Compression for Multi-Arm Bandits

Osama A. Hanna, Lin F. Yang, *Member, IEEE*, and Christina Fragouli, *Fellow, IEEE*

***Abstract*—The multi-armed bandit (MAB) problem is one of the most well-known active learning frameworks. The aim is to select the best among a set of actions by sequentially observing rewards that come from an unknown distribution. Recently, a number of distributed bandit applications have become popular over wireless networks, where agents geographically separated from a learner collect and communicate the observed rewards. In this paper we propose a compression scheme, that compresses the rewards collected by the distributed agents. By providing nearly matching upper and lower bounds, we tightly characterize the *number of bits* needed per reward for the learner to accurately learn without suffering additional regret. In particular, we establish a generic reward quantization algorithm, *QuBan*, that can be applied on top of any (no-regret) MAB algorithm to form a new communication-efficient counterpart. *QuBan* requires only a few (converging to as low as 3 bits as the number of iterations increases) bits to be sent per reward while preserving the *same* regret bound as uncompressed rewards. Our lower bound is established via constructing hard instances from a subGaussian distribution. Our theory is further corroborated by numerical experiments.**

***Index Terms*—Distributed multi-armed bandits, contextual bandits, compression, communication constraints.**

I. Introduction

MULTI-ARMED bandit (MAB) is an active learning framework that finds application in diverse domains, such as recommendation systems, clinical trials, and adaptive routing [2]. MAB systems in areas such as mobile healthcare, social decision-making and spectrum allocation have also already been implemented in a distributed manner, using limited bandwidth wireless links and simple sensors with low computational power [3], [4], [5], [6], [7], [8]. Motivated from such communication constrained environments, in this paper we explore compression schemes tailored to distributed MAB applications.

In the classical MAB problem formulation, a learner interacts with an environment by pulling an arm from a set of arms, each of which, if played, gives a scalar reward, sampled from an unknown but fixed distribution. The goal of the learner is to find the arm with the highest mean reward using the minimum number of pulls. The performance of a learner is measured in terms of regret, that captures the expected difference between the observed rewards and rewards drawn from the best arm. Work on MAB algorithms and their applications span several decades, cultivating a rich literature that considers a variety of models and algorithmic approaches [9]. MAB algorithms include explore-then-commit [10], [11], ϵ-greedy [12], Thomson sampling [13], and the upper confidence bound (UCB) [12], [14], to name a few. Under some assumptions on the reward distribution, the explore-then-commit and ϵ-greedy algorithms achieve a regret bound $\propto O(\sqrt{n})$, where n is number of steps the learner plays, for the worst-case but known minimum reward gap,[1] while Thomson sampling and UCB achieve a regret bound $\propto O(\sqrt{n\log(n)})$ without knowledge of the minimum means gap.[2] However, all these works assume that the rewards can be communicated to the learner at full precision which can be costly in communication-constrained setups. In this paper we ask: *is it possible to perform efficient and effective bandit learning with only a few bits communicated per reward?*

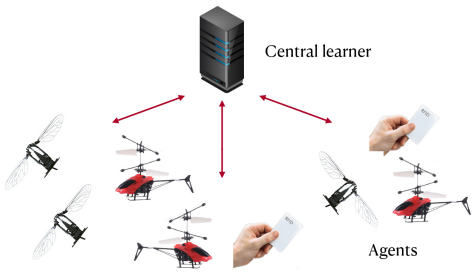


Fig. 1. A central learner collects rewards from a set of agents. The agents can join and leave at any time and hence can be different and unaware of the historical rewards, (i.e., they are memoryless).

In particular, in this paper we consider reward compression for the setup illustrated in Fig. 1, where a central learner can directly communicate with a set of agents. We assume that agents can observe and communicate rewards - but do not keep memory of past rewards. There are several use cases where this setup can apply: the agent may be mobile (e.g., the central learner is a "traffic policeman" that directs passing by small drones to perform manoeuvres and searches for best current policies); they may be low complexity sensors (e.g., swarms of tiny robots such as RoboBees and RoboFlies [17], wearable -inside and outside the body- sensors, backscatterer and RFID networks, IoT and embedded systems); or they may

Manuscript received 14 April 2022; revised 23 September 2022 and 2 February 2023; accepted 17 March 2023. Date of publication 22 March 2023; date of current version 13 June 2023. This work was supported in part by NSF under Grant 2007714 and Grant 2221871; in part by DARPA under Grant HR00112190130; and in part by the Army Research Laboratory under Cooperative Agreement W911NF-17-2-0196. This work was presented in part at the International Conference on Artificial Intelligence and Statistics (AISTATS), 2022 (Hanna, 2022). *(Corresponding author: Osama A. Hanna.)*

The authors are with the Electrical and Computer Engineering Department, University of California at Los Angeles, Los Angeles, CA 90095 USA (e-mail: ohanna@ucla.edu; linyang@ucla.edu; christina.fragouli@ucla.edu).

Digital Object Identifier 10.1109/JSAIT.2023.3260770

[1]The reward gap is defined to be the difference between the reward means of the best and second best arm.

[2]Variants of the UCB [15], [16] can achieve regret $\propto O(\sqrt{n})$, but can be worse than UCB in some regimes [9].

simply be willing to report a few rewards but not to perform more involved cooperative operations.

Our main contribution is a set of upper and lower bounds on the required number of bits to achieve the same (as in unquantized communication) regret bound up to a small constant factor. In particular, our lower bound states that it is necessary to send at least 1.9 bits per reward to achieve a regret bound within a factor of 1.5 from the regret bound of unquantized algorithms. Our upper bounds state that, on average, 3.4 bits are sufficient to maintain a regret bound within a factor of 1.5 from the unquantized regret bound.

The upper bounds are proved constructively using a novel quantization scheme, that we term *QuBan*, that is tailored to compressing MAB rewards, and can be applied on top of existing MAB algorithms. *QuBan* only cares to maintain what matters to the MAB algorithm operation, namely the ability to decide which is the best arm. At a high level, *QuBan* maps rewards to quantization levels chosen to be dense around an estimate of the arm's mean values and sparse otherwise. *QuBan* employs a stochastic correction term that enables to convey an unbiased estimate of the rewards with a small variance. It also introduces a simple novel rounding trick to guarantee that the quantization error is conditionally independent of the history given the current pulled arm index. This maintains the Markov property which is crucial in the analysis of bandit algorithms and enables reusing the same analysis methods as for unquantized rewards to bound the regret after quantization. Finally, *QuBan* encodes the reward values that occur more frequently with shorter representations, in order to reduce the number of bits communicated. Numerical results corroborate that *QuBan*, applied on top of MAB algorithms such as UCB and ϵ-greedy, uses a few bits (as small as 3) to achieve the same regret as unquantized communication.

Related Work: To the best of our knowledge, for the distributed dynamic model we consider, no scheme in the literature can be used to solve the problem of maintaining a regret bound that matches the unquantized regret bound, up to a small constant factor, while using a few bits of communication. In the following, we distinguish our work from a representative sample of existing literature.

MAB algorithms: There is a long line of research in the literature about MAB algorithms. For instance, explore-then-commit [10], [11], ϵ-greedy [12], Thomson sampling [13], the upper confidence bound [12], [14] and its variant for contextual bandits [18], [19]. Under the assumption that the reward distributions are 1-subGaussian, these algorithms provide a worst-case regret bound that is almost $O(\sqrt{n})$. The explore-then-commit regret is upper bounded by $C\sqrt{n}$ for bandits with 2 arms and known minimum means gap, while the regret of ϵ-greedy is upper bounded by $C'\sqrt{kn}$ for k-armed bandits with knowledge of the minimum means gap, where C, C' are constants that do not depend on k, n [9]. Thomson sampling and UCB achieve a regret with upper bound $C\sqrt{kn\log(n)}$ for k-armed bandits, where C is a constant that does not depend on k, n [12], [20], [21], [22], [23], [24]. For contextual linear bandits, the best known (frequentist) regret upper bound is $\tilde{O}(d\sqrt{n})$, where d is the dimension of an unknown system parameter, achieved by LinUCB [9], [18]. This matches a lower bound (for any algorithm) provided in [25] up to log factors. If focusing on Contextual Thompson sampling algorithm, the best known frequentist regret upper bound is $\tilde{O}(d^{3/2}\sqrt{n})$ [26], [27], and the best known Bayesian regret upper bound is $\tilde{O}(d\sqrt{n})$ [20]. These algorithms assume access to a full precision reward at each iteration. Our goal is not to replace the existing MAB algorithms to deal with quantized rewards; instead, we are interested in a general quantization framework that can be applied on top of any existing (or future) MAB algorithm.

Compression for ML and distributed optimization: There is a number of research results targeting reducing the communication cost of learning systems using compression. For instance, compression is applied on gradient updates [28], [29], [30], [31]. Recent work has also looked at compression for classification tasks [32]. However, compression schemes tailored to active learning, such as MAB problems, have not been explored. Our quantization scheme can be understood as a reward compression scheme that reduces the communication complexity for MABs. The main difference between the quantization for MABs and for distributed learning is that the later targets reducing the dependency of the number of bits and performance on the dimensionality of bounded training data, which can be in the order of tens of millions. In contrast, the rewards of MABs are scalars. The main challenge of our setting is to deal with a reward distribution that is either unbounded or the upper bound on the reward is much larger than the noise variance, which are typical in many MAB applications. This can be done by exploiting the fact that the rewards are more likely to be picked from the arm that appear to be best. Such a property is not applicable in the general distributed optimization setup and comes with new challenges as will discussed later.

Sample complexity: Compression is related to sample complexity [33], [34]: indeed, sending a small number of samples, reduces the overall communication load. However, the question we ask is different (and complementary): sample complexity asks how many (full precision) samples from each distribution do we need to draw; we are asking, how many bits of each sample do we really need to transmit, when we only care to decide the best arm and not to reconstruct the samples.

Distributed multi-agent MAB: Researchers have explored the distributed multi-agent MAB problem with a single [35] or multiple [36] decision makers; in these settings, distributed agents pick arms under some constraints (all agents pick the same arm [36], at most one agent can pick the same arm at a time otherwise no reward is given [3] and other constraints [37]). The agents cooperate to aggregate their observed rewards so as to jointly make a more informed decision on the best arm. Most of the works do not take into account communication constraints, and rather focus on cooperation/coordination schemes. Our setup is different: we have a single learner (central server) and simple agents who do not learn (do not keep memory) but simply observe and transmit rewards, one at a time. Our scheme can be potentially applied to these settings to reduce communication cost.

Within this previous framework, some work considers "batched" rewards, where agents keep their observed rewards in memory and communicate them infrequently, potentially summarizing their findings and thus reducing the communication load [38], [39], [40]. Such schemes require agents to be present for the whole duration of learning, and can also not be implemented by memory/computation limited agents.

Independently and in parallel to ours, the work in [41] also considered MAB learning with reduced number of bits, restricted in their case to UCB policies. Their main result shows that for rewards supported on $[0, 1]$, one bit of communication is sufficient; our work recovers this result using a much simpler approach as a special case of Section III. Additionally, our work applies on top of any MAB algorithm, and for unbounded rewards.

Paper Organization: Section II presents our model and notation; Section III looks at a special case; Section IV describes *QuBan*; Section V presents our main theorems and Section VI provides numerical evaluation.

II. Model and Notation

MAB Framework: We consider a multi-armed bandit (MAB) problem over a horizon of size n [10]. At each iteration $t = 1, \ldots, n$, a learner chooses an arm (action) A_t from a set of arms $\mathcal{A}_t$ and receives a random reward r_t distributed according to an unknown reward distribution with mean μ_{A_t}. The reward distributions are assumed to be σ^2-subGaussian [42]. The arm selected at time t depends on the previously selected arms and observed rewards $\mathcal{A}_1, A_1, r_1, \ldots, \mathcal{A}_{t-1}, A_{t-1}, r_{t-1}, \mathcal{A}_t$. The learner is interested in minimizing the expected regret $R_n = \mathbb{E}[R'_n]$, where R'_n is the regret defined as

$$R'_n = \Sigma_{t=1}^{n}\left(\mu_t^* - r_t\right), \tag{1}$$

where $\mu_t^* = \max_{A\in\mathcal{A}_t} \mu_A$. The expected regret captures the difference between the expected total reward collected by the learner over n iterations and the reward if we would collect if we play the arm with the maximum mean (optimal arm).

Notation: When the set of arms $\mathcal{A}_t$ is finite and does not depend on t: we denote the number of arms by $k = |\mathcal{A}_t|$, the best arm mean by μ^*, and the gap between the best arm and the arm-i mean by $\Delta_i := \mu^* - \mu_i$. If X, Y are random variables, we refer to the expectation of X, variance of X, conditional expectation of X given Y, and conditional variance of X given Y as $\mathbb{E}[X]$, $\sigma^2(X)$, $\mathbb{E}[X|Y]$, and $\sigma^2(X|Y)$ respectively.

Popular MAB algorithms for the case where the set of actions is fixed over time, $\mathcal{A} = \mathcal{A}_t$, and $\mathcal{A}$ is finite include explore-then-commit [10], [11], ϵ-greedy [12], Thomson sampling [13], and UCB [12], [14]. In addition to this case we also consider an important class of bandit problems, contextual bandits [26], [43]. In this case, before picking an action, the learner observes a side information, the context. Specifically we consider the widely used stochastic linear bandits model [44], where the contexts are modeled by changing the action set $\mathcal{A}_t$ across time. In this model, at iteration t, the learner chooses an action A_t from a given set $\mathcal{A}_t \subseteq \mathbb{R}^d$ and gets a reward

$$r_t = \langle \theta_*, A_t \rangle + \eta_t, \tag{2}$$

where $\theta_* \in \mathbb{R}^d$ is an unknown parameter, and η_t is a noise. Conditioned on $\mathcal{A}_1, A_1, r_1, \ldots, \mathcal{A}_t, A_t, r_t, \mathcal{A}_{t+1}, A_{t+1}$, the noise η_{t+1} is assumed to be zero mean and σ^2-subGaussian. Popular algorithms for this case include LinUCB [18], explore-then-commit strategy [45], and contextual Thomson sampling [26].

System Setup: We are interested in a distributed setting, where a learner asks at each time a potentially different agent to play the arm A_t; the agent observes the reward r_t and conveys it to the learner over a communication constrained channel, as depicted in Fig. 1. In our setup, each agent needs to immediately communicate the observed reward (with no memory), using a quantization scheme to reduce the communication cost. As learning progresses, the learner is allowed to refine the quantization scheme by broadcasting parameters to the agents they may need. We do not count these broadcast (downlink) transmissions in the communication cost since the learner has no restrictions in its power. We stress again that the agents cannot store information of the reward history since they may *join and leave* the system at any time. We thus opt to use a setting where the agents have no memory. This setting allows to support applications with simple agents (e.g., RFID applications and embedded systems).

Quantization: A quantizer consists of an encoder $\mathcal{E} : \mathbb{R} \to \mathcal{S}$ that maps $\mathbb{R}$ to a countable set $\mathcal{S}$, and a decoder $D : \mathcal{S} \to \mathbb{R}$. At each time t, the agent that observes the reward r_t transmits a finite length binary sequence representing $\mathcal{E}(r_t)$ to the learner which in turn decodes it using the decoder D to obtain the quantized reward $\hat{r}_t = D(\mathcal{E}(r_t))$. The range of a decoder is referred to as the set of quantization levels; the encoding and decoding operation of a quantizer maps the reward to a quantization level. We next describe a specific quantization module that we will use.

Stochastic Quantization (SQ): A stochastic quantizer that uses quantization levels in a set $\mathcal{L}$, which is a form of dithering [29], [46], consists of a randomized encoder $\mathcal{E}_\mathcal{L}$ and decoder $D_\mathcal{L}$ modules that can be described as following. The encoder $\mathcal{E}_\mathcal{L}$, that uses the set of quantization levels $\mathcal{L} = \{\ell_i\}_{i=1}^{2^B}$, takes as input a value x in $[\ell_1, \ell_{2^B}]$; it maps x to a level index described by B bits. The decoder, that uses the set of quantization levels $\mathcal{L} = \{\ell_i\}_{i=1}^{2^B}$, takes as input an index in $\{1, \ldots, 2^B\}$, and outputs the corresponding level value. Precisely,

$$\begin{aligned} i(x) &= \max\left\{j | \ell_j \le x \text{ and } j < 2^B\right\}, \\ \mathcal{E}_\mathcal{L}(x) &= \begin{cases} i(x) & \text{with probability } \frac{\ell_{i(x)+1}-x}{\ell_{i(x)+1}-\ell_{i(x)}} \\ i(x)+1 & \text{with probability } \frac{x-\ell_{i(x)}}{\ell_{i(x)+1}-\ell_{i(x)}} \end{cases}, \\ D_\mathcal{L}(j) &= \ell_j, j \in \left\{1, \ldots, 2^B\right\}. \end{aligned} \tag{3}$$

That is, if x is such that $\ell_i \le x < \ell_{i+1}$, then the index i is transmitted with probability $\frac{\ell_{i+1}-x}{\ell_{i+1}-\ell_i}$ (and x is decoded to be ℓ_i) while the index $i+1$ is transmitted with probability $\frac{x-\ell_i}{\ell_{i+1}-\ell_i}$ (and x is decoded to be ℓ_{i+1}).

The analysis of bandit algorithms leverages the fact that conditioned on A_t, the communicated reward r_t is an unbiased estimate of the mean μ_{A_t}. It is not difficult to see that SQ

preserves this property, namely conditioned on A_t, it conveys to the learner an unbiased estimate of μ_{A_t}.

Performance Metric $B_n, \bar{B}(n)$: Among the schemes that achieve a regret matching the unquantized regret, up to a fixed small constant factor, our performance metrics are the *instantaneous* and *average* number of communication bits per reward B_n, and $\bar{B}(n)$ respectively. Let B_t be the number of bits used to transmit $\hat{r}_t$, and define the average number of bits after n iterations of the algorithm as $\bar{B}(n) = \frac{\sum_{t=1}^{n} B_t}{n}$. Our goal is to design quantization schemes that achieve expected regret matching the expected regret of unquantized communication (up to a small constant factor) while using a small number of bits B_n, and $\bar{B}(n)$.

III. A Case Where 1 Bit Is Sufficient

In this section we show that there exist some "easy" cases where we can use just one bit per reward and a very simple quantization scheme. Note that one bit is a trivial lower bound on the instantaneous number of bits communicated B_n, since each agent needs to respond to the learner for each observed reward. We also note that by definition of the average number of bits $\bar{B}_n$ as the average of $B_1, \ldots, B_n$, one bit is also a lower bound on the average number of bits. Consider the case where the rewards are supported on [0, 1] and all reward distributions have the same variance σ (but different means). Since $r_t \in [0, 1]$, its variance is upper bounded by $\frac{1}{4}$; we will here assume this worst case variance $\sigma^2(r_t|A_t) \approx 1/4$.[3] We will use 1-bit Stochastic Quantization (SQ), as in (3). The stochastic 1 bit quantizer takes r_t as input and interprets it as probability: outputs 1 with probability r_t and 0 with probability $1 - r_t$. Let $\hat{r}_t$ be the (binary) quantized reward, we then have that

$$\mathbb{E}\big[\hat{r}_t|A_t\big] = \mathbb{E}\big[\mathbb{E}\big[\hat{r}_t|r_t, A_t\big]|A_t\big] = \mathbb{E}[r_t|A_t] = \mu_{A_t}. \tag{4}$$

Recall that for bandit algorithms the expected regret scales linearly with the variance. For example, the UCB algorithm (c.f., [9]) with unquantized rewards, achieves $R_n \leq C\sigma\sqrt{nk\log(n)}$ for a constant C that does not depend on k, n.

Proposition 1: UCB with $r_t \in [0, 1]$ that uses 1-bit SQ achieves a regret $R_n \leq C\sqrt{nk\log(n)}$.

Proof: The proof follows directly from the case of reward distributions that are supported on $[0, 1]$ in [12]. It follows a standard analysis based on confidence intervals by bounding the regret conditioning on the good event $\mathcal{G} = \{|\frac{\sum_{i=1}^{t} r_i \mathbf{1}\{A_i=A_t\}}{\sum_{i=1}^{t} \mathbf{1}\{A_i=A_t\}} - \mu_{A_t}| \leq \frac{C\log(n)}{\sqrt{\sum_{i=1}^{t} \mathbf{1}\{A_i=A_t\}}} \forall t = 1, \ldots, n\}$ which is shown to hold with probability at least $1 - \frac{1}{n}$. Assuming $\mathcal{G}$, it can be shown that the total number of pulls for an arm with gap Δ_i, according to the UCB rule, is $O(\frac{\log(n)}{\Delta_i^2})$ resulting in a regret that is bounded as

$$\begin{aligned} R_n &\leq \mathbb{E}\big[R'_n|\mathcal{G}\big] + \mathbb{E}\Big[R'_n|\mathcal{G}^C\Big]\mathbb{P}\Big[\mathcal{G}^C\Big] \\ &\leq n\Delta + \sum_{i:\Delta_i>\Delta} \frac{C\log(n)}{\Delta_i} + n\frac{1}{n} \leq n\Delta + \frac{Ck\log(n)}{\Delta} + 1. \end{aligned} \tag{5}$$

[3] A similar analysis extends, showing that UCB with 1-bit SQ achieves a regret within a small constant factor from the unquantized regret, when the variances differ but they are all close to $\frac{1}{4}$.

The result follow by maximizing over Δ. ∎

Simulation results, in this section and Section VI, verify that, for $r_t \in [0, 1]$, 1-bit SQ performs very close to unquantized rewards.

To motivate our general quantization scheme, we consider a case where 1-bit SQ results in a potentially large performance loss. Assume that the variance, σ, is much smaller than the range of $r_t : r_t \in [-\lambda, \lambda]$ and $\sigma = 1$, where $\lambda \gg 1$ is a parameter known to the learner. The 1-bit SQ maps r_t to either λ or $-\lambda$; it is not difficult to see that we still have $\mathbb{E}[\hat{r}_t|A_t] = \mu_{A_t}$, but $R_n \leq C\lambda\sqrt{kn\log(n)}$, where C is a constant that does not depend on n, k [12].[4] This can be seen by observing that the expected regret can be written as $R_n = \sum_{t=1}^{n} \mathbb{E}(\mu_t^* - r_t) = \sum_{t=1}^{n} \mathbb{E}(\mu_t^* - \hat{r}_t) = 2\lambda \sum_{t=1}^{n} \mathbb{E}(\frac{(\mu_t^*+\frac{1}{2})-(\hat{r}_t+\frac{1}{2})}{2\lambda})$, which transforms the problem to one with reward distributions supported on [0, 1]. Thus the expected regret bound grows linearly with λ, which can be arbitrarily large. In contrast, without quantization UCB achieves $C'\sqrt{kn\log(n)}$, where C' is another constant of the same order of C.

Simulation results verify that the convergence to the unquantized case can be slow. Fig. 2 shows the regret of unquantized and 1-bit SQ with the UCB algorithm for the setup described in Section III with $\sigma = 1$ and clipped reward distributions that have support only on an interval $[-\lambda, \lambda]$, for $\lambda = 1$ and 100 respectively. As discussed, we observe a regret penalty when $\lambda \gg \sigma$.

We take away the following observations:

- If the range λ is of the same order as the variance σ, 1-bit SQ is sufficient to preserve the regret bound up to a small constant factor.
- If the range λ is much larger than σ, 1-bit SQ leads to a regret penalty proportional to $\frac{\lambda}{\sigma}$; thus we may want to only perform stochastic quantization within intervals of size similar to σ.
- In our discussion up to now we assumed that the rewards r_t are bounded almost surely. This is not true in general; we would like an algorithm that uses a small average number of bits even when the reward distributions are unbounded.

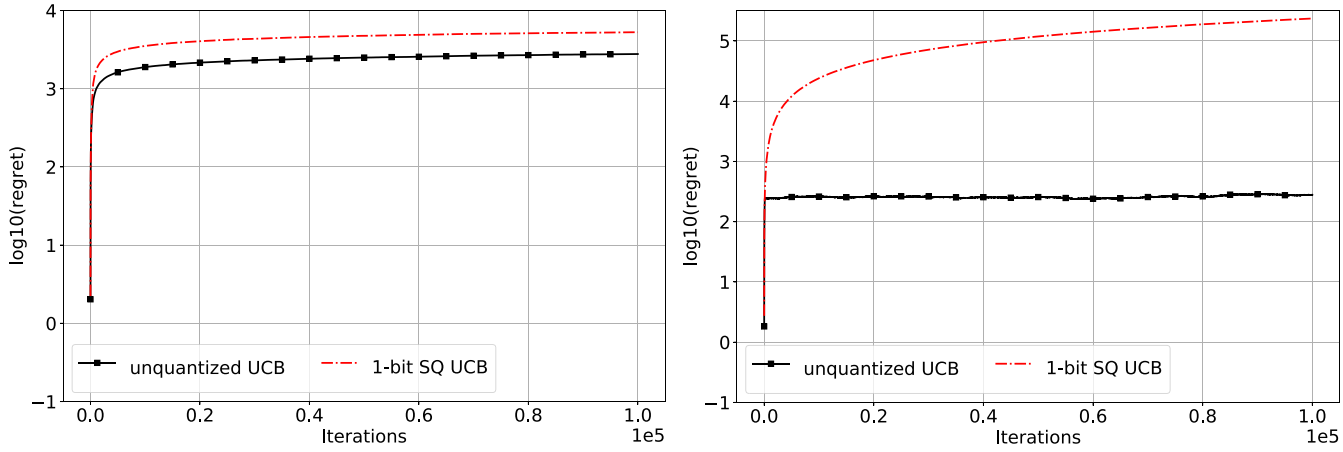
In the next section we introduce *QuBan*, that achieves a small average number of bits in all of the above cases.

IV. *QuBan*: A General Quantizer for Bandit Rewards

In this section, we propose *QuBan*, an adaptive quantization scheme that can be applied on top of any MAB algorithm. Our scheme maintains attractive properties (in particular, the Markov property, unbiasedness, and bounded variance) for the quantized rewards that enable to retain the same regret bound as unquantized communication for the vast majority of MAB algorithms, while using a few bits for communication (simulation results show convergence to ~ 3 bits per iteration for n that is sufficiently large, see Section VI).

QuBan uses ideas that include: (i) centering the quantization scheme around a value that is believed to be close to the

[4] We note that this bound cannot be improved using techniques in [12], since it is possible that $\sigma^2(\hat{r}_t|A_t) = \lambda^2$ (e.g., if $r_t = 0$ almost surely).



(a) $\lambda = 1$. (b) $\lambda = 100$.

Fig. 2. Regret versus number of iterations. We use $\sigma = 1$.

picked arm mean in the majority of iterations; (ii) maintaining a quantization error that is conditionally independent on previously observed rewards given the arm selection, which is achieved by choosing the quantization center to be an integer value (illustrated in more detail in the proof of Theorem 2); (iii) assigning shorter codes to the values near the quantization center and otherwise longer codes to maintain a finite expected number of bits even if the reward distribution has infinite support; and (iv) using stochastic quantization to convey an unbiased estimate of the reward. We next describe *QuBan* in more detail.

A. QuBan Centers the Quantization Around a Value $\hat{\mu}(t)$

Recall that at time t the learner selects an action A_t and needs to convey the observed reward r_t. As we expect r_t to be close to the mean μ_{A_t}, we would like to use quantization levels that are dense around μ_{A_t} and sparse in other areas. Since μ_{A_t} is unknown, we estimate it using some function of the observed rewards that we term $\hat{\mu}(t)$; we can think of $\hat{\mu}(t)$ as specifying a "point" on the real line around which we want to provide denser quantization.

Choices for $\hat{\mu}(t)$: In this work, we analyze the following three choices for $\hat{\mu}(t)$, the first two applying to MAB with a finite fixed set of arms, while the third to linear bandits.

- *Average arm point (Avg-arm-pt):* $\hat{\mu}(t) = \hat{\mu}_{A_t}(t-1)$. We use $\hat{\mu}_{A_t}(t-1)$, the average of the samples picked from arm A_t up to time $t-1$, as an estimate of μ_{A_t}.
- *Average point (Avg-pt):* $\hat{\mu}(t) = \frac{1}{t-1}\sum_{j=1}^{t-1} \hat{r}_j$ (the average over all observed rewards). Here we can think of $\frac{1}{t-1}\sum_{j=1}^{t-1} \hat{r}_j$ as an estimate of the mean of the best arm. Indeed, the average reward of a well behaved algorithm will converge to the best mean reward.

These two choices of $\hat{\mu}(t)$ give us flexibility to fit different regimes of MAB systems. In particular, we expect the avg-arm-pt to be a better choice for a small number of arms k and MAB algorithms that achieve good estimates of μ_{A_t} (explore all arms sufficiently so that $\hat{\mu}_{A_t}(t-1)$ approaches μ_{A_t}). However, since the first time an arm is pulled we do not have an estimate of its mean, this results in possibly larger number of bits for the first pull; this penalty increases with the number of arms k. As our analysis also shows (see Section V), if k is large, acquiring good estimates for all arms may be costly and not what good algorithms necessarily pursue; instead, the avg-pt has a simpler implementation, as it only requires to keep track of a single number, and still enables to distinguish well in the neighborhood of the best arm (connecting the number of bits to the regret), which is essentially what we mostly want.

- *Contextual bandit choice:* $\hat{\mu}(t) = \langle \theta_t, A_t \rangle$. Consider the widely used stochastic linear bandits model in Section II. We observe that linear bandit algorithms, such as contextual Thomson sampling and LinUCB, choose a parameter θ_t believed to be close to the unknown parameter θ_*, and pick an action based on θ_t. For example, LinUCB [18] chooses a confidence set $\mathcal{C}_t$ with center θ_t believed to contain θ_* and picks an action $A_t = \arg\max_{a \in \mathcal{A}_t} \max_{\theta \in \mathcal{C}_t} \langle \theta, a \rangle$. Accordingly, we propose to use $\hat{\mu}(t) = \langle \theta_t, A_t \rangle$. We note that our intuition for the avg-pt choice does not work for contextual bandits as it relies on that $\max_{a \in \mathcal{A}_t} \langle \theta_*, a \rangle$ is the same for all t, which might not hold in general. Likewise, the avg-arm-pt choice will not work as the set of actions $\mathcal{A}_t$ can be infinite or change with time.

We underline that the estimator $\hat{\mu}(t)$ is only maintained at the learner's side and is broadcasted to the agents. As discussed before, this downlink communication is not counted as communication cost.

B. QuBan Components

As discussed, at iteration t, *QuBan* centers its quantization around the value $\hat{\mu}(t)$. It then quantizes the normalized reward $\bar{r}_t = r_t/M_t - \lfloor \hat{\mu}(t)/M_t \rfloor$ to one of the two values $\lfloor \bar{r}_t \rfloor$, $\lceil \bar{r}_t \rceil$, where $M_t = \epsilon \sigma X_t$,[5] ϵ is a parameter to control the regret vs number of bits trade-off as will be illustrated later in this section, and $\{X_t\}_{i=1}^n$ are independent samples from a $\frac{1}{4}$-subGaussian distribution satisfying $|X_t| \geq 1$ almost surely, e.g., we can use

[5]The case where σ is unknown is discussed in Section V.

Algorithm 1 Learner Operation With Input MAB Algorithm Λ

Initialize: $\hat{\mu}(1) = 0$
for $t = 1, \ldots, n$ **do**
  Choose an action A_t based on the bandit algorithm Λ and ask the next agent to play it
  Send M_t[8], $\hat{\mu}(t)$ to an agent
  Receive the encoded reward $(b_t, I_t, \mathcal{E}_{\mathcal{L}_t}(e_t))$ (see Algorithm 2)
  Decode $\hat{r}_t$**:**
  if length$(b_t) \leq 3$ **then**
    $\hat{r}_t$ can be decoded using a lookup table
  else
    Decode the sign, s_t, of $\hat{\bar{r}}_t$ from b_t
    Set ℓ_t to be the I_t-th element in the set $\{0, 2^0, ...\}$
    Set $\mathcal{L}_t = \{\ell_t, \ell_t + 1, \ldots, \max\{2\ell_t, \ell_t + 1\}\}$
    Define $e_t^{(q)} = D_{\mathcal{L}_t}(\mathcal{E}_{\mathcal{L}_t}(e_t))$
    $\hat{r}_t = (s_t(e_t^{(q)} + \ell_t + 3.5s_t + 0.5) + \lfloor \hat{\mu}(t)/M_t \rfloor)M_t$
  Calculate $\hat{\mu}(t+1)$ (using one of the discussed choices)
  Update the parameters required by Λ

Algorithm 2 Distributed Agent Operation

Inputs: r_t, $\hat{\mu}(t)$ and M_t
Set $L = \{\lfloor \bar{r}_t \rfloor, \lceil \bar{r}_t \rceil\}$, $\hat{\bar{r}}_t = D_L(\mathcal{E}_L(\bar{r}_t))$, where $\bar{r}_t = r_t/M_t - \lfloor \hat{\mu}(t)/M_t \rfloor$.
Set b_t with three bits to distinguish between the 8 cases: $\hat{\bar{r}}_t < -2, \hat{\bar{r}}_t > 3, \hat{\bar{r}}_t = i, i \in \{-2, -1, 0, 1, 2, 3\}$. This implicitly encodes the sign of $\hat{\bar{r}}_t$, which we denote s_t.
if $|\hat{\bar{r}}_t| > |a|$ and $\hat{\bar{r}}_t a > 0$, $a \in \{-2, 3\}$ **then**
  Augment b_t with an extra one bit to indicate if $|\hat{\bar{r}}_t| = |a| + 1$ or $|\hat{\bar{r}}_t| > |a| + 1$.
  if $|\hat{\bar{r}}_t| > |a| + 1$ **then**
    Let $L' = \{0, 2^0, \ldots\}$
    Set $\ell_t = \max\{j \in L' | j \leq |\bar{r}_t| - (|a| + 1)\}$
    Encode ℓ_t by $I_t - 1$ zeros followed by a one (unary coding), where I_t is the index of ℓ_t in the set L'.
    Let $e_t = |\bar{r}_t| - (|a| + 1) - \ell_t$
    Set $\mathcal{L}_t = \{\ell_t, \ell_t + 1, \ldots, \max\{2\ell_t, \ell_t + 1\}\}$
    Encode e_t using SQ to get $\mathcal{E}_{\mathcal{L}_t}(e_t)$
Transmit $(b_t, I_t, \mathcal{E}_{\mathcal{L}_t}(e_t))$

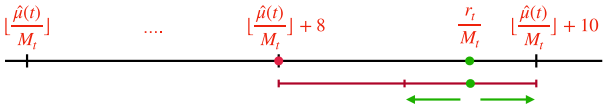


Fig. 3. Illustration of *QuBan*. In the shown example, r_t is mapped to a value of the red dot (conveyed with the index $I_t = 4$), and stochastically to one of the two nearest quantization levels depicted on the red line.

$X_t = 1$ almost surely.[6] If X_t is allowed to take larger values with some probability, it will result in coarser quantization with some probability, and a smaller number of bits. This introduces an error in estimating $\bar{r}_t$ that is bounded by 1, which results in error of at most M_t in estimating $r_t = M_t(\bar{r}_t + \lfloor \hat{\mu}(t)/M_t \rfloor)$. This quantization is done in a randomized way to convey an unbiased estimate of r_t. The encoding of $\hat{r}_t$ is a composition of: sign of $\bar{r}_t$, a unary encoding of the least power of 2 below $|\bar{r}_t|$ (denoted by 2^{I_t}), and SQ for $|\bar{r}_t| - 2^{I_t}$.[7] The unary encoding of I_t consists of I_t zeros followed by 1 one. Both the unary encoding and the SQ use $O(\log(\bar{r}_t))$ bits. We recall that $\frac{\hat{\mu}(t)}{M_t}$ is believed to be close to $\frac{r_t}{\sigma}$ in the majority of iterations resulting in small values for $\log(\bar{r}_t)$.

The precise learner and agent operations used for *QuBan* are presented in pseudo-code in Algorithms 1 and 2 (see Fig. 3 for an example), respectively. The learner at each iteration broadcasts $\hat{\mu}(t)$ and asks one of the agents available at time t to play an action A_t. Initially, since we have no knowledge about μ_i, the learner assumes that $\hat{\mu}(0) = 0$. The agent that plays the action uses the observed r_t together with $\hat{\mu}(t)$ it has received to transmit three values we term (b_t, I_t, e_t), to the learner, as described in Algorithm 2 using $O(\log(|\bar{r}_t|))$ bits.

Rounding of $\hat{\mu}(t)/M_t$: the reason for choosing the quantization to be centered around $\lfloor \hat{\mu}(t)/M_t \rfloor$ instead of $\hat{\mu}(t)/M_t$ is to guarantee that the distance between r_t and the two closest quantization levels is independent of $\hat{\mu}(t)$[9] (which is dependent on $\hat{r}_1, \ldots, \hat{r}_{t-1}$). As we discuss in the following section, this preserves the Markov property (given A_t, the quantized reward $\hat{r}_t$ is conditionally independent on the history $A_1, \hat{r}_1, \ldots, A_{t-1}, \hat{r}_{t-1}$), a property that is exploited in the analysis of bandit algorithms to guarantee that $|\sum_{t=1}^{n} \hat{r}_t - \mu_{A_t}/n|$ approaches zero in some probabilistic sense as n increases.

Sending the least power of 2 below $\bar{r}_t$: For simplicity we consider the case where $\bar{r}_t \geq 0$. We note that since it is possible for the decoded reward to take any value in the set $\{\lfloor \frac{\hat{\mu}}{M_t} \rfloor, \lfloor \frac{\hat{\mu}}{M_t} \rfloor + 1, \lfloor \frac{\hat{\mu}}{M_t} \rfloor + 2 \cdots\}$ (to guarantee the uniform upper bound on $|\hat{\bar{r}}_t - \bar{r}_t|$), every value in that set needs to be encoded. A good encoding strategy assigns shorter codes to the levels that are close to $\lfloor \frac{\hat{\mu}}{M_t} \rfloor$ as they are expected to occur more often. Hence, the best we can hope for is to encode r_t using $O(\log(\frac{r_t}{M_t} - \lfloor \frac{\hat{\mu}}{M_t} \rfloor)$ bits as it is quantized to either $\lfloor \frac{r_t}{M_t} \rfloor$ or $\lceil \frac{r_t}{M_t} \rceil$ and the quantization level at $\lfloor \frac{r_t}{M_t} \rfloor$ is encoded using the largest number of bits among the levels in the set $\{\lfloor \frac{\hat{\mu}}{M_t} \rfloor, \lfloor \frac{\hat{\mu}}{M_t} \rfloor + 1, \lfloor \frac{\hat{\mu}}{M_t} \rfloor + 2, \ldots, \lfloor \frac{r_t}{M_t} \rfloor\}$. As can be seen in Appendix B, sending the greatest power of 2 below $\bar{r}_t$ then quantizing the difference using SQ gives that r_t is encoded using $O(\log(\frac{r_t}{M_t} - \lfloor \frac{\hat{\mu}}{M_t} \rfloor)$ bits. This is achieved since I_t is $O(\log(\frac{r_t}{M_t} - \lfloor \frac{\hat{\mu}}{M_t} \rfloor)$ and the SQ uses $2^{I_t} + 1$ quantization levels. An alternative way to encode I_t is using integer compression, or recursively applying our scheme by using unary coding to transmit the largest $I_t^{(2)}$ with $2^{I_t^{(2)}} \leq I_t$ and then encode the difference $I_t - 2^{I_t^{(2)}}$ using $\log(1 + 2^{I_t^{(2)}})$ bits noting that $I_t - 2^{I_t^{(2)}} \leq 2^{I_t^{(2)}}$. This results in using $O(\log(\log(\frac{r_t}{\sigma} - \lfloor \frac{\hat{\mu}}{\sigma} \rfloor)))$ bits to encode I_t. We keep the unary

[6]For our proofs we set $X_t = 1$ for simplicity; more sophisticated choices can further improve the upper bounds such as X_t picked from a Gaussian distribution.

[7]Note that $0 \leq |\bar{r}_t| - 2^{I_t} \leq 2^{I_t}$.

[8]If X_t is chosen to be 1, then sending M_t is not required.

[9]As will be shown in Appendix A, centering the quantization around any integer value implies that the two closest quantization levels to $\frac{r_t}{M_t}$ are $\lfloor \frac{r_t}{M_t} \rfloor, \lceil \frac{r_t}{M_t} \rceil$.

coding for I_t for simplicity and since it does not dominate the average number of bits.

Preserving regret bounds: The main reasons *QuBan* preserves existing regret bounds is that it does not destroy the Markov property (as we prove in Appendix B) and it provides that $|\hat{\bar{r}}_t - \bar{r}_t|$ is uniformly upper bounded. The later property implies that if given A_t, r_t is conditionally integrable, sub-exponential, sub-Gaussian, or almost surely bounded, then M_t can be chosen such that given A_t, $\hat{r}_t$ is conditionally integrable, sub-exponential, sub-Gaussian, or almost surely bounded respectively. A widely used assumption is that given A_t, r_t is conditionally sub-Gaussian.

We observe that all the operations in *QuBan* can be performed in a constant time except steps 9, 13 in Algorithm 2 which require $O(B_t)$ running time. As shown in Sections V, VI, B_t is only a few bits on average resulting in a linear amortized running time.

V. Upper and Lower Bounds

In this section we present an upper bound on the number of bits used by *QuBan* and show that it provides properties for the quantized reward that result in a regret within a small constant factor from the unquantized regret. We also present a lower bound, within 1.5 bits from the upper bound, on the number of bits needed to satisfy the required properties. Before stating the results, we state our assumptions.

Assumption 1: We assume that we are given:

(i) a MAB instance with σ^2-subGaussian[10] rewards where the Markov property holds: conditioned on the action at time t, the current reward is conditionally independent on the history (past actions and rewards).

(ii) a MAB algorithm Λ such that for any instance with σ^2-subGaussian rewards, and time horizon n, the algorithm's expected regret (with unquantized rewards) is upper-bounded by R_n^U.

We note that assumption (*i*) is standard for the analysis of MAB algorithms, while assumption (*ii*) essentially only introduces notation.

A. Upper Bounds

The following proposition gives an upper bound on the regret after quantization, and shows that for $\epsilon = 1$, the regret is within a factor of 1.5 from the regret of the unquantized case. The proof is provided in Appendix A.

Proposition 2: Suppose Assumption 1 holds. Then, when we apply *QuBan*, the following hold:

1) Conditioned on A_t, the quantized reward $\hat{r}_t$ is $((1+\frac{\epsilon}{2})\sigma)^2$-subGaussian, conditionally independent on the history $A_1, \hat{r}_1, \ldots, A_{t-1}, \hat{r}_{t-1}$ (Markov property), and satisfies $\mathbb{E}[\hat{r}_t|A_t] = \mu_{A_t}$, $|\hat{r}_t - r_t| \le M_t$ almost surely $(t = 1, \ldots, n)$.

2) The expected regret R_n is bounded as $R_n \le (1+\frac{\epsilon}{2})R_n^U$, where ϵ is a parameter to control the regret vs number of bits trade-off.

In the following we provide an upper bound on the expected average number of bits. We also provide a high-probability upper bound on the instantaneous number of bits. For simplicity we only consider the case where $\epsilon = 1$ and discuss the other case in Appendix B. The proof is given in Appendix B. At a high level, to upper bound the regret after quantization we show that *QuBan* maintains a number of desirable properties for the quantized rewards, namely, unbiasedness, and the fact that the quantized rewards are $(1.5\sigma)^2$-subGaussian and satisfy the Markov property. To upper bound the expected number of bits we use the fact that *QuBan* assigns short representations for the rewards around an estimate of the mean, which we expect to see more frequently.

Theorem 1: Suppose Assumption 1 holds. Let $\epsilon = 1$. There is a universal constant C such that:

1) For *QuBan* with $\hat{\mu}(t) = \hat{\mu}_{A_t}(t-1)$ (avg-arm-pt), the average number of bits communicated satisfies that $\mathbb{E}[\bar{B}(n)] \le 3.4 + (C/n)\sum_{i=1}^{k} \log(1 + |\mu_i|/\sigma) + C/\sqrt{n}$.

2) For *QuBan* with $\hat{\mu}(t) = \frac{1}{t-1}\sum_{j=1}^{t-1}\hat{r}_j$ (avg-pt), the average number of bits communicated satisfies $\mathbb{E}[\bar{B}(n)] \le 3.4 + \frac{C}{n}(1 + \log(1 + \frac{|\mu^*|}{\sigma}) + \frac{R_n}{\sigma} + \sum_{t=1}^{n-1}\frac{R_t}{(\sigma t)}) + C/\sqrt{n}$.

3) For *QuBan* with $\hat{\mu}(t) = \langle \theta_t, A_t \rangle$ (stochastic linear bandit), the average number of bits communicated satisfies that $\mathbb{E}[\bar{B}](n) \le 3.4 + C\mathbb{E}[\sum_{t=1}^{n} |\langle \theta_t - \theta_*, A_t \rangle|]/(\sigma n)$.

In Appendix B we also provide almost surely bounds on the asymptotic average number of bits, namely, $\lim_{n\to\infty}(1/n)\sum_{t=1}^{n} B_t \le 3.4$ almost surely.

In the following we provide a high probability bound on the number of bits that *QuBan* uses in each iteration. We analyze the performance for avg-arm-pt only; the other choices for $\hat{\mu}(t)$ can be handled similarly.

Theorem 2: For a MAB instance with σ^2-subGaussian rewards, *QuBan* with $\epsilon = 1$, $\hat{\mu}(t) = \hat{\mu}_{A_t}(t-1)$ (avg-arm-pt), satisfies that for t with $T_t(A_t) > 0$, where $T_t(i)$ is the number of pulls for arm i prior to iteration t, with probability at least $1 - \frac{1}{n}$ it holds that $\forall t \le n$:

$$B_t \le 4 + \lceil \log(4\log(n)) \rceil + \lceil \log\log(4\log(n)) \rceil. \tag{6}$$

The proof is provided in Appendix C.

Remark 1: Using the previous lemma we can modify *QuBan* to have that (6) is satisfied almost surely, by sending a random 1 bit when (6) is not satisfied. This will only add at most $n\sum_{i=1}^{k}\Delta_i$ regret with probability at most $\frac{1}{n}$. Hence, the expected regret is increased by at most a factor of 2.

Remark 2: Throughout the paper, we assume a known upper bound on the noise variance. However, it is not difficult to see that a variance estimate within a constant factor would suffice. Running *QuBan* with an estimate σ' that is possibly different from the true σ results in a degradation in the regret by a factor of $\max\{1, \frac{\sigma'}{\sigma}\}$ and increase in the communication by $2\log(\frac{\sigma}{\sigma'})$ bits. An optimistic estimate of the noise $\sigma' < \sigma$ results in finer quantization, hence, no degradation in the regret at the cost of increasing the communication by $2\log(\frac{\sigma}{\sigma'})$ bits.

B. Lower Bound

In this subsection we provide a lower bound showing that an average number of 1.9 bits per iteration are required to

[10]This is a standard assumption used for simplicity but is not required for our main results.

maintain a sublinear regret and a $(\frac{\sigma}{2})^2$-subGaussian quantization error, $\hat{r}_t - r_t$. We also show that the instantaneous number of bits cannot be almost surely bounded by a constant. In our lower bound, we focus on prefix free codes [47]; a similar analysis can be performed for non-singular codes leading to different constants. We also note that our achievable scheme (Algorithms 1, 2) provides a prefix-free code. We first state the following lemma which shows that for the quantizer to preserve the sublinear regret of the bandit algorithm, Q needs to satisfy that $\mathbb{E}[Q(r_t)|r_t] = c_1 r_t + c_2$, where c_1, c_2 are constants. Hence, by a proper shifting and scaling, the quantizer Q can be made unbiased, i.e., satisfying $\mathbb{E}[Q(r_t)|r_t] = r_t$.

Lemma 1: Let ALG be any algorithm for multi-arm bandits with sublinear regret and Q be (a possibly randomized) quantizer. Let R_n be the worst-case expected regret of ALG when using rewards $Q(r_1), \ldots, Q(r_t)$. If R_n is sublinear in n, then Q satisfies

$$\mathbb{E}[Q(r_t)|r_t] = c_1 r_t + c_2,$$

where c_1, c_2 are constants.

By the previous lemma, it suffices to only consider unbiased quantizers. We next state our lower bound theorem.

Theorem 3: Any (possibly randomized) quantizer Q that uses prefix-free encoding and satisfies:

1) **(Unbiased Property)** $\mathbb{E}[Q(r_t)|r_t] = r_t$,

2) **(SubGaussian Property)** Conditioned on r_t, $Q(r_t) - r_t$ is $(\frac{\sigma}{2})^2$-subGaussian $(t = 1, \ldots, n)$,

we have that there exist $(4\sigma)^2$-subGaussian reward distributions for which:

1) $(\forall b \in \mathbb{N})$ $(\exists t, \delta > 0)$ such that $\mathbb{P}[B_t > b] > \delta$.

2) $(\forall t > 0)$ $(\exists n > t)$ such that $\mathbb{E}[\bar{B}(n)] \geq 1.9$ bits.

The proofs are given in Appendix D.

C. Application to UCB, ϵ-Greedy, and LinUCB

We here leverage Theorem 1 to derive bounds for three widely used MAB algorithms. We highlight that although the regret bounds hide constant factors, these constants are within 1.5 of the unquantized constants according to Theorem 1. The proofs are in Appendix E for Corollaries 1 and 2 and in Appendix F for Lemma 3.

Corollary 1: Assume we use *QuBan* with avg-pt on top of **UCB** [12] with σ^2-subGaussian reward distributions. Then there is a constant C that does not depend on n and k such that $R_n \leq C\sigma\sqrt{nk\log(n)}$, $\mathbb{E}[\bar{B}(n)] \leq 3.4 + C\sqrt{k\log(n)/n}$.

Corollary 2: Assume we use *QuBan* with avg-pt on top of ϵ-greedy [12] with σ^2-subGaussian reward distributions and constant gaps Δ_i $\forall i$. Let $\epsilon_t = \min\{1, Ck/(t\Delta_{\min}^2)\}$, where $\Delta_{\min} = \min_i\{\Delta_i|\Delta_i > 0\}$ and $C > 0$ is a sufficiently large universal constant. Then there exists a constant C' that does not depend on n and k such that $R_n \leq C'\sigma k \log(1 + n/k)$, $\mathbb{E}[\bar{B}(n)] \leq 3.4 + C'(k\log^2(n)/n + 1/\sqrt{n})$.

To simplify the expressions, we include the dependency on μ^* and Δ_i in the constant C for Corollary 1 and respectively C' for Corollary 2.

Corollary 3: Assume we use *QuBan* on top of **LinUCB** [18], then there is a constant C that does not depend on n and d such that $R_n \leq Cd\sqrt{n}\log(n)$, $\mathbb{E}[\bar{B}(n)] \leq 3.4 + C\frac{d\log(n)}{\sqrt{n}}$.

VI. Numerical Evaluation

We here present our numerical results.

Quantization Schemes: We compare *QuBan* against the baseline schemes described next.

Unquantized: Rewards are conveyed using the standard 32 bits representation.

r-bit SQ: We implement r-bit stochastic quantization, by using the quantizer described in Section II, with 2^r levels uniformly dividing a range $[-\lambda, \lambda]$.

QuBan: We implement *QuBan* with $\epsilon = X_t = 1$.

MAB Algorithms: We use quantization on top of:

(i) the UCB implementation in [9, ch. 8]. The UCB exploration constant is chosen to be σ_q, an estimate of the standard deviation of the quantized reward distribution.

(ii) the ϵ-greedy algorithm in [9, ch. 6], where ϵ_t is set to be $\epsilon_t = \min\{1, \frac{C\sigma_q k}{t\Delta_{\min}^2}\}$.

(iii) the LinUCB algorithm for stochastic linear bandits in [9, ch. 19].

Synthetic Dataset:

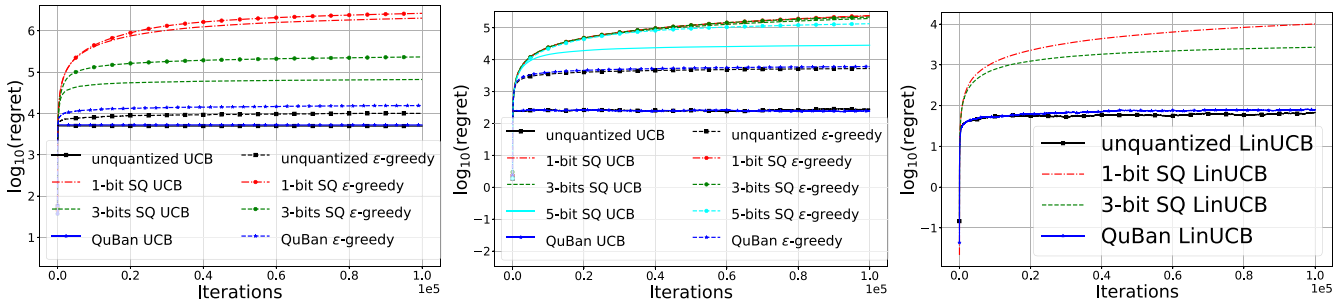
MAB Setup: We simulate three cases. In each case we average over 10 runs of each experiment. The parameters σ_q, C are determined by the underlying MAB algorithm we use. In our simulations, we set σ_q to the variance of the quantized rewards (or best known upper bound), while C is chosen to be the value resulting in best regret among $\{0, \lambda/100, \ldots, \lambda\}$.

• *Setup 1 (Figs. 4-6(a)):* We use $k = 100$, $\lambda = 100$, $C = 10$, the arms' means are picked from a Gaussian distribution with mean 0 and standard deviation 10 and the reward distributions are conditionally Gaussian given the actions A_t with variance 0.1. The parameter σ_q is set to be 0.1 for *QuBan* and $200/2^r - 1$ for the r-bit SQ.

• *Setup 2 (Figs. 4-6(b)):* This differs from the previous only in that the means are picked from a Gaussian distribution with mean 95 and standard deviation 1 (leading to smaller Δ_i).

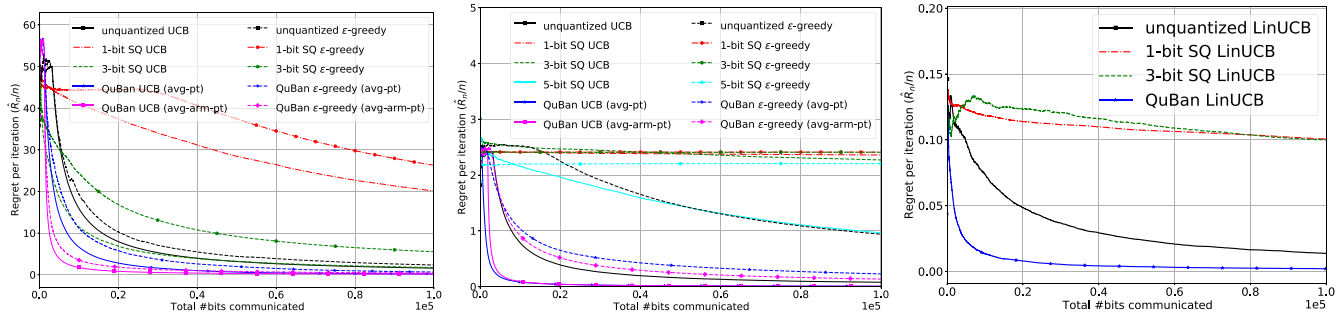
• *Setup 3 (Figs. 4-6(c)):* This is our contextual bandit setup. We use $d = 20$ dimensions, θ_* picked uniformly at random from the surface of a radius 1 ball centered at the origin, and the noise η_t is picked from a Gaussian distribution with zero mean and 0.1 variance. At each time t we construct the actions set $\mathcal{A}_t$ by sampling 5 actions uniformly at random from the surface of a radius 0.5 ball centered at the origin independently of the previously sampled actions. We evaluate the regret and the average number of bits used by *QuBan* as well as the 3 and 1 bit stochastic quantizers in the interval $[-10, 10]$ (the interval in which we observe the majority of rewards). These quantization schemes are used on top of the LinUCB algorithm. The LinUCB exploration constant is chosen to be σ_q, where σ_q is set to be 0.1 for *QuBan* and $\frac{20}{2^r-1}$ for the r-bit SQ.

Results: Fig. 4 plots the regret R'_n in (1) vs. the number of iterations, Fig. 5 plots $\frac{\hat{R}_n}{n}$, the regret per iteration, vs. the total number of bits communicated, Fig. 6 plots the regret versus number of iterations, and Fig. 7 plots the average number of bits versus iterations. We find that:



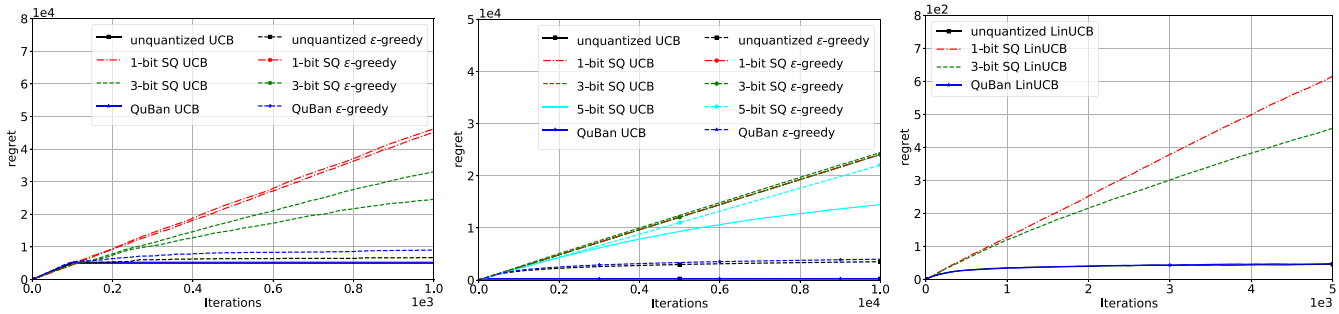
(a) Setup 1 (larger Δ_i values). (b) Setup 2 (smaller Δ_i values). (c) Setup 3 (linear bandits).

Fig. 4. Regret versus number of iterations.



(a) Setup 1 (larger Δ_i values). (b) Setup 2 (smaller Δ_i values). (c) Setup 3 (linear bandits).

Fig. 5. Total number of bits versus regret per iteration.



(a) Setup 1 (larger Δ_i values). (b) Setup 2 (smaller Δ_i values). (c) Setup 3 (linear bandits).

Fig. 6. Regret versus number of iterations.

• *QuBan* in all three setups offers minimal or no regret increase compared to the unquantized rewards regret and achieves savings of tens of thousands of bits as compared to unquantized communication.

• 1-bit SQ significantly diverges in most cases; 3-bit and 5-bit SQ show better performance yet still not matching *QuBan* with a performance gap that increases when the arms means are closer (Δ_i smaller), and hence, more difficult to distinguish.

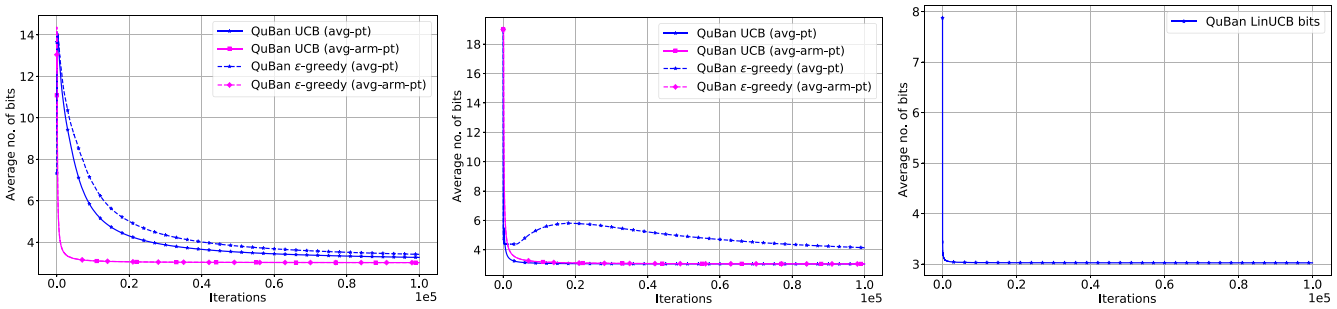
• *QuBan* allows for more than 10x saving in the number of bits over the unquantized case to achieve the same regret. In all three setups *QuBan* achieves $\mathbb{E}[\bar{B}(n)] \approx 3$ (see Fig. 7).

• Both *QuBan* avg-pt and avg-arm-pt achieve the same regret (they are not distinguishable in Fig. 4 and thus we use a common legend), yet avg-arm-pt uses a smaller number of bits when the means of the arms tend to be well separated (Fig. 5(a)) while avg-pt uses a smaller number of bits when they tend to be closer together (Fig. 5(b)). We also observe that the avg-pt tends to perform better for a well-behaved bandit scheme, while the avg-arm-pt performs better when the algorithm picks sub-optimal arms for many iterations (e.g., ϵ-greedy in Fig. 7(b)).

Cryptocurrency Returns Dataset:

MAB Setup: In this part we compare the performance of our scheme against 3-bit SQ using multiple cryptocurrencies prices from binance.com in October 2021, where the reward is the investment return. The action represents which cryptocurrency, among {Bitcoin, Ethereum, Dogecoin, and Litecoin}, to buy then sell on the next day. The return for each currency is samples uniformly at random from the daily returns in the month of October 2021.

Results: In Fig. 8 we plot the regret (daily return - optimal average return) versus the number of iterations for the UCB algorithm using our quantization scheme and 3-bit SQ. We



(a) Setup 1 (larger Δ_i values). (b) Setup 2 (smaller Δ_i values). (c) Setup 3 (linear bandits).

Fig. 7. Average number of bits versus iterations.

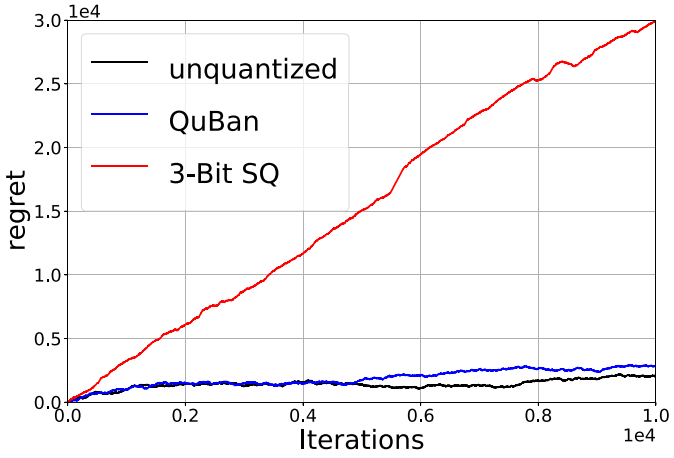


Fig. 8. Regret versus number of iterations for the cryptocurrency prices dataset.

observe that the performance of *QuBan* almost matches the unquantized performance (using ≈ 3 bits) while the regret of 3-bit SQ is linear for the used number of iterations.

VII. Conclusion and Future Work

In this paper we provide a generic framework, *QuBan*, to quantize rewards for MAB problems. This framework can be used on top of nearly all the existing and future MAB algorithms, making them attractive for distributed learning applications where communication can become a bottleneck. We have demonstrated that, both in theory and by numerical experiments, *QuBan* can provide very significant savings in terms of communication and barely affects the learning performance.

We identify several future research directions: (1) How to exploit memory? In the setup we consider, the remote agents are changing over time, and thus they are essentially memoryless, i.e., a new agent does not know the history information of previous agents. (2) How to deal with heavy tailed noise? (3) How to convey contexts in the contextual bandit setting if these are not implicitly conveyed? Resolving such questions can offer additional benefits for communication-sensitive bandit learning setups.

Appendix A
Proof of Proposition 1

Proof: We start by proving that $\hat{r}_t$ is an unbiased estimate of μ_{A_t}. If $-3 \leq r_t \leq -4$, we have that $\hat{r}_t$ takes the value $\lceil r_t \rceil$ with probability $r_t - \lfloor r_t \rfloor$, and the value $\lfloor r_t \rfloor$ with probability $\lceil r_t \rceil - r_t$. Hence, $\mathbb{E}[\hat{r}_t | r_t] = r_t$. For all the other cases we have that

$$\begin{aligned}
\mathbb{E}\big[\hat{r}_t|r_t\big] &= \mathbb{E}\Bigg[M_t\bigg(s_t\Big(e_t^{(q)} + \ell_t + 3.5s_t + 0.5\Big) + \left\lfloor \frac{\hat{\mu}(t)}{M_t} \right\rfloor\bigg)|r_t\Bigg] \\
&= \mathbb{E}\Bigg[M_t\mathbb{E}\Bigg[s_t\Big(e_t^{(q)} + \ell_t + 3.5s_t + 0.5\Big) + \left\lfloor \frac{\hat{\mu}(t)}{M_t} \right\rfloor \\
&\qquad + s_t\ell_t | r_t, \hat{\mu}(t), M_t\Bigg]|r_t\Bigg] \\
&\overset{(i)}{=} \mathbb{E}\Bigg[M_t\bigg(\frac{r_t}{M_t} - \left\lfloor \frac{\hat{\mu}(t)}{M_t} \right\rfloor + s_t(-[3.5s_t + 0.5 + \ell_t] \\
&\qquad + s_t\ell_t + 3.5s_t + 0.5) + \left\lfloor \frac{\hat{\mu}(t)}{M_t} \right\rfloor\bigg)|r_t\Bigg] \\
&= r_t, \qquad (7)
\end{aligned}$$

where (i) follows from the fact that the stochastic quantization (SQ) that we use gives an unbiased estimate of the input. We note that from Algorithm 2, e_t encodes $|\bar{r}_t| - (|a| + 1) - \ell_t$, where $|a|+1 = 3$ when $s_t = -1$, and $|a|+1 = 4$ when $s_t = 1$, i.e., $|a| + 1 = 3.5s_t + 0.5$. Hence, in all cases we have that

$$\begin{aligned}
\mathbb{E}\big[\hat{r}_t|A_t\big] &= \mathbb{E}\big[\mathbb{E}\big[\hat{r}_t|r_t, A_t\big]|A_t\big] = \mathbb{E}\big[\mathbb{E}\big[\hat{r}_t|r_t\big]|A_t\big] \\
&= \mathbb{E}[r_t|A_t] = \mu_{A_t} \qquad (8)
\end{aligned}$$

The bound on $|r_t - \hat{r}_t|$ follows from the fact that the distance between the quantization levels for which we use the randomized quantization is 1, hence, in all cases we have that

$$1 \geq |s_t e_t^{(q)} - \left(\frac{r_t}{M_t} - \left\lfloor \frac{\hat{\mu}(t)}{M_t} \right\rfloor - s_t\ell_t\right)| = \frac{|\hat{r}_t - r_t|}{M_t}. \qquad (9)$$

We note that this implies

$$\begin{aligned}
\mathbb{E}\Big[|\hat{r}_t - \mu_{A_t}|^2|A_t\Big] &= \mathbb{E}\Big[|\hat{r}_t - r_t + r_t - \mu_{A_t}|^2|A_t\Big] \\
&= \mathbb{E}\Big[|\hat{r}_t - r_t|^2|A_t\Big] + \mathbb{E}\Big[|r_t - \mu_{A_t}|^2|A_t\Big] \\
&\quad + 2\mathbb{E}\big[\big(r_t - \mu_{A_t}\big)\big(\hat{r}_t - r_t\big)|A_t\big] \\
&\leq \big(1 + \epsilon^2\big)\sigma^2 \\
&\quad + 2\mathbb{E}\big[\big(r_t - \mu_{A_t}\big)\mathbb{E}\big[\big(\hat{r}_t - r_t\big)|A_t, r_t\big]|A_t\big] \\
&= \big(1 + \epsilon^2\big)\sigma^2. \qquad (10)
\end{aligned}$$

To see that conditioned on A_t, $\hat{r}_t$ is conditionally independent on the history $A_1, \hat{r}_1, \ldots, A_{t-1}, \hat{r}_{t-1}$, we notice that since we replace $\frac{\hat{\mu}(t)}{M_t}$ by an integer, $\lfloor \frac{\hat{\mu}(t)}{M_t} \rfloor$ and since the distance

between the quantization levels is 1, we have that the two nearest quantization levels to $\frac{r_t}{M_t}$ are at $\lfloor \frac{r_t}{M_t} \rfloor, \lceil \frac{r_t}{M_t} \rceil$. Hence, conditioned on M_t, $\hat{r}_t$ takes the value $M_t \lceil \frac{r_t}{M_t} \rceil$ with probability $\frac{r_t}{M_t} - \lfloor \frac{r_t}{M_t} \rfloor$, and the value $M_t \lfloor \frac{r_t}{M_t} \rfloor$ with probability $\lceil \frac{r_t}{M_t} \rceil - \frac{r_t}{M_t}$. This shows that despite the fact that the encoding of $\hat{r}_t$ is a function of $r_1, \ldots, r_t$, the value of $\hat{r}_t$ is a function of r_t only, since M_t is generated independently of the history. As a result, given A_t, $\hat{r}_t$ is conditionally independent on the history $A_1, \hat{r}_1, \ldots, A_{t-1}, \hat{r}_{t-1}$.

The fact that $\hat{r}_t$ is subGaussian can be proven by Cauchy-Schwartz

$$\begin{aligned}\mathbb{E}\Big[e^{\lambda(\hat{r}_t-\mu_{A_t})}|A_t\Big] &= \mathbb{E}\Big[e^{\lambda(\hat{r}_t-r_t+r_t-\mu_{A_t})}|A_t\Big] \\ &\le \mathbb{E}\Big[e^{p\lambda(\hat{r}_t-r_t)}|A_t\Big]^{\frac{1}{p}} \\ &\quad \mathbb{E}\Big[e^{(1-p)\lambda(r_t-\mu_{A_t})}|A_t\Big]^{\frac{1}{1-p}} \\ &\le e^{\lambda^2 \frac{\sigma^2(1+\frac{\epsilon}{2})^2}{2}}, \end{aligned} \tag{11}$$

where $p = 1 + \frac{2}{\epsilon}$. To bound the expected regret after quantization we observe that $R_n = \sum_{t=1}^n \mathbb{E}(\mu_t^* - r_t) = \sum_{t=1}^n \mathbb{E}(\mu_t^* - \hat{r}_t) = (1+\frac{\epsilon}{2}) \sum_{t=1}^n \mathbb{E}(\frac{\mu_t^* - \hat{r}_t}{1+\frac{\epsilon}{2}})$. We have that $\frac{\hat{r}_t}{(1+\frac{\epsilon}{2})}$ is σ^2-subGaussian. Applying the bandit algorithm using $\frac{\hat{r}_t}{(1+\frac{\epsilon}{2})}$ results in $\sum_{t=1}^n \mathbb{E}(\frac{\mu_t^* - \hat{r}_t}{(1+\frac{\epsilon}{2})}) \le R_n^U(\{\Delta_i/(1+\frac{\epsilon}{2})\})$, hence

$$R_n \le \Big(1+\frac{\epsilon}{2}\Big) R_n^U\Big(\Big\{\Delta_i/\Big(1+\frac{\epsilon}{2}\Big)\Big\}\Big). \tag{12}$$

■

APPENDIX B
PROOF OF THEOREM 1

Proof: We have that B_t can be bounded as

$$\begin{aligned} B_t &\le 3 + \mathbf{1}\Big[\frac{r_t}{M_t} - \Big\lfloor \frac{\hat{\mu}(t)}{M_t} \Big\rfloor > 3\Big] + \mathbf{1}\Big[\Big\lfloor \frac{\hat{\mu}(t)}{M_t} \Big\rfloor - \frac{r_t}{M_t} > 2\Big] \\ &\quad + 2\Big(\mathbf{1}\Big[\frac{r_t}{M_t} - \Big\lfloor \frac{\hat{\mu}(t)}{M_t} \Big\rfloor > 4\Big]\Big[\log\Big(\frac{r_t}{M_t} - \Big\lfloor \frac{\hat{\mu}(t)}{M_t} \Big\rfloor - 3\Big)\Big]\Big) \\ &\quad + 2\Big(\mathbf{1}\Big[\Big\lfloor \frac{\hat{\mu}(t)}{M_t} \Big\rfloor - \frac{r_t}{M_t} > 3\Big]\Big[\log\Big(\Big\lfloor \frac{\hat{\mu}(t)}{M_t} \Big\rfloor - \frac{r_t}{M_t} - 2\Big)\Big]\Big) \\ &\le 3 + \mathbf{1}\Big[|\frac{r_t}{M_t} - \frac{\hat{\mu}(t)}{M_t}| > 2\Big] + 2\Big(\mathbf{1}\Big[|\frac{r_t}{M_t} - \frac{\hat{\mu}(t)}{M_t}| > 3\Big]\Big) \\ &\quad + 2\Big(\mathbf{1}\Big[|\frac{r_t}{M_t} - \frac{\hat{\mu}(t)}{M_t}| > 3\Big]\log\Big(|\frac{r_t}{M_t} - \frac{\hat{\mu}(t)}{M_t}| - 2\Big)\Big). \end{aligned} \tag{13}$$

Hence for each $\delta > 0$, we have

$$\begin{aligned} B_t &\le 3 + \mathbf{1}\Big[|\frac{r_t - \mu_{A_t}}{\sigma}| > 2(1-\delta)\Big] + \mathbf{1}\Big[|\frac{\mu_{A_t} - \hat{\mu}(t)}{\sigma}| > 2\delta\Big] \\ &\quad + 2\Big(\mathbf{1}\Big[|\frac{r_t - \mu_{A_t}}{\sigma}| > 3(1-\delta)\Big] + \mathbf{1}\Big[|\frac{\mu_{A_t} - \hat{\mu}(t)}{\sigma}| > 3\delta\Big]\Big) \\ &\quad + 2\Big(\mathbf{1}\Big[|\frac{r_t - \mu_{A_t}}{\sigma}| > 3\Big]\Big)\log\Big(|\frac{r_t - \hat{\mu}(t)}{\sigma}| - 2\Big). \end{aligned} \tag{14}$$

Taking the expectation of both sides, we get that

$$\begin{aligned} \mathbb{E}[B_t] &\le 3 + \mathbb{P}\Big[|\frac{r_t - \mu_{A_t}}{\sigma}| > 2(1-\delta)\Big] \\ &\quad + \mathbb{P}\Big[|\frac{\mu_{A_t} - \hat{\mu}(t)}{\sigma}| > 2\delta\Big] \\ &\quad + 2\Big(\mathbb{P}\Big[|\frac{r_t - \mu_{A_t}}{\sigma}| > 3(1-\delta)\Big] \\ &\quad + \mathbb{P}\Big[|\frac{\mu_{A_t} - \hat{\mu}(t)}{\sigma}| > 3\delta\Big]\Big) \\ &\quad + 2\mathbb{E}\Big[\Big(\mathbf{1}\Big[|\frac{r_t - \mu_{A_t}}{\sigma}| > 3\Big]\Big)\log\Big(|\frac{r_t - \hat{\mu}(t)}{\sigma}| - 2\Big)\Big]. \end{aligned} \tag{15}$$

Hence, there are universal constants C, C' such that

$$\begin{aligned} \mathbb{E}[B_t] &\le 3.32 + C'\mathbb{E}\Big[|\frac{\mu_{A_t} - \hat{\mu}(t)}{\sigma}|\Big] \\ &\quad + 2\mathbb{E}\Big[\mathbf{1}\Big[|\frac{r_t}{M_t} - \frac{\hat{\mu}(t)}{M_t}| > 3\Big]\Big(|\frac{r_t}{M_t} - \frac{\hat{\mu}(t)}{M_t}| - 3\Big)\Big] \\ &\le 3.32 + C'\mathbb{E}\Big[|\frac{\mu_{A_t} - \hat{\mu}(t)}{\sigma}|\Big] \\ &\quad + 2\mathbb{E}\Big[\mathbf{1}\Big[|\frac{r_t - \mu_{A_t}}{\sigma}| > 3(1-\delta)\Big]||\frac{r_t - \mu_{A_t}}{\sigma}| - 3|\Big] \\ &\quad + 2\mathbb{E}\Big[\mathbf{1}\Big[|\frac{\mu_{A_t} - \hat{\mu}(t)}{\sigma}| > 3\delta\Big]||\frac{r_t - \mu_{A_t}}{\sigma}| - 3|\Big] \\ &\quad + 2\mathbb{E}\Big[|\frac{\mu_{A_t} - \hat{\mu}(t)}{\sigma}|\Big] \\ &\le 3.32 + \big(C' + 2\big)\mathbb{E}\Big[|\frac{\mu_{A_t} - \hat{\mu}(t)}{\sigma}|\Big] \\ &\quad + 2\sum_{i=3}^{\infty} |i(1-\delta) - 3|\mathbb{P}\Big[|\frac{\mu_{A_t} - \hat{\mu}(t)}{\sigma}| > i(1-\delta)\Big] \\ &\quad + 2\mathbb{E}\Big[\mathbf{1}\Big[|\frac{\mu_{A_t} - \hat{\mu}(t)}{\sigma}| > 3\delta\Big]\Big]\mathbb{E}\Big[||\frac{r_t - \mu_{A_t}}{\sigma}| - 3|\Big] \\ &\quad + 2\mathbb{E}\Big[|\frac{\mu_{A_t} - \hat{\mu}(t)}{\sigma}|\Big] \\ &\le 3.4 + C\mathbb{E}\Big[|\frac{\mu_{A_t} - \hat{\mu}(t)}{\sigma}|\Big] \end{aligned} \tag{16}$$

From (14), $\mathbb{E}[|r_t - \mu_{A_t}|^2|A_t] \le \sigma^2$, Markov property and the strong law of large numbers for martingales, we also have that there is a universal constant C such that

$$\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^n B_t \le 3.4 + \lim_{n\to\infty} \tfrac{C}{n}\textstyle\sum_{t=1}^n |\frac{\mu_{A_t} - \hat{\mu}(t)}{\sigma}| \text{ almost surely.} \tag{17}$$

It then remains to analyze $|\mu_{A_t} - \hat{\mu}(t)|$ for the three proposed choices of $\hat{\mu}(t)$.

- **avg-pt** ($\hat{\mu}(t) = \frac{1}{t-1}\sum_{j=1}^{t-1} \hat{r}_j$)**:**

We have that for $t > 1$

$$\begin{aligned} \frac{|\mu_{A_t} - \hat{\mu}(t)|}{\sigma} &\le \frac{|\mu_{A_t} - \mu^*|}{\sigma} + \frac{|\mu^* - \hat{\mu}(t)|}{\sigma} \\ &= \frac{\Delta_{A_t}}{\sigma} + |\frac{\sum_{j=1}^{t-1} \mu^* - \mu_{A_j} + \mu_{A_j} - \hat{r}_j}{(t-1)\sigma}| \\ &\le \frac{\Delta_{A_t}}{\sigma} + |\frac{\sum_{j=1}^{t-1} \mu^* - \mu_{A_j}}{(t-1)\sigma}| \\ &\quad + |\frac{\sum_{j=1}^{t-1} \mu_{A_j} - \hat{r}_j}{(t-1)\sigma}| \\ &= \frac{\Delta_{A_t}}{\sigma} + \frac{\sum_{i=1}^{k} \Delta_i T_i(t-1)}{(t-1)\sigma} \\ &\quad + |\frac{\sum_{j=1}^{t-1} \mu_{A_j} - \hat{r}_j}{(t-1)\sigma}|. \end{aligned} \tag{18}$$

For $t=1$ we have

$$\frac{|\mu_{A_t}-\hat{\mu}(t)|}{\sigma} \le \frac{|\mu_{A_t}-\mu^*|}{\sigma}+\frac{|\mu^*-\hat{\mu}(t)|}{\sigma} = \frac{\Delta_{A_1}}{\sigma}+\frac{|\mu^*|}{\sigma}. \tag{19}$$

We then have that

$$\begin{aligned}\frac{1}{n}\sum_{t=1}^{n}\log\Big(1+|\frac{\mu_{A_t}-\hat{\mu}(t)}{\sigma}|\Big) &\le \frac{\log\big(1+\frac{|\mu^*|}{\sigma}\big)}{n\sigma} \\ &+\frac{1}{n}\sum_{t=1}^{n}\log\Big(1+\frac{\Delta_{A_t}}{\sigma}\Big) \\ &+\frac{1}{n}\sum_{t=2}^{n}\log\Big(1+\frac{\sum_{i=1}^{k}\Delta_i T_i(t-1)}{(t-1)\sigma}\Big) \\ &+\log\Big(1+|\frac{\sum_{j=1}^{t-1}\mu_{A_j}-\hat{r}_j}{(t-1)\sigma}|\Big) \\ &\le \frac{\log\big(1+\frac{|\mu^*|}{\sigma}\big)}{n\sigma}+\frac{1}{n}\Bigg(\frac{\sum_{i=1}^{k}\Delta_i T_i(n)}{\sigma} \\ &+\sum_{t=1}^{n-1}\frac{\sum_{i=1}^{k}\Delta_i T_i(t)}{t\sigma}+|\frac{\sum_{j=1}^{t}\mu_{A_j}-\hat{r}_j}{t\sigma}|\Bigg).\end{aligned} \tag{20}$$

We have that since $\mathbb{E}[|r_t-\mu_{A_t}|^2|A_t]\le\sigma^2$, and Markov property, then by the strong law of large numbers for martingales $\lim_{t\to\infty}\frac{\sum_{j=1}^{t-1}\mu_{A_j}-\hat{r}_j}{(t-1)\sigma}=0$ almost surely. We then have that if the limit of average regret is 0 almost surely (or in probability), then from (17) and (20) we get that

$$\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}B_t\le 3.4 \text{ almost surely (or in probability).} \tag{21}$$

By observing that we can generate a long sequence of rewards from each arm before the process starts and since $\mathbb{E}[|r_t-\mu_{A_t}|^2|A_t]\le\sigma^2$, then by the triangle inequality we have that

$$\begin{aligned}\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\Bigg[|\frac{\sum_{j=1}^{t-1}\mu_{A_j}-\hat{r}_j}{(t-1)\sigma}|\Bigg] &\overset{(i)}{\le} \frac{2}{n}\sum_{t=1}^{n}\frac{1}{\sqrt{t}} \\ &= \frac{2}{n}\sum_{t=1}^{n}\frac{1}{\sqrt{t}} \\ &\le \frac{2}{n}\Big(1+\int_{t=1}^{n}\frac{1}{\sqrt{t}}dt\Big) \\ &\le \frac{4}{\sqrt{n}},\end{aligned} \tag{22}$$

where (i) follows from the fact that $\mu_{A_j}-\hat{r}_j, \mu_{A_i}-\hat{r}_i$ are uncorrelated for all $i<j$ since

$$\begin{aligned}&\mathbb{E}\big[\big(\mu_{A_j}-\hat{r}_j\big)\big(\mu_{A_i}-\hat{r}_i\big)\big] \\ &\quad=\mathbb{E}\big[\mathbb{E}\big[\big(\mu_{A_j}-\hat{r}_j\big)\big(\mu_{A_i}-\hat{r}_i\big)|A_j,A_i,\hat{r}_i\big]\big]=0.\end{aligned} \tag{23}$$

We conclude that there is a universal constant C such that

$$\hat{B}(n)\le 3.4+(C/n)\Big(1+\log\big(1+|\mu^*|/\sigma\big)+R_n/\sigma+\sum_{t=1}^{n-1}R_t/(\sigma t)\Big)+C/\sqrt{n} \tag{24}$$

- **avg-arm-pt** ($\hat{\mu}(t)=\hat{\mu}_{A_t}(t-1)$)**:**

We have that for $T_{A_t}(t-1)>0$

$$\frac{|\mu_{A_t}-\hat{\mu}(t)|}{\sigma}=\Bigg|\frac{\sum_{j=1}^{t-1}\big(\mu_{A_t}-\hat{r}_j\big)\mathbf{1}\big(A_j=A_t\big)}{T_{A_t}(t-1)\sigma}\Bigg|. \tag{25}$$

For $T_{A_t}(t-1)=0$, we have that $\hat{\mu}(t)=0$. Then

$$\begin{aligned}&\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\Bigg[\log\Big(1+\frac{|\mu_{A_t}-\hat{\mu}(t)|}{\sigma}\Big)\Bigg] \\ &\overset{(i)}{\le}\frac{1}{n}\sum_{i=1}^{k}\log\Big(1+\frac{|\mu_i|}{\sigma}\Big)+\frac{2}{n}\sum_{t=1}^{n}\frac{1}{\sqrt{t}} \\ &\overset{(ii)}{\le}\frac{1}{n}\sum_{i=1}^{k}\log\Big(1+\frac{|\mu_i|}{\sigma}\Big)+\frac{4}{\sqrt{n}}\end{aligned} \tag{26}$$

where (ii) is as in (22), and (i) can be seen by observing that we can generate a long sequence of rewards from each arm before the process starts, from the fact that $\hat{r}_j-\mu_{A_j}, \hat{r}_i-\mu_{A_i}$ are uncorrelated for all $i\ne j$ and since $\mathbb{E}[|r_t-\mu_{A_t}|^2|A_t]\le\sigma^2$.

We conclude that there is a universal constant C such that

$$\hat{B}(n)\le 3.4+\frac{C}{n}\sum_{i=1}^{k}\log\Big(1+\frac{|\mu_i|}{\sigma}\Big)+\frac{C}{\sqrt{n}}. \tag{27}$$

The fact that $\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}B_t\le 3.4$ almost surely, can be seen using the strong law of large numbers by observing that we can generate a long sequence of rewards from each arm before the process starts, the number of arms is finite, and if $\lim_{n\to\infty}T_i(n)<\infty$ then the contribution of arm i in the number of bits decays to zero almost surely as $n\to\infty$.

- **stochastic linear bandits** ($\hat{\mu}(t)=\langle\theta_t,A_t\rangle$)**:**

The results follow directly from (14), (16), (17) and choice of $\hat{\mu}(t)$.

For the case where $\epsilon\ne 1$, it is easy to see that for small values of ϵ, the number of transmitted bits increases by $2\log(\frac{1}{\epsilon})$ bits. This can be further decreased to $\log(\frac{1}{\epsilon})+\log(\log(\frac{1}{\epsilon}))$ bits using the encoding in Section IV. ■

Appendix C
Proof of the High Probability Bound

From Section IV, we have that

$$\begin{aligned}B_t &\le 3+\mathbf{1}\Big[\frac{r_t}{M_t}-\Big\lfloor\frac{\hat{\mu}(t)}{M_t}\Big\rfloor>3\Big]+\mathbf{1}\Big[\Big\lfloor\frac{\hat{\mu}(t)}{M_t}\Big\rfloor-\frac{r_t}{M_t}>2\Big] \\ &+\mathbf{1}\Big[\frac{r_t}{M_t}-\Big\lfloor\frac{\hat{\mu}(t)}{M_t}\Big\rfloor>4\Big]\Bigg(\Bigg\lceil\log\Big(\frac{r_t}{M_t}-\Big\lfloor\frac{\hat{\mu}(t)}{M_t}\Big\rfloor-3\Big)\Bigg\rceil \\ &+\Bigg\lceil\log\Big(\log\Big(\frac{r_t}{M_t}-\Big\lfloor\frac{\hat{\mu}(t)}{M_t}\Big\rfloor-3\Big)\Big)\Bigg\rceil\Bigg) \\ &+\mathbf{1}\Big[\Big\lfloor\frac{\hat{\mu}(t)}{M_t}\Big\rfloor-\frac{r_t}{M_t}>3\Big]\Bigg(\Bigg\lceil\log\Big(\Big\lfloor\frac{\hat{\mu}(t)}{M_t}\Big\rfloor-\frac{r_t}{M_t}-2\Big)\Bigg\rceil \\ &+\Bigg\lceil\log\Big(\log\Big(\Big\lfloor\frac{\hat{\mu}(t)}{M_t}\Big\rfloor-\frac{r_t}{M_t}-2\Big)\Big)\Bigg\rceil\Bigg) \\ &\le 4+\log\Big(\frac{\hat{\mu}(t)}{\sigma}-\frac{r_t}{\sigma}-2\Big)+\log\Big(\log\Big(\frac{\hat{\mu}(t)}{\sigma}-\frac{r_t}{\sigma}-2\Big)\Big).\end{aligned} \tag{28}$$

Let the event G be that $\forall t \in \{1,\dots,n\} : |r_t - \mu_{A_t}| \leq \sigma\sqrt{4\log(n)}$. From the subGaussian assumption and applying the union bound we have that

$$\mathbb{P}[G] > 1 - \sum_{t=1}^{n} e^{-2\log(n)}. \tag{29}$$

We have that if G holds then for t with $T_t(A_t) > 0$, we have that $|\hat{\mu}(t) - \mu_{A_t}| \leq \sigma$, $|r_t - \mu_{A_t}| \leq \sigma$. Hence, $|\hat{\mu}(t) - r_t| \leq 2\sigma$. Substituting in (28), we get the desired result.

Appendix D
Proof of Lower Bound (Lemma 1 and Theorem 3)

A. Proof of Lemma 1

Proof: To simplify notation, we omit the time index t and only mention it when it is necessary.

Let P, P' denote reward distributions with means μ_1 and μ_2, respectively. We have that, for any given quantizer Q, either:

Case 1: $\forall\ P, P'$ with $\mu_1 \neq \mu_2$, we have that $\mathbb{E}_P[Q(r)] \neq \mathbb{E}_{P'}[Q(r)]$; or

Case 2: $\exists\ P, P'$ with $\mu_1 \neq \mu_2$, and $\mathbb{E}_P[Q(r)] = \mathbb{E}_{P'}[Q(r)]$.

We will first show that any quantizer Q satisfying **Case 1** must saisfy $\mathbb{E}[Q(r)|r] = c_1 r + c_2$ for some constants c_1, c_2. To do so, we first construct distributions P and P' as follows. Let $\{x_i, p_i, p_i'\}_{i=1}^3$ be real values such that $x_1 \neq x_2$, $\sum_{i=1}^3 p_i = \sum_{i=1}^3 p_i' = 1$ and $p_i,\ p_i' \geq 0,\ \forall\ i \in \{1,2,3\}$. We design P to be the distribution of a random variable that takes the value x_i with probability p_i, and P' be the distribution of a random variable that takes the value x_i with probability p_i' for $i = 1, 2, 3$.

For Case 1, it is necessary that $\mathbb{E}_P[Q(r)] = \mathbb{E}_{P'}[Q(r)]$ *only if* $\sum_{i=1}^3 p_i x_i = \sum_{i=1}^3 p_i' x_i$. Or equivalently,

$$\sum_{i=1}^{3}(p_i - p_i')\mathbb{E}[Q(r)|r = x_i] = 0 \text{ only if } \sum_{i=1}^{3}(p_i - p_i')x_i = 0. \tag{30}$$

This implies that the right null space of the matrix

$$\mathbf{E} = \begin{bmatrix} \mathbb{E}[Q(r)|r=x_1] & \mathbb{E}[Q(r)|r=x_2] & \mathbb{E}[Q(r)|r=x_3] \\ 1 & 1 & 1 \end{bmatrix}$$

is subset of the right null space of the matrix

$$\mathbf{X} = \begin{bmatrix} x_1 & x_2 & x_3 \\ 1 & 1 & 1 \end{bmatrix}$$

(note that $\sum_i (p_i - p_i') = 0$). This is because for any vector $\mathbf{a}$ in the nullspace of $\mathbf{E}$, there exist vectors $\mathbf{p}, \mathbf{p}'$ such that $\mathbf{p}, \mathbf{p}' \geq 0$, $\mathbf{1}^\top \mathbf{p} = \mathbf{1}^\top \mathbf{p}' = 1$, and $\mathbf{a} = c(\mathbf{p} - \mathbf{p}')$ for some constant c; in particular, $\mathbf{p} = \frac{\mathbf{a}^+}{\mathbf{1}^\top \mathbf{a}^+}$, $\mathbf{p}' = \frac{|\mathbf{a}^-|}{\mathbf{1}^\top |\mathbf{a}^-|}$, where $\mathbf{a}^+$ is the same as $\mathbf{a}$ with the negative entries replaced by zeros, while in $\mathbf{a}^-$ the positive entries of $\mathbf{a}$ are replaced by zeros. Note that $\mathbf{1}^\top \mathbf{a}^+ = \mathbf{1}^T |\mathbf{a}^-|$ since $\mathbf{a}$ is in the right null space of $\mathbf{E}$, hence, $\mathbf{1}^\top \mathbf{a} = 0$. Thus, by (30), the same vector $\mathbf{a}$ also belongs in the nullspace of $\mathbf{X}$.

We also observe that since $x_1 \neq x_2$ and $\mathbb{E}[Q(r)|r = x_1] \neq \mathbb{E}[Q(r)|r = x_2]$ (as we assumed in **Case 1**); hence the ranks of $\mathbf{E}$ and $\mathbf{X}$ equal to 2. Therefore the dimension of the right null space of each of the matrices $\mathbf{E}, \mathbf{X}$ is exactly one. This, together with the fact that the right null space of $\mathbf{E}$ is a subset of the right null space of $\mathbf{X}$, imply that the right null spaces of these two matrices are *exactly the same* (and one-dimensional). We note that the right null space of $\mathbf{X}$ includes the vector

$$\mathbf{a} = \begin{bmatrix} \frac{x_3+x_2}{x_1-x_2} \\ \frac{x_3+x_1}{x_2-x_1} \\ 1 \end{bmatrix}.$$

Hence, we have that $\mathbf{E}\mathbf{a} = \mathbf{0}$ which implies that (from the first row of $\mathbf{E}\mathbf{a} = \mathbf{0}$)

$$\mathbb{E}[Q(r)|r = x_3] = \left(\frac{\mathbb{E}[Q(r)|r=x_1] - \mathbb{E}[Q(r)|r=x_2]}{x_2 - x_1}\right)x_3 + \frac{x_2\mathbb{E}[Q(r)|r=x_1] - x_1\mathbb{E}[Q(r)|r=x_2]}{x_2 - x_1}$$

As x_3 was arbitrary, we have that, for all $x \in \mathbb{R}$

$$\mathbb{E}[Q(r)|r = x] = c_1 x + c_2,$$

where $c_1 = \frac{\mathbb{E}[Q(r)|r=x_1] - \mathbb{E}[Q(r)|r=x_2]}{x_1 - x_2}$, $c_2 = \frac{x_2\mathbb{E}[Q(r)|r=x_1] - x_1\mathbb{E}[Q(r)|r=x_2]}{x_1 - x_2}$. This completes the proof for Case 1.

For **Case 2**, if we consider a MAB instance with two arms with distributions P, P' that witness the property in **Case 2**, then even if we have infinite samples from the quantization scheme we cannot achieve better than $O(|\mu_1 - \mu_2|n)$ regret. ∎

B. Proof of Theorem 3

Proof: To simplify notation, we omit the time index t and only mention it when it is necessary. Normalizing the rewards by σ, it suffices to consider the case where $\sigma = 1$.

We first show that it suffices to consider schemes with deterministic quantization levels. Let us consider a quantizer Q with encoder $\mathcal{E} : \mathbb{R} \to \mathbb{N}$ and decoder $D : \mathbb{N} \to \mathbb{R}$, where $\mathcal{E}, D$ can both be random. We note that as D is allowed to be random, the set of quantization levels is now random. Let us consider a new decoder D' defined as

$$D'(i) = \mathbb{E}[D(i)]. \tag{31}$$

We now consider the quantizer Q' defined by $\mathcal{E}, D'$ as an encoder-decoder pair. We note that the decoder D' is a deterministic function, hence, the set of quantization levels for the quantizer Q' is deterministic. We will show that: (a) $\mathbb{E}[Q(r_t)|r_t] = \mathbb{E}[Q'(r_t)|r_t]$ and (b) if Q results in sub-Gaussian quantized rewards conditioned on r_t, then Q' also results in sub-Gaussian quantized rewards conditioned on r_t with the same sub-Gaussian parameter as Q. Properties (a) and (b) will allow us to switch D with D' in the rest of our proofs without affecting the encoder $\mathcal{E}$ (hence without affecting the number of bits). To show $\mathbb{E}[Q(r_t)|r_t] = \mathbb{E}[Q'(r_t)|r_t]$, we observe that

$$\begin{aligned}\mathbb{E}[Q(r_t)|r_t] &= \mathbb{E}[D(\mathcal{E}(r_t))|r_t] = \mathbb{E}[\mathbb{E}[D(i)|r_t, \mathcal{E}(r_t) = i]|r_t] \\ &= \mathbb{E}\big[D'(\mathcal{E}(r_t))|r_t\big] = \mathbb{E}\big[Q'(r_t)|r_t\big].\end{aligned} \tag{32}$$

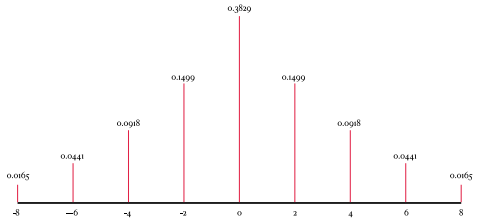


Fig. 9. Illustration of reward distribution.

To show the second property we observe that

$$\begin{aligned}
&\mathbb{E}\big[\exp\left(\lambda(Q(r_t)-\mathbb{E}[Q(r_t)|r_t])\right)|r_t,\mathcal{E}(r_t)\big]\\
&\overset{(i)}{=}\mathbb{E}\big[\exp\left(\lambda\big(Q'(r_t)-\mathbb{E}[Q(r_t)|r_t]\big)\right)|r_t,\mathcal{E}(r_t)\big]\\
&\mathbb{E}\big[\exp\left(\lambda\big(Q(r_t)-Q'(r_t)\big)\right)|r_t,\mathcal{E}(r_t)\big]\\
&\overset{(ii)}{\geq}\mathbb{E}\big[\exp\left(\lambda\big(Q'(r_t)-\mathbb{E}[Q(r_t)|r_t]\big)\right)|r_t,\mathcal{E}(r_t)\big]\\
&\exp\left(\lambda\mathbb{E}\big[\big(Q(r_t)-Q'(r_t)\big)|r_t,\mathcal{E}(r_t)\big]\right)\\
&=\mathbb{E}\big[\exp\left(\lambda\big(Q'(r_t)-\mathbb{E}[Q(r_t)|r_t]\big)\right)|r_t,\mathcal{E}(r_t)\big]\\
&\exp\left(\lambda\mathbb{E}\big[\big(D(\mathcal{E}(r_t))-D'(\mathcal{E}(r_t))\big)|r_t,\mathcal{E}(r_t)\big]\right)\\
&\overset{(iii)}{=}\mathbb{E}\big[\exp\left(\lambda\big(Q'(r_t)-\mathbb{E}[Q(r_t)|r_t]\big)\right)|r_t,\mathcal{E}(r_t)\big]\\
&\overset{(iv)}{=}\mathbb{E}\big[\exp\left(\lambda\big(Q'(r_t)-\mathbb{E}\big[Q'(r_t)|r_t\big]\big)\right)|r_t,\mathcal{E}(r_t)\big] \qquad (33)
\end{aligned}$$

where (i) follows by the fact that $Q'(r_t)=D'(\mathcal{E}(r_t))$ is a deterministic function of $\mathcal{E}(r_t)$, (ii) follows by Jensen's inequality and non-negativity of the exp function, (iii) follows by definition of D', and (iv) follows from (32). By taking the conditional expectation given r_t of both sides in (33) we get that

$$\begin{aligned}
&\mathbb{E}\big[\exp\left(\lambda(Q(r_t)-\mathbb{E}[Q(r_t)|r_t])\right)|r_t\big]\\
&\quad\geq\mathbb{E}\big[\exp\left(\lambda\big(Q'(r_t)-\mathbb{E}\big[Q'(r_t)|r_t\big]\big)\right)|r_t\big]. \qquad (34)
\end{aligned}$$

a) Proof of Statetment 1 & 2 in Theorem 3: To prove 1, 2, we consider the following distribution that takes values on $2z\forall z\in\mathbb{Z}$ with

$$\mathbb{P}[r_t=2z]=\begin{cases}\mathbb{P}\big[\mathcal{N}\big(0,(4\sigma)^2\big)\in[2z,2(z+1)]\big] & \text{if } z>0\\ \mathbb{P}\big[\mathcal{N}\big(0,(4\sigma)^2\big)\in[2(z-1),2z]\big] & \text{if } z<0\\ \mathbb{P}\big[\mathcal{N}\big(0,(4\sigma)^2\big)\in[-2,2]\big] & \text{if } z=0,\end{cases} \qquad (35)$$

where $\mathcal{N}(0,(4\sigma)^2)$ is a random variable with Gaussian distribution with zero mean and standard deviation 4σ. An illustration of the distribution is depicted in Fig. 9.

By construction of the distribution, we have that r_t is $(4\sigma)^2$-subGaussian, since every value in the Gaussian distribution is mapped to one that is closer to the mean in the distribution of r_t. We next prove 1. Suppose towards a contradiction that $\exists b,t$ such that $\mathbb{P}[B_n\leq b]=1\forall n>t$. Pick n arbitrary large, we have that b can describe at most 2^b quantization levels. We note that the maximum distance between any consecutive quantization levels cannot exceed 4, lest there is a reward r, that is in the middle of the two quantization levels, mapped to $\hat{r}$ with $|\hat{r}-r|\geq 2$ almost surely which violates the fact that $\mathbb{P}[|\hat{r}-r|\geq 2|r=z]\leq\exp(-\frac{2^2}{2(\sigma/2)^2})<1$ for some z given by the subGaussian concentration of assumption (ii). Hence, either the interval $(-\infty,-4(2^b+1)]$ or the interval $[4(2^b+1),\infty)$ will have no quantization levels. We assume without loss of generality that the interval $[4(2^b+1),\infty)$ has no quantization levels. Hence, all the values in that interval will be mapped to values in $(-\infty,4(2^b+1))$. We notice that for the described reward, the interval $(-\infty,4(2^b+1))$ has non-zero probability. This contradicts assumption (i) (unbiasedness).

We next prove 2. Let G_t be the event that $|Q(r_t)-r_t|\leq 1$. We observe that by assumption (ii), since $\hat{r}_t-r_t$ is $(\frac{\sigma}{2})^2$-subGaussian, we have that $\mathbb{P}[G_t|r_t]\geq 1-2e^{-2}\geq 0.729$. Let us consider the intervals of the form $[2i-1,2i+1]\forall i\in\mathbb{Z}$. As we proved above, it is sufficient to consider quantization schemes with deterministic quantization levels. Let $\ell_i(t)$ be the minimum length of a quantization level in the interval $[2i-1,2i+1]$. We have that

$$\begin{aligned}
\mathbb{E}[B_t]&=\sum_{i\in\mathbb{Z}}\mathbb{P}[r_t=2i]\mathbb{E}[B_t|r_t=2i]\\
&\geq\sum_{i\in\mathbb{Z}}\mathbb{P}[r_t=2i]\mathbb{P}[G_t|r_t=2i]\mathbb{E}[B_t|r_t=2i,G_t]\\
&\geq\sum_{i\in\mathbb{Z}}\mathbb{P}[r_t=2i]\mathbb{P}[G_t|r_t=2i]\ell_i\\
&\geq\sum_{|i|\leq 4,i\in\mathbb{Z}}\mathbb{P}[r_t=i]\mathbb{P}[G_t|r_t=2i]\ell_i\\
&\geq 0.729\min_{\{\ell_i\}}\sum_{|i|\leq 4,i\in\mathbb{Z}}\mathbb{P}[r_t=2i]\ell_i. \qquad (36)
\end{aligned}$$

We also notice that as the code is prefix free, then if we restrict the code over a subset of quantization levels, it is still prefix free. It follows that the lengths ℓ_i need to satisfy the tree inequality [47], namely, $\sum_{|i|\leq 4,i\in\mathcal{Z}}2^{-\ell_i}\leq 1$. Hence, we have that the code that minimizes (36) is Huffman code [47]. Performing Huffman code with the weights in (36) gives the following code lengths for $\ell_{-4},\ldots,\ell_4$ respectively: $6,5,4,3,1,3,4,4,6$. Substituting in (36) gives $\mathbb{E}[B_t]\geq 0.729\min_{\{\ell_i\}}\sum_{|i|\leq 4,i\in\mathbb{Z}}\mathbb{P}[r_t=i]\ell_i\geq 1.9$. ■

Appendix E
Proofs of Corollaries 1 and 2

The expected regret bounds follow directly from Theorem 1. To bound the average number of bits used for the avg-pt, we only need to bound the decay rate of $\frac{1}{n}\sum_{t=1}^{n-1}\frac{R_t}{\sigma t}$.

Corollary 1: From Theorem 1 and [12], we have that for *QuBan* with UCB, there is a constant C such that $R_n\leq C\sigma\sqrt{kn\log(n)}$. Then,

$$\begin{aligned}
\frac{1}{n}\sum_{t=1}^{n-1}\frac{R_t}{\sigma t}&\leq C\frac{1}{n}\sum_{t=1}^{n}\frac{\sqrt{kt\log(t)}}{t}\\
&\leq\frac{C\sqrt{k\log(n)}}{n}\sum_{t=1}^{n}\frac{1}{\sqrt{t}}\\
&\leq\frac{C\sqrt{k\log(n)}}{n}\left(1+\int_{t=1}^{n}\frac{1}{\sqrt{t}}\right)\\
&\leq C\sqrt{k\log(n)n}. \qquad (37)
\end{aligned}$$

Corollary 2: From Theorem 1 and [12], we have that for *QuBan* with ϵ-greedy, there is a constant C such that $R_n\leq$

$C\sigma k \log(1+\frac{n}{k})$. Then,

$$\begin{aligned}\frac{1}{n}\sum_{t=1}^{n-1}\frac{R_t}{\sigma t} &\le \frac{Ck}{n}\sum_{t=1}^{n-1}\frac{\log(1+t)}{t}\\ &\le \frac{Ck\log(1+n)}{n}\sum_{t=1}^{n-1}\frac{1}{t}\\ &\le \frac{Ck\log(1+n)}{n}\left(1+\int_1^{n-1}\frac{1}{t}\right)\\ &\le \frac{Ck(\log(1+n))^2}{n}. \end{aligned} \tag{38}$$

Appendix F
Proof of Corollary 3

We observe that the LinUCB parameters, β_t, can be chosen such that $\max_{t\in\{1,\dots,n\}}\sup_{a\in\mathcal{A}_t}\langle\theta_t-\theta_*,a\rangle \le \sqrt{\beta_n}$. Hence, by Cauchy–Schwarz we have that

$$\begin{aligned}\sum_{t=1}^{n}|\langle\theta_t-\theta_*,A_t\rangle| &\le \sqrt{n\sum_{t=1}^{n}|\langle\theta_t-\theta_*,A_t\rangle|^2}\\ &\le \sqrt{n\sum_{t=1}^{n}\min\left\{\beta_n,\langle\theta_t-\theta_*,A_t\rangle^2\right\}}. \end{aligned} \tag{39}$$

The proof of the expected regret and average number of bits bounds then follows as in [9, Th. 19.2] using Theorem 1.

References

[1] O. A. Hanna, L. F. Yang, and C. Fragouli, "Solving multi-arm bandit using a few bits of communication," in *Proc. Int. Conf. Artif. Intell. Stat.*, 2022, pp. 1–6.

[2] D. Bouneffouf and I. Rish, "A survey on practical applications of multi-armed and contextual bandits," 2019, *arXiv:1904.10040*.

[3] A. Anandkumar, N. Michael, A. K. Tang, and A. Swami, "Distributed algorithms for learning and cognitive medium access with logarithmic regret," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 4, pp. 731–745, Apr. 2011.

[4] S. Buccapatnam, A. Eryilmaz, and N. B. Shroff, "Multi-armed bandits in the presence of side observations in social networks," in *Proc. 52nd IEEE Conf. Decis. Control*, 2013, pp. 7309–7314.

[5] S. Buccapatnam, A. Eryilmaz, and N. B. Shroff, "Stochastic bandits with side observations on networks," in *Proc. ACM Int. Conf. Meas. Model. Comput. Syst.*, 2014, pp. 289–300.

[6] J. Mary, R. Gaudel, and P. Preux, "Bandits and recommender systems," in *Proc. Int. Workshop Mach. Learn. Optim. Big Data*, 2015, pp. 325–336.

[7] L. Song, C. Fragouli, and D. Shah, "Recommender systems over wireless: Challenges and opportunities," in *Proc. IEEE Inf. Theory Workshop (ITW)*, 2018, pp. 1–5.

[8] K. Ding, J. Li, and H. Liu, "Interactive anomaly detection on attributed networks," in *Proc. 12th ACM Int. Conf. Web Search Data Min.*, 2019, pp. 357–365.

[9] T. Lattimore and C. Szepesvári, *Bandit Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2020.

[10] H. Robbins, "Some aspects of the sequential design of experiments," *Bull. Trans. Amer. Math. Soc.*, vol. 58, no. 5, pp. 527–535, 1952.

[11] F. Anscombe, "Sequential medical trials," *J. Amer. Stat. Assoc.*, vol. 58, no. 302, pp. 365–383, 1963.

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

[13] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, nos. 3–4, pp. 285–294, 1933.

[14] T. L. Lai, "Adaptive treatment allocation and the multi-armed bandit problem," *Ann. Stat.*, vol. 15, no. 3, pp. 1091–1114, 1987.

[15] J.-Y. Audibert and S. Bubeck, "Minimax policies for adversarial and stochastic bandits," in *Proc. COLT*, vol. 7, 2009, pp. 1–122.

[16] R. Degenne and V. Perchet, "Anytime optimal algorithms in stochastic multi-armed bandits," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1587–1595.

[17] R. Wood, R. Nagpal, and G.-Y. Wei, "Flight of the RoboBees," *Sci. Amer.*, vol. 308, no. 3, pp. 60–65, 2013.

[18] V. Dani, T. P. Hayes, and S. M. Kakade, "Stochastic linear optimization under bandit feedback," in *Proc. 21st Annu. Conf. Learn. Theory*, 2008, pp. 355–366.

[19] L. Li, W. Chu, J. Langford, and R. E. Schapire, "A contextual-bandit approach to personalized news article recommendation," in *Proc. 19th Int. Conf. World Wide Web*, 2010, pp. 661–670.

[20] D. Russo and B. Van Roy, "Learning to optimize via posterior sampling," *Math. Oper. Res.*, vol. 39, no. 4, pp. 1221–1243, 2014.

[21] S. Agrawal and N. Goyal, "Analysis of Thompson sampling for the multi-armed bandit problem," in *Proc. Conf. Learn. Theory*, 2012, pp. 1–26.

[22] S. Agrawal and N. Goyal, "Further optimal regret bounds for Thompson sampling," in *Proc. Artif. Intell. Stat.*, 2013, pp. 99–107.

[23] M. N. Katehakis and H. Robbins, "Sequential choice from several populations," *Proc. Nat. Acad. Sci.*, vol. 92, no. 19, p. 8584, 1995.

[24] R. Agrawal, "Sample mean based index policies with O(log n) regret for the multi-armed bandit problem," *Adv. Appl. Probab.*, vol. 27, no. 4, pp. 1054–1078, 1995.

[25] Y. Li, Y. Wang, and Y. Zhou, "Nearly minimax-optimal regret for linearly parameterized bandits," in *Proc. Conf. Learn. Theory*, 2019, pp. 2173–2174.

[26] S. Agrawal and N. Goyal, "Thompson sampling for contextual bandits with linear payoffs," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 127–135.

[27] M. Abeille and A. Lazaric, "Linear Thompson sampling revisited," in *Proc. Artif. Intell. Stat.*, 2017, pp. 176–184.

[28] F. Seide, H. Fu, J. Droppo, G. Li, and D. Yu, "1-bit stochastic gradient descent and its application to data-parallel distributed training of speech DNNs," in *Proc. 15th Annu. Conf. Int. Speech Commun. Assoc.*, 2014, pp. 1058–1062.

[29] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, "QSGD: Communication-efficient SGD via gradient quantization and encoding," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1709–1720.

[30] P. Mayekar and H. Tyagi, "RATQ: A universal fixed-length quantizer for stochastic optimization," in *Proc. Int. Conf. Artif. Intell. Stat.*, 2020, pp. 1399–1409.

[31] O. A. Hanna, Y. H. Ezzeldin, C. Fragouli, and S. Diggavi, "Quantization of distributed data for learning," *IEEE J. Sel. Areas Inf. Theory*, vol. 2, no. 3, pp. 987–1001, Sep. 2021.

[32] O. A. Hanna, Y. H. Ezzeldin, T. Sadjadpour, C. Fragouli, and S. Diggavi, "On distributed quantization for classification," *IEEE J. Sel. Areas Inf. Theory*, vol. 1, no. 1, pp. 237–249, May 2020.

[33] E. Even-Dar, S. Mannor, and Y. Mansour, "PAC bounds for multi-armed bandit and Markov decision processes," in *Proc. Int. Conf. Comput. Learn. Theory*, 2002, pp. 255–270.

[34] S. Mannor and J. N. Tsitsiklis, "The sample complexity of exploration in the multi-armed bandit problem," *J. Mach. Learn. Res.*, vol. 5, pp. 623–648, Jun. 2004.

[35] V. Anantharam, P. Varaiya, and J. Walrand, "Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part I: IID rewards," *IEEE Trans. Autom. Control*, vol. IT-32, no. 11, pp. 968–976, Nov. 1987.

[36] S. Shahrampour, A. Rakhlin, and A. Jadbabaie, "Multi-armed bandits in multi-agent networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2017, pp. 2786–2790.

[37] P. Landgren et al., "Distributed multi-agent multi-armed bandits," Ph.D. dissertation, Dept. Mech. Aerosp. Eng., Princeton Univ., Princeton, NJ, USA, 2019.

[38] V. Perchet, P. Rigollet, S. Chassang, and E. Snowberg, "Batched bandit problems," *Ann. Stat.*, vol. 44, no. 2, pp. 660–681, 2016.

[39] H. Esfandiari, A. Karbasi, A. Mehrabian, and V. Mirrokni, "Regret bounds for batched bandits," 2019, *arXiv:1910.04959*.

[40] E. Even-Dar, S. Mannor, Y. Mansour, and S. Mahadevan, "Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems," *J. Mach. Learn. Res.*, vol. 7, pp. 1079–1105, Jun. 2006.

[41] D. Vial, S. Shakkottai, and R. Srikant, "One-bit feedback is sufficient for upper confidence bound policies," 2020, *arXiv:2012.02876*.

[42] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford, U.K.: Oxford Univ. Press, 2013.
[43] J. Langford and T. Zhang, "Epoch-greedy algorithm for multi-armed bandits with side information," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 20, 2007, p. 1.
[44] N. Abe and P. M. Long, "Associative reinforcement learning using linear probabilistic concepts," in *Proc. ICML*, 1999, pp. 3–11.
[45] P. Rusmevichientong and J. N. Tsitsiklis, "Linearly parameterized bandits," *Math. Oper. Res.*, vol. 35, no. 2, pp. 395–411, 2010.
[46] R. M. Gray and T. G. Stockham, "Dithered quantizers," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 805–812, May 1993.
[47] T. M. Cover, *Elements of Information Theory*. New York, NY, USA: Wiley, 1999.

Osama A. Hanna received the B.S. degree in electrical engineering from the Faculty of Engineering, Cairo University in 2014, and the M.S. degree in electrical engineering from Nile University, Egypt, in 2018. He is currently pursuing the Ph.D. degree with the Electrical and Computer Engineering Department, University of California at Los Angeles, Los Angeles (UCLA). His research interests are machine learning, information theory, and algorithms. He received the Award of Excellence from Cairo University in 2014. He received the Masters Fellowship and a Graduate Research Assistantship from Nile University in 2014–2018. He received the Electrical and Computer Engineering Department Fellowship from UCLA in 2018–2019.

Lin F. Yang (Member, IEEE) received the two Ph.D. degrees (in computer science and in physics and astronomy) from Johns Hopkins University. He is an Assistant Professor with the Electrical and Computer Engineering Department, University of California at Los Angeles, Los Angeles. He was a Postdoctoral Fellow with Princeton University. His research focuses on developing and applying fast algorithms for machine learning and data science. He is keen on understanding the fundamental theory and computation limits of optimization for different machine learning problems. His current research focus is on reinforcement learning theory and applications, learning for control, nonconvex optimization, and streaming algorithms. He was a recipient of the Simons' Research Fellowship, Dean Robert H. Roy Fellowship, and JHU MINDS Best Dissertation Award.

Christina Fragouli (Fellow, IEEE) received the B.S. degree in electrical engineering from the National Technical University of Athens, Athens, Greece, and the M.Sc. and Ph.D. degrees in electrical engineering from ULCA. She is a Professor with the Electrical and Computer Engineering Department, University of California at Los Angeles (ULCA). She has worked with the Information Sciences Center, AT&T Labs, Florham Park, NJ, USA, and the National University of Athens. She also visited Bell Laboratories, Murray Hill, NJ, USA, and DIMACS, Rutgers University. From 2006 to 2015, she was an Assistant and an Associate Professor with the School of Computer and Communication Sciences, EPFL, Switzerland. Her current research interests include compression for machine learning applications, coding techniques, wireless networks, and network security. She has served as the 2022 IEEE Information Theory Society President, an Information Theory Society Distinguished Lecturer, and an Associate Editor for IEEE COMMUNICATIONS LETTERS, *Journal on Computer Communication* (Elsevier), IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE TRANSACTIONS ON INFORMATION THEORY, and IEEE TRANSACTIONS ON MOBILE COMMUNICATIONS. She has served in multiple IEEE committees, and received awards for her work.