It Doesn't Get Better and Here's Why: A Fundamental Drawback in Natural Extensions of UCB to Multi-agent Bandits

Udari Madhushani
Princeton University
udarim@princeton.edu

Naomi Ehrich Leonard
Princeton University
naomi@princeton.edu

Abstract

We identify a fundamental drawback of natural extensions of Upper Confidence Bound (UCB) algorithms to the multi-agent bandit problem in which multiple agents facing the same explore-exploit problem can share information. We provide theoretical guarantees that when agents use a natural extension of the UCB sampling rule, sharing information about the optimal option *degrades* their performance. For K the number of agents and T the time horizon, we prove that when agents share information only about the optimal option they suffer an expected group cumulative regret of $O(K \log T + K \log K)$, whereas when they do not share any information they only suffer a group regret of $O(K \log T)$. Further, while information sharing about all options yields much better performance than with no information sharing, we show that including information about the optimal option is not as good as sharing information only about suboptimal options.

1 Introduction

Sequential decision making in multi-agent systems is a subclass of collective learning problems with a wide variety of real-world applications ranging from ecology to machine learning systems [17, 4, 5]. A crucial part of developing high performance algorithms for these systems is modifying individual decision-making strategies to improve performance by leveraging shared information. [16, 3]. Natural extensions of Upper Confidence Bound (UCB) algorithms constitute a widely used class of algorithms in multi-agent sequential decision making. Generally, we expect any well designed multi-agent algorithm to provide better or comparable performance when more information is shared about any and all decisions. This gives rise to the following question.

Does more information always provide improved or comparable performance?

We answer this question by proving that performance of any natural extension of UCB algorithms degrades when information about the optimal decision is shared. UCB algorithms, initially proposed by [2], are used in the stochastic multi-armed bandit (MAB) problem [15, 8, 1], a mathematical model that captures the trade-off between exploring and exploiting in sequential decision making under random payoffs. Consider an agent repeatedly faced with the problem of choosing an option from the same fixed set of options. At each time step, the agent chooses an option and receives a numerical reward drawn from a probability distribution associated with its chosen option. The goal of the agent is to maximize the cumulative reward at the end of the decision-making process. To realize this goal, the agent should choose the optimal option, i.e., the option with the maximum expected reward. However, the agent is unaware of the statistical properties of probability distributions associated with the options. So, instead, the agent needs both to choose options with seemingly high expected rewards (exploit) and to choose lesser known options to learn the true expected rewards (explore).

1st I Can't Believe It's Not Better Workshop (ICBINB@NeurIPS 2020), Vancouver, Canada.

In the UCB algorithm, at each time step t for each option i, the agent constructs an objective function

$$Q_i(t) = \widehat{\mu}_i(t) + C_i(t), \quad \text{where } C_i(t) = \sigma_i \sqrt{\frac{2(\xi+1)\log t}{n_i(t)}}.$$

Here $\widehat{\mu}_i(t)$ is the agent's estimate of the expected reward of option i and $n_i(t)$ the number of samples taken from option i at time t. σ_i is the variance of the reward probability distribution associated with option i and ξ is a constant. The term $C_i(t)$ denotes the uncertainty associated with the estimate. At each time step, the agent chooses the option that maximizes the objective function: the corresponding sequence of choices addresses the trade-off between exploiting and exploring. In natural extensions of UCB algorithms to multi-agent bandit settings in which agents can share information, each agent uses the information it receives from other agents, in addition to the information obtained through its own sampling, to compute the objective functions. Thus, through information sharing, agents reduce the uncertainty associated with estimates of expected reward of the options.

To meet their goal, agents sample the optimal option much more frequently than they do suboptimal options. This implies that if they are sharing information about the optimal option, uncertainty associated with the estimate of the optimal option reduces significantly as compared to uncertainty associated with the estimates of suboptimal options. Thus, the $C_i(t)$ term corresponds to the optimal option becomes significantly smaller than the $C_i(t)$ terms correspond to suboptimal options. Since the objective function consists of the sum of estimate and uncertainty, this reduces the number of times agents sample the optimal option as compared to the case in which agents do not share information about the optimal option. As a result, agents perform better when they do not share any information as compared to the case in which they share information only about the optimal option. Likewise, while sharing information about all options is better than no sharing, agents perform better when they share information only about suboptimal options.

Related work With a few exceptions, the work related to UCB type sampling rules for multi-agent bandits use natural extensions [3, 6, 10–14]. The papers [10, 11] use a running consensus algorithm to update estimates and provide graph-structure-dependent performance measures that predict the relative performance of agents and networks. The paper [11] also addresses the case of a constrained reward model in which agents that choose the same option at the same time step receive no reward. In [14] the authors propose an accelerated consensus procedure in the case that agents know the spectral gap of the communication graph and design a decentralized UCB algorithm based on delayed rewards. The paper [12] considers the case where agents observe their neighbors according to a modified Erdős Rényi communication graph. In [3] the authors consider that at each time step, agents decide either to sample an option or to broadcast the last received reward to the entire group. In [13], each agent broadcasts its reward and action to its neighbors when it is exploring. The paper [6] considers multi-agent bandits with collision where agents that choose the same option at the same time receive zero reward.

The papers that propose exceptions to natural extensions [7, 9] consider sampling protocols where agents imitate the actions of others. A strategy where agents observe the rewards and choices of their neighbors according to a leader-follower setting is considered in [9]. The paper [7] considers a case where agents copy the actions of their nearest leader. However, this line of work suffers from the limitation that in order to imitate, each agent is required to possess knowledge about the sampling rules of the agents it is imitating.

Key contributions Our paper is the *first work* to highlight a fundamental drawback of natural extensions of UCB algorithms to multi-agent setting. Our contributions include the following:

- We show that performance of a group of agents faced with the stochastic bandit problem degrades when they exchange information about the optimal option.
- We prove that agents suffer an expected cumulative group regret of $O(K \log T + K \log K)$, where expected regret is defined as the expected loss agents suffer by sampling suboptimal options instead of the optimal option. Here K is the number of agents and T is the time horizon of the decision-making process.
- We show that, while sharing information about all options is significantly better than sharing no information, performance of the group further improves when they share information only about suboptimal options.

2 Mathematical Formulation

We consider the distributed MAB problem with N options and K agents. The reward of each option $i \in \{1, \ldots, N\}$ is modeled with a sub-Gaussian random variable X_i, which includes widely used distributions such as Bernoulli, Gaussian, and bounded rewards. Let μ_i and σ_i^2 be the expected reward and variance proxy associated with option i, respectively. In this paper we assume that the variance proxy of each option is known to the agents. Let i^* denote the optimal option, i.e., the option with highest expected mean $\mu_{i^*} = \max\{\mu_1, \ldots, \mu_N\}$. Let $\Delta_i = \mu_{i^*} - \mu_i$ be the expected reward gap between option i^* and option i. Let φ_t^k denote the option chosen by agent k at time t. Define $\mathbb{I}_{\{\varphi_t^k = i\}}$ to be an indicator random variable that takes value 1 if agent k chooses option i at time t and 0 otherwise.

Assumption 1 *All the agents know the variance proxy σ_i^2 of the rewards associated with each option.*

Remark 1 *Assumption 1 can be relaxed by allowing that the agents only know an upper bound for the variance proxy of all the options. All results presented in this paper can be generalized to this case by replacing σ_i with the upper bound.*

Let $G(\mathcal{V}, \mathcal{E})$ be an undirected graph that encodes the fixed communication network graph among agents. $\mathcal{V}$ is a set of K vertices and $\mathcal{E}$ is the set of edges between pairs of vertices. Without loss of generality, let vertex $k \in \mathcal{V}$ correspond to agent k. If there is a communication link $e(k, j) \in \mathcal{E}$ between agents k and j, then agents k and j are neighbors. Define d_k as the degree of vertex k. This is same as the number of neighbors of agent k. Define d_{avg} as the average degree of the graph. Let $\mathbb{I}_{\{k,j\}}$ be an indicator variable that takes value 1 if agent j is a neighbor of agent k and 0 otherwise. Let $\mathbb{I}_{\{k,j\}}^t$ be an indicator random variable that takes value 1 if agent k receives information, i.e. chosen option and the value of the obtained reward, from agent j at time t and 0 otherwise. And let $\mathbb{I}_{\{k,k\}}^t = 1$, for all t, k.

Assumption 2 *When more than one agent chooses the same option at the same time, the agents receive rewards independently and identically drawn from the probability distribution associated with the chosen option.*

Let $n_i^k(t)$ and $N_i^k(t)$ denote the number of times agent k sampled option i and observed rewards from option i, respectively until time t:

$$n_i^k(t) = \sum_{\tau=1}^{t} \mathbb{I}_{\{\varphi_\tau^k = i\}}, \quad N_i^k(t) = \sum_{\tau=1}^{t} \sum_{j=1}^{K} \mathbb{I}_{\{\varphi_\tau^j = i\}} \mathbb{I}_{\{k,j\}}^\tau.$$

Note that the number of observations $N_i^k(t)$ is the sum of the number of samples taken by agent k of option i and the number of times agent k received reward values for option i from its neighbors. Each agent reduces the uncertainty associated with estimates of expected reward of options by leveraging the information received from neighbors.

Let $S_i^k(t)$ be the sum of reward values until time t received by agent k from sampling option i and receiving reward values from neighbors obtained from option i. Each agent estimates the expected reward of option i by averaging. Let $\widehat{\mu}_i^k(t)$ be the estimated expected reward of option i for agent k. Then we have

$$S_i^k(t) = \sum_{\tau=1}^{t} \sum_{j=1}^{K} X_i \mathbb{I}_{\{\varphi_\tau^j = i\}} \mathbb{I}_{\{k,j\}}^\tau, \qquad \widehat{\mu}_i^k(t) = \frac{S_i^k(t)}{N_i^k(t)}.$$

The natural extension of UCB sampling rule to multi-agent bandit problem can be stated as follows.

Definition 1 *(***Natural Extension of UCB Sampling Rule***)* *The sampling rule $\{\varphi_t^k\}_1^T$ for agent k at time step $t \in \{1, \ldots, T\}$ is defined as*

$$\mathbb{I}_{\{\varphi_{t+1}^k = i\}} = \begin{cases} 1 & , \quad i = \arg\max\{Q_1^k(t), \cdots, Q_N^k(t)\} \\ 0 & , \quad \text{o.w.} \end{cases}$$

with

$$Q_i^k(t) = \widehat{\mu}_i^k(t) + C_i^k(t)$$
$$C_i^k(t) = \sigma_i \sqrt{\frac{2(\xi+1)\log t}{N_i^k(t)}}$$

where $\xi > 1$.

We proceed to show that sharing information about the optimal option degrades performance by considering two cases as follows. In the first case agents do not communicate any information. This is equivalent to K agents using the single-agent UCB algorithm individually. In the second case we consider an oracle communication rule where agents broadcast reward values obtained by sampling the optimal option to their neighbors. We state the oracle communication protocol in Definition 2.

Definition 2 (Oracle Communication Rule) *The communication rule for agents* $k, j \in \{1, \ldots, K\}$ *such that* $e(k, j) \in \mathcal{E}$*, at time step* $t \in \{1, \ldots, T\}$ *is defined as*

$$\mathbb{I}_{\{k,j\}}^{t+1} = \begin{cases} 1 & , \quad \varphi_t^j = i^* \\ 0 & , \quad \text{o.w.} \end{cases}$$

3 Theoretical Results

In this section we analyze the performance of natural extension of UCB sampling rule (Definition 1), with the no communication protocol and with the oracle communication protocol defined in Definition 2. Following convention, as a performance measure, we use the expected regret agents incur by sampling suboptimal options. Let $R(t)$ be the cumulative regret of the group at time t. Then the expected group cumulative regret at time t can be given as

$$\mathbb{E}(R(t)) = \sum_{i=1}^{N} \sum_{k=1}^{K} (\mu_{i^*} - \mu_i) \mathbb{E}\left(n_i^k(t)\right).$$

We provide upper bounds for the expected cumulative group regret under the two communication protocols.

Theorem 1 (Upper Bound on Group Regret) *Consider a multi-agent stochastic bandit problem with* T *time steps and* K *agents that communicate over network graph* G*. All agents sample according to the rule defined in Definition 1. Then the expected group cumulative regret satisfies:*

1. *With no communication*
$$\mathbb{E}(R(T)) = O(K \log T);$$
2. *With oracle communication given in Definition 2*
$$\mathbb{E}(R(T)) = O(K \log T + +K \log(d_{avg} + 1)).$$

The proof of Theorem 1 is given in Appendix A. In the following corollary, we specialize this result to the case where agents broadcast reward values received from the optimal option to the group. This corresponds to the case where G is a complete graph.

Corollary 1 (Specialization to complete graph) *Let the communication graph* G *be complete. With no communication agents suffer a group regret of* $\mathbb{E}(R(T)) = O(K \log T)$ *and with oracle communication given in Definition 2 they suffer a group regret of* $\mathbb{E}(R(T)) = O(K \log T + +K \log K)$.

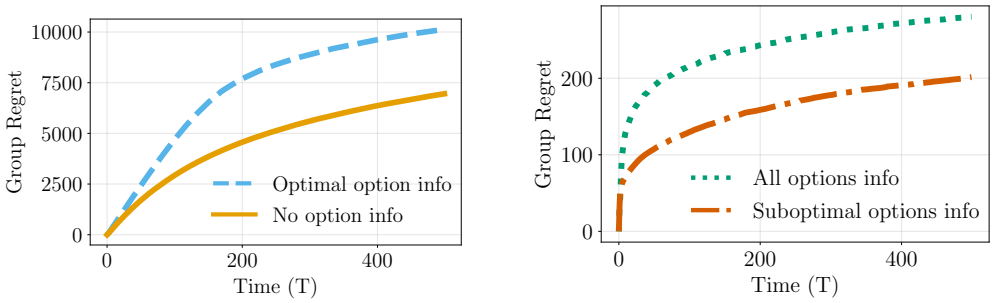
Proof of Corollary 1 When agents do not communicate the group is equivalent to K individual agents facing the same bandit problem independently. Thus expected cumulative group regret is the same irrespective of the graph structure. When the communication graph is complete each agent has $K - 1$ number if neighbors. Thus $d_{avg} = K - 1$. The result follows from Theorem 1.

Remark 2 *Additional group regret incurred by sharing information about the optimal option is independent of the time horizon of the decision making process. The additional regret increases with the average degree of the communication graph.*

From Theorem 1 we see that performance of agents degrade when they broadcast reward values received from the optimal option.

4 Simulation Results

In this section we provide numerical simulations illustrating the results and validating our theoretical claim. For all the simulations presented in this section, we consider a group of 50 agents ($K = 50$) and 10 options ($N = 10$) with Gaussian reward distributions. We consider a challenging problem in which we let the expected reward value of the optimal option be 11, the expected reward of all other options be 10, and the variance of all options be 1. We let the communication network graph G be complete. We provide results with 500 time steps ($T = 500$) using 1000 Monte Carlo simulations.



(a) Regret of the group when agents do not share information and when agents share information only about the optimal option as per Definition 2.

(b) Regret of the group when agents share information about all the options and when agents share information only about suboptimal options.

Figure 1: Expected group cumulative regret of 50 agents using a natural extension of UCB sampling rule given in Definition 1.

The orange solid line in Figure 1(a) shows expected cumulative group regret when agents do not share any information with their neighbors. This corresponds to the case where K agents are independently faced with the same bandit problem. The blue dashed line corresponds to the case where agents share rewards values received from the optimal option only. The figure illustrates that agents perform better when they do not communicate as compared to the case where they share information about only the optimal option. Figure 1(b) shows group regret for the cases where agents share information about reward values received from all options (green dotted line) and agents share information about reward values received only from suboptimal options (red dash-dot line). The figure illustrates that agents perform better when they only share information about suboptimal options as compared to the case where they share information about all options. Comparing Figures 1(a) and 1(b) note also that agents perform much better with sharing than without sharing as long they share information on suboptimal options, whether or not they also share information on optimal options.

5 Discussion and Conclusion

The main result presented in this paper proves that when sampling options according to a natural extension of UCB algorithm, sharing information about the optimal option degrades performance of the agents. Our work has two major implications for distributed sequential learning algorithms that use natural extensions of UCB. The first implication is the importance of developing communication protocols that minimize the amount of information shared about the optimal option. Recall that under any efficient sampling rule agents sample suboptimal options only logarithmically in time. Thus, in cases where communication is costly such protocols have the added advantage of incurring a significantly smaller communication cost. The second implication is that adversaries can decrease the performance of agents by intentionally only sharing information about the optimal option.

In this work we shed light on an inherent drawback of natural extensions of UCB algorithms to multi-agents setting. We proved that when agent sample options according to a natural extension of UCB, group performance degrades when they share information about the optimal option. We provided numerical results illustrating that agents perform better when they do not communicate as compared to the case where they share information only about the optimal option. Similarly, agents perform better when they share information only about suboptimal options as compared to the case where they share information about all options.

Acknowledgment

This research has been supported in part by ONR grants N00014-18-1-2873 and N00014-19-1-2556 and ARO grant W911NF-18-1-0325.

References

[1] V. Anantharam, P. Varaiya, and J. Walrand. 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.

[2] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

[3] M. Chakraborty, K. Y. P. Chua, S. Das, and B. Juba. Coordinated versus decentralized exploration in multi-agent multi-armed bandits. In *IJCAI*, pages 164–170, 2017.

[4] A. Durand, C. Achilleos, D. Iacovides, K. Strati, G. D. Mitsis, and J. Pineau. Contextual bandits for adapting treatment in a mouse model of de novo carcinogenesis. In *MLHC*, 2018.

[5] R. Féraud, R. Alami, and R. Laroche. Decentralized exploration in multi-armed bandits. *arXiv preprint arXiv:1811.07763*, 2018.

[6] D. Kalathil, N. Nayyar, and R. Jain. Decentralized learning for multiplayer multiarmed bandits. *IEEE Transactions on Information Theory*, 60(4):2331–2345, 2014.

[7] R. K. Kolla, K. Jagannathan, and A. Gopalan. Collaborative learning of stochastic bandits over a social network. *IEEE/ACM Transactions on Networking*, 26(4):1782–1795, 2018.

[8] T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[9] P. Landgren, V. Srivastava, and N. E. Leonard. Social imitation in cooperative multiarmed bandits: partition-based algorithms with strictly local information. In *2018 IEEE Conf. Decision and Control*, pages 5239–5244, 2018.

[10] P. Landgren, V. Srivastava, and N. E. Leonard. Distributed cooperative decision-making in multiarmed bandits: Frequentist and Bayesian algorithms. *arXiv:1606.00911v3*, 2019.

[11] P. Landgren, V. Srivastava, and N. E. Leonard. Distributed cooperative decision making in multi-agent multi-armed bandits. *arXiv:2003.01312*, 2020.

[12] U. Madhushani and N. E. Leonard. Heterogeneous stochastic interactions for multiple agents in a multi-armed bandit problem. In *18th European Control Conf.*, pages 3502–3507, 2019.

[13] U. Madhushani and N. E. Leonard. Distributed learning: Sequential decision making in resource-constrained environments. *arXiv:2004.06171*, 2020.

[14] D. Martínez-Rubio, V. Kanade, and P. Rebeschini. Decentralized cooperative stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 4531–4542, 2019.

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

[16] C. J. Torney, A. Berdahl, and I. D. Couzin. Signalling and the evolution of cooperative foraging in dynamic environments. *PLoS Comput Biol*, 7(9):e1002194, 2011.

[17] C. J. Torney, T. Lorenzi, I. D. Couzin, and S. A. Levin. Social information use and the evolution of unresponsiveness in collective systems. *Journal of the Royal Society Interface*, 12(103): 20140893, 2015.

A Proof of Theorem 1

We begin by proving a couple of lemmas that are useful in proving Theorem 1.

Lemma 1 (Concentration Bound for Estimate) *Let d_k be the degree of agent k. For any option i for any $\vartheta > 0$ and some $\zeta > 1$ we have*

$$\mathbb{P}\left(\left|\widehat{\mu}_i(t) - \mu_i\right| \geq \sqrt{\frac{\vartheta}{N_i(t)}}\right) \leq \frac{\log((d_k+1)t)}{\log \zeta} \exp(-2\kappa\vartheta)$$

where $\kappa = \frac{1}{\sigma_i^2\left(\zeta^{\frac{1}{4}}+\zeta^{-\frac{1}{4}}\right)^2}$.

Proof of Lemma 1 Since X_i is a sub-Gaussian random variable with variance proxy σ_i^2, we have

$$E\left(\exp(\lambda(X_i - \mu_i))\right) \leq \exp\left(\frac{\lambda^2\sigma_i^2}{2}\right). \tag{1}$$

Define a new random variable such that $\forall \tau > 0$. Then,

$$Y_i^k(\tau) = (X_i - \mu_i)\sum_{j=1}^{K} \mathbb{I}_{\{\varphi_\tau^j = i\}}\mathbb{I}_{\{k,j\}}^{\tau}. \tag{2}$$

Note that $\mathbb{I}_{\{\varphi_\tau^j=i\}}, \mathbb{I}_{\{k,j\}}^{\tau}$ are pre-visible variables and hence $E(Y_i^k(\tau)) = E(Y_i^k(\tau)|\mathcal{F}_{\tau-1}) = 0$. Let $Z_i^k(t) = \sum_{\tau=1}^{t} Y_i^k(\tau)$. For any $\lambda > 0$ from (1) and (2)

$$E\left(\exp(\lambda Y_i^k(\tau))|\mathcal{F}_{\tau-1}\right) \leq \exp\left(\frac{\lambda^2\sigma_i^2}{2}\sum_{j=1}^{K}\mathbb{I}_{\{\varphi_\tau^j=i\}}\mathbb{I}_{\{k,j\}}^{\tau}\right).$$

Since $\mathbb{I}_{\{k,j\}}^{\tau}, \mathbb{I}_{\{\varphi_\tau^j=i\}}$ are $\mathcal{F}_{\tau-1}$ measurable random variables,

$$E\left(\exp\left(\lambda Y_i^k(\tau) - \frac{\lambda^2\sigma_i^2}{2}\sum_{j=1}^{K}\mathbb{I}_{\{\varphi_\tau^j=i\}}\mathbb{I}_{\{k,j\}}^{\tau}\right)\Big|\mathcal{F}_{\tau-1}\right) \leq 1.$$

Using the properties of conditional expectations we obtain

$$E\left(\exp\left(\lambda Z_i^k(t) - \frac{\lambda^2\sigma_i^2}{2}N_i^k(t)\right)\Big|\mathcal{F}_{t-1}\right) \leq \exp\left(\lambda Z_i^k(t-1) - \frac{\lambda^2\sigma_i^2}{2}N_i^k(t-1)\right).$$

Thus, we get

$$E\left(\exp\left(\lambda Z_i^k(t) - \frac{\lambda^2\sigma_i^2}{2}N_i^k(t)\right)\right) \leq 1.$$

Note that we have

$$\begin{aligned}\mathbb{P}\left(\exp\left(\lambda Z_i^k(t) - \frac{\lambda^2\sigma_i^2}{2}N_i^k(t)\right) \geq \exp\left(2\kappa\vartheta\right)\right) &= \mathbb{P}\left(\lambda Z_i^k(t) - \frac{\lambda^2\sigma_i^2}{2}N_i^k(t) \geq 2\kappa\vartheta\right)\\ &= \mathbb{P}\left(\frac{Z_i^k(t)}{\sqrt{N_i^k(t)}} \geq \frac{2\kappa\vartheta}{\lambda}\sqrt{\frac{1}{N_i^k(t)}} + \frac{\sigma_i^2}{2}\lambda\sqrt{N_i^k(t)}\right).\end{aligned} \tag{3}$$

Let $\zeta > 1$. Then $1 \leq N_i^k(t) \leq \zeta^{D_t}$ where $D_t = \frac{\log((d_k+1)t)}{\log\zeta}$. For $\lambda_j = \frac{2}{\sigma_i}\sqrt{\frac{\kappa\vartheta}{\zeta^{j-1/2}}}$ and $\zeta^{j-1} \leq N_i^k(t) \leq \zeta^j$ we have

$$\frac{2\kappa\vartheta}{\lambda_j}\sqrt{\frac{1}{N_i^k(t)}} + \frac{\sigma_i^2}{2}\lambda_j\sqrt{N_i^k(t)} = \sigma_i\sqrt{\kappa\vartheta}\left(\sqrt{\frac{\zeta^{j-1/2}}{N_i^k(t)}} + \sqrt{\frac{N_i^k(t)}{\zeta^{j-1/2}}}\right) \leq \sqrt{\vartheta}, \tag{4}$$

where $\kappa = \frac{1}{\sigma_i^2\left(\zeta^{\frac{1}{4}}+\zeta^{-\frac{1}{4}}\right)^2}$.

Recall from the Markov inequality that $\mathbb{P}(Y \geq a) \leq \frac{\mathbb{E}(Y)}{a}$ for any positive random variable Y. Thus, from (3) and (4) we have

$$\mathbb{P}\left(\frac{Z_i(t)}{\sqrt{N_i^k(t)}} \geq \sqrt{\vartheta}\right) \leq \cup_{j=1}^{D_T} \exp(-2\kappa\vartheta).$$

Then, we get

$$\mathbb{P}\left(\frac{Z_i^k(t)}{N_i^k(t)} \geq \sqrt{\frac{\vartheta}{N_i(t)}}\right) \leq \cup_{j=1}^{D_T} \exp(-2\kappa\vartheta).$$

Substituting $Z_i^k(t) = \sum_{\tau=1}^{t}\sum_{j=1}^{K}(X_i - \mu_i)\,\mathbb{I}_{\{\varphi_\tau^j = i\}}\mathbb{I}_{\{k,j\}}^{\tau}$ and from symmetry of the reward distribution we get

$$\mathbb{P}\left(\left|\widehat{\mu}_i(t) - \mu_i\right| \geq \sqrt{\frac{\vartheta}{N_i(t)}}\right) \leq \frac{\log((d_k+1)t)}{\log\zeta}\exp(-2\kappa\vartheta)$$

This concludes the proof of Lemma 1.

Lemma 2 *Let d_k be the degree of agent k. For any option i for some $\zeta > 1$ we have*

$$\mathbb{P}\left(\left|\widehat{\mu}_i^k(t) - \mu_i\right| > C_i^k(t)\right) \leq \frac{1}{\log\zeta}\frac{\log\left((1+d_k)t\right)}{t^{\xi+1}}.$$

Proof of Lemma 2 Substituting $\vartheta = 2\sigma_i^2(\xi+1)\log t$ to Lemma 1 we get

$$\mathbb{P}\left(\left|\widehat{\mu}_i(t) - \mu_i\right| > \sigma_i\sqrt{\frac{2(\xi+1)\log t}{N_i^k(t)}}\right) \leq \frac{1}{\log\zeta}\frac{\log((d_k+1)t)}{t^{\xi+1}}.$$

Since $C_i^k(t) = \sqrt{\frac{2(\xi+1)\log t}{N_i^k(t)}}$ this concludes the proof of Lemma 2.

Lemma 3 *Let $\alpha, \beta > 1$ and $t \in \{1, \cdots, T\}$. Then,*

$$\sum_{t=1}^{T}\frac{\log\alpha t}{t^{\beta+1}} \leq \frac{\beta^2\log\alpha + \beta\log\alpha + 1}{\beta^2}.$$

Proof of Lemma 3 Note that

$$\sum_{t=1}^{T}\frac{\log\alpha t}{t^{\beta+1}} \leq \log\alpha + \int_1^T \frac{\log\alpha t}{t^{\beta+1}}dt$$

Let $z = \alpha t$. Then we have

$$\int_1^T \frac{\log\alpha t}{t^{\beta+1}}dt = \alpha^\beta\int_\alpha^{\alpha T}\frac{\log z}{z^{\beta+1}}dz = \alpha^\beta\left[-\frac{\log z}{\beta z^\beta} - \frac{1}{\beta^2 z^\beta}\right]_\alpha^{\alpha T}.$$

Thus,

$$\sum_{t=1}^{T}\frac{\log\alpha t}{t^{\beta+1}} \leq \log\alpha + \left[\frac{\log\alpha}{\beta} + \frac{1}{\beta^2} - \frac{\log(\alpha T)}{\beta T^\beta} - \frac{1}{\beta^2 T^\beta}\right]. \tag{5}$$

The proof of Lemma 3 follows from (5).

Proof of Theorem 1 Recall that $n_i^k(t)$ and $N_i^k(t)$ denote the number of samples and number of observations taken from option i by agent k until time t. Further, under both no communication and oracle communication protocols, agents do not share information about suboptimal options. Thus for each suboptimal option i we have $n_i^k(t) = N_i^k(t)$, for all k, t, and

$$\sum_{k=1}^{K} \mathbb{E}\left(n_i^k(T)\right) \leq \sum_{k=1}^{K}\sum_{t=1}^{T} \mathbb{P}\left(\varphi_t^k = i, n_i^k(t) \leq \eta_i(t)\right) + \sum_{k=1}^{K}\sum_{t=1}^{T} \mathbb{P}\left(Q_i^k(t) > Q_{i^*}^k(t), n_i^k(t) > \eta_i(t)\right).$$

Here $\eta_i(t)$ is a non negative non decreasing function. Note that we have

$$\left\{Q_i^k(t) \geq Q_{i^*}^k(t)\right\} \subseteq \left\{\mu_{i^*} < \mu_i + 2C_i^k(t)\right\} \cup \left\{\widehat{\mu}_{i^*}^k(t) \leq \mu_{i^*} - C_{i^*}^k(t)\right\} \cup \left\{\widehat{\mu}_i^k(t) \geq \mu_i + C_i^k(t)\right\}.$$

When $\eta_i(t) = \left(\frac{8(\xi+1)\sigma_i^2}{\Delta_i^2}\right)\log t$ we see that

$$\mathbb{P}\left(\mu_{i^*} < \mu_i + 2C_i^k(t), n_i^k(t) \geq \eta_i(t)\right) = 0.$$

Thus,

$$\begin{aligned}\sum_{k=1}^{K} \mathbb{E}\left(n_i^k(T)\right) \leq & \sum_{k=1}^{K}\sum_{t=1}^{T} \mathbb{P}\left(\varphi_t^k = i, n_i^k(t) \leq \eta_i(t)\right) + \sum_{k=1}^{K}\sum_{t=1}^{T} \mathbb{P}\left(\widehat{\mu}_{i^*}^k(t) \leq \mu_{i^*} - C_{i^*}^k(t)\right) \\ & + \sum_{k=1}^{K}\sum_{t=1}^{T} \mathbb{P}\left(\widehat{\mu}_i^k(t) \geq \mu_i + C_i^k(t)\right). \end{aligned} \tag{6}$$

Now we proceed to upper bound the last two summation terms on the right hand side of (6). First we provide bounds for the case where agents do not communicate. No communication is equivalent to having no neighbors. Thus we have $d_k = 0$, for all k. Then from Lemma 2 for all options i we have

$$\mathbb{P}\left(\left|\widehat{\mu}_i(t) - \mu_i\right| > \sigma_i\sqrt{\frac{2(\xi+1)\log t}{N_i^k(t)}}\right) \leq \frac{1}{\log\zeta}\frac{\log t}{t^{\xi+1}}.$$

Thus when agents do not communicate we have

$$\sum_{k=1}^{K}\sum_{t=1}^{T} \mathbb{P}\left(\widehat{\mu}_{i^*}^k(t) \leq \mu_{i^*} - C_{i^*}^k(t)\right) + \sum_{k=1}^{K}\sum_{t=1}^{T} \mathbb{P}\left(\widehat{\mu}_i^k(t) \geq \mu_i + C_i^k(t)\right) \leq 2\sum_{k=1}^{K}\sum_{t=1}^{T}\frac{1}{\log\zeta}\frac{\log t}{t^{\xi+1}}.$$

Using the results from Lemma 3 we get

$$\sum_{k=1}^{K}\sum_{t=1}^{T} \mathbb{P}\left(\widehat{\mu}_{i^*}^k(t) \leq \mu_{i^*} - C_{i^*}^k(t)\right) + \sum_{k=1}^{K}\sum_{t=1}^{T} \mathbb{P}\left(\widehat{\mu}_i^k(t) \geq \mu_i + C_i^k(t)\right) \leq \frac{2K}{\xi^2\log\zeta}. \tag{7}$$

Now we proceed to upper bound the last two summations of (6) when agents communicate according to the oracle communication protocol. Since agents do not share information about suboptimal options for any suboptimal option i we have

$$\mathbb{P}\left(\left|\widehat{\mu}_i(t) - \mu_i\right| > \sigma_i\sqrt{\frac{2(\xi+1)\log t}{N_i^k(t)}}\right) \leq \frac{1}{\log\zeta}\frac{\log t}{t^{\xi+1}}.$$

Recall that d_k is the degree of agent k. Then for the optimal option we have

$$\mathbb{P}\left(\left|\widehat{\mu}_{i^*}(t) - \mu_{i^*}\right| > \sigma_{i^*}\sqrt{\frac{2(\xi+1)\log t}{N_{i^*}^k(t)}}\right) \leq \frac{1}{\log\zeta}\frac{\log((d_k+1)t)}{t^{\xi+1}}.$$

Thus when agents only share information about the optimal option we have

$$\begin{aligned}\sum_{k=1}^{K}\sum_{t=1}^{T} \mathbb{P}\left(\widehat{\mu}_{i^*}^k(t) \leq \mu_{i^*} - C_{i^*}^k(t)\right) + & \sum_{k=1}^{K}\sum_{t=1}^{T} \mathbb{P}\left(\widehat{\mu}_i^k(t) \geq \mu_i + C_i^k(t)\right) \\ & \leq \sum_{k=1}^{K}\sum_{t=1}^{T}\frac{1}{\log\zeta}\frac{2\log t + \log(d_k+1)}{t^{\xi+1}}. \end{aligned} \tag{8}$$

Since $\log$ is a concave function,

$$\sum_{k=1}^{K} \log(d_k + 1) \leq K \log(d_{avg} + 1). \tag{9}$$

Using the results from (8), (9) and Lemma 3 we get

$$\sum_{k=1}^{K}\sum_{t=1}^{T} \mathbb{P}\left(\widehat{\mu}_{i^*}^k(t) \leq \mu_{i^*} - C_{i^*}^k(t)\right) + \sum_{k=1}^{K}\sum_{t=1}^{T} \mathbb{P}\left(\widehat{\mu}_{i}^k(t) \geq \mu_{i} + C_{i}^k(t)\right)$$
$$\leq \frac{K}{\xi^2 \log \zeta}\left[\xi^2 \log(d_{avg}+1) + \xi \log(d_{avg}+1) + 2\right]. \tag{10}$$

Recall that $\mathbb{E}(R(T)) = \sum_{i=1}^{N}\sum_{k=1}^{K} \Delta_i \mathbb{E}\left(n_i^k(T)\right)$. Using this result from (7) and substituting for $\eta_i(T)$ when agents do not communicate we get

$$\mathbb{E}(R(T)) \leq \sum_{i=1}^{N} \frac{8(\xi+1)\sigma_i^2}{\Delta_i} K \log T + \sum_{i=1}^{N} \Delta_i \frac{2K}{\xi^2 \log \zeta}. \tag{11}$$

From (10) and substituting for $\eta_i(T)$ when agents only share information about the optimal option,

$$\mathbb{E}(R(T)) \leq \sum_{i=1}^{N} \frac{8(\xi+1)\sigma_i^2}{\Delta_i} K \log T$$
$$+ \sum_{i=1}^{N} \Delta_i \frac{K}{\xi^2 \log \zeta}\left[\xi^2 \log(d_{avg}+1) + \xi \log(d_{avg}+1) + 2\right]. \tag{12}$$

Thus, from (11) and (12) we have, with no communication $\mathbb{E}(R(T)) = O(K \log T)$ and with oracle communication defined in Definition 2, $\mathbb{E}(R(T)) = O(K \log T + K \log(d_{avg}+1))$. This concludes the proof of Theorem 1.