applied to existing confidence-bound based learning algorithms such as Successive Elimination. We analyze the performance of ICQ applied to Successive Elimination and show that the overall algorithm, named ICQ-SE, has the order-optimal sample complexity as that of the (unquantized) SE algorithm. Moreover, it requires only an exponentially sparse frequency of communication between the learner and the agents, thus requiring considerably fewer bits than existing quantization schemes to successfully identify the best arm. We validate the performance improvement offered by ICQ with other quantization methods through numerical experiments.

Contents

1 Introduction

The *multi-armed bandit* (MAB) problem is a sequential decision-making model involving a learner and an environment, where in each round, the learner chooses from K actions or arms. Each arm is associated with a reward distribution that is *a priori* unknown to the learner. The learner then receives a random reward drawn from the distribution of the chosen arm. We consider the *pure exploration* variant [1] of this problem, where the learner is required to identify the best arm, i.e., the arm with the highest mean reward, with accuracy better than a prescribed confidence level. The learner is evaluated based on the number of samples (*sample complexity*) it requires to identify the best arm. Thus a 'desirable' algorithm is one that identifies the best arm with a smaller sample complexity, for a given confidence level. The pure exploration setting is well-studied [2–4] under the assumption that the learner gets to observe the reward values with full precision.

We consider a distributed variant of the pure exploration MAB setup where the learner cannot observe the reward samples directly, but through intermediate agents that act as an interface for each arm. Unlike the traditional pure exploration MAB setup, the learner no longer has access to the rewards with full precision, i.e., the agents observe the rewards obtained and communicate aggregated information to the learner over noiseless, bit-constrained channels. A key point is that each agent 'represents' a single arm, i.e., each agent pulls and observes rewards from only one fixed arm associated with it.

Motivation. Such distributed learning setups with limited communication arise in many real-world systems. For example, in wireless networks with bandwidth constraints involving remote and low-complexity agents, the cost of communicating the rewards could become a performance bottleneck. Reducing the number of bits transmitted would result in lower power consumption and wireless interference. This is particularly significant in IoT networks where devices are typically resource-constrained and battery-powered [5–8]. It may then be easier to have the agents process/compress the observations locally, before passing on information in a more condensed format. Finally, when learning from privacy-sensitive data, quantization can be used to hide the exact rewards obtained in such a way that the specific contents remain unintelligible to the learner, but there is still enough information to carry out the overall learning task, as in [9].

Contributions. To overcome constraints on the precision with which information can be sent from the agents to the learner, we develop a quantization scheme called *Inflating Confidence for Quantization* (ICQ) for the best-arm identification problem. Our quantization scheme allows agents to communicate the reward information with fewer bits while still allowing the learner to extract enough information to identify the best arm. The key idea behind the scheme is to generate a high-probability range for the mean reward estimator at the agent that is smaller than the actual range of the rewards to reduce the range over which quantization needs to be done. This is done using appropriately defined confidence intervals for the quantized values.

While we build our scheme on top of the successive elimination framework proposed for the standard best-arm identification problem [1], and develop an algorithm called ICQ-SE, a key feature of our proposed quantization strategy is that it can be used in conjunction with a broad class of alternate schemes (such as LUCB [10] and lil'UCB [11], to obtain corresponding algorithms ICQ-LUCB and ICQ-lil'UCB, and so on). This 'universality' is clearly desirable and draws inspiration from [12] where the proposed quantization strategy had a similar feature with relation to a broad class of regret-minimization algorithms. Our algorithm ICQ-SE has the following features (shown in Section 5):

1. the learner needs to communicate with the agents only exponentially sparsely;
2. requires only $B \geq 1$ number of bits for each round of communication for bounded rewards;
3. ensures order-optimal sample complexity compared to the distributed setup with no bit constraints; and
4. can be easily modified to be used with other confidence bound based algorithms (see Remark 1).

In Section 6, through simulations, we show that this scheme performs better than other quantization schemes in the MAB literature, for both bounded and unbounded rewards.

2 Related work

MAB problems are well explored in the literature in various settings like expected regret minimization, simple regret minimization, and Best Arm Identification (BAI) [13]. For BAI problems [1–4, 10, 11, 14], the goal is to identify the best arm either with high confidence within a given budget (*fixed budget*) or with fewer samples for a given threshold on the probability of making a mistake (*fixed confidence*). The algorithms developed for these settings assume that the learner has access to samples from the reward distributions with full precision. However, this assumption need not hold in federated setups [15] where the learners and the agents need to exchange information, and any bottleneck in communication needs to be taken into account. Some recent works have addressed such issues by modeling communication bottlenecks as capacity constraints [12, 16, 17] or as a limited resource that comes at an additional cost [18].

Our work is closer to [12, 16, 17], which propose quantization methods to improve the performance of learning algorithms under channel capacity constraints. In [12], the authors propose a quantization scheme, named QuBan, that can be used over a large class of MAB algorithms in the regret minimization setting to achieve order-optimal regret. QuBan differs from our method as it does not make use of confidence bounds for the mean reward estimators. In [16], the authors propose an adaptive quantization scheme and a decision-making policy for the Linear Stochastic Bandit setting and show that $B = \mathcal{O}(d)$ bits guarantees order-optimal regret, where d is the dimension of the arm set. Their scheme is the closest to ours, where confidence bounds for the mean reward estimators are used to find a smaller range to quantize on, albeit for the regret minimization setting. Moreover, similar to us, they show that using 1 bit for each transmission from the agent to the learner is sufficient to ensure order-optimal regret compared to the unquantized setting when the reward distributions have bounded support.

In [17], a pure exploration setting is considered, where a learner and multiple clients identify the best arm together, with each client being allotted a disjoint subset of the arms. Like us, they also propose a quantization scheme for communicating rewards between the clients and the learner; the difference being that their scheme only works with bounded rewards. Also, it is hard to control the number of bits used by their algorithm whereas our proposed scheme ICQ-SE provably allows for a trade-off between the amount of communication and performance. See Section 6 for a more detailed comparison.

3 Problem setup

In this section, we define notation that will be used throughout the paper and formalize our system model. The overall setup has been illustrated in Figure 1. Broadly, we study a distributed multi-armed bandit (MAB) problem where the decision-making and observing entities are separated. They must communicate their 'results' (decisions or observations) to each other over a noiseless channel using a finite number of bits, with the overall goal of performing a learning task.

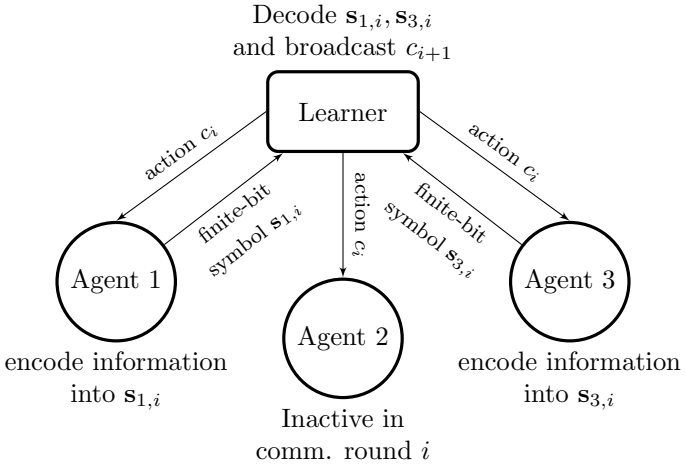

Figure 1: Block diagram illustrating the overall setup, shown here for the case with 3 agents, i.e., $K = 3$.

The distributed MAB. There is a central learner, connected to each of the K distributed agents, via a noiseless channel that is bit-constrained from the agent to the learner. The learner and the agents work together to solve a fixed confidence stochastic MAB problem consisting of K arms. Agent $i \in \{1, \ldots, K\} \triangleq [K]$ has access to arm i of the MAB, which is associated with a reward distribution ν_i. We assume that the distributions $\{\nu_i\}_{i=1}^K$ are σ^2-subgaussian[1] and bounded on $[a, b]$. For $i \in [K]$, let μ_i denote the mean of the distribution ν_i and $r_{i,t}$ denote the t^{th} reward sample drawn from ν_i. For each arm i, $\{r_{i,t} : t \geq 1\}$ is an i.i.d. process; furthermore the reward samples are independent across different arms. We also assume that arm 1 is the best arm for notational convenience, i.e., $\mu_1 \geq \mu_i$ where $i \neq 1$. Also, define the suboptimality gaps $\Delta_i = \mu_1 - \mu_i$ for $i \in [K]$.

The broad objective for the learner is to identify the arm with the highest mean reward by sequentially selecting arms and sampling from their associated reward distributions.

The communication model. Our communication model is summarized in Figure 1. Communication between the learner and the agents happens in rounds. At the beginning of *(communication) round* i, based on the information the learner has seen till $i-1$, it broadcasts an action c_i to all the K agents, where c_i encodes the information about the actions to be taken by each agent at round i. The agents respond to the learner after a fixed synchronized duration and the learner updates its estimate of the best arm based on the information it has received from the agents at round i. This constitutes one round. Each agent communicates at most once in each round.

We assume that each agent is only capable of collecting samples from its associated arm reward distribution, aggregating the information from the samples it has seen so far and transmitting it to the central learner. The agents cannot share information between each other and can communicate only with the central learner. This is commonly the case with low-complexity and resource-constrained devices such as drones and sensors. Moreover, they are only allowed to use a finite number of bits for each transmission to the learner. We assume that these transmissions are completed without any errors or erasures. Note that we do not assume that communication from the learner to the agents is bit-constrained and also do not put any computational restrictions on it; this is true for several application settings and similar assumptions are also common in the literature [12, 16, 17].

Performance metrics. We consider the *fixed confidence* BAI problem for the distributed MAB setting described above. The goal here is to find *sound* strategies that find the optimal arm in a finite number of rounds for a given confidence level δ. Formally, if the strategy stops after using τ_δ samples and outputs arm J_{τ_δ}, a sound strategy ensures that $\mathbb{P}(\tau_\delta < \infty, J_{\tau_\delta} \neq 1) \leq \delta$. Since we are dealing with a communication-constrained setting, we also use the metrics of *(communication) round complexity* $\tau_{r,\delta}$ (the number of communication rounds needed) and *communication complexity* B_δ (the total number of bits used by the algorithm) to study the performance of any sound strategy.

An online learning algorithm for this distributed MAB setup has the following components: (1) a *sampling rule* at the learner that at each time considers the history of communicated messages received from the agents thus far, and prescribes the arm(s) to be pulled in the next round; (2) a *communication rule* at each agent that prescribes the rounds at which the agent will communicate with the learner and also specifies the content of the message; (3) a *stopping rule* at the learner that specifies when the learner will stop sampling arms any further; and (4) a *recommendation rule* at the learner that specifies its estimate for the best arm after the algorithm terminates.

Objective. We develop learning algorithms for the distributed MAB setting outlined above where the agents need to limit the amount of communication with the learner. This is done via restricting both the frequency of communication (in terms of rounds) as well as the sizes of the messages exchanged (in terms of bits). Using too little communication can of course lead to a large penalty in terms of the sample complexity, and therein lies the main technical challenge of designing communication and quantization strategies which can effectively navigate this trade-off.

In the following sections, we propose a class of policies parameterized by the frequency of communication and the number of bits used in each message, and then explicitly characterize the impact of these parameters on the sample complexity.

[1] A random variable X is said to be σ^2-subgaussian if for any $t > 0$, $\mathbb{P}\left(|X - \mathbb{E}[X]| > t\right) \leq 2\exp\left(-t^2/2\sigma^2\right)$

4 Proposed quantization scheme and algorithm

This section is devoted to the description of our proposed quantization scheme *Inflating Confidence for Quantization* (ICQ) and its application to Successive Elimination as ICQ-SE. Algorithm 1 describes the actions to be taken by the learner while Algorithm 2 describes the agent operation (definitions of the additional notation involved can be found in (1), (4) and (5)). We provide more details below.

Algorithm 1 ICQ-SE algorithm (learner-side)

procedure ICQ-SE-LEARNER($K, \delta, B, \{b_i\}$)
  Let $S \leftarrow \{1, \ldots, K\}$
  For $1 \leq j \leq K$, let $\tilde{\mu}_{j,0}$ be sampled uniformly from $[a, b]$
  for $1 \leq i < \infty$ **do**
    for $j \in S$ **do**
      Instruct agent j to sample b_i times
      Receive quantized value $\mathbf{s}_{j,i}$ from agent j
      $L_{i,j} \leftarrow [\mathrm{LCB}(j, i-1, \delta) - U'(i, \delta),\ \mathrm{UCB}(j, i-1, \delta) + U'(i, \delta)]$
      Decode $\tilde{\mu}_{j,i} = \mathsf{dec}(\mathbf{s}_{j,i}, B, L_{i,j})$
    end for
    $S \leftarrow S \setminus \{m \in S : \max_{j \in [K]} \mathrm{LCB}(j, i, \delta) \geq \mathrm{UCB}(m, i, \delta)\}$
    STOP if $|S| = 1$
  end for
  return only element in S
end procedure

Algorithm 2 ICQ-SE algorithm (agent-side)

procedure ICQ-SE-AGENT$_j$($\delta, B, i, \tilde{\mu}_{j,i-1}$)
  Pull arm j b_i times
  $L_{i,j} \leftarrow [\mathrm{LCB}(j, i-1, \delta) - U'(i, \delta),\ \mathrm{UCB}(j, i-1, \delta) + U'(i, \delta)]$
  Send $\mathbf{s}_{j,i} = \mathsf{enc}(\hat{\mu}_{j,i}, B, L_{i,j})$
  return quantized value $\mathbf{s}_{j,i}$
end procedure

Successive Elimination. The broad idea behind the *Successive Elimination* (SE) framework [1] for a classical MAB setting (where there is only a learner observing full-precision rewards and no intermediate agents) is to characterize high-confidence bounds for the means of the distributions of each arm. One derives confidence widths $U'(i, \delta)$ such that the empirical mean $\hat{\mu}_{j,i}$ of arm j at round i lies in the interval $[\hat{\mu}_{j,i} - U'(i, \delta), \hat{\mu}_{j,i} + U'(i, \delta)]$ around the actual mean μ_j with a 'high' probability (we make this formal later). The upper limit is called the Upper Confidence Bound (UCB) and the lower limit is called the Lower Confidence Bound (LCB). The learner constantly keeps track of a set S of *active arms*, i.e., the set of arms still in contention to be the best arm. The set S is initialized to be the set of all arms $[K]$. At the end of a round, if the UCB of any arm k lies below the LCB of any other arm j, then arm k is removed from the active set. Thus, under the high probability event that these confidence bounds contain the actual reward means, removing arms whose UCBs lie below the LCB of some other arm would guarantee that the algorithm is removing only suboptimal arms. The algorithm makes a mistake only when these high probability events do not occur.

High-level description of the algorithm. In our setting, the learner no longer has access to full-precision rewards, and instead receives quantized estimates from the agents associated with each arm. In line with the SE framework, we also refer to the agents corresponding to the active arms as *active agents*. As we would like to reduce the number of bits used, it is inefficient for the agent to communicate each sample that it sees. We thus consider a batched approach that ensures that communication happens in a sparse manner. The learner pulls active arms (through the agents) in batches; in particular, during (communication) round

i, the agents pull and observe rewards from their associated arms b_i times before sending (a summary of) the results to the learner. We also define t_i to be the cumulative sum of arm pulls for each active arm till round i, i.e., $t_i = \sum_{j=1}^{i} b_j$. We show our results for an *exponentially-sparse* communication framework similar to [18], i.e., $t_i = \alpha^i$ for some $\alpha > 1$.

At the beginning of round i, the learner instructs each agent in S, the set of active agents, to sample from their associated reward distributions b_i times. At the end of round i, each active agent must have made a total of t_i cumulative arm pulls and sends to the learner a quantized estimate of the empirical mean of the rewards obtained. The learner first decodes the quantized estimate, then decides which arms will remain in the active set using the SE framework. This update requires defining new confidence intervals that account for quantization; details will be provided later. This marks the end of a communication round. The algorithm terminates when there is only one arm left in the active set, which is the recommended arm. Before describing the working of the algorithm in more detail, it is instructive to look at the quantization part separately.

Quantization scheme. Each agent calculates the empirical mean of the observed rewards, which must first be quantized and encoded into a bit string to be sent over the bit-constrained channel. Similarly, at the learner, we must be able to obtain a decoded estimate of the empirical mean from the encoded bit string. This is achieved as follows. We first fix an interval that we 'expect' the empirical mean to belong to with high probability (this will become clear later), divide it into 2^B equal bins, then transmit a bit string that will be decoded at the learner as the midpoint of the bin. We formalize this below.

Let $[\alpha, \beta]$ be the 'expected' real interval as described above. First, divide $[\alpha, \beta]$ into 2^B bins of equal width $\frac{\beta-\alpha}{2^B}$, and associate with the midpoint of each bin, a B-length bit string. Then, the encoder $\mathsf{enc}(x, B, [\alpha, \beta])$ returns the bit string $\mathbf{s}$ associated with its nearest bin midpoint (even if $x \notin [\alpha, \beta]$), and the decoder $\mathsf{dec}(\mathbf{s}, B, [\alpha, \beta])$ returns the midpoint (a real number in $[\alpha, \beta]$) corresponding to the bit string $\mathbf{s}$. We may simply refer to them as $\mathsf{enc}(x)$ and $\mathsf{dec}(\mathbf{s})$ when B and the interval are clear from context. The quantization scheme is summarized in Figure 2. Note that this quantization scheme uses B bits for each transmission. Another point to note is that if $x \in [\alpha, \beta]$, the quantization error between x and the decoded quantized value $\mathsf{dec}(\mathsf{enc}(x))$ is at most $\frac{\beta-\alpha}{2 \cdot 2^B}$, i.e., if x is known to lie within the interval, then the quantization error is at most a factor of $\frac{1}{2^{B+1}}$ times the width of the interval.

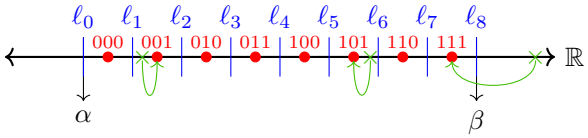

Figure 2: Illustration of the quantization scheme when $B = 3$ — the blue lines given by ℓ_i mark the separation between the 2^B equal bins, the red points denote the midpoints of these bins, and the green '$\times$'s (the values to be quantized) get mapped to their nearest midpoints.

Details of the algorithm. At round i, consider the agent $j \in [K]$ making its k^{th} cumulative arm pull, where $1 \le k \le t_i$. It observes a reward $r_{j,k}$. At the end of this round, it calculates the empirical mean of the rewards from arm j observed over all rounds upto and including round i, $\hat{\mu}_{j,i} = \frac{1}{t_i} \sum_{k=1}^{t_i} r_{j,k}$. By defining the confidence width

$$U'(i, \delta) = \sigma \sqrt{\frac{2 \log(4Kt_i^2/\delta)}{t_i}}, \tag{1}$$

for each round i and arbitrary $\delta > 0$, it can be shown (as in the proof of Lemma 3) from the subgaussian concentration inequality [13] and a union bound that

$$\mathbb{P}\left(\cup_{i \ge 1} \cup_{j \in [K]} |\hat{\mu}_{j,i} - \mu_j| > U'(i, \delta)\right) \le \delta. \tag{2}$$

Thus, at any round i, the actual mean of arm j lies in the interval $[\hat{\mu}_{j,i} - U'(i, \delta), \hat{\mu}_{j,i} + U'(i, \delta)]$ w.h.p. However, since the communication channel is bit-constrained, the agents cannot simply transmit the infinite precision real number $\hat{\mu}_{j,i}$ as is — they instead transmit a quantized version of $\hat{\mu}_{j,i}$ as described above. Let $\tilde{\mu}_{j,i}$ be the decoded estimate of the mean of arm j that the learner recovers at the end of round i, i.e., $\tilde{\mu}_{j,i} = \mathsf{dec}(\mathsf{enc}(\hat{\mu}_{j,i}))$.

To account for the potential increase in error due to the quantization necessitated by the bit-constrained channel, we introduce a slack in the confidence interval through a different confidence width $U(i,\delta)$. The goal is to obtain a concentration bound for $\tilde{\mu}_{j,i} - \mu_j$ in terms of $U(i,\delta)$, in a form similar to (2). We now provide an intuitive explanation to motivate an expression of $U(i,\delta)$ that achieves exactly this. (Note that this is not meant to be a proof; that this does indeed work will be shown in Section 5). First note that a straightforward application of the triangle inequality gives

$$|\tilde{\mu}_{j,i} - \mu_j| \leq |\tilde{\mu}_{j,i} - \hat{\mu}_{j,i}| + |\hat{\mu}_{j,i} - \mu_j|. \tag{3}$$

The first term corresponds to the quantization error and the second term corresponds to the error in the empirical mean itself. An interval in which the latter lies w.h.p. is taken care of by the bound (2), so it is enough to establish such an interval for the quantization error. Recall from the 'Quantization scheme' description before Figure 2 that the quantization error is at most $\frac{1}{2^{B+1}}$ times the width of the interval if the empirical mean $\hat{\mu}_{j,i}$ is known to originally lie in the interval. Thus, our task is to find an appropriate interval in which $\hat{\mu}_{j,i}$ lies w.h.p. to perform the quantization.

As the latest estimate of the mean that the learner has access to at the beginning of round i is the quantized estimate at round $i-1$, i.e., $\tilde{\mu}_{j,i-1}$, we construct the interval to be centered around $\tilde{\mu}_{j,i-1}$. We also want that this interval contain the new empirical mean at round i, i.e., $\hat{\mu}_{j,i}$, w.h.p. We now find a recursive expression for $U(i,\delta)$ that inductively satisfies these properties. We first make the inductive assumption that $\tilde{\mu}_{j,i-1}$ lies in $[\mu_j - U(i-1,\delta), \mu_j + U(i-1,\delta)]$ w.h.p. Combined with the knowledge that $\hat{\mu}_{j,i}$ lies in $[\mu_j - U'(i,\delta), \mu_j + U'(i,\delta)]$ w.h.p. from (2), we have that the interval $[\tilde{\mu}_{j,i-1} - U'(i,\delta) - U(i-1,\delta), \tilde{\mu}_{j,i-1} + U'(i,\delta) + U(i-1,\delta)]$ contains $\hat{\mu}_{j,i}$ w.h.p. This is illustrated in Figure 3. Note that we have found an interval that we 'expect' the empirical mean to belong to, as promised when defining the quantization scheme relative to an interval. Thus, we perform the quantization over an interval of width $2[U'(i,\delta) + U(i-1,\delta)]$ centered at $\tilde{\mu}_{j,i-1}$, and hence, the quantization error is at most $\frac{1}{2^B}[U'(i,\delta) + U(i-1,\delta)]$.

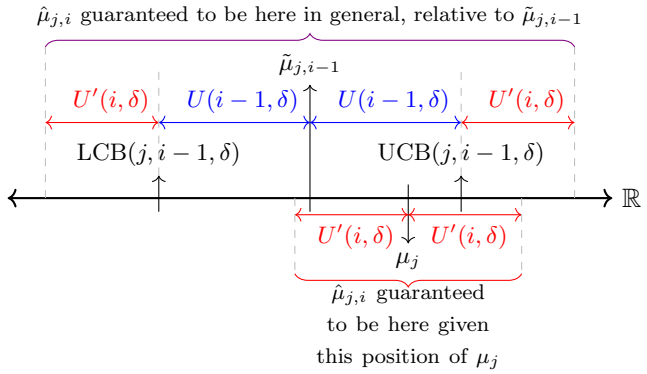

Figure 3: Motivation for defining $U(i,\delta)$ as (4) – Start with $\tilde{\mu}_{j,i-1}$, then by the inductive hypothesis, μ_j should lie within $U(i-1,\delta)$ of $\tilde{\mu}_{j,i-1}$ (blue interval). From (2), $\hat{\mu}_{j,i}$ lies within $U'(i,\delta)$ of μ_j (red interval). Putting them together, $\hat{\mu}_{j,i}$ lies within $U(i-1,\delta) + U'(i,\delta)$ of $\tilde{\mu}_{j,i-1}$.

Motivated by the above and (3), define, for $i \geq 1, j \in [K]$,

$$U(i,\delta) = \frac{1}{2^B}\left[U'(i,\delta) + U(i-1,\delta)\right] + U'(i,\delta), \tag{4}$$

where we set $U(0,\delta) = b - a$ to ensure that the induction hypothesis (namely, that $\tilde{\mu}_{j,i-1}$ satisfies its confidence bounds) holds for the index $i-1 = 0$. Now that we have an appropriate confidence width $U(i,\delta)$ that we expect (heuristically, from the above discussion; this will be proved in Section 5) to provide a similar confidence bound for $\tilde{\mu}_{j,i}$ as (2) does for $\hat{\mu}_{j,i}$, we define the appropriate LCB and UCB for our algorithm, as

$$\begin{aligned} \text{LCB}(j,i,\delta) &= \tilde{\mu}_{j,i} - U(i,\delta), \\ \text{UCB}(j,i,\delta) &= \tilde{\mu}_{j,i} + U(i,\delta). \end{aligned} \tag{5}$$

5 Analysis of the ICQ-SE algorithm

The following result can be shown for the sample complexity of SE run on the distributed MAB setup outlined above if the channel from the agent to the learner is not bit-constrained:

Theorem 1. *SE is a sound algorithm. Moreover, with probability at least $1-\delta$, it successfully identifies the best arm using at most*

$$\mathcal{O}\left(\sum_{j\neq 1}\frac{102\alpha\sigma^2}{\Delta_j^2}\ln\left(\frac{64\sigma^2\sqrt{4K/\delta}}{\Delta_j^2}\right)+1\right)$$

samples.

We now show that ICQ-SE is a sound algorithm and also analyze its sample complexity. We restrict our attention to ICQ-SE just to provide concrete results, but a similar analysis can be carried out for other confidence bound-based algorithms as well (see Remark 1). We do so by a sequence of lemmas and theorems, similar to a standard analysis of Successive Elimination-type algorithms, such as in [1] and [18]. The main novelties in our work are Lemma 2 and Theorem 6, where we relate the confidence intervals $U(i,\delta)$ and $U'(i,\delta)$ at the agents and the learner respectively, thereby allowing us to prove similar results as for vanilla Successive Elimination. Proofs of all lemmas and theorems stated below can be found in the Appendix.

Recall that μ_j is the mean of arm $j\in[K]$, $\hat{\mu}_{j,i}$ is the empirical mean of arm j at round i at the agent (which will then be encoded and sent to the learner), and $\tilde{\mu}_{j,i}$ is the decoded estimate of the mean of arm j at round i at the learner. Note that the only concentration bound we know about these quantities *a priori* is (2), which relates $\hat{\mu}_{j,i}$ and $U'(i,\delta)$. The learner, however, observes $\tilde{\mu}_{j,i}$ and constructs confidence intervals of width $U(i,\delta)$. Our goal is for the learner to be able to identify w.h.p. the best arm in $[K]$. To this end, it is desirable to have a concentration bound on the quantities available at the learner, namely, $\tilde{\mu}_{j,i}$ and $U(i,\delta)$.

Lemma 2 below relates the following — (1) the event that in some round i (hence the union over rounds), the estimated mean of arm j, $\hat{\mu}_{j,i}$, falls outside the confidence interval of width $U'(i,\delta)$ centered about the true mean μ_j, and (2) the identical event at the learner, except with the decoded mean $\tilde{\mu}_{j,i}$ and confidence width $U(i,\delta)$.

Lemma 2. *For all arms $1\leq j\leq K$, and any $\delta>0$,*

$$\bigcup_{i=1}^{\infty}\left\{|\tilde{\mu}_{j,i}-\mu_j|>U(i,\delta)\right\}\subseteq\bigcup_{i=1}^{\infty}\left\{|\hat{\mu}_{j,i}-\mu_j|>U'(i,\delta)\right\}.$$

Using Lemma 2, we obtain the desired confidence bound on the decoded means at the learner, stated formally below.

Lemma 3. *For any $\delta>0$, define the event*

$$\mathcal{E}:=\bigcup_{j=1}^{K}\bigcup_{i=1}^{\infty}\left\{|\tilde{\mu}_{j,i}-\mu_j|>U(i,\delta)\right\},$$

then $\mathbb{P}(\mathcal{E})\leq\delta$.

The above lemma establishes that the event $\mathcal{E}^c$, wherein for each arm j and at every round i, the decoded estimate at the learner $\hat{\mu}_{j,i}$ is sufficiently close to the actual mean μ_j, occurs with large probability. It then follows that with high probability, any arm that is eliminated during the successive elimination procedure must be suboptimal, as stated in Theorem 4. We now show that ICQ-SE is a sound algorithm and provide an upper bound for its sample complexity in the exponentially sparse regime (i.e., with $t_i=\alpha^i$) in Theorem 6.

Theorem 4. *With probability $\geq 1-\delta$, the best arm remains in the active set S until termination.*

To prove our main result in Theorem 6, we require a technical lemma that relates the confidence widths $U(i,\delta)$ and $U'(i,\delta)$.

Lemma 5. *Consider $B \geq 1$ and let $t_i = \alpha^i$ where $\alpha \in \mathbb{N}$ such that $\alpha < 2^{2B}$. Then for $i \geq 1$,*

$$U(i, \delta) \leq 2c\, U'(i, \delta),$$

where

$$c = \left(1 + \frac{2}{2^B}\right) \frac{2^B}{2^B - \sqrt{\alpha}}.$$

Theorem 6. *Consider $B \geq 1$ and let $t_i = \alpha^i$ where $\alpha \in \mathbb{N}$ and $1 < \alpha < 2^{2B}$. With probability at least $1 - \delta$, ICQ-SE will terminate and successfully identify the best arm after using*

$$\mathcal{O}\left(\sum_{j \neq 1} \frac{410 \alpha c^2 \sigma^2}{\Delta_j^2} \ln\left(\frac{256 c^2 \sigma^2 \sqrt{4K/\delta}}{\Delta_j^2}\right) + 1\right)$$

samples,

$$\mathcal{O}\left(\sum_{j \neq 1} \log_\alpha \left(\frac{410 \alpha c^2 \sigma^2}{\Delta_j^2} \ln\left(\frac{256 c^2 \sigma^2 \sqrt{4K/\delta}}{\Delta_j^2}\right) + 1\right)\right)$$

rounds, and

$$\mathcal{O}\left(B \sum_{j \neq 1} \log_\alpha \left(\frac{410 \alpha c^2 \sigma^2}{\Delta_j^2} \ln\left(\frac{256 c^2 \sigma^2 \sqrt{4K/\delta}}{\Delta_j^2}\right) + 1\right)\right)$$

bits, where c is as defined in Lemma 5.

Comparing Theorem 6 with the equivalent result for Successive Elimination with no quantization in Theorem 1, we see that the upper bound for the sample complexity is worse only by a constant factor, i.e., we have order-optimal performance.

Additionally, note that c depends on the choice of B and α, and in fact decreases as B and α increase. Combined with the upper bound on sample complexity in Theorem 6, we thus see a trade-off between the performance of the algorithm and the number of bits that it is allowed to use. In addition to the above theoretical result, we also investigate this trade-off via numerical simulations in Section 6.

Remark 1. *The quantization scheme ICQ can in fact be used for any other algorithm that uses confidence bounds, such as LUCB [10] and lil'UCB [11], to obtain algorithms ICQ-LUCB and ICQ-lil'UCB. This is because the crux of this scheme is the recursive definition for $U(\cdot, \cdot)$ that we obtain by separating the quantization error and the error in the empirical mean itself, via (3). A similar analysis can be carried out for ICQ-LUCB and ICQ-lil'UCB too, that we omit here for brevity.*

Remark 2. *We use the boundedness of the rewards only in the definition of $U(0, \delta)$. Recall that the algorithm proceeds with an initial random guess for the mean of each arm $\{\tilde{\mu}_{j,0}\}_{j=1}^K$. As $U(i, \delta)$ is defined in a recursive fashion, for the induction in the proof of Lemma 2 to hold, $U(0, \delta)$ needs to be such that $\{\tilde{\mu}_{j,0}\}_{j=1}^K$ are good guesses for the actual means with high probability satisfying Lemma 2. For $[a, b]$-bounded rewards, this holds trivially by taking $U(0, \delta) = (b - a)$, as $\{|\tilde{\mu}_{j,0} - \mu_j| > U(0, \delta)\} = \{|\tilde{\mu}_{j,0} - \mu_j| > (b - a)\} = \emptyset$. For unbounded rewards, however, we can no longer use a constant number of bits in each round. Specifically, in round 1, it is not possible to obtain a high-probability guess for the mean by using a bounded number of bits. Nonetheless, we can use a different quantization scheme just for the first round to ensure that the quantization error is bounded. The problem is then reduced to that with bounded rewards from round 2, whence we can use ICQ-SE as is. In Section 6, we demonstrate this on Gaussian rewards by running QuBan [12] (designed for unbounded rewards) in the first round.*

6 Numerical experiments

In this section, we present results of numerical experiments comparing the performance of ICQ-SE with other quantization algorithms proposed for multi-armed bandits in the literature. In addition to the unquantized setting as a baseline, we compare ICQ-SE with the quantization schemes QuBan [12] and Fed-SEL [17]. Each of these schemes is implemented on top of the same batched Successive Elimination algorithm that ICQ-SE uses to highlight the difference between the quantization schemes. The algorithms are compared based on their (expected) sample complexity $E[\tau_\delta]$, (expected) round complexity $E[\tau_{r,\delta}]$ and (expected) communication complexity $E[B_\delta]$. In all our experiments, the performance of each algorithm was averaged over 4000 iterations.

QuBan [12] was proposed for the regret minimization setting where each sample that the agent observes is quantized and sent to the learner. A key feature of QuBan is that the agent uses shorter codewords to quantize samples close to the current estimate of the mean at the learner, while reward samples which are farther away are assigned longer codewords. This helps to minimize the expected number of bits used at each round. While this is a sound approach, it could result in a higher number of bits being used unnecessarily for our framework. QuBan has a parameter $\epsilon > 0$ that provides a trade-off between the number of bits used and the performance of the algorithm (a smaller value of ϵ provides a smaller regret using a higher number of bits).

In Fed-SEL [17], in each round i, the entire interval $[a, b]$ is divided into bins of length $U'(i, \delta)$ and the empirical mean is quantized to the midpoint of one of these bins. The main drawback of this approach is that the number of bits used at each round is inversely proportional to the confidence bound at each round, i.e., the number of bits used per round grows with the number of rounds, making it hard to control the cumulative number of bits that the algorithm uses.

The first set of experiments in Figures 5a, 5b, 5c, 5d and 5e analyze the dependence of the performance of ICQ-SE on the parameters α (controlling the sparsity of communication) and B (the number of bits in each transmission). We consider a five-armed multi-armed bandit instance where each arm is associated with a Beta$(\gamma, 1-\gamma)$ distribution, with γ generated uniformly at random from $[0, 1]$. In Figures 5a and 5b, we observe that the number of communication rounds used by ICQ-SE to converge decreases with α while the number of samples used increases with α. This is expected because increasing α results in sparser communication between the agents and the learner reducing the round complexity while the number of samples used increases. Figure 5c shows the dependence of the cumulative number of bits used by the algorithm to converge (which we call *communication complexity*) on α. The decrease in the communication complexity with α is a natural consequence of the decrease in the communication round complexity.

Figures 4a, 4b, and 4c compare the performance of ICQ-SE, QuBan and Fed-SEL. We consider a five-armed multi-armed bandit instance where each arm is associated with a bounded support reward distribution, in particular the Beta$(\gamma, 1-\gamma)$ distribution with γ generated uniformly at random from $[0, 1]$. We observe that ICQ-SE with $B = 3$ performs comparably with QuBan ($\epsilon = 0.5$), and better than QuBan ($\epsilon = 2$) and Fed-SEL in terms of sample and round complexity, while using a much lesser number of bits than all of them.

In Figures 4d, 4e, and 4f, we compare the performance of ICQ-SE and QuBan when the reward distributions associated with the arms are Gaussian (and hence unbounded; recall Remark 2). We consider a five-armed multi-armed bandit instance where each arm is associated with a Gaussian reward distribution of standard deviation 0.125 whose means are generated uniformly at random from the interval $[0, N]$, where N is a sample drawn from a Gaussian distribution $\mathcal{N}(0, 9)$. We use QuBan with $\epsilon = 2$ for the first round of ICQ-SE as discussed in Remark 2. We again observe that ICQ-SE with $B = 3$ performs comparably with the others in terms of sample and round complexity while using a much lesser number of bits. We make no comparison with Fed-SEL in this case because it is unclear how to extend the scheme to unbounded rewards, since it starts by dividing the finite-size reward range into bins of length $U'(i, \delta)$.

The final set of experiments in Figures 4g, 4h and 4i compare the dependence of the performance of ICQ-SE and QuBan on the hardness of the underlying bandit instance when the reward distributions associated with the arms are Gaussian. We consider a five-armed multi-armed bandit instance where each arm is associated with a Gaussian distribution of standard deviation 0.125. Four of the arms have mean 0 and the remaining arm has mean $\Delta \in [0, 1]$. A lower Δ implies that the mean of the optimal arm is closer to that of the non-optimal arms, resulting in a harder instance. As expected, the performance of all the algorithms improves as the hardness of the instance decreases. Moreover, we see the same trend with ICQ-SE ($B = 3$)

as earlier, especially on the harder instances.

Finally, in Figure 5, we also numerically analyze the impact of varying α (controlling the sparsity of communication) and B (the number of bits in each transmission) on the performance of ICQ-SE.

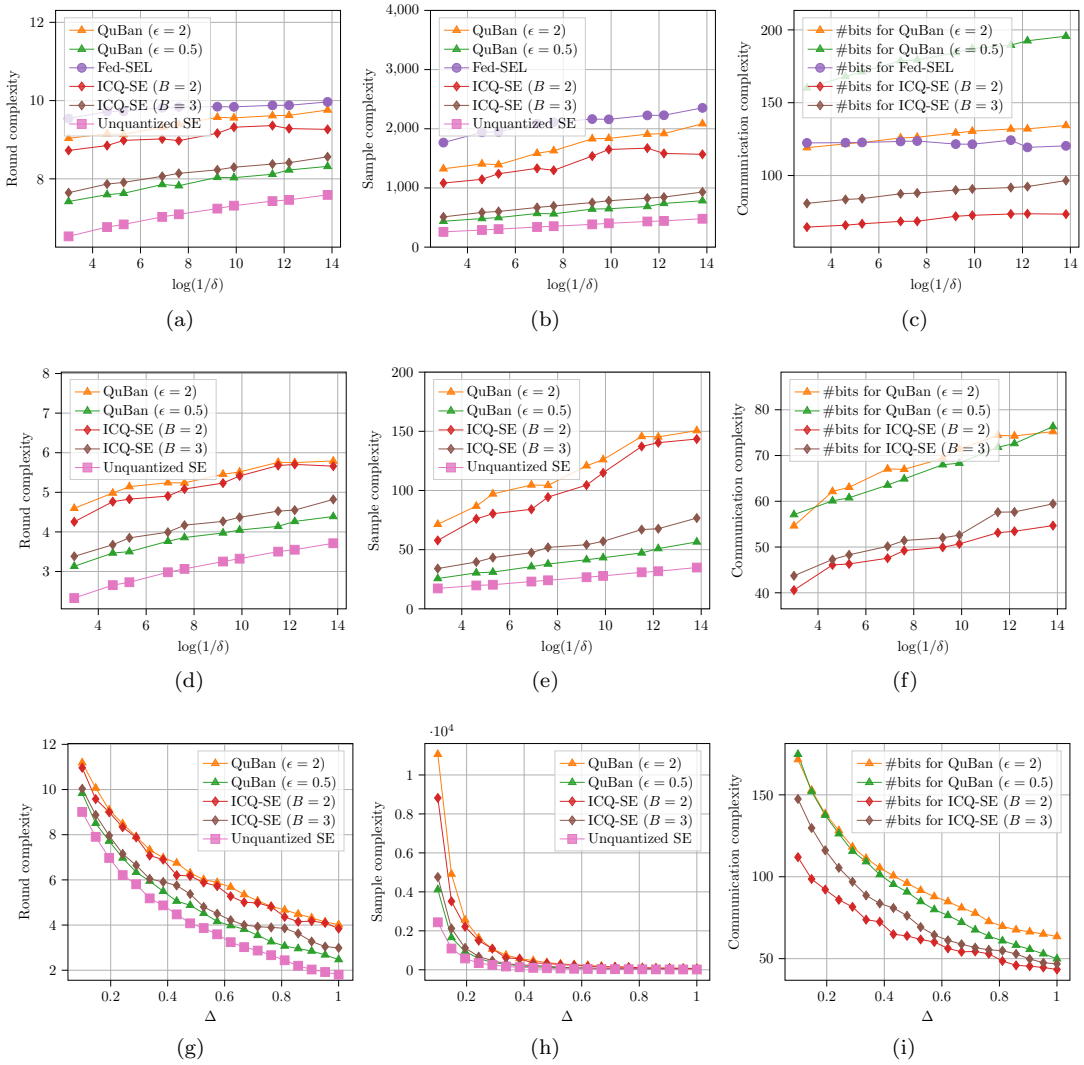

Figure 4: Figures 5a, 5b and 5c demonstrate the dependence of the performance of ICQ-SE on α. Figures 5d and 5e demonstrate the dependence on β. Figures 4a, 4b and 4c compare ICQ-SE with QuBan [12] and Fed-SEL [17] for bounded rewards while Figures 4d, 4e and 4f compare ICQ-SE with QuBan for unbounded rewards. Finally, Figures 4g, 4h and 4i compare the dependence of ICQ-SE and QuBan on the hardness of the underlying instance.

7 Conclusion

We propose ICQ, a novel quantization scheme for the distributed best-arm identification problem where the learner does not have access to full-precision rewards, and analyze ICQ-SE, which is the application of ICQ to the Successive Elimination algorithm for this setting. Future lines of work include: (1) using a variable-length and adaptive quantization scheme in each round to reduce the communication complexity; for example, a Lloyd-Max quantizer based on the empirical distribution, (2) characterizing a lower bound

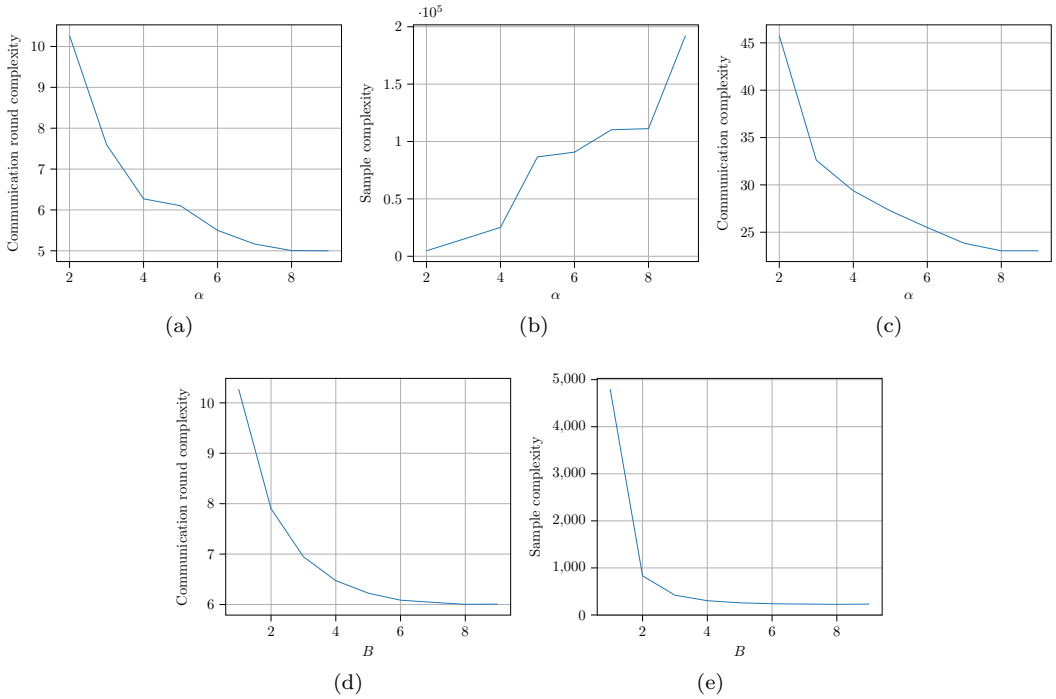

Figure 5: Figures 5a, 5b and 5c demonstrate the dependence of the performance of ICQ-SE on α. Figures 5d and 5e demonstrate the dependence on β.

on the communication complexity required to ensure a certain sample/round complexity, and (3) developing quantization schemes for the fixed budget variant of the best-arm identification problem.

References

[1] E. Even-Dar, S. Mannor, and Y. Mansour, "PAC bounds for multi-armed bandit and markov decision processes," in *Computational Learning Theory*. Springer, 2002.

[2] J. Audibert, S. Bubeck, and R. Munos, "Best arm identification in multi-armed bandits," in *COLT*, 2010.

[3] K. Jamieson and R. Nowak, "Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting," in *CISS*, 2014.

[4] S. Bubeck, R. Munos, and G. Stoltz, "Pure exploration in finitely-armed and continuous-armed bandits," *Theoretical Computer Science*, 2011.

[5] P. Dames, P. Tokekar, and V. Kumar, "Detecting, localizing, and tracking an unknown number of moving targets using a team of mobile robots," *The International Journal of Robotics Research*, 2017.

[6] V. Savic, H. Wymeersch, and S. Zazo, "Belief consensus algorithms for fast distributed target tracking in wireless sensor networks," *Signal Processing*, 2014.

[7] L. Song, C. Fragouli, and D. Shah, "Recommender systems over wireless: Challenges and opportunities," in *IEEE Information Theory Workshop (ITW)*, 2018.

[8] K. Ding, J. Li, and H. Liu, "Interactive anomaly detection on attributed networks," in *Proceedings of the 12th ACM international conference on web search and data mining (WSDM)*, 2019.

[9] V. Gandikota, D. Kane, R. K. Maity, and A. Mazumdar, "vqSGD: Vector quantized stochastic gradient descent," *IEEE Transactions on Information Theory*, 2022.

[10] S. Kalyanakrishnan, A. Tewari, P. Auer, and P. Stone, "PAC subset selection in stochastic multi-armed bandits." in *ICML*, 2012.

[11] K. Jamieson, M. Malloy, R. Nowak, and S. Bubeck, "lil'ucb: An optimal exploration algorithm for multi-armed bandits," in *COLT*. PMLR, 2014.

[12] O. A. Hanna, L. Yang, and C. Fragouli, "Solving multi-arm bandit using a few bits of communication," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022.

[13] T. Lattimore and C. Szepesvári, *Bandit algorithms*. Cambridge University Press, 2020.

[14] A. Garivier and E. Kaufmann, "Optimal best arm identification with fixed confidence," in *COLT*. PMLR, 2016.

[15] C. Shi and C. Shen, "Federated multi-armed bandits," *Proceedings of the AAAI Conference on Artificial Intelligence*, May 2021.

[16] A. Mitra, H. Hassani, and G. J. Pappas, "Linear stochastic bandits over a bit-constrained channel," 2022. [Online]. Available: https://arxiv.org/abs/2203.01198

[17] A. Mitra, H. Hassani, and G. Pappas, "Exploiting heterogeneity in robust federated best-arm identification," 2021. [Online]. Available: https://arxiv.org/abs/2109.05700

[18] K. S. Reddy, P. N. Karthik, and V. Y. F. Tan, "Almost cost-free communication in federated best arm identification," *AAAI (to appear)*, 2023. [Online]. Available: https://arxiv.org/abs/2208.09215

[19] F. Alzahrani and A. Salem, "Sharp bounds for the Lambert W function," *Integral Transforms and Special Functions*, 2018.

A Proofs

Proof of Lemma 2. We wish to show that for all arms $1 \leq j \leq K$ and any $\delta > 0$,

$$\bigcup_{i=1}^{\infty} \{|\tilde{\mu}_{j,i} - \mu_j| > U(i,\delta)\} \subseteq \bigcup_{i=1}^{\infty} \{|\hat{\mu}_{j,i} - \mu_j| > U'(i,\delta)\}.$$

First observe that the sets on the left-hand side are empty when $i = 0$, as

$$\{|\tilde{\mu}_{j,0} - \mu_j| > U(0,\delta)\} = \{|\tilde{\mu}_{j,0} - \mu_j| > (b-a)\} = \emptyset, \tag{6}$$

so it is sufficient to show that

$$\bigcup_{i=0}^{\infty} \{|\tilde{\mu}_{j,i} - \mu_j| > U(i,\delta)\} \subseteq \bigcup_{i=1}^{\infty} \{|\hat{\mu}_{j,i} - \mu_j| > U'(i,\delta)\}.$$

We do so by instead proving the following equivalent statement — for each $0 \leq i < \infty$,

$$\{|\tilde{\mu}_{j,i} - \mu_j| > U(i,\delta)\} \subseteq \bigcup_{k=1}^{\infty} \{|\hat{\mu}_{j,k} - \mu_j| > U'(k,\delta)\}.$$

The proof is by induction over i. The base case, $i = 0$, holds trivially as the left-hand side is $\emptyset$. Assume now that this property holds for index $i - 1 \geq 0$, i.e.,

$$\{|\tilde{\mu}_{j,i-1} - \mu_j| > U(i-1,\delta)\} \subseteq \bigcup_{k=1}^{\infty} \{|\hat{\mu}_{j,k} - \mu_j| > U'(k,\delta)\}.$$

Using (3) and (4) respectively, we have

$$\begin{aligned}
&\{|\tilde{\mu}_{j,i} - \mu_j| > U(i,\delta)\} \\
\subseteq &\{|\tilde{\mu}_{j,i} - \hat{\mu}_{j,i}| + |\hat{\mu}_{j,i} - \mu_j| > U(i,\delta)\}, \\
\subseteq &\{|\hat{\mu}_{j,i} - \mu_j| > U'(i,\delta)\} \cup \\
&\left\{|\tilde{\mu}_{j,i} - \hat{\mu}_{j,i}| > \frac{1}{2^B}U'(i,\delta) + \frac{1}{2^B}U(i-1,\delta)\right\},
\end{aligned}$$

where the second step follows as $U(i,\delta) = U'(i,\delta) + \frac{1}{2^B}U'(i,\delta) + \frac{1}{2^B}U(i-1,\delta)$ and $a+b > x+y$ requires at least one of $a > x$ or $b > y$.

We now bound the second term in the union on the right-hand side. The estimated mean is encoded using enc on the interval $[\mathrm{LCB}(j,i-1,\delta) - U'(i,\delta), \mathrm{UCB}(j,i-1,\delta) + U'(i,\delta)]$. As long as $\hat{\mu}_{j,i}$ lies in this interval, the learner can decode $\tilde{\mu}_{j,i}$ to an error within $\frac{1}{2} \cdot \frac{1}{2^B}$ times the width of this interval, given by $2\left(U'(i,\delta) + U(i-1,\delta)\right)$. Hence we have

$$\begin{aligned}
&\left\{|\tilde{\mu}_{j,i} - \hat{\mu}_{j,i}| > \frac{1}{2^B}U'(i,\delta) + \frac{1}{2^B}U(i-1,\delta)\right\} \\
\subseteq &\{|\tilde{\mu}_{j,i-1} - \hat{\mu}_{j,i}| > U'(i,\delta) + U(i-1,\delta)\} \\
\subseteq &\{|\tilde{\mu}_{j,i-1} - \mu_j| + |\hat{\mu}_{j,i} - \mu_j| > U'(i,\delta) + U(i-1,\delta)\} \\
\subseteq &\{|\tilde{\mu}_{j,i-1} - \mu_j| > U(i-1,\delta)\} \cup \{|\hat{\mu}_{j,i} - \mu_j| > U'(i,\delta)\} \\
\subseteq &\bigcup_{k=1}^{\infty} \{|\hat{\mu}_{j,k} - \mu_j| > U'(k,\delta)\}, \qquad (7)
\end{aligned}$$

where the last step follows from the induction hypothesis. □

Proof of Lemma 3. This follows easily by a union bound, as

$$\begin{aligned}
\mathbb{P}(\mathcal{E}) &= \mathbb{P}\left(\bigcup_{j=1}^{K}\bigcup_{i=1}^{\infty}\{|\tilde{\mu}_{j,i} - \mu_j| > U(i,\delta)\}\right) \\
&\overset{(a)}{\leq} \mathbb{P}\left(\bigcup_{j=1}^{K}\bigcup_{i=1}^{\infty}\{|\hat{\mu}_{j,i} - \mu_j| > U'(i,\delta)\}\right) \\
&\leq \mathbb{P}\left(\bigcup_{j=1}^{K}\bigcup_{i=1}^{\infty}\left\{|\hat{\mu}_{j,i} - \mu_j| > \sigma\sqrt{\frac{2\log(4Kt_i^2/\delta)}{t_i}}\right\}\right) \\
&\leq \sum_{j=1}^{K}\sum_{i=1}^{\infty}\mathbb{P}\left(\left|\frac{1}{t_i}\sum_{k=1}^{t_i} r_{j,k} - \mu_j\right| > \sigma\sqrt{\frac{2\log(4Kt_i^2/\delta)}{t_i}}\right) \\
&\leq \sum_{j=1}^{K}\sum_{i=1}^{\infty}\mathbb{P}\left(\left|\frac{1}{t_i}\sum_{k=1}^{t_i} r_{j,k} - \mu_j\right| > \sigma\sqrt{\frac{2\log(4Kt_i^2/\delta)}{t_i}}\right) \\
&\overset{(b)}{\leq} \sum_{j=1}^{K}\sum_{i=1}^{\infty} 2\exp\left(-\log(4Kt_i^2/\delta)\right) = \sum_{j=1}^{K}\sum_{i=1}^{\infty} 2\left(\frac{\delta}{4Kt_i^2}\right) \\
&\overset{(c)}{<} \sum_{j=1}^{K}\frac{\delta}{K} = \delta,
\end{aligned}$$

where (a) follows from Lemma 2, (b) from the definition of σ^2-subgaussianity, and (c) from $\sum_{i=1}^{\infty}\frac{1}{t_i^2} \leq \sum_{i=1}^{\infty}\frac{1}{i^2} < 2$. □

Proof of Theorem 4. For arm k to be dropped from the active set S at round i, there must exist an arm l such that

$$\begin{aligned} &\text{LCB}(l,i,\delta) \geq \text{UCB}(k,i,\delta) \\ \iff &\tilde{\mu}_{l,i} - U(i,\delta) \geq \tilde{\mu}_{k,i} + U(i,\delta). \end{aligned}$$

On the set $\mathcal{E}^c$ (with $\mathcal{E}$ as defined in Lemma 3), we have for all arms j at round i, $|\tilde{\mu}_{j,i} - \mu_j| \leq U(i,\delta)$. In particular, for arms k and l, we have

$$\mu_l + U(i,\delta) \geq \tilde{\mu}_{l,i} \text{ and } \tilde{\mu}_{k,i} \geq \mu_k - U(i,\delta).$$

Combining the above inequalities, we have that arm k can be dropped from S only if there exists an arm l such that

$$\begin{aligned} \mu_l + U(i,\delta) - U(i,\delta) &\geq \mu_k - U(i,\delta) + U(i,\delta) \\ \iff \mu_l &\geq \mu_k. \end{aligned}$$

We thus have that the best arm will always remain in the active set under $\mathcal{E}^c$, which completes the proof. $\square$

Proof of Lemma 5. Let $i \geq 1$. From the definition of $U(i,\delta)$ and using the result that for $j \leq i$, $t_j \leq t_i$, it follows that

$$\begin{aligned} U(i,\delta) &= U'(i,\delta) + \frac{1}{2^B}\left[U'(i,\delta) + U(i-1,\delta)\right] \\ &= \left(1+\frac{1}{2^B}\right)\sum_{j=1}^{i}\left(\frac{1}{2^B}\right)^{i-j} U'(j,\delta) + \left(\frac{1}{2^B}\right)^i (b-a) \\ &= \left(1+\frac{1}{2^B}\right)\sum_{j=1}^{i}\left(\frac{1}{2^B}\right)^{i-j} \sigma\sqrt{\frac{2\log\left(4Kt_j^2/\delta\right)}{t_j}} + \left(\frac{1}{2^B}\right)^i (b-a) \\ &\leq \sigma\left(1+\frac{1}{2^B}\right)\sqrt{2\log\left(4Kt_i^2/\delta\right)}\sum_{j=1}^{i}\left(\frac{1}{2^B}\right)^{i-j}\frac{1}{\sqrt{t_j}} + \left(\frac{1}{2^B}\right)^i (b-a). \end{aligned}$$

Now, using $t_i = \alpha^i$ gives us that

$$U(i,\delta) \leq \sigma\left(1+\frac{1}{2^B}\right)\sqrt{2\log\left(4Kt_i^2/\delta\right)}\left(\frac{1}{2^B}\right)^i\sum_{j=1}^{i}\left(\frac{2^B}{\sqrt{\alpha}}\right)^j + \left(\frac{1}{2^B}\right)^i (b-a).$$

As $\sqrt{\alpha} < 2^B$, it follows that

$$\begin{aligned} U(i,\delta) &\leq \sigma\left(1+\frac{1}{2^B}\right)\sqrt{2\log\left(4Kt_i^2/\delta\right)}\left(\frac{1}{2^B}\right)^i \frac{2^B}{\sqrt{\alpha}}\frac{\left(\frac{2^B}{\sqrt{\alpha}}\right)^i - 1}{\frac{2^B}{\sqrt{\alpha}} - 1} + \left(\frac{1}{2^B}\right)^i (b-a) \\ &\leq \sigma\left(1+\frac{1}{2^B}\right)\sqrt{2\log\left(4Kt_i^2/\delta\right)}\left(\frac{1}{2^B}\right)^i \frac{2^B}{\sqrt{\alpha}}\frac{\left(\frac{2^B}{\sqrt{\alpha}}\right)^i}{\frac{2^B}{\sqrt{\alpha}} - 1} + \left(\frac{1}{2^B}\right)^i (b-a) \\ &= \sigma\left(1+\frac{1}{2^B}\right)\sqrt{2\log\left(4Kt_i^2/\delta\right)}\frac{2^B}{\sqrt{t_i}}\frac{1}{2^B-\sqrt{\alpha}} + \left(\frac{1}{2^B}\right)^i (b-a) \\ &= \left(1+\frac{1}{2^B}\right)\frac{2^B}{2^B-\sqrt{\alpha}}U'(i,\delta) + \left(\frac{1}{2^B}\right)^i (b-a). \end{aligned}$$

To show that

$$U(i,\delta) \leq 2\left(1+\frac{1}{2^B}\right)\frac{2^B}{2^B-\sqrt{\alpha}}U'(i,\delta),$$

it suffices to show that

$$\left(\frac{1}{2^B}\right)^i (b-a) \leq \left(1+\frac{1}{2^B}\right)\frac{2^B}{2^B-\sqrt{\alpha}}U'(i,\delta).$$

As

$$\left(1+\frac{1}{2^B}\right)\frac{2^B}{2^B-\sqrt{\alpha}} \geq 1,$$

for $i \geq 1$, and the denominator of $U'(i,\delta)$ satisfies the following result:

$$\sqrt{t_i} = \left(\sqrt{\alpha}\right)^i < \left(2^B\right)^i,$$

it suffices to show that $\sigma\sqrt{2\log\left(4Kt_i^2/\delta\right)} \geq (b-a)$. This will hold as long as

$$i \geq \log_\alpha\left(\sqrt{\frac{\delta}{4K}\exp\left(\frac{(b-a)^2}{2\sigma^2}\right)}\right).$$

For a small enough δ, this will be satisfied trivially. If

$$\delta < 4K\alpha^2\exp\left(-(b-a)^2/(2\sigma^2)\right),$$

then we have that for all $i \geq 1$,

$$U(i,\delta) \leq 2\left(1+\frac{1}{2^B}\right)\frac{2^B}{2^B-\sqrt{\alpha}}U'(i,\delta).$$

□

Proof of Theorem 6. Assume w.l.o.g. that the optimal arm is arm 1. Similar to the proof of Theorem 4, a suboptimal arm j will be removed if

$$\mathrm{LCB}(1,i,\delta) > \mathrm{UCB}(j,i,\delta). \tag{8}$$

On $\mathcal{E}^c$, we have that

$$\tilde{\mu}_{1,i} \geq \mu_1 - U(i,\delta),$$
$$\tilde{\mu}_{j,i} \leq \mu_j + U(i,\delta).$$

The equation (8) is guaranteed to occur if

$$\mu_1 - 2U(i,\delta) \geq \mu_j + 2U(i,\delta).$$

Define $\Delta_j = \mu_1 - \mu_j$ to be the suboptimality gap of arm j for $j \in \{2,\ldots,K\}$. The sample complexity of arm j is the number of samples needed to remove the suboptimal arm j from S. Define T_j to be the smallest value of t_i that satisfies

$$\Delta_j \geq 4U(i,\delta).$$

Then αT_j is an upper bound on the sample complexity of arm j. The factor of α is needed here as we communicate exponentially sparsely only. For $t_i = \alpha^i$, we have, from Lemma 5,

$$U(i,\delta) \leq 2c\,U'(i,\delta),$$

where

$$c = \left(1+\frac{1}{2^B}\right)\frac{2^B}{2^B-\sqrt{\alpha}}.$$

If T_j' is the smallest value of t_i satisfying

$$U'(i,\delta) \leq \frac{\Delta_j}{8c},$$

then $\alpha T_j'$ is an upper bound on αT_j and thus an upper bound on the sample complexity of suboptimal arm j. Define $a = \frac{\Delta_j^2}{256c^2\sigma^2}$ and $b = \ln(4K/\delta)/2$. Then the smallest value of i that is a solution to

$$U'(i,\delta) = \frac{\Delta_j}{8c}$$

satisfies

$$-at_i e^{-at_i} = -ae^{-b}.$$

Let δ be small enough such that

$$ae^{-b} = \frac{\Delta_j^2}{256c^2\sigma^2} e^{-\ln(4K/\delta)/2} = \frac{\Delta_j^2}{256c^2\sigma^2}\sqrt{\frac{\delta}{4K}} < \frac{1}{e}.$$

Then the solution to this equation would be $\lceil \frac{-1}{a} W_{-1}(-ae^{-b}) \rceil$, where $W_{-1}(y)$, $\frac{-1}{e} < y < 0$ is the smallest value of $x < 0$ satisfying $xe^x = y$. There are in fact two solutions, $W_0(y)$ and $W_{-1}(y)$, out of which $W_{-1}(y)$ is smaller (more negative) and belongs to $(-\infty, -1)$. Using Theorem 3.1 of [19], we have that $W_{-1}(y) > \frac{e}{e-1}\ln(-y)$. It follows that

$$\begin{aligned}
\alpha T_j &\leq \frac{-\alpha}{a}\frac{e}{e-1}\ln(ae^{-b}) + 1 \\
&= \frac{\alpha e}{e-1}\frac{256c^2\sigma^2}{\Delta_j^2}(b - \ln a) + 1 \\
&\leq \frac{410\alpha c^2\sigma^2}{\Delta_j^2}\ln\left(\frac{256c^2\sigma^2\sqrt{4K/\delta}}{\Delta_j^2}\right) + 1.
\end{aligned}$$

□