Solving Multi-Arm Bandit Using a Few Bits of Communication

Osama A. Hanna [1] **Lin F. Yang** [1] **Christina Fragouli** [1]

Abstract

The multi-armed bandit (MAB) problem is an active learning framework that aims to select the best among a set of actions by sequentially observing rewards. Recently, it has become popular for a number of applications over wireless networks, where communication constraints can form a bottleneck. Yet existing works usually fail to address this issue and can become infeasible in certain applications. In this paper, we propose $QuBan$, a generic reward quantization algorithm that applies to any (no-regret) multi-armed bandit algorithm. The modified algorithm requires on average a few (as low as 3) bits to be sent per iteration, yet preserving the same regret as the original algorithm. Our upper bounds apply under mild assumptions on the reward distributions over all current (and future) MAB algorithms, including those used in contextual bandits. We also numerically evaluate the application of $QuBan$ to widely used algorithms such as UCB and ϵ-greedy.

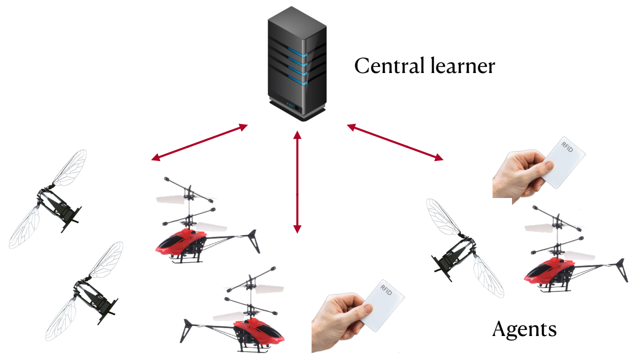


Figure 1. A central learner collects rewards from a set of agents. The agents can join and leave at any time and hence can be different and unaware of the historical rewards, i.e., memoryless.

1. Introduction

Multi-armed bandit (MAB) is an active learning framework that finds applications in diverse domains, including recommendation systems, clinical trials, adaptive routing, and so on (Bouneffouf & Rish, 2019). In a MAB problem, 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 using a 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 spans several decades, cultivating a rich literature that considers a variety of models and algorithmic approaches (Lattimore & Szepesvári, 2020; Robbins, 1952; Anscombe, 1963; Auer et al., 2002a; Thompson, 1933; Lai, 1987). 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?*

Understanding how many bits of communication are really needed, is not only interesting from a theoretical viewpoint, but can also enable the MAB framework to support learning applications in settings that were challenging before. Consider for instance swarms of tiny robots (such as RoboBees and RoboFlies (Wood et al., 2013)), wearable (inside and outside the body) sensors, backscatterer and RFID networks, IoT and embedded systems; generally whenever low complexity sensors cooperate, the communication cost can fast become a performance bottleneck. MAB systems in areas such as mobile healthcare, social decision-making and spectrum allocation have already been implemented in a distributed manner, using limited bandwidth wireless links and simple sensors with low computational power (Anandkumar et al., 2011; Buccapatnam et al., 2013; 2014; Mary et al., 2015; Song et al., 2018; Ding et al., 2019); reducing the number of bits communicated directly translates to reduced power consumption and wireless interference for these systems.

---

*Equal contribution [1]Electrical and Computer Engineering Department, University of California at Los Angeles, Los Angeles, CA 90095 USA. Correspondence to: Osama A. Hanna <ohanna@ucla.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

In this paper we consider the common setup illustrated in Fig. 1, where a central learner can directly communicate with a set of agents. We assume that the agents may change from time to time (e.g., are mobile), but that each agent can pull whichever arm the learner requests it to, observe the reward, and immediately communicate the reward to the learner. For example, the learner could be a "traffic policeman" for small drones that searches best current policies; or a base-station that helps low-capability sensors achieve spectrum sharing. For many existing systems, the learner may have already implemented a MAB algorithm to handle the learning task. Hence our goal is to design a communication protocol such that the rewards are communicated with only a few bits and yet the performance of the original MAB algorithm does not degrade.

Our main contribution is a novel quantization scheme, that we term $QuBan$, tailored to compressing MAB rewards. $QuBan$ 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. $QuBan$ introduces a simple novel rounding trick to guarantee that the quantization error is conditionally independent on 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 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. We provide a set of upper bounds on the average number of bits $\hat{B}_n$ that $QuBan$ needs to achieve the same learning performance as using unquantized rewards. We find that if applied on top of a MAB algorithm with sub-linear regret, then $\hat{B}_n$ is a small constant (as small as 3). We provide empirical studies for a number of MAB algorithms, e.g., UCB and ϵ-greedy. Numerical results corroborate our theoretical findings.

To the best of our knowledge, the proposed model is novel and no scheme from the literature can be used to solve the problem of maintaining a regret that matches the unquantized regret while using a few bits of communication. A review of the literature is provided in App. A.

2. Model and Notation

MAB Framework. We consider a multi-armed bandit (MAB) problem over a horizon of size n (Robbins, 1952). At each iteration $t = 1, ..., n$, 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 P_{A_t} with mean μ_{A_t}. The reward distributions, P_{A_t}, are assumed to be σ^2-subgaussian (Boucheron et al., 2013)[1]. Throughout the paper, we assume a known σ. However, an estimate of σ within a constant factor would suffice. The arm selected at time t depends on the previously selected arms and observed rewards $A_1, r_1, ..., A_{t-1}, r_{t-1}$. The learner is interested in minimizing the expected regret $R_n = \mathbb{E}[\hat{R}_n]$, where $\hat{R}_n$ is the regret defined as

$$\hat{R}_n = \Sigma_{t=1}^{n}(\mu_t^* - r_t), \tag{1}$$

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 selected 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 $\Delta_i := \mu^* - \mu_i$.

In addition to the case where the set of actions is fixed over time, we also consider an important class of bandit problems, contextual bandits (Auer et al., 2002b; Langford & Zhang, 2007; Agrawal & Goyal, 2013b). 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 (Abe & Long, 1999), 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 $\mathcal{A}_1, A_1, r_1, ..., \mathcal{A}_t, A_t, r_t$, the noise η_{t+1} is assumed to be zero mean and σ^2-subgaussian.

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 the 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 can have 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

[1]This assumption is not required for our main results, however, it allows to provide regret bounds for popular MAB algorithms.

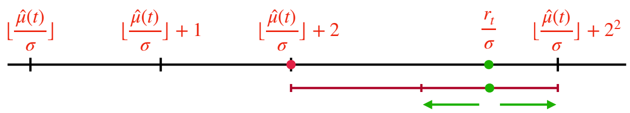


Figure 2. Illustration of $QuBan$. The reward r_t is mapped to a value 2 (conveyed with the index I_t), and stochastically to one of the two nearest quantization levels depicted on the red line.

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 end-to-end operation of a quantizer maps the reward to a quantization level.

Performance Metric $\hat{B}(n)$. Among the schemes that achieve a regret matching the unquantized regret, our performance metric is the average number of communication bits $\hat{B}(n)$ used per reward after n iterations. 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 average number of bits $\hat{B}(n)$.

3. $QuBan$: A MAB Reward Quantizer

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 (such that 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 indicate convergence to ~ 3 bits per iteration for n that is sufficiently large, see App. E).

$QuBan$ builds on the following observations. 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.

3.1. 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 thus 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, a well behaved algorithm will converge to selecting the best arm for the majority of times.

These two choices of $\hat{\mu}(t)$ give us flexibility to fit different regimes of MAB systems as discussed in App. B.

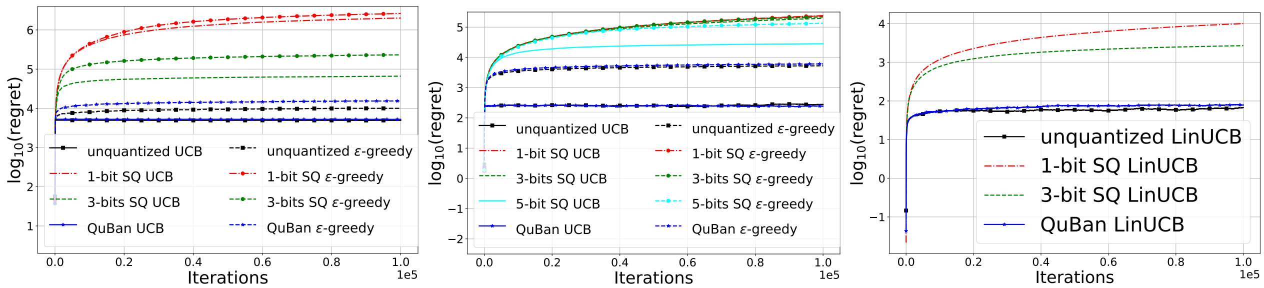
- **Contextual bandit choice**: $\hat{\mu}(t) = \langle\theta_t, A_t\rangle$. Consider the widely used stochastic linear bandits model in Section 2. We observe that linear bandit algorithms, such as contextual Thomson sampling and LinUCB, choose a parameter θ_t believed to be close to the unknown parameter θ_*, and pick an action based on θ_t. Accordingly, we propose to use $\hat{\mu}(t) = \langle\theta_t, A_t\rangle$.

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.

3.2. $QuBan$ Components

At iteration t, $QuBan$ centers its quantization around the value $\hat{\mu}(t)$. It then quantizes the normalized regret $\bar{r}_t = r_t/\sigma - \lfloor\hat{\mu}(t)/\sigma\rfloor$ to one of the two values $\lfloor\bar{r}_t\rfloor, \lceil\bar{r}_t\rceil$. This introduces an error in estimating $\bar{r}_t$ that is bounded by 1, which results in error of at most σ in estimating $r_t = \sigma(\bar{r}_t + \lfloor\hat{\mu}(t)/\sigma\rfloor)$. This quantization is done in a randomized way to convey an unbiased estimate of r_t. More precisely, $QuBan$ transmits the sign of $\bar{r}_t$, and the greatest power of 2 below $|\bar{r}_t|$, call it 2^{I_t} (the handling of the case where $|\bar{r}_t| \le 1$ can be seen in App. B). Then, it quantizes $|\bar{r}_t| - 2^{I_t}$ using a randomized quantizer with levels that are 1 distance apart in the interval $[0, 2^{I_t}]$[2] (see Fig. 2 for an example). The sign of $\bar{r}_t$ is transmitted using one bit, I_t is transmitted with unary coding using $O(\log(\bar{r}_t))$ bits, and the randomized quantizer uses $2^{I_t}+1$ levels, hence $O(\log(\bar{r}_t))$ bits. An estimated value of r_t is obtained from the quantized $\bar{r}_t$ by a proper shift and scaling. We recall that $\hat{\mu}(t)$ is believed to be close to r_t in the majority of iterations resulting in small values for $\log(\bar{r}_t)$. An illustration of the algorithm is provided in Fig. 2. The pseudo-code of the algorithm is given in App. B together with intuition on the used techniques.

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



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

Figure 3. Regret versus number of iterations.

3.3. $QuBan$ Performance

Our main theorem in App. C provides an upper bound on the regret and the average number of bits communicated, when $QuBan$ is used on top of any MAB algorithm. At a high level the theorem states that if $QuBan$ is applied on top of a MAB algorithm with sublinear regret, it requires an average number of bits asymptotically bounded by 7 (3 bits in our numerical results). The theorem also shows that $QuBan$ maintains properties for the quantized reward, that include the Markov property, unbiasdness, and bounded variance, which result in achieving the same regret bound as the unquantized case up to a factor of $\sqrt{5/4}$.

4. Numerical Evaluation

We here present representatives of our numerical results; additional plots are included in App. E.

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 App. B, with 2^r levels uniformly dividing a range $[-M, M]$.
$QuBan$. We implement a minor variation of $QuBan$ described in App. D. The variant maintains the same quantized value and only changes its encoding in the neighborhood of $\hat{\mu}(t)$. We implemented the avg-pt, the avg-arm-pt and the contextual reward choice for $\hat{\mu}(t)$ (described in Section 3).

MAB Algorithms. We use quantization on top of:
(i) the UCB implementation in Lattimore & Szepesvári, 2020, chapter 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 Lattimore & Szepesvári, 2020, chapter 6, where ϵ_t is set to be $\epsilon_t = \min\{1, \frac{C\sigma_q k}{t\Delta^2_{\text{min}}}\}$.
(iii) the LinUCB algorithm for stochastic linear bandits in Lattimore & Szepesvári, 2020, chapter 19.

MAB Setup. We simulate three cases. In each case we average over 10 runs of each experiment.
• **Setup 1: (Figs 3(a)).** We use $k = 100, M = 100, C = 10$, 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 3(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 3(c)).** This is our contextual bandit setup with parameters included in App. E. 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. 3 plots the regret $\hat{R}_n$ in (1) vs. the number of iterations. In App. E, we also plot the number of bits required to achieve a certain average regret. We find that:
• $QuBan$ 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.
• Both $QuBan$ avg-pt and avg-arm-pt achieve the same regret (they are not distinguishable in Fig. 3 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. 4(a) in App. E) while avg-pt uses a smaller number of bits when they tend to be closer together (Fig. 4(b) in App. E).
• 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.
• $QuBan$ in all three setups achieves $\hat{B}_n \approx 3$ (plots are provided in App. E).

References

Abe, N. and Long, P. M. Associative reinforcement learning using linear probabilistic concepts. In *ICML*, pp. 3–11. Citeseer, 1999.

Abeille, M. and Lazaric, A. Linear thompson sampling revisited. In *Artificial Intelligence and Statistics*, pp. 176–184. PMLR, 2017.

Agrawal, R. Sample mean based index policies with o (log n) regret for the multi-armed bandit problem. *Advances in Applied Probability*, pp. 1054–1078, 1995.

Agrawal, S. and Goyal, N. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pp. 39–1. JMLR Workshop and Conference Proceedings, 2012.

Agrawal, S. and Goyal, N. Further optimal regret bounds for thompson sampling. In *Artificial intelligence and statistics*, pp. 99–107. PMLR, 2013a.

Agrawal, S. and Goyal, N. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pp. 127–135. PMLR, 2013b.

Alistarh, D., Grubic, D., Li, J., Tomioka, R., and Vojnovic, M. Qsgd: Communication-efficient sgd via gradient quantization and encoding. In *Advances in Neural Information Processing Systems*, pp. 1709–1720, 2017.

Anandkumar, A., Michael, N., Tang, A. K., and Swami, A. Distributed algorithms for learning and cognitive medium access with logarithmic regret. *IEEE Journal on Selected Areas in Communications*, 29(4):731–745, 2011.

Anantharam, V., Varaiya, P., and Walrand, J. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part i: Iid rewards. *IEEE Transactions on Automatic Control*, 32(11):968–976, 1987.

Anscombe, F. Sequential medical trials. *Journal of the American Statistical Association*, 58(302):365–383, 1963.

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

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002b.

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

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

Buccapatnam, S., Eryilmaz, A., and Shroff, N. B. Multi-armed bandits in the presence of side observations in social networks. In *52nd IEEE Conference on Decision and Control*, pp. 7309–7314. IEEE, 2013.

Buccapatnam, S., Eryilmaz, A., and Shroff, N. B. Stochastic bandits with side observations on networks. In *The 2014 ACM international conference on Measurement and modeling of computer systems*, pp. 289–300, 2014.

Dani, V., Hayes, T. P., and Kakade, S. M. Stochastic linear optimization under bandit feedback. *21st Annual Conference on Learning Theory*, pp. 355–366, 2008.

Ding, K., Li, J., and Liu, H. Interactive anomaly detection on attributed networks. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pp. 357–365, 2019.

Even-Dar, E., Mannor, S., and Mansour, Y. Pac bounds for multi-armed bandit and markov decision processes. In *International Conference on Computational Learning Theory*, pp. 255–270. Springer, 2002.

Gray, R. M. and Stockham, T. G. Dithered quantizers. *IEEE Transactions on Information Theory*, 39(3):805–812, 1993.

Hanna, O. A., Ezzeldin, Y. H., Sadjadpour, T., Fragouli, C., and Diggavi, S. On distributed quantization for classification. *IEEE Journal on Selected Areas in Information Theory*, 1(1):237–249, 2020.

Katehakis, M. N. and Robbins, H. Sequential choice from several populations. *Proceedings of the National Academy of Sciences of the United States of America*, 92 (19):8584, 1995.

Lai, T. L. Adaptive treatment allocation and the multi-armed bandit problem. *The Annals of Statistics*, pp. 1091–1114, 1987.

Landgren, P. et al. *Distributed Multi-agent Multi-armed Bandits*. PhD thesis, Princeton University, 2019.

Langford, J. and Zhang, T. Epoch-greedy algorithm for multi-armed bandits with side information. *Advances in Neural Information Processing Systems (NIPS 2007)*, 20: 1, 2007.

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

Li, L., Chu, W., Langford, J., and Schapire, R. E. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pp. 661–670, 2010.

Mannor, S. and Tsitsiklis, J. N. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5(Jun):623–648, 2004.

Mary, J., Gaudel, R., and Preux, P. Bandits and recommender systems. In *International Workshop on Machine Learning, Optimization and Big Data*, pp. 325–336. Springer, 2015.

Mayekar, P. and Tyagi, H. Ratq: A universal fixed-length quantizer for stochastic optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 1399–1409. PMLR, 2020.

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

Russo, D. and Van Roy, B. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39 (4):1221–1243, 2014.

Seide, F., Fu, H., Droppo, J., Li, G., and Yu, D. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech dnns. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

Shahrampour, S., Rakhlin, A., and Jadbabaie, A. Multi-armed bandits in multi-agent networks. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2786–2790. IEEE, 2017.

Song, L., Fragouli, C., and Shah, D. Recommender systems over wireless: Challenges and opportunities. In *2018 IEEE Information Theory Workshop (ITW)*, pp. 1–5. IEEE, 2018.

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

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

Wood, R., Nagpal, R., and Wei, G.-Y. Flight of the robobees. *Scientific American*, 308(3):60–65, 2013.