Communication-Constrained Bandits under Additive Gaussian Noise

Prathamesh Mayekar [1] **Jonathan Scarlett** [1] **Vincent Y. F. Tan** [1]

Abstract

We study a distributed stochastic multi-armed bandit where a client supplies the learner with communication-constrained feedback based on the rewards for the corresponding arm pulls. In our setup, the client must encode the rewards such that the second moment of the encoded rewards is no more than P, and this encoded reward is further corrupted by additive Gaussian noise of variance σ^2; the learner only has access to this corrupted reward. For this setting, we derive an information-theoretic lower bound of $\Omega\left(\sqrt{\frac{KT}{\mathtt{SNR}\wedge 1}}\right)$ on the minimax regret of any scheme, where $\mathtt{SNR} := \frac{P}{\sigma^2}$, and K and T are the number of arms and time horizon, respectively. Furthermore, we propose a multi-phase bandit algorithm, UE-UCB++, which matches this lower bound to a minor additive factor. UE-UCB++ performs uniform exploration in its initial phases and then utilizes the *upper confidence bound* (UCB) bandit algorithm in its final phase. An interesting feature of UE-UCB++ is that the coarser estimates of the mean rewards formed during a uniform exploration phase help to refine the encoding protocol in the next phase, leading to more accurate mean estimates of the rewards in the subsequent phase. This positive reinforcement cycle is critical to reducing the number of uniform exploration rounds and closely matching our lower bound.

1. Introduction

The *stochastic multi-armed bandit* (SMAB) is a classic sequential decision-making problem with decades of research (see Lattimore & Szepesvári (2020) for a recent survey). The SMAB problem crucially captures the trade-off between exploring actions to find the optimal one or exploiting the action believed to be the best, but possibly sub optimal action, lying at the heart of many practical sequential decision-making problems. In its simplest form, the SMAB problem is a sequential game between the learner and the environment. The game is played over a finite number of rounds. In each round, the learner plays an arm from a given set of arms, and then the environment reveals a reward corresponding to the played arm. The reward for a particular arm is iid across rounds, with the reward distribution unknown to the learner. The learner's goal is to maximize the expected cumulative reward over the entirety of the game.

In another direction, in modern distributed optimization frameworks such as Federated Learning (Konečný et al., 2016), a central *learner* trains a global machine learning model by leveraging updates from remote clients (typically, stochastic gradients computed on local data). However, a major bottleneck in such settings is the *uplink* communication cost – the cost of sending remote client's updates to the learner. This has spurred a recent line of work that seeks to minimize the amount of communication by a remote client while still maintaining the accuracy of the global model (see Kairouz et al. (2021) for a recent survey).

An emerging abstraction to understand the trade-off between communication and accuracy in distributed optimization is to understand the one-client setting, where one remote client supplies the learner with gradient updates over a rate-limited channel (see, for instance, Acharya et al. (2021); Gandikota et al. (2019); Lin et al. (2021); Mayekar & Tyagi (2020)). A similar abstraction has been studied for distributed SMAB problems in two interesting recent works: Hanna et al. (2022); Mitra et al. (2022). Our work is inspired by these two studies.

In particular, Hanna et al. (2022) studies the distributed SMAB problem where a client supplies the learner with reward feedback corresponding to its arm pull. To account for the uplink cost from the client to the learner, at the tth round, the client must quantize the reward feedback to r_t-bits, and only the quantized reward feedback is available to the learner for further decision-making. Mitra et al. (2022) study a linear distributed SMAB problem where at each round, the remote client supplies the learner with a r-

[1]National University of Singapore. Correspondence to: Prathamesh Mayekar, Jonathan Scarlett, Vincent Y. F. Tan <pratha22@nus.edu.sg, scarlett@comp.nus.edu.sg, vtan@nus.edu.sg>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

bit quantized version of its estimate of the d-dimensional linear model parameter. These studies investigate similar questions:

Q1a Hanna et al. (2022): *What is the minimum average bits $\frac{1}{T}\sum_{t=1}^{T} r_t$ needed to attain the regret of the centralized setting in the distributed bandit game?*

Q1b Mitra et al. (2022): *What is the minimum r needed to attain the regret of the centralized setting in the distributed bandit game?*

Hanna et al. (2022) shows that $\mathcal{O}(1)$ bits of average communication is sufficient to attain the regret of centralized setting, and Mitra et al. (2022) shows that $r = \mathcal{O}(d)$ is sufficient[1] to attain the regret of the centralized setting.

In many practical distributed learning scenarios, the updates from the client to the learner are sent over a noisy wireless network. To model this, recent works on distributed optimization have considered communication over noisy channels; see, for instance, Amiri & Gündüz (2020); Jha et al. (2022) and the references therein. Unfortunately, the results for distributed optimization over noisy channels do not apply to the distributed SMAB problem over noisy channels. Moreover, the model for communication constraints in the recent works on distributed SMAB does not accurately capture communication over noisy channels. To bridge this gap, we study a distributed SMAB problem where the remote client codes the reward feedback to ensure that the encoded reward has a second moment less than P, and then the client transmits the encoded reward over an (AWGN) channel with noise variance σ^2 (this is equivalent to corrupting the encoded reward with an additive Gaussian noise of mean zero and variance σ^2); only the channel's output is available to the learner for pulling the arm at the next round. We define the signal-to-noise ratio in our problem as

$$\texttt{SNR} := \frac{P}{\sigma^2}.$$

We remark that the restriction of the second moment of the encoded messages to P and transmission over an AWGN channel is a ubiquitous communication model in information theory and communication and dates back to Shannon's seminal work; see, for instance, Shannon (1948), Cover & Thomas (2006, Chapter 9), and Tse & Viswanath (2005, Chapter 8). The second moment constraint is used to model that most edge devices are restricted to transmitting up to a particular transmission power due to physical constraints. The AWGN channel is often used as a first approximation for various communication settings.

Coming back to comparisons with Hanna et al. (2022) and Mitra et al. (2021), the counterpart of **Q1** in our setting is the following question:

Q2: *What is the minimum* SNR *needed to attain the regret of the centralized setting in the distributed bandit game over an AWGN channel?*

In this work, we resolve the following more general question upto multiplicative log and minor additive factors:

Q3: *How does the minimax regret scale as a function of* SNR*?*

1.1. Our Contributions

We show that any bandit algorithm employed by the learner and any encoding strategy employed by the client must incur a worst-case expected cumulative regret of at least[2]

$$\Omega\left(\sqrt{\frac{KT}{\texttt{SNR} \wedge 1}} + KB\right),$$

where K is the number of bandit arms, T is time horizon, and B is upper bound on the L_2-norm[3] of rewards. A technical difficulty in deriving this lower bound is to account for the loss of information due to transmission over the AWGN channel. We overcome this difficulty by leveraging classic information-theoretic results for transmission over an AWGN channel and combining these techniques with the standard minimax lower bound proof for SMAB. An interesting consequence of our proof technique is that, up to constant factors, we can recover the lower bound in Hanna et al. (2022) with an arguably simpler proof; see Appendix for details.

Unlike Hanna et al. (2022) and Mitra et al. (2022), we don't quantize our updates. In our setup, quantization and transmitting over an AWGN channel would lead to decoding errors at every round, which can pile up and result in linear regret unless SNR grows with the time horizon T; such a scheme would not allow us to resolve **Q3**. Instead, we use a *center and scale* encoding protocol CAS, where the centering and scaling factors are learned over time. We also design a new bandit algorithm UE-UCB++ which, along with the encoding protocol CAS, bounds the worst-case expected cumulative regret by a factor of[4]

$$\mathcal{O}\left(\sqrt{\left(\frac{1}{\texttt{SNR}} + 1\right) \cdot KT \ln T} + \frac{KB \log B}{\texttt{SNR} \wedge 1}\right).$$

Thus, we resolve **Q3** up to a multiplicative factor of $\sqrt{\ln T}$ and an additive factor of $\frac{KB \log B}{\texttt{SNR} \wedge 1}$.

UE-UCB++ performs uniform exploration for multiple initial sub-phases and then employs the classic upper con-

[1] The discrepancy between results is because of different setting: In Hanna et al. (2022), the scaler reward feedback needs to be compressed, whereas in Mitra et al. (2022) the d-dimensional linear model parameter needs to be compressed.

[2] $a \wedge b := \min\{a, b\}$; $a \vee b := \max\{a, b\}$.

[3] The L_2 norm of a random variable X is $\sqrt{\mathbb{E}\left[X^2\right]}$.

[4] $\log(\cdot)$ refers to the logarithm to the base 2; $\ln$ refers to the logarithm to the base e.

fidence bound UCB bandit algorithm. The multiple initial uniform exploration sub-phases allow us to learn CAS's optimal centering and scaling factors. An interesting feature of UE-UCB++ is that we start with coarser mean reward estimates for all arms in a given sub-phase, which helps us refine the communication scheme for the subsequent sub-phases. The refined communication scheme, in turn, leads to better mean rewards estimates in the subsequent sub-phase. Due to this positive reinforcement cycle, the variance of the mean reward estimates decays exponentially with respect to the number of uniform exploration sub-phases. As a result, we can learn an almost optimal encoding protocol after relatively few rounds of uniform exploration.

We also design a private[5] scheme where the learner does not send any information to the client. We show that for such a scheme, the worst-case expected cumulative regret can be bounded by

$$\mathcal{O}\left(\sqrt{\frac{B^2}{\mathtt{SNR}}+1}\cdot\sqrt{KT\ln T}+KB\right).$$

This simple scheme nearly matches our lower bound when $B=\mathcal{O}(1)$.

2. Setup and Preliminaries

We now proceed to describe our setup; see Figure 1 for a pictorial description.

The Distributed SMAB Problem over an AWGN channel:

1. At round $t\in[T]$,[6] the learner asks a client to pull an arm A_t. The learner also sends some side information S_t based on the history of the game until time $t-1$ to aid the communication at the client.

2. Upon receiving this information, the client pulls arm A_t and gets a reward X_{t,A_t} corresponding to the arm A_t. The client then encodes this reward using an encoding function $C_t\colon\mathbb{R}\times\mathcal{S}\to\mathbb{R}$ to form $C_t(X_{t,A_t},S_t)$, where $\mathcal{S}$ is the domain of the information S_t sent by the learner. The encoding function must satisfy the power constraint

$$\mathbb{E}\left[C_t(X_{t,A_t},S_t)^2\right]\le P \tag{1}$$

for some $P>0$.

3. The client sends the encoded reward over the AWGN channel. The output of the channel Y_t is received by the learner, where

$$Y_t=C_t(X_{t,A_t},S_t)+Z_t \tag{2}$$

and $\{Z_t\}_{t=1}^{\infty}$ is sequence of *i.i.d.* Gaussian random variables with mean 0 and variance σ^2.

4. The learner decodes Y_t using a decoding function $D_t\colon\mathbb{R}\times\mathcal{S}\to\mathbb{R}$ to form a reward estimate $D_t(Y_t,S_t)$. The learner then uses a bandit algorithm π to decide on arm A_{t+1}, the side information S_{t+1}, and decoding function D_{t+1} using history of the game up to round t, $H_t\coloneqq(A_1,Y_1,\ldots,A_t,Y_t)$.

Remark 1. In most distributed learning settings, the communication from the learner to client (downlink communication) is not expensive. Therefore, we allow for schemes where the learner can, in principle, send entire history of the game up to a particular round as side information for the next round. However, our proposed scheme will send only a single real number as side information at each round, which is more desirable in practical settings.

Remark 2. The client is 'memoryless' in our setup; the client cannot use rewards from previous arm pulls to encode the current reward. This is motivated by the fact that, as observed in Hanna et al. (2022), in a typical distributed sequential decision-making problem, the client observing the reward is a low memory sensor. Additionally, in some applications, the client at any given round may differ from those at the previous rounds.

Denote by Π_T be the set of all possible T-round bandit algorithms as described above and denote by $\mathcal{C}$ the set of all possible encoding strategies satisfying (1). We note that Π_T contains randomized algorithms, so our lower bounds hold for such algorithms, too. However, we will use a deterministic algorithm for our upper bounds.

Reward Distributions: We make the following standard assumptions on the reward distributions. We assume that the reward sequence $\{X_{t,i}\}_{t>1}$ is *i.i.d.* for all arms i in $[K]$ and the rewards for different arms are independent of each other. Denote by μ_i the mean reward of arm i. We assume that $X_{1,i}-\mu_i$ is subgaussian with variance factor 1,[7] where a subgaussian random variable with a variance factor σ is as defined below.

Definition 2.1. (Boucheron et al., 2013) *We say a random variable X is subgaussian with a variance factor α^2, if for all $\lambda\in\mathbb{R}$, we have*

$$\ln[\exp(\lambda X)]\le\frac{\lambda^2\alpha^2}{2}.$$

[5] As opposed to the popular privacy notion of differential privacy Dwork et al. (2014), we emphasize that here we refer to a much more rudimentary privacy notion, where the learner does not send information to the remote clients at any of the rounds.

[6] For an integer n, $[n]\coloneqq\{1,\ldots,n\}$.

[7] The variance factor 1 is without loss of generality; all our regret bounds scale linearly with this factor.

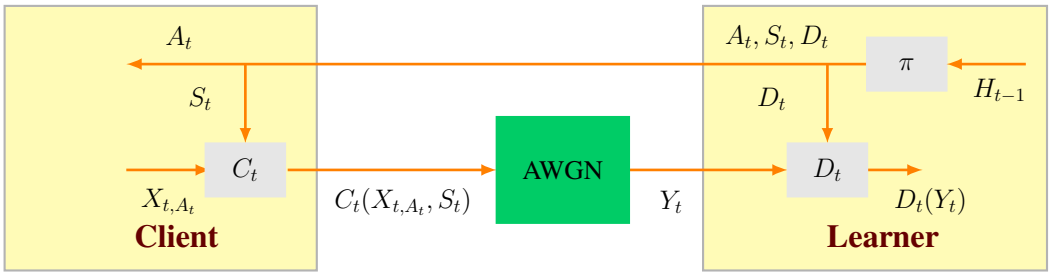

Figure 1. The Distributed Bandit Game over an AWGN channel

Furthermore, for a subgaussian random variable with variance factor α^2, we have $\mathbb{E}[X] = 0$ and $\mathbb{E}[X^2] \leq \alpha^2$.

Finally, for $B \geq 1$,[8] we assume for all $i \in [K]$, that

$$\mathbb{E}\left[X_{1,i}^2\right] \leq B^2. \tag{3}$$

Denote by $\mathcal{P}$ the set of all subgaussian distributions with variance factor 1 whose second moment satisfies (3). Then, $\mathcal{E} := \mathcal{P}^K$ describes the set of all possible bandit instances considered in our setup.

Regret: We define the regret for a bandit instance $\nu \in \mathcal{E}$, when a bandit algorithm $\pi \in \Pi$ and encoding strategies $C_1^T := (C_t)_{t\in[T]} \in \mathcal{C}^T$ are employed, as

$$R_T(\nu, \pi, C_1^T) := T\mu_{i^*} - \mathbb{E}\left[\sum_{t=1}^{T} X_{t,A_t}\right],$$

where i^* is the arm with the largest mean for the bandit instance ν. We are interested in characterizing the minimax regret

$$R_T^*(\texttt{SNR}) := \inf_{\pi\in\Pi_T} \inf_{C_1^T\in\mathcal{C}^T} \sup_{\nu\in\mathcal{E}} R_T(\nu, \pi, C_1^T) \tag{4}$$

as function of $\texttt{SNR}$, K, B, and T.

UCB bandit algorithm: We now describe the classical *Upper Confidence Bound* (UCB) bandit algorithm (Auer et al., 2002). We will build on the UCB algorithm to come up with our algorithm. Recall that at each round UCB plays the arm with the largest upper confidence bound. For subgaussian rewards with variance factor α^2, the upper confidence bound for an arm $k \in [K]$ after t rounds is defined as

$$\texttt{UCB}_k(t;\eta) = \begin{cases} \infty & \text{if} \quad N_k(t) = 0 \\ \hat{\mu}_k(t) + \sqrt{\frac{4\eta \ln T}{N_k(t)}} & \text{otherwise,} \end{cases}$$

where $N_k(t)$ is the number of times the arm k has been played up to round t, and $\hat{\mu}_k(t)$ is the empirical mean estimate of arm k. It is not difficult to prove the following regret guarantees for the UCB algorithm (see, for instance, Lattimore & Szepesvári (2020, Theorem 7.2)).

Lemma 2.2. *Suppose we have a classical bandit game over T rounds and K arms with the following reward feedback: (i), the reward feedback corresponding to all arms is subgaussian with a variance factor of α^2; (ii), the second moment of the reward feedback is at most B. Then, the regret of the* UCB *algorithm with parameter $\eta = \alpha^2$ satisfies*

$$R_T \leq 8\sqrt{\alpha^2 KT \ln T} + 6KB.$$

3. Information-Theoretic Lower Bound

Our lower bound is stated as follows.

Theorem 3.1. *For universal constants c_1 and c_2, where $0 < c_1 < 1$ and $c_2 > 0$, we have*

$$R_T^*(\texttt{SNR}) \geq c_1\left(\sqrt{KT} \cdot \frac{1}{\sqrt{\frac{1}{2}\log(1+\texttt{SNR}) \wedge 1}} \wedge c_2 T + KB\right)$$

for all $T \geq \frac{c_2 K}{\frac{1}{2}\log(1+\texttt{SNR})\wedge 1}$.

The proof of this lower bound adopts a similar high-level approach to the lower bound proofs in the standard setting (see, for instance, Lattimore & Szepesvári (2020, Chapter 15)). However, a crucial difference from the standard proofs is that we now have to account for the loss in information due to the rewards being sent over the Gaussian channel. We defer the proof to the Appendix.

Remark 3. In information theory, the term $\frac{1}{2}\log(1+\texttt{SNR})$ is the capacity of the Gaussian channel. It refers to the maximum rate at which information can be transmitted with arbitrarily small probability of error; see Cover & Thomas (2006) for details.

[8] Since $X_{1,i} - \mu_i$ is subgaussian with variance factor 1, its second moment is bounded by 1, so we may assume that $B \geq 1$.

The following corollary of Theorem 3.1 will be useful to derive matching upper bounds.

Corollary 3.2. *For universal constants c_1 and c_2, where $0 < c_1 < 1$ and $c_2 > 0$, we have*

$$R_T^*(\texttt{SNR}) \geq c_1 \left(\sqrt{KT} \cdot \frac{1}{\sqrt{\texttt{SNR} \wedge 1}} + KB \right)$$

for all $T \geq \frac{c_2 K}{\frac{1}{2}\log(1+\texttt{SNR}) \wedge 1}$.

Proof. The proof simply follows by combining Theorem 3.1 with the fact that $\ln(1+x) \leq x$ for $x \geq 0$. □

We note while $\texttt{SNR} \wedge 1$ and $\frac{1}{2}\log(1+\texttt{SNR}) \wedge 1$ may appear quite different at first glance, they are, in fact, the same up to constant factors, since $\frac{1}{2}\log(1+\texttt{SNR})$ is both upper and lower bounded by a constant times SNR for $\texttt{SNR} \in [0, 3]$, whereas for $\texttt{SNR} > 3$ both terms take the value 1.

4. Bandit Algorithms over an AWGN Channel

4.1. Our Encoding Algorithm: CAS

A crucial component of all our schemes is an encoding strategy we refer to as *centered amplitude scaling* (CAS). CAS simply centers the input X around side information S and multiplies it by θ to get

$$C_{\texttt{CAS}}(X, S; \theta) := \theta(X - S),$$

which is transmitted over the AWGN channel. If the input and side information are such that $\mathbb{E}\left[(X-S)^2\right] \leq \frac{P^2}{\theta^2}$, then $C(X, S)$ satisfies the power constraint (1).

We now build up to our main algorithm in the following two subsections.

4.2. Warm-up: UCB0 bandit algorithm

We begin by describing a simple warm-up which uses CAS with parameter $\theta = \frac{\sqrt{P}}{B}$ as the encoding algorithm at each round, and the UCB0 bandit algorithm described below.

The UCB0 bandit algorithm is a standard bandit algorithm that sends no side information at all rounds (equivalent to sending zero as the side information at all rounds). Additionally, UCB0 simply scales channel output Y_t by the inverse of the scaling parameter used by CAS, $\frac{1}{\theta} = \frac{B}{\sqrt{P}}$, to get

$$\tilde{X}_{t,A(t)} = \frac{B}{\sqrt{P}} Y_t,$$

which it uses as the reward estimate for round t. UCB0 along with its interaction with CAS is described in Algorithm 1.

Parameters: $\eta = \frac{B^2}{\texttt{SNR}} + 1$, $\theta = \frac{\sqrt{P}}{B}$
for $t \in [T]$ **do**
  At Learner
    $A_t = \arg\max_{k \in K} \texttt{UCB}_k(t-1; \eta)$ and $S_t = 0$
    Send A_t, S_t to Client
  At Client
    Play arm A_t and observe X_{t,A_t}
    Send $C_{\texttt{CAS}}(X_{t,A_t}, 0; \theta)$ over the AWGN channel
  At Learner
    Observe AWGN Output $Y_t = C_{\texttt{CAS}}(X_{t,A_t}, 0) + Z_t$
    Use $\tilde{X}_{t,A_t} = \frac{Y_t}{\theta}$ as the tth round reward
    Update UCB index for all arms

Figure 1. UCB0 bandit algorithm along with CAS encoding algorithm

Theorem 4.1. *For all bandit instances $\nu \in \mathcal{E}$, the* UCB0 *algorithm along with* CAS *encoding satisfies the power constraint* (1) *at all rounds and has the following regret guarantee:*

$$R_T(\nu, \texttt{UCB0}, \texttt{CAS}) \leq 8\sqrt{\frac{B^2}{\texttt{SNR}} + 1} \cdot \sqrt{KT \ln T} + 6KB.$$

Proof. Using (3), we have that $\mathbb{E}\left[\left(\frac{\sqrt{P}}{B}X\right)^2\right] \leq P$. We will show that the reward estimates $\tilde{X}_{t,A_t}$ used by the UCB algorithm are unbiased and subgaussian with a variance factor of $\frac{B^2}{\texttt{SNR}} + 1$. First, notice that $\tilde{X}_{t,A_t} = \frac{BY_t}{\sqrt{P}}$ and $Y_t = \frac{\sqrt{P}X_{t,A_t}}{B} + Z_t$. Combining these, we get

$$\tilde{X}_{t,A_t} = X_{t,A_t} + \frac{BZ_t}{\sqrt{P}}.$$

Note that X_{t,A_t} and Z_t are independent and subgaussian with variance factors 1 and σ^2, respectively. Therefore, the centered reward estimate, $\tilde{X}_{t,A_t} - \mu_{A_t}$, is subgaussian with variance factor $\frac{B^2}{\texttt{SNR}} + 1$.

Hence, this algorithm reduces our problem to playing a bandit game with subgaussian rewards that have a variance factor of $\alpha^2 = \frac{B^2}{\texttt{SNR}} + 1$. Since the learner uses the UCB algorithm with parameter $\eta = \alpha^2$, we can bound the regret by using Lemma 2.2. □

Remark 4. The regret bound of UCB0 matches our lower bound when $B = \mathcal{O}(1)$. However, for larger values of B, the regret bound of UCB0 is far from our lower bound. While this is a significant weakness of UCB0, it is still appealing in settings where the learner does not want to send any side information to the client due to privacy concerns.

4.3. UE-UCB bandit algorithm

The regret bound in Theorem 4.1 is off by a multiplicative factor of $\mathcal{O}(B)$ from our lower bound. This gap results from our encoding algorithm, which leads to subgaussian rewards with a variance factor of $\frac{B^2}{\texttt{SNR}}$. On the other hand,

to match the lower bound in Theorem 3.1, one requires a strategy such that the encoded rewards are subgaussian with a variance factor of $\mathcal{O}\left(\frac{1}{\texttt{SNR}}\right)$. We take the first step in arriving at such an algorithm in our second scheme, in which we combine a two-phase bandit algorithm, *Uniform Exploration-Upper Confidence Bound* (UE-UCB), with the CAS encoding algorithm.

In the first phase, UE-UCB performs uniform exploration, and in the second phase, UE-UCB uses the UCB algorithm to select the arm. We will now describe in detail the UE-UCB algorithm and the encoding strategy at the client.

Phase 1: In the first phase, the learner uniformly explores all arms, and sends no side information to the clients. The communication strategy in this phase is the same as in the previous scheme. In particular, we use CAS with parameter $\theta_1 = \frac{\sqrt{P}}{B}$ for encoding at the client. For decoding at the learner, UE-UCB scales the channel output Y_t by $1/\theta_1$ to get

$$\tilde{X}_{t,A(t)} = \frac{B}{\sqrt{P}} Y_t.$$

We run the phase one of UE-UCB for $K\tau$ rounds, where[9] $\tau := \frac{B^2}{\texttt{SNR}} + 1$.

Phase 2: For all $k \in [K]$, denote by $\hat{\mu}_k$ the mean estimates formed for all arms at the end of the first phase. Then, in the second phase, the learner runs the classic UCB algorithm treating $\tilde{X}_{t,A_t}$ as the reward to decide on an arm A_t. It then informs the client about arm A_t and the mean estimate formed for this arm after phase one, $\hat{\mu}_{A_t}$.

On the client's side, the reward corresponding to A_t is encoded by using CAS with parameter $\theta_2 = \sqrt{\frac{P}{2}}$ to get

$$C_{\texttt{CAS}}(X_{t,A_t}, \mu_{A_t}; \theta_2) := \sqrt{\frac{P}{2}}(X_{t,A_t} - \hat{\mu}_{A_t}).$$

The decoding at the learner is simply done by scaling Y_t by $\frac{1}{\theta_2}$ and adding the mean estimate $\hat{\mu}_{A_t}$ to get

$$\tilde{X}_{t,A_t} = \sqrt{\frac{2}{P}} Y_t + \hat{\mu}_{A_t}.$$

UE-UCB along with its interaction with CAS is described in Algorithm 2.

Remark 5. We will soon show in the proof of Theorem 4.2 that $\mathbb{E}\left[\left(X_{t,A_t} - \hat{\mu}_{A_t}\right)^2\right] \leq 2$. Thus our choice of θ_2 ensures that the power constraint is satisfied. Moreover, for $B \geq 2$, θ_2 is greater than θ_1, which essentially means that the client can scale the centered reward in the second phase by a bigger factor as compared to the scaling factor used in Algorithm 1. In both algorithms, since the learner multiplies Y_t by the inverse of the scaling factor used at the client, the greater value of scaling factor results in smaller variances for the decoded rewards than in Algorithm 1.

[9]For simplicity of notation, we assume that $\frac{B^2}{\texttt{SNR}}$, $\frac{1}{\texttt{SNR}}$, and $\log B$ are integers throughout the paper. If not, the rounding only slightly impacts the constant factors in our results.

Parameters: $\theta_1 = \frac{\sqrt{P}}{B}$, $\tau = \frac{B^2}{\texttt{SNR}} + 1$

Phase 1

for $k \in [K]$ **do**
  for $t \in \{(k-1)\tau + 1, \ldots, K\tau\}$ **do**
    At Learner
      $A_t = k$ and $S_t = 0$
      Send A_t, S_t to Client
    At Client
      Play arm A_t and observe X_{t,A_t}
      Send $C_{\texttt{CAS}}(X_{t,A_t}, 0; \theta_1)$ over the AWGN channel
    At Learner
      Observe AWGN Output $Y_t = C_{\texttt{CAS}}(X_{t,A_t}, 0; \theta_1) + Z_t$
      Use $\tilde{X}_{t,A_t} = \frac{Y_t}{\theta_1}$ as the tth round reward
  At Learner
    Form a mean estimate for arm k as

$$\hat{\mu}_k = \frac{1}{\tau} \cdot \sum_{t=(k-1)\tau+1}^{k\tau} \tilde{X}_{t,k}$$

Parameters: $B_2^2 = 2$, $\eta_2 = \frac{2}{\texttt{SNR}} + 1$, $\theta_2 = \sqrt{\frac{P}{2}}$

Phase 2

for $t \in \{K\tau + 1, \ldots, T\}$ **do**
  At Learner
    $A_t = \arg\max_{k \in K} \texttt{UCB}_k(t-1; \eta_2)$ and $S_t = \hat{\mu}_{A_t}$
    Send A_t, S_t to Client
  At Client
    Play arm A_t and observe X_{t,A_t}
    Send $C_{\texttt{CAS}}(X_{t,A_t}, S_t; \theta_2)$ over the AWGN channel
  At Learner
    Observe AWGN Output $Y_t = C_{\texttt{CAS}}(X_{t,A_t}, S_t; \theta_2) + Z_t$
    Use $\tilde{X}_{t,A_t} = \frac{Y_t}{\theta_2} + S_t$ as the tth round reward
    Update UCB index for all arms

Figure 2. UE-UCB Algorithm along with the CAS encoding algorithm

Theorem 4.2. *For all bandit instances $\nu \in \mathcal{E}$, the* UE-UCB *algorithm along with* CAS *encoding as described in Algorithm (2) satisfies the power constraint (1) at all rounds and has the following regret guarantee:*

$$R_T(\nu, \texttt{UE-UCB}, \texttt{CAS}) \leq \frac{2KB^3}{\texttt{SNR}} + 8KB + 8\sqrt{\frac{2}{\texttt{SNR}} + 1}\sqrt{KT\ln T}.$$

Proof. In the first phase, we incur regret linear regret in the time horizon. Specifically, since the time horizon for the first phase is $K\tau = K(B^2/\texttt{SNR} + 1)$ and the gap between the mean of the optimal arm and the suboptimal arm can at most be $2B$, we have that the regret incurred in the first phase is $(KB^2/\texttt{SNR} + K) \cdot 2B$. Moreover, since the encoding of rewards on the client's side and decoding of the

reward on the learner's side in the first phase is same as in Algorithm 1, by the same arguments, the power constraint is satisfied, and for all $t \in \{(k-1)\tau+1, \ldots, k\tau\}$,

$$\mathbb{E}\left[(\tilde{X}_{t,k} - \mu_k)^2\right] \leq \frac{B^2}{\texttt{SNR}} + 1. \tag{5}$$

For the second phase, we will first show that the power constraint is satisfied. Towards proving this, we will bound the variance of the mean estimates formed after the first phase. We have from the description of Algorithm 2 that for all $k \in K$,

$$\mathbb{E}\left[(\hat{\mu}_k - \mu_k)^2\right] = \frac{1}{\tau^2} \sum_{t=(k-1)\tau+1}^{k\tau} \mathbb{E}\left[(\tilde{X}_{t,k} - \mu)^2\right] \leq 1,$$

where the inequality follows from (5) and the fact that $\tau = \frac{B^2}{\texttt{SNR}} + 1$.

Now, note that for any round t in phase 2, we have

$$\begin{aligned}&\mathbb{E}\left[(X_{t,A_t} - \hat{\mu}_{A_t})^2\right] \\ &= \mathbb{E}\left[(X_{t,A_t} - \mu_{A_t})^2\right] + \mathbb{E}\left[(\hat{\mu}_{A_t} - \mu_{A_t})^2\right] \leq 2,\end{aligned}$$

where the equality follows from the independence of X_{t,A_t} and $\hat{\mu}_{A_t}$, and the inequality follows from the fact that both $(X_{t,A_t} - \mu_{A_t})$ and $(\hat{\mu}_{A_t} - \mu_{A_t})$ are subgaussian with a variance factor of 1.

Therefore, scaling the centered reward $X_{t,k} - \hat{\mu}_k$ by $\frac{\sqrt{P}}{\sqrt{2}}$ satisfies the power constraint.

The reward estimate used by the learner in phase 2 is

$$\begin{aligned}\tilde{X}_{t,A_t} &= \frac{\sqrt{2}Y_t}{\sqrt{P}} + \hat{\mu}_{A_t} \\ &= X_{t,A_t} + \frac{\sqrt{2}Z_t}{\sqrt{P}}.\end{aligned}$$

This implies that the centered reward estimate, $\tilde{X}_{t,A_t} - \mu_{A_t}$, is subgaussian with a variance factor of $2/\texttt{SNR} + 1$ in the second phase. The proof is complete by upper bounding the regret in the second phase by using Lemma 2.2 and summing the regret in the two phases. □

4.4. UE-UCB++ bandit algorithm

Algorithm 2 improves over Algorithm 1 in the sense that the T-dependent term in its regret upper bound does not have the multiplicative $\mathcal{O}(B)$ dependence. This term matches the lower bound in Theorem 3.1 up to logarithmic factors in T. However, this comes at the cost of an additional additive term of $\mathcal{O}(\frac{KB^3}{\texttt{SNR}})$ in the regret bound of Algorithm 2. The strong dependence on B in this additive term means that for large values of B, Algorithm 2 can be far from optimal. In this section, we present our main scheme, which improves over this shortcoming of Algorithm 2, and is only off by a minor additive term from our lower bound. Our scheme will combine an improved version of the UE-UCB algorithm, UE-UCB++, and the CAS encoding algorithm.

Parameters: $L = 2\log B$, $\tau = \frac{2}{\texttt{SNR}}$, $B_1^2 = B^2$

for $\ell \in [L]$ **do**
  $\theta_\ell = \frac{\sqrt{P}}{B_\ell}$
  $B_{\ell+1}^2 = \left(\frac{\frac{B_\ell^2}{\texttt{SNR}}+1}{\tau} + 1\right)$

for $k \in [K]$ **do** $\hat{\mu}_{0,k} = 0$

Phase 1

for $\ell \in [L]$ **do**
  for $k \in [K]$ **do**
    for $t \in \{(\ell-1)K\tau + (k-1)\tau + 1, \ldots, (\ell-1)K\tau + k\tau\}$ **do**
      At Learner
        $A_t = k$ and $S_t = \hat{\mu}_{\ell-1,k}$
        A_t, S_t to the client
      At Client
        Play arm A_t and observe X_{t,A_t}
        Send $C_{\texttt{CAS}}(X_{t,A_t}, S_t; \theta_\ell)$ over the AWGN channel
      At Learner
        Observe AWGN Output $Y_t = C_{\texttt{CAS}}(X_{t,A_t}, S_t; \theta_\ell) + Z_t$
        Use $\tilde{X}_{t,A_t} = \frac{Y_t}{\theta_t} + S_t$ as the tth round reward
    At Learner
      Form a mean estimate for arm k

$$\hat{\mu}_{\ell,k} = \frac{1}{\tau} \cdot \sum_{t=(\ell-1)K\tau+(k-1)\tau+1}^{(\ell-1)K\tau+k\tau} \tilde{X}_{t,k}$$

Parameters: $\theta_{L+1} = \frac{\sqrt{P}}{B_{L+1}}$, $\eta_{L+1} = \frac{B_{L+1}^2}{\texttt{SNR}} + 1$

Phase 2

for $t \in \{K\tau + 1, \ldots, T\}$ **do**
  At Learner
    $A_t = \arg\max_{k \in K} \texttt{UCB}_k(t-1; \eta_{L+1})$ and $S_t = \hat{\mu}_{L,A_t}$
    Send A_t, S_t to Client
  At Client
    Play arm A_t and observe X_{t,A_t}
    Send $C_{\texttt{CAS}}(X_{t,A_t}, S_t; \theta_{L+1})$ over the AWGN channel
  At Learner
    Observe AWGN Output $Y_t = C_{\texttt{CAS}}(X_{t,A_t}, S_t; \theta_{L+1}) + Z_t$
    Use $\tilde{X}_{t,A_t} = \frac{Y_t}{\theta_{L+1}} + S_t$ as the tth round reward
    Update UCB index for all arms

Figure 3. UE-UCB++ Algorithm along with the CAS encoding algorithm

Like UE-UCB, UE-UCB++ is also a two-phase bandit algorithm, that performs uniform exploration in the first phase and runs the UCB algorithm in the second phase. The second phase of the algorithm proceeds in the same manner as that of the UE-UCB algorithm, and the improvement over UE-UCB is in the phase 1 of the algorithm.

Phase 1: UE-UCB++, like UE-UCB, does uniform exploration in all rounds of phase 1. However, phase 1 is divided

into L different sub-phases, where the encoding of rewards is different in each sub-phase. The mean estimates formed in the previous sub-phase are used as side information in the current sub-phase and sent to the client for better encoding.

In more detail, we have $L = 2\log B$ sub-phases in total and each sub-phase runs for $K\tau$ rounds, where

$$\tau = \frac{2}{\texttt{SNR}} \vee 2. \tag{6}$$

Thus, phase 1 runs for $L \cdot K\tau \leq \frac{4K\log B}{\texttt{SNR}}$ rounds.

Remark 6. Phase 1 in UE-UCB++ runs for $\mathcal{O}\big(\frac{K\log B}{\texttt{SNR}}\big)$ rounds as opposed to $\mathcal{O}\big(\frac{KB^2}{\texttt{SNR}}\big)$ rounds used in the phase 1 of UE-UCB. We will show that despite performing uniform exploration for significantly fewer rounds for all arms, we will end up with mean estimates with the same variance as the means estimates formed after phase 1 of UE-UCB because of a superior encoding protocol in this phase.

The first sub-phase proceeds in exactly the same manner as phase 1 of Algorithm 2. To describe the algorithm in the subsequent sub-phases, we set up some notation: For all $k \in [K]$, denote by $\hat{\mu}_{\ell,k}$ the mean estimates formed for all arms at the end of the ℓth sub-phase. For a round t in ℓth sub-phase, the learner informs the client about the arm A_t it wants to play and the mean estimate formed for this arm in the $(\ell-1)$th sub-phase, $\hat{\mu}_{\ell-1,A_t}$.

On the client's side, the reward corresponding to A_t is encoded by using CAS with parameter $\theta_\ell = \frac{\sqrt{P}}{B_\ell}$ to get

$$C_{\texttt{CAS}}(X_{t,A_t}, \mu_{A_t}; \theta_\ell) := \frac{\sqrt{P}}{B_\ell}(X_{t,A_t} - \hat{\mu}_{\ell-1,A_t}).$$

Here B_ℓ, $\ell \in [L]$, will be shown to be an upper bound on the L_2-norm of random variable $X_{t,A_t} - \hat{\mu}_{\ell-1,A_t}$ (see the proof of Theorem 4.3) and is described by the recursion

$$B_{\ell+1}^2 = \frac{\frac{B_\ell^2}{\texttt{SNR}} + 1}{\tau} + 1 \quad \text{and} \quad B_1^2 = B^2. \tag{7}$$

The decoding at the learner is simply done by scaling Y_t by $\frac{1}{\theta_\ell}$ and adding the mean estimate $\hat{\mu}_{\ell,A_t}$ to get

$$\tilde{X}_{t,A_t} = \frac{B_\ell}{\sqrt{P}} Y_t + \hat{\mu}_{A_t}.$$

Phase 2: Phase 2 proceeds in the same manner as the phase 2 of Algorithm 2, with θ_{L+1} used for encoding at CAS and the parameter η_{L+1} of the UCB algorithm set to $\frac{B_{L+1}^2}{\texttt{SNR}} + 1$.

Remark 7. Equation (7) sheds some light on the reasons behind UE-UCB++ requiring fewer uniform exploration rounds relative to UE-UCB to come up with mean estimates of similar variances. Essentially, the centered reward in the ℓth sub-phase has a much smaller L_2-norm than the centered reward in the $(\ell-1)$th sub-phase. This smaller second moment allows the client to scale the centered reward by a larger factor, leading to smaller variances for the decoded reward. This, in turn, leads to mean estimates formed in that sub-phase to have smaller variances or, in other words, are more accurate. Then, in the $(\ell+1)$th sub-phase, since we center the rewards around these more accurate mean estimates, the centered rewards have even smaller L_2 norms. This positive reinforcement cycle leads to the variance of the mean estimates decreasing exponentially with respect to sub-phases, which, in turn, leads to much fewer uniform exploration rounds to come up with mean estimates with the same variance as those after phase 1 of UE-UCB.

UE-UCB++, along with its interaction with CAS, is described in Algorithm 3.

Theorem 4.3. *For all bandit instances $\nu \in \mathcal{E}$, the* UE-UCB++ *bandit algorithm along with* CAS *encoding algorithm as described in Algorithm* (3) *satisfies the power constraint* (1) *at all rounds and has the following regret guarantee:*

$$R_T(\nu, \texttt{UE-UCB++}, \texttt{CAS}) \leq \frac{8KB\log B}{\texttt{SNR} \wedge 1} + 6KB + 8\sqrt{\left(\frac{4}{\texttt{SNR}} + 1\right)} \cdot \sqrt{KT\ln T}.$$

The key to the proof of Theorem 4.3 is to show that the second moment of the centered rewards $X_{t,A_t} - \hat{\mu}_{\ell-1,A_t}$ is less than B_ℓ, $\ell \in [L]$, when round t belongs to the ℓth sub-phase of phase 1, and when round t belongs to phase 2, it is less than B_{L+1}. Similar arguments as in the proof of Theorem 4.2 can show this. We defer the details to the Appendix.

Remark 8. The $\ln T$ factor in the third term of Theorem 4.3 can be removed by using MOSS (Audibert & Bubeck, 2009) instead of the UCB algorithm. Our work focuses on understanding how the regret grows as a function of SNR; we are not as concerned with logarithmic factors in T. Hence, we chose the more popular UCB algorithm.

5. Concluding Remarks

The near-optimal Algorithm 3 use a highly adaptive encoding protocol, where the history of the game was used to refine the encoding in a particular round. On the other hand, Acharya et al. (2021) showed for distributed convex, Lipschitz optimization under information constraints, a fixed, non-adaptive encoding protocol gives optimal performance. This suggests an interesting distinction between encoding for distributed sequential decision-making problems and distributed optimization problems.

An interesting open problem is closing the additive $\frac{8KB\log B}{\texttt{SNR}\wedge 1}$ gap between our upper and lower bound. Furthermore, the refinement to our model where the learner-to-client communication is also over an AWGN channel and the multi-client setting are interesting future research directions.

Acknowledgement

This work is supported by the Singapore National Research Foundation (NRF) Fellowship programme (Grant Numbers: A-0005077-01-00 and A-0008064-00-00) and an MOE Tier 1 Grant (Grant Number: A-0009042-01-00).

References

Acharya, J., Canonne, C., Mayekar, P., and Tyagi, H. Information-constrained optimization: can adaptive processing of gradients help? *Advances in Neural Information Processing Systems*, 34:7126–7138, 2021.

Amiri, M. M. and Gündüz, D. Machine learning at the wireless edge: Distributed stochastic gradient descent over-the-air. *IEEE Transactions on Signal Processing*, 68:2155–2169, 2020.

Audibert, J.-Y. and Bubeck, S. Minimax policies for adversarial and stochastic bandits. In *COLT*, volume 7, pp. 1–122, 2009.

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

Boucheron, S., Lugosi, G., and Massart, P. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Cover, T. M. and Thomas, J. A. *Elements of Information Theory. 2nd edition*. John Wiley & Sons Inc., 2006.

Csiszár, I., Shields, P. C., et al. Information theory and statistics: A tutorial. *Foundations and Trends® in Communications and Information Theory*, 1(4):417–528, 2004.

Dragomir, S. Upper and lower bounds for csiszár's f-divergence in terms of the kullback-leibler distance and applications. *Inequalities for the Csiszár's*, 2000.

Dwork, C., Roth, A., et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Gandikota, V., Kane, D., Maity, R. K., and Mazumdar, A. vqsgd: Vector quantized stochastic gradient descent. *ICML Workshop on Security and Privacy in Machine Learning)*, 2019.

Hanna, O. A., Yang, L., and Fragouli, C. Solving multi-arm bandit using a few bits of communication. In *International Conference on Artificial Intelligence and Statistics*, pp. 11215–11236. PMLR, 2022.

Jha, S. K., Mayekar, P., and Tyagi, H. Fundamental limits of over-the-air optimization: Are analog schemes optimal? *IEEE Journal on Selected Areas in Information Theory*, 3(2):217–228, 2022.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

Konečný, J., McMahan, H. B., Yu, F. X., Richtarik, P., Suresh, A. T., and Bacon, D. Federated learning: Strategies for improving communication efficiency. *NIPS Workshop on Private Multi-Party Machine Learning*, 2016.

Lattimore, T. and Szepesvári, C. *Bandit Algorithms*. Cambridge University Press, 2020.

Lin, C.-Y., Kostina, V., and Hassibi, B. Differentially quantized gradient descent. In *IEEE International Symposium on Information Theory (ISIT)*, pp. 1200–1205, 2021.

Mayekar, P. and Tyagi, H. RATQ: A universal fixed-length quantizer for stochastic optimization. *IEEE Transactions on Information Theory*, 2020.

Mitra, A., Jaafar, R., Pappas, G. J., and Hassani, H. Linear convergence in federated learning: Tackling client heterogeneity and sparse gradients. *Advances in Neural Information Processing Systems*, 34:14606–14619, 2021.

Mitra, A., Hassani, H., and Pappas, G. J. Linear stochastic bandits over a bit-constrained channel. *arXiv preprint arXiv:2203.01198*, 2022.

Shannon, C. E. A mathematical theory of communication. *The Bell System Technical Journal*, 27(4):623–656, 1948. doi: 10.1002/j.1538-7305.1948.tb00917.x.

Tse, D. and Viswanath, P. *Fundamentals of wireless communication*. Cambridge University Press, 2005.

A. Proof of Theorem 4.3 (Regret bound of UE-UCB++)

We will first show that the centered rewards transmitted by the client satisfy the power constraint. Note that the centered reward coded by the client at any round in the ℓth sub-phase of phase 1 is $X_{t,A_t} - \hat{\mu}_{\ell-1,A_t}$. Furthermore, the client scales the reward by a factor of $\frac{\sqrt{P}}{B_\ell}$. Thus, we need to show that $\mathbb{E}\left[(X_{t,A_t} - \hat{\mu}_{\ell-1,A_t})^2\right] \leq B_\ell^2$. Towards proving this, we write

$$\begin{aligned}
&\mathbb{E}\left[(X_{t,A_t} - \hat{\mu}_{\ell-1,A_t})^2\right] \\
&= \mathbb{E}\left[(X_{t,A_t} - \mu_{A_t})^2\right] + \mathbb{E}\left[(\hat{\mu}_{\ell-1,A_t} - \mu_{A_t})^2\right] \\
&\leq 1 + \mathbb{E}\left[(\hat{\mu}_{\ell-1,A_t} - \mu_{A_t})^2\right].
\end{aligned}$$

From the description of Algorithm 3, for $k \in [K]$, we have

$$\hat{\mu}_{\ell-1,k} = \frac{1}{\tau} \sum_{t=(\ell-2)K\tau+(k-1)\tau+1}^{(\ell-2)K\tau+k\tau} \left(X_{t,k} + \frac{B_{\ell-1}}{\sqrt{P}} Z_t \right).$$

Therefore, we have

$$\mathbb{E}\left[(\hat{\mu}_{\ell,A_t} - \mu_{A_t})^2\right] \leq \frac{\frac{B_{\ell-1}^2}{\texttt{SNR}} + 1}{\tau}.$$

Using the preceding inequalities and definition of B_ℓ in (7), we get

$$\mathbb{E}\left[(X_{t,A_t} - \mu_{\ell-1,A_t})^2\right] \leq 1 + \frac{\frac{B_{\ell-1}^2}{\texttt{SNR}} + 1}{\tau} = B_\ell^2.$$

To prove that the power constraint is satisfied in the second phase, note that the centered reward transmitted by the client in the second phase is $X_{t,A_t} - \hat{\mu}_{L,A_t}$. Furthermore, client scales the reward by a factor of $\frac{\sqrt{P}}{B_{L+1}}$. By precisely the same arguments as above, we can deduce that $\mathbb{E}\left[(X_{t,A_t} - \hat{\mu}_{L,A_t})^2\right] \leq B_{L+1}^2$.

To bound the regret, first note that we incur linear regret in phase 1 of UE-UCB++. Specifically, since the time horizon for phase 1 is $\frac{4K\log B}{\texttt{SNR}\wedge 1}$ and each round's regret is at the most $2B$, we incur a total regret of $\frac{8KB\log B}{\texttt{SNR}\wedge 1}$ in the first phase.

To bound the regret in the second phase, note that the centered reward estimate $\tilde{X}_{t,A_t} - \mu_{A_t}$ in the second phase is subgaussian with a variance factor of $\frac{B_{L+1}^2}{\texttt{SNR}} + 1$. Substituting the value of τ from (6) in (7), we obtain

$$\begin{aligned}
B_{\ell+1}^2 &= \frac{\frac{B_\ell^2}{\texttt{SNR}} + 1}{\tau} + 1 \\
&\leq \frac{B_\ell^2}{2} + \frac{\texttt{SNR}\wedge 1}{2} + 1.
\end{aligned}$$

Rearranging the terms, we have

$$B_{\ell+1}^2 - 2\left(\frac{\texttt{SNR}\wedge 1}{2} + 1\right) \leq \frac{B_\ell^2 - 2\left(\frac{\texttt{SNR}\wedge 1}{2} + 1\right)}{2}.$$

Recursively applying the above identity, it follows that

$$B_{L+1}^2 - 2\left(\frac{\texttt{SNR}\wedge 1}{2} + 1\right) \leq \frac{B^2 - 2\left(\frac{\texttt{SNR}\wedge 1}{2} + 1\right)}{2^L}.$$

Substituting $L = 2\log B$, we get B_{L+1}^2 less than 4.

The proof is then completed by using Lemma 2.2 to bound the regret in the second phase and adding the regret from the first phase.

B. Proof of Theorem 3.1 (Lower Bound)

We will first show that for $T \geq c_2 K$ and a universal constant $c_1 \in (0,1)$, we have

$$R_T^*(\texttt{SNR}) \geq c_1 K B.$$

From the definition of $R_T^*(\texttt{SNR})$ in (4), it is enough to prove a lower bound on the average regret when the problem instances are uniformly chosen from a subset of $\mathcal{E}$. We choose that finite subset to be the instances where a single arm has deterministic reward B, and the others deterministic $-B$. Therefore, we have K different instances with the probability of each instance occurring being $1/K$. Now, observe that regardless of the algorithm, the probability of finding the B-reward arm is at the most $\frac{1}{2}$ in the first $\lfloor K/2 \rfloor$ arm pulls. Hence the regret in the first $\lfloor K/2 \rfloor$ arm pulls when averaged over the K instances must be at least $\frac{1}{2} \cdot \lfloor K/2 \rfloor \cdot 2B = c_1 K B$.

In the rest of proof we will show that for $T \geq \frac{c_2 K}{\frac{1}{2}\log(1+\texttt{SNR})\wedge 1}$, we have

$$R_T^*(\texttt{SNR}) \geq c_1 \left(\sqrt{KT} \cdot \frac{1}{\sqrt{\texttt{SNR} \wedge 1}} \right). \tag{8}$$

Then, noting that both the bounds hold for $T \geq \frac{c_2 K}{\frac{1}{2}\log(1+\texttt{SNR})\wedge 1}$ and using the fact that $a \vee b \geq \frac{a+b}{2}$ completes the proof.

Consider the bandit instance ν where rewards for each arm pull are either -1 or 1 and the arms have the following means: $\mu_1 = \Delta$, and $\mu_i = 0$ for all $i \in \{2, \ldots, K\}$, where Δ is some parameter in $(0, 1/4)$ whose precise value we will set later. Using the fact that rewards for all arms lie in $[-1, 1]$ and by Hoeffding's lemma (Boucheron et al., 2013, Lemma 2.2), it is easy to see that the centered rewards for all arms are subgaussian with a variance factor of 1. Therefore, it follows that $\nu \in \mathcal{E}$. Now, fix a bandit algorithm $\pi \in \Pi_T$ and an encoding scheme $C^T \in \mathcal{C}^T$. Denote by P^T the probability distribution of the sequence $(A_1, Y_1, A_2, Y_2, \ldots, A_T, Y_T)$ induced by the algorithm π and encoding strategy C^T on the instance ν.

Denote by $N_i(T)$ the number of times arm i has been played up to and including round T. Let

$$j = \arg\min_{i>1} \mathbb{E}_{P^T}\left[N_i(T)\right].$$

That is, j is the arm pulled least amount of times from the arms $\{2, \ldots, K\}$. Trivially,

$$\mathbb{E}_{P^T}\left[N_i(T)\right] \leq \frac{T}{K-1}. \tag{9}$$

Now consider another bandit instance ν' where again the rewards are either -1 or 1 with the following means: $\mu_1 = \Delta$, $\mu_j = 2\Delta$, $\mu_i = 0$ for all $i \in \{2, \ldots, K\} \setminus \{j\}$. Let Q^T be the the probability distribution of the sequence $(A_1, Y_1, A_2, Y_2, \ldots, A_T, Y_T)$ induced by the algorithm π and encoding strategy C on the instance ν'.

By a simple calculation, we have

$$R_T(\nu, \pi, C) \geq \frac{T\Delta}{2} P^T(N_1(T) \leq T/2) \quad \text{and} \quad R_T(\nu', \pi, C) \geq \frac{T\Delta}{2} Q^T(N_1(T) > T/2).$$

Combining the two inequalities, using the variational formula for total variation distance, and applying Pinsker's inequality, we get

$$\begin{aligned}
R_T(\nu, \pi, C) + R_T(\nu', \pi, C) &\geq \frac{T\Delta}{2}\left(P^T(N_1(T) \leq T/2) + Q^T(N_1(T) > T/2)\right) \\
&= \frac{T\Delta}{2}\left(1 - (P^T(N_1(T) > T/2) - Q^T(N_1(T) > T/2))\right) \\
&\geq \frac{T\Delta}{2}\left(1 - \mathrm{d}_{\mathrm{TV}}(P^T, Q^T)\right) \\
&\geq \frac{T\Delta}{2}\left(1 - \sqrt{\frac{1}{2} D(P^T \,||\, Q^T)}\right).
\end{aligned} \tag{10}$$

Note that P^T and Q^T differ only when arm j is played. Recall that the dependence of Y on an encoding strategy C and reward X is given by (2). For encoding scheme C, used in a round t, denote by $C(\cdot \mid x)$ the conditional distribution of Y given that the reward is $X = x$. Let $p_j \circ C$ and $q_j \circ C$ denote the distribution of Y when arm j is played at round t. Recalling the reward distribution for arm j, it follows that,

$$(p_j \circ C)(Y = y) = \frac{C(y \mid 1)}{2} + \frac{C(y \mid -1)}{2} \quad \text{and}$$
$$(q_j \circ C)(Y = y) = \frac{C(y \mid 1)}{2} + \frac{C(y \mid -1)}{2} + \Delta(C(y \mid 1) - C(y \mid -1)).$$

Moreover, using similar arguments as in Lattimore & Szepesvári (2020, Lemma 15.1), we can prove the following lemma.

Lemma B.1.

$$D(P^T \mid\mid Q^T) \leq \mathbb{E}_{P^T}\left[N_j(T)\right] \cdot \sup_{C \in \mathcal{C}} D(p_j \circ C \mid\mid q_j \circ C)$$

We defer the proof the Lemma to Appendix C and proceed with proving the lower bound. Combining Lemma B.1 with (9), we have

$$D(P^T \mid\mid Q^T) \leq \frac{T}{K-1} \sup_{C \in \mathcal{C}} D(p_j \circ C \mid\mid q_j \circ C).$$

We will now use the following fact.

Fact 1. Recall that $d_{\chi^2}(p, q) := \int \frac{(p(x)-q(x))^2}{q(x)} \mathrm{d}x$ denotes the χ^2 distance between probability measures p and q when they are absolutely continuous with respect to the Lebesgue measure $\mathrm{d}x$. Then, it holds that

$$D(p \mid\mid q) \leq d_{\chi^2}(p, q).$$

Fact 1 is well known (see, for instance Csiszár et al. (2004)]), but we provide a short proof in Appendix C for completeness.

Using this fact, $D(p_j \circ C \mid\mid q_j \circ C)$ can be bounded as follows:

$$\begin{aligned}
\mathrm{D}(p_i \circ C \mid\mid q_i \circ C) &\leq d_{\chi^2}(p_i \circ C, q_i \circ C) \\
&= \int \frac{(p_i \circ C - q_i \circ C)^2}{q_i \circ C} \mathrm{d}y \\
&= \Delta^2 \int \frac{(C(y \mid 1) - C(y \mid -1))^2}{\frac{C(y|1)}{2} + \frac{C(y|-1)}{2} + \Delta(C(y \mid 1) - C(y \mid -1))} \mathrm{d}y
\end{aligned}$$

Now note that for $\Delta \leq 1/4$, we have $\frac{C(y|1)}{2} + \frac{C(y|-1)}{2} + \Delta(C(y \mid 1) - C(y \mid -1)) \geq \frac{C(y|1)}{4} + \frac{C(y|-1)}{4}$. Using this fact, we get

$$\begin{aligned}
\mathrm{D}(p_i \circ C \mid\mid q_i \circ C) &\leq 2\Delta^2 \int \frac{(C(y \mid 1) - C(y \mid -1))^2}{\frac{C(y|1)}{2} + \frac{C(y|-1)}{2}} \mathrm{d}y \\
&= 2\Delta^2 \Bigg(\frac{1}{2} d_{\chi^2}(C(\cdot \mid 1), C(\cdot \mid 1)/2 + C(\cdot \mid -1)/2) \\
&\qquad + \frac{1}{2} d_{\chi^2}(C(\cdot \mid -1), C(\cdot \mid 1)/2 + C(\cdot \mid -1)/2) \Bigg) \\
&\leq 4\Delta^2 \Bigg(\frac{1}{2} \mathrm{D}(C(\cdot \mid 1) \mid\mid C(\cdot \mid 1)/2 + C(\cdot \mid -1)/2) \\
&\qquad + \frac{1}{2} \mathrm{D}(C(\cdot \mid -1) \mid\mid C(\cdot \mid 1)/2 + C(\cdot \mid -1)/2) \Bigg),
\end{aligned}$$

where the equality is easily verified by substituting the definition d_{χ^2} and the last inequality uses the following fact.

To proceed further, we will use the following fact.

Fact 2. If $\frac{\mathrm{d}P}{\mathrm{d}Q}(x) \le c \quad \forall x \in \mathbb{R}$, then

$$d_{\chi^2}(P, Q) \le cD(P \,||\, Q).$$

Fact 2 is also known (Dragomir, 2000), but we provide a proof in Appendix C for completeness.

Denote by V a hypothetical reward that takes values -1 and 1 with equal probability, and by Y the output of the AWGN channel when V is further coded to $C(V)$ to satisfy (1) and transmitted over the AWGN channel. Then,

$$\frac{1}{2}\mathrm{D}(C(\cdot \mid 1) \,||\, C(\cdot \mid 1)/2 + C(\cdot \mid -1)/2) + \frac{1}{2}\mathrm{D}(C(\cdot \mid -1) \,||\, C(\cdot \mid 1)/2 + C(\cdot \mid -1)/2)$$

is mutual information between V and Y, denoted by $I(V;Y)$. To see this, note that $I(W_1;W_2) \coloneqq D(P_{W_1,W_2} \,||\, P_{W_1} \times P_{W_2}) = \mathbb{E}_{P_{W_1}}\left[D(P_{W_2|W_1} \,||\, P_{W_2})\right]$.

We have therefore shown that

$$D(P^T \,||\, Q^T) \le \frac{4\Delta^2 T}{K-1} \sup_{C \in \mathcal{C}} I(V;Y). \tag{11}$$

Note that $I(V;Y) \le H(V) \le \ln 2$. Since $V - C(V) - Y$ forms a Markov chain in this order, by the data processing inequality (Cover & Thomas, 2006, Chapter 2), we have that

$$I(V;Y) \le I(C(V);Y).$$

Then, using the fact that $I(C(V);Y)$ can be further upper bounded by the capacity of an AWGN channel (Cover & Thomas, 2006, Chapter 9), we have

$$I(C(V);Y) \le \frac{1}{2}\ln(1+\texttt{SNR}) = \frac{1}{2}\log(1+\texttt{SNR}) \cdot \ln 2.$$

Thus, we have shown that

$$D(P^T \,||\, Q^T) \le \frac{4\ln 2 \cdot \Delta^2 T}{K-1}\left(\frac{1}{2}\log(1+\texttt{SNR}) \wedge 1\right).$$

Substituting this into (10), we get

$$R_T(\nu,\pi,C) + R_T(\nu',\pi,C) \ge \frac{T\Delta}{2}\left(1 - \sqrt{\frac{4\ln 2 \cdot \Delta^2 T}{K-1}\left(\frac{1}{2}\log(1+\texttt{SNR}) \wedge 1\right)}\right).$$

Setting $\Delta^2 = \frac{K-1}{16\ln 2 \cdot T\left(\frac{1}{2}\log(1+\texttt{SNR})\wedge 1\right)}$ completes the proof. (Note that we can only set such a Δ and ensure that $\Delta \le \frac{1}{4}$, if $T \ge \frac{K-1}{\ln 2\left(\frac{1}{2}\log(1+\texttt{SNR})\wedge 1\right)}$, which is the reason we need this condition in the statement of our lower bound.)

We will now show that for $T \ge \frac{K-1}{\ln 2\left(\frac{1}{2}\log(1+\texttt{SNR})\wedge 1\right)} \ge K$ and for a universal constant c, we have

$$R_T^*(\texttt{SNR}) \ge cKB.$$

In fact, we show the stronger bound

$$R_K^*(\texttt{SNR}) \ge cKB.$$

Then, using the fact that $a \vee b \ge \frac{a+b}{2}$ completes the proof.

We now choose our instance ν where rewards for each arm pull are Gaussian with variance 1 and the following means: $\mu_1 = 1, \mu_i = 0$ for all $i \in \{2,\dots,K\}$. Once again, let j be the arm pulled least amount of times. Then define the alternate instance ν' where rewards for each arm pull are Gaussian with variance 1 and the following means: $\mu_1 = B/2, \mu_j = B-1, \mu_i = 0$ for all $i \in \{2,\dots,K\}/j$. All the arm rewards have Gaussian distribution with the variance 1 and means stated above. Proceeding as earlier, we have

$$R_T(\nu,\pi,C) + R_T(\nu',\pi,C) \ge \frac{T\Delta}{2}\left(1 - \mathrm{d}_{\mathrm{TV}}(P^T, Q^T)\right)$$

Using the fact that $T \geq \frac{K-1}{\ln 2\left(\frac{1}{2}\log(1+\mathtt{SNR})\wedge 1\right)} \geq c(K-1)$ and the fact that for our chosen distributions $1 - \mathrm{d}_{\mathrm{TV}}(P^T, Q^T)$ is a constant completes the proof.

Remark 9. (*Communication-constrained setting of Hanna et al. (2022)*) We now remark on possibility of recovering the lower bound in Hanna et al. (2022) with our proof technique, albeit with weaker constants. Note that (10) also holds for the setting considered in Hanna et al. (2022), where the output Y is restricted to r bits. Therefore, $I(V;Y) \leq H(Y) \wedge H(V) = (\ln 2)(r \wedge 1)$. Proceeding, in the rest of the proof, *mutatis mutandis*, we get that the regret in this setting is bounded by $\Omega(\sqrt{KT} \cdot \frac{1}{r\wedge 1})$. Such a lower bound suggests that even when using code of length 1 to encode our rewards, it might be possible to achieve the regret of the centralized setting. This is indeed shown to be possible by the scheme in Hanna et al. (2022).

C. Proofs of Intermediary Results in the Proof of Theorem 3.1

Proof of Lemma B.1

We prove the inequality by similar arguments as in Lattimore & Szepesvári (2020, Lemma 15.1).

Recall that $H_t = \{A_1, Y_1, \ldots, A_t, Y_t\}$ denotes the history of the game till time t. Also recall that the hint S_t is a function of history of the game till time $t-1$. Therefore the encoding scheme at time t depends on H_{t-1}, so we write it as $C_{H_{t-1}}$. We additionally denote by p_{A_t} and q_{A_t} the distribution of reward X_{t,A_t} under instances ν and ν', respectively. Therefore, $p_{A_t} \circ C_{H_{t-1}}$ and $q_{A_t} \circ C_{H_{t-1}}$ denote distribution of Y_t for an arm pull A_t and history H_{t-1} under instances ν and ν', respectively.

Using the definition of KL divergence, we have

$$\begin{aligned} D(P^T \mid\mid Q^T) &= \mathbb{E}_{P^T}\left[\log\left(\frac{P^T(A_1, Y_1, \ldots, A_T, Y_T)}{Q^T(A_1, Y_1, \ldots, A_T, Y_T)}\right)\right] \\ &= \sum_{t=1}^{T} \mathbb{E}_{P^T}\left[\log \frac{P^T(A_t, Y_t \mid H_{t-1})}{Q^T(A_t, Y_t \mid H_{t-1})}\right]. \end{aligned} \tag{12}$$

By the definition of the bandit algorithm π, the distribution over all actions to be played remains the same for different problem instances as long as the history is the same. Therefore,

$$\begin{aligned} \frac{P^T(A_t, Y_t \mid H_{t-1})}{Q^T(A_t, Y_t \mid H_{t-1})} &= \frac{P^T(A_t \mid H_{t-1})}{Q^T(A_t \mid H_{t-1})} \cdot \frac{P^T(Y_t \mid A_t, H_{t-1})}{Q^T(Y_t \mid A_t, H_{t-1})} \\ &= \frac{P^T(Y_t \mid A_t, H_{t-1})}{Q^T(Y_t \mid A_t, H_{t-1})}. \end{aligned}$$

Substituting the above identity in (12), we have

$$\begin{aligned} D(P^T \mid\mid Q^T) &= \sum_{t=1}^{T} \mathbb{E}_{P^T}\left[\log \frac{P^T(Y_t \mid A_t, H_{t-1})}{Q^T(Y_t \mid A_t, H_{t-1})}\right] \\ &= \sum_{t=1}^{T} \mathbb{E}_{P^T}\left[\mathbb{E}_{P^T}\left[\log \frac{P^T(Y_t \mid A_t, H_{t-1})}{Q^T(Y_t \mid A_t, H_{t-1})}\middle| A_t, H_{t-1}\right]\right] \\ &= \sum_{t=1}^{T} \mathbb{E}_{P^T}\left[D(p_{A_t} \circ C_{H_{t-1}} \mid\mid q_{A_t} \circ C_{H_{t-1}})\right], \end{aligned}$$

where the second last inequality uses the law of iterated expectations and the final one uses the definition of KL divergence.

Since p_{A_t} and q_{A_t} differ only when arm $A_t = j$, we have

$$D(P^T \mid\mid Q^T) = \sum_{t=1}^{T} \mathbb{E}_{P^T}\left[\mathbf{1}_{A_t=j} D(p_j \circ C_{H_{t-1}} \mid\mid q_j \circ C_{H_{t-1}})\right]$$

$$
\begin{aligned}
&\leq \sum_{t=1}^{T} \mathbb{E}_{P^T}\left[\mathbf{1}_{A_t=j} \sup_{C\in\mathcal{C}} D(p_j \circ C \,||\, q_j \circ C)\right] \\
&\leq \mathbb{E}_{P^T}[N_j(T)] \cdot \sup_{C\in\mathcal{C}} D(p_j \circ C \,||\, q_j \circ C) \\
&\leq \frac{T}{K-1} \sup_{C\in\mathcal{C}} D(p_j \circ C \,||\, q_j \circ C).
\end{aligned}
$$

□

C.1. Proof of Fact 1

Using the fact that $\ln x \leq x - 1$ and $\int \big(p(x) - q(x)\big)\mathrm{d}x = 0$, we have

$$
\begin{aligned}
D(p \,||\, q) &= \int p(x) \ln \frac{p(x)}{q(x)} \mathrm{d}x - \int (p(x) - q(x))\,\mathrm{d}x \\
&\leq \int p(x) \left(\frac{p(x)}{q(x)} - 1\right) \mathrm{d}x - \int (p(x) - q(x))\,\mathrm{d}x \\
&= \int \frac{(p(x) - q(x))^2}{q(x)} \mathrm{d}x \\
&= d_{\chi^2}(p, q).
\end{aligned}
$$

□

C.2. Proof of Fact 2

We have

$$
\begin{aligned}
d_{\chi^2}(p, q) &= \int \frac{(p(x) - q(x))^2}{q(x)} \mathrm{d}x \\
&= \int p(x) \left(\frac{p(x)}{q(x)} - 1\right) \mathrm{d}x - \int (p(x) - q(x))\,\mathrm{d}x \\
&\leq \int p(x) \cdot \frac{p(x)}{q(x)} \cdot \ln \frac{p(x)}{q(x)} \mathrm{d}x \\
&\leq c \int p(x) \ln \frac{p(x)}{q(x)} \mathrm{d}x = cD(p \,||\, q),
\end{aligned}
$$

where the fist inequality follows by using the fact that $\frac{x-1}{x} \leq \ln x \quad \forall x \geq 0$ and $\int (p(x) - q(x))\,\mathrm{d}x = 0$ and the second inequality follows from the fact that $\frac{\mathrm{d}P}{\mathrm{d}Q}(x) \leq c \quad \forall x \in \mathbb{R}$.

□