DISTRIBUTED LEARNING: SEQUENTIAL DECISION MAKING IN RESOURCE-CONSTRAINED ENVIRONMENTS

Udari Madhushani
Princeton University
udarim@princeton.edu

Naomi Ehrich Leonard
Princeton University
naomi@princeton.edu

ABSTRACT

We study cost-effective communication strategies that can be used to improve the performance of distributed learning systems in resource-constrained environments. For distributed learning in sequential decision making, we propose a new cost-effective partial communication protocol. We illustrate that with this protocol the group obtains the same order of performance that it obtains with full communication. Moreover, we prove that under the proposed partial communication protocol the communication cost is $O(\log T)$, where T is the time horizon of the decision-making process. This improves significantly on protocols with full communication, which incur a communication cost that is $O(T)$. We validate our theoretical results using numerical simulations.

1 INTRODUCTION

In resource-constrained environments, the difficulty in constructing and maintaining large-scale infrastructure limits the possibility of developing a centralized learning system that has access to global information, resources for effectively processing that information, and the capacity to make all the decisions. Consequently, developing cost-efficient distributed learning systems, i.e., groups of units that collectively process information and make decisions with minimal resource, is an essential step towards making machine learning practical in such constrained environments. In general, most distributed learning strategies allow individuals to make decisions using locally available information (Kalathil et al., 2014; Landgren et al., 2016a; Madhushani & Leonard, 2019), i.e., information that they observe or is communicated to them from their neighbors. However, the performance of such systems is strongly dependent on the underlying communication structure. Such dependence inherently leads to a trade-off between communication cost and performance. Our goal is to develop high performance distributed learning systems with minimal communication cost.

We focus on developing cost-effective distributed learning techniques for sequential decision making under stochastic outcomes. Our work is motivated by the growing number of real-world applications such as clinical trials, recommender systems, and user-targeted online advertising. For example, consider a set of organizations networked to recommend educational programs to online users under high demand. Each organization makes a series of sequential decisions about which programs to recommend according to the user feedback (Warlop et al., 2018; Féraud et al., 2018). As another example, consider a set of small pharmaceutical companies conducting experimental drug trials (Tossou & Dimitrakakis, 2016; Durand et al., 2018). Each company makes a series of sequential decisions about the drug administration procedure according to the observed patient feedback. In both examples, feedback received by the decision maker is stochastic, i.e., the feedback is associated with some uncertainty. This is due to the possibility that at different time steps online users (patients) can experience the same program (drug) differently due to internal and external factors, including their own state of mind and environmental conditions.

Performance of distributed learning in these systems can be significantly improved by establishing a communication network that facilitates *full communication*, whereby each organization shares all feedback immediately with others. However, communication can often be expensive and time-consuming. Under full communication, the amount of communicated data is directly proportional

to the time horizon of the decision-making process. In a resource-constrained environment, when the decision-making process continues for a long time, the full communication protocol becomes impractical. We address this problem by proposing a *partial communication* strategy that obtains the *same order of performance* as the full communication protocol, while using a *significantly smaller amount of data communication*.

To derive and analyze our proposed strategy, we make use of the bandit framework, a mathematical model that has been developed to model sequential decision making under stochastic outcomes (Lai & Robbins, 1985; Robbins, 1952). Consider a group of agents (units) making sequential decisions in an uncertain environment. Each agent is faced with the problem of repeatedly choosing an option from the same fixed set of options (Kalathil et al., 2014; Landgren et al., 2016a;b; 2020; Martínez-Rubio et al., 2019). After every choice, each agent receives a numerical reward drawn from a probability distribution associated with its chosen option. The objective of each agent is to maximize its individual cumulative reward, thereby contributing to maximizing the group cumulative reward.

The best strategy for an agent in this situation is to repeatedly choose the optimal option, i.e., the option that provides the maximum average reward. However, agents do not know the expected reward values of the options. Each individual is required to execute a combination of *exploiting actions*, i.e., choosing the options that are known to provide high rewards, and *exploring actions*, i.e., choosing the lesser known options in order to identify options that might potentially provide higher rewards.

This process is sped up through distributed learning that relies on agents exchanging their reward values and actions with their neighbors (Madhushani & Leonard, 2019; 2020). The protocols in these works determine when an agent observes (samples) the reward values and actions of its neighbors. Our proposed protocol instead determines only when an agent shares (broadcasts). A key result is that this seemingly altruistic action (sharing) provably benefits the individual as well as the group.

We define *exploit-based communication* to be information sharing by agents only when they execute exploiting actions. Similarly, we define *explore-based communication* to be information sharing by agents only when agents execute exploring actions. Thus, for information sharing, we have that

$$\textit{full communication} = \textit{exploit-based communication} + \textit{explore-based communication}.$$

We propose a new partial communication protocol that uses only explore-based communication. We illustrate that explore-based communication obtains the same order of performance as full communication, while incurring a significantly smaller communication cost.

Key Contributions In this work, we study cost-efficient, information-sharing, communication protocols in sequential decision making. Our contributions include the following:

- We propose a new cost-effective partial communication protocol for distributed learning in sequential decision making. The communication protocol determines information sharing.
- We illustrate that with this protocol the group obtains the same order of performance as it obtains with full communication.
- We prove that under the proposed partial communication protocol, the communication cost is $O(\log T)$, where T is the number of decision making steps; whereas under full communication protocols, the communication cost is $O(T)$.

Related Work Previous works (Kalathil et al., 2014; Landgren et al., 2016a;b; 2018; 2020; Martínez-Rubio et al., 2019) have considered the distributed bandit problem without a communication cost. Landgren et al. (2016a;b; 2020) use a running consensus algorithm to update estimates and provide graph-structure-dependent performance measures that predict the relative performance of agents and networks. Landgren et al. (2020) also address the case of a constrained reward model in which agents that choose the same option at the same time step receive no reward. Martínez-Rubio et al. (2019) 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. Szörényi et al. (2013) consider a P2P communication where an agent is only allowed to communicate with two other agents at each time step. In Chakraborty et al. (2017), at each time step, agents decide either to sample an option or to broadcast the last obtained reward to the entire group. In this setting, each agent suffers from an opportunity cost whenever it decides to broadcast the last obtained reward. A communication strategy where agents observe the rewards and choices of their neighbors according

to a leader-follower setting is considered in Landgren et al. (2018). Decentralized bandit problems with communication costs are considered in the works of Tao et al. (2019); Wang et al. (2020). Tao et al. (2019) consider the pure exploration bandit problem with a communication cost equivalent to the number of times agents communicate. Wang et al. (2020) propose an algorithm that achieves near-optimal performance with a communication cost equivalent to the amount of data transmitted. Madhushani & Leonard (2020) propose a communication rule where agents observe their neighbors when they execute an exploring action.

2 METHODOLOGY

2.1 PROBLEM FORMULATION

In this section we present the mathematical formulation of the problem. Consider a group of K agents faced with the same N-armed bandit problem for T time steps. In this paper we use the terms arms and options interchangeably. Let X_i be a sub-Gaussian random variable with variance proxy σ_i^2, which denotes the reward of option $i \in \{1, 2, \ldots, N\}$. Define $\mathbb{E}(X_i) = \mu_i$ as the expected reward of option i. We define the option with maximum expected reward as the optimal option $i^* = \arg\max\{\mu_1, \ldots, \mu_N\}$. Let $\Delta_i = \mu_{i^*} - \mu_i$ be the expected reward gap between option i^* and option i. Let $\mathbb{I}_{\{\varphi_t^k = i\}}$ be the indicator random variable that takes value 1 if agent k chooses option i at time t and 0 otherwise.

We define the communication network as follows. Let $G(\mathcal{V}, \mathcal{E})$ be a fixed nontrivial graph that defines neighbors, where $\mathcal{V}$ denotes the set of agents and $e(k, j) \in \mathcal{E}$ denotes the communication link between agents k and j. Let $\mathbb{I}^t_{\{\cdot,k\}}$ be the indicator variable that takes value 1 if agent k shares its reward value and choice with its neighbors at time t. Since agents can send reward values and choices only to their neighbors, it holds that $\mathbb{I}^t_{\{j,k\}} = 0, \forall k, j, t$, such that $e(j, k) \notin \mathcal{E}$.

2.2 OUR ALGORITHM

Let $\widehat{\mu}_i^k(t)$ be the estimated mean of option i by agent k at time t. Let $n_i^k(t)$ and $N_i^k(t)$ denote the number of samples of option i and the number of observations of option i, respectively, obtained by agent k until time t. $N_i^k(t)$ is equal to $n_i^k(t)$ plus the number of observations of option i that agent k received from its neighbors until time t. So, by definition

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

Assumption 1 Initially, all the agents are given a reward value for one sample from each option.

The initial given reward values are used as the empirical estimates of the mean values of the options. Let $X_i^k(0)$ denote the reward received initially by agent k for option i. The estimated mean value is calculated by taking the average of the total reward observed for option i by agent k until time t:

$$\widehat{\mu}_i^k(t) = \frac{S_i^k(t) + X_i^k(0)}{N_i^k(t) + 1}$$

where $S_i^k(t) = \sum_{\tau=1}^{t} \sum_{j=1}^{k} X_i \mathbb{I}_{\{\varphi_\tau^j = i\}} \mathbb{I}^\tau_{\{k,j\}}$.

The goal of each agent is to maximize its individual cumulative reward, thereby contributing to maximizing the group cumulative reward. We assume known variance proxy as follows.

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

Assumption 3 When more than one agent chooses the same option at the same time they receive rewards independently drawn from the probability distribution associated with the chosen option.

To realize the goal of maximizing cumulative reward, agents are required to minimize the number of times they sample sub-optimal options. Thus, each agent employs an agent-based strategy that

captures the trade-off between exploring and exploiting by constructing an objective function that strikes a balance between the estimation of the expected reward and the uncertainty associated with the estimate (Auer et al., 2002). Each agent samples options according to the following rule.

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

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

with

$$Q_i^k(t) \triangleq \widehat{\mu}_i^k(t) + C_i^k(t), \quad C_i^k(t) \triangleq \sigma_i \sqrt{\frac{2(\xi+1)\log(t)}{N_i^k(t)+1}}, \quad \textit{and} \quad \xi > 1.$$

$C_i^k(t)$ represents agent k's uncertainty of the estimated mean of option i. When the number of observations of option i is high, the uncertainty associated with the estimated mean of option i will be low; this is reflected in the inverse relation between $C_i^k(t)$ and $N_i^k(t)$.

An exploiting action corresponds to choosing the option with maximum estimated mean value. This occurs when the option with maximum objective function value is the same as the option with maximum estimated mean value. An exploring action correspond to choosing an option with high uncertainty. This occurs when the option with maximum objective function value is different from the option with maximum estimated mean value. Each agent can reduce the number of samples it takes from sub-optimal options by leveraging communication to reduce the uncertainty associated with the estimates of sub-optimal options. Thus, in resource-constrained environments, it is desirable to communicate reward values obtained from sub-optimal options only. Exploring actions often lead to taking samples from sub-optimal options. So, we define a partial communication protocol such that agents share their reward values with their neighbors only when they execute an exploring action.

Definition 2 (Communication Rule) *The communication rule for agent* k *at time* $t \in \{1, \ldots, T\}$ *is*

$$\mathbb{I}_{\{\cdot,k\}}^{t+1} = \begin{cases} 1 & , \quad \varphi_{t+1}^k \neq \arg\max\{\widehat{\mu}_1^k(t), \cdots, \widehat{\mu}_N^k(t)\} \\ 0 & , \quad \text{o.w.} \end{cases}$$

3 RESULTS

The goal of maximizing cumulative reward is equivalent to minimizing cumulative regret, which is the loss incurred by the agent when sampling sub-optimal options. We analyze the performance of the proposed algorithm using expected cumulative regret and expected communication cost.

For a group of K agents facing the N-armed bandit problem for T time steps, the expected group cumulative regret can be expressed as

$$\mathbb{E}\left(R(T)\right) = \sum_{i=1}^{N} \sum_{k=1}^{K} \Delta_i \mathbb{E}\left(n_i^k(T)\right).$$

Thus, minimizing the expected group cumulative regret can be achieved by minimizing the expected number of samples taken from sub-optimal options.

Communication Cost Since communication refers to agents sharing their reward values and actions with their neighbors, each communicated message has the same length. We define communication cost as the total number of times the agents share their reward values and actions during the decision-making process. Let $L(T)$ be the group communication cost up to time T. Then, we have that

$$L(T) = \sum_{k=1}^{K} \sum_{t=1}^{T} \mathbb{I}_{\{\cdot,k\}}^{t}. \tag{1}$$

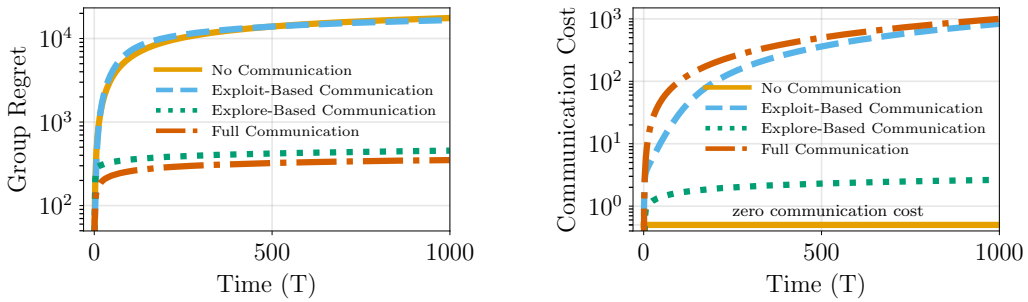
Under full communication, expected communication cost is $O(T)$. We now proceed to analyze the expected communication cost under the proposed partial communication protocol.

Lemma 1 *Let $\mathbb{E}(L(T))$ be the expected cumulative communication cost of the group under the communication rule given in Definition 2. Then, we have that*

$$\mathbb{E}(L(T)) = O(\log T).$$

The proof of Lemma 1 follows from Lemma 3 in the paper Madhushani & Leonard (2020). A detailed proof is provided in Appendix A.

Experimental Results We provide numerical simulation results illustrating the performance of the proposed sampling rule and the communication rule. For all the simulations presented in this section, we consider a group of 100 agents ($K = 100$) and 10 options ($N = 10$) with Gaussian reward distributions. 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 1000 time steps ($T = 1000$) using 1000 Monte Carlo simulations with $\xi = 1.01$.



(a) Expected cumulative group regret of 100 agents with sampling rule from Definition 1 under full communication, explore-based communication, exploit-based communication and no communication.

(b) Expected cumulative communication cost per agent for a group of 100 agents under full communication, explore-based communication, exploit-based communication and no communication.

Figure 1: Performance of a group of 100 agents using the sampling rule given in Definition 1 under different communication protocols.

Figure 1(a) presents expected cumulative group regret for 1000 time steps. The curves illustrate that both full communication and explore-based communication significantly improve the performance of the group as compared to the case of no communication. Further, group performance with explore-based communication is of the same order as group performance with full communication. Group performance improvement obtained with exploit-based communication is insignificant as compared to the case of no communication. Figure 1(b) presents the results for expected cumulative communication cost per agent for 1000 time steps. The curves illustrate that communication cost incurred by explore-based communication is significantly smaller than the cost incurred by full communication and by exploit-based communication. In fact, the cost incurred by exploit-based communication is quite close to the cost incurred by full communication. Overall, the results illustrate that our proposed explore-based communication protocol, in which agents share their reward values and actions only when they execute an exploring action, incurs only a small communication cost while significantly improving group performance.

4 Discussion and Conclusion

The development of cost-effective communication protocols for distributed learning is desirable in resource-constrained environments. We proposed a new partial communication protocol for sharing (broadcasting) in the distributed multi-armed bandit problem. We showed that the proposed communication protocol has a significantly smaller communication cost as compared to the case of full communication while obtaining the same order of performance. An important future extension of our work is to analyze and improve the performance of the proposed communication protocol under random communication failures.

ACKNOWLEDGEMENT

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

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.

Mithun Chakraborty, Kai Yee Phoebe Chua, Sanmay Das, and Brendan Juba. Coordinated versus decentralized exploration in multi-agent multi-armed bandits. In *IJCAI*, pp. 164–170, 2017.

Audrey Durand, Charis Achilleos, Demetris Iacovides, Katerina Strati, Georgios D. Mitsis, and Joelle Pineau. Contextual bandits for adapting treatment in a mouse model of de novo carcinogenesis. In *MLHC*, 2018.

Raphaël Féraud, Réda Alami, and Romain Laroche. Decentralized exploration in multi-armed bandits. *arXiv preprint arXiv:1811.07763*, 2018.

Dileep Kalathil, Naumaan Nayyar, and Rahul Jain. Decentralized learning for multiplayer multiarmed bandits. *IEEE Transactions on Information Theory*, 60(4):2331–2345, 2014.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

Peter Landgren, Vaibhav Srivastava, and Naomi Ehrich Leonard. On distributed cooperative decision-making in multiarmed bandits. In *European Control Conference (ECC)*, pp. 243–248, 2016a.

Peter Landgren, Vaibhav Srivastava, and Naomi Ehrich Leonard. Distributed cooperative decision-making in multiarmed bandits: Frequentist and Bayesian algorithms. In *IEEE Conference on Decision and Control (CDC)*, pp. 167–172, 2016b.

Peter Landgren, Vaibhav Srivastava, and Naomi Ehrich Leonard. Social imitation in cooperative multiarmed bandits: Partition-based algorithms with strictly local information. In *IEEE Conference on Decision and Control (CDC)*, pp. 5239–5244, 2018.

Peter Landgren, Vaibhav Srivastava, and Naomi Ehrich Leonard. Distributed cooperative decision making in multi-agent multi-armed bandits. *arXiv preprint arXiv:2003.01312*, 2020.

Udari Madhushani and Naomi Ehrich Leonard. Heterogeneous stochastic interactions for multiple agents in a multi-armed bandit problem. In *European Control Conference (ECC)*, pp. 3502–3507, 2019.

Udari Madhushani and Naomi Ehrich Leonard. A dynamic observation strategy for multi-agent multi-armed bandit problem. In *European Control Conference (ECC)*, 2020.

David Martínez-Rubio, Varun Kanade, and Patrick Rebeschini. Decentralized cooperative stochastic bandits. In *Advances in Neural Information Processing Systems*, pp. 4531–4542, 2019.

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

Balázs Szörényi, Róbert Busa-Fekete, István Hegedűs, Róbert Ormándi, Márk Jelasity, and Balázs Kégl. Gossip-based distributed stochastic bandit algorithms. In *Journal of Machine Learning Research Workshop and Conference Proceedings*, volume 2, pp. 1056–1064. International Machine Learning Society, 2013.

Chao Tao, Qin Zhang, and Yuan Zhou. Collaborative learning with limited interaction: Tight bounds for distributed exploration in multi-armed bandits. In *IEEE Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 126–146, 2019.

Aristide CY Tossou and Christos Dimitrakakis. Algorithms for differentially private multi-armed bandits. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

Yuanhao Wang, Jiachen Hu, Xiaoyu Chen, and Liwei Wang. Distributed bandit learning: Near-optimal regret with efficient communication. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SJxZnR4YvB.

Romain Warlop, Alessandro Lazaric, and Jérémie Mary. Fighting boredom in recommender systems with linear reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 1757–1768, 2018.

A Expected Communication Cost

Lemma 1 *Let $\mathbb{E}(L(T))$ be the expected cumulative communication cost of the group under the communication rule given in Definition 2. Then, we have that*

$$\mathbb{E}(L(T)) = O(\log T).$$

Proof of Lemma 1 The expected communication cost can be given as

$$\mathbb{E}\left(L(T)\right) = \sum_{k=1}^{K}\sum_{t=1}^{T}\mathbb{P}\left(\varphi_k^t = i, \widehat{\mu}_i^k(t) \neq \max\{\widehat{\mu}_1^k(t), \cdots, \widehat{\mu}_N^k(t)\}\right).$$

To analyze the expected number of exploring actions, we use

$$\begin{aligned}\mathbb{P}\left(\varphi_k^t = i, \widehat{\mu}_i^k(t) \neq \max\{\widehat{\mu}_1^k(t), \cdots, \widehat{\mu}_N^k(t)\}\right) = \mathbb{P}\left(\varphi_k^t = i^*, \widehat{\mu}_{i^*}^k(t) \neq \max\{\widehat{\mu}_1^k(t), \cdots, \widehat{\mu}_N^k(t)\}\right)\\ +\mathbb{P}\left(\varphi_k^t = i, \widehat{\mu}_i^k(t) \neq \max\{\widehat{\mu}_1^k(t), \cdots, \widehat{\mu}_N^k(t)\}, i \neq i^*\right).\end{aligned}$$

We first upper bound the expected number of times agent k broadcasts rewards and actions to its neighbors until time T when it samples a sub-optimal option:

$$\sum_{t=1}^{T}\mathbb{P}\left(\varphi_k^t = i, \widehat{\mu}_i^k(t) \neq \max\{\widehat{\mu}_1^k(t), \cdots, \widehat{\mu}_N^k(t)\}, i \neq i^*\right) \leq \sum_{t=1}^{T}\sum_{i=1}^{N}\mathbb{E}\left(\mathbb{I}_{\{\varphi_k^t = i\}}\right) = O(\log T). \tag{2}$$

This follows from the fact that we only sample sub-optimal options logarithmically with time (see Lemma 2 in Madhushani & Leonard (2020)).

Next we analyze the expected number of times agent k broadcasts rewards and actions to its neighbors until time T when it samples the optimal option. Note that $\forall i, k, t$ we have

$$\begin{aligned}\{\varphi_k^t = i^*, \widehat{\mu}_{i^*} \neq \max\{\widehat{\mu}_i^k(t), \cdots, \widehat{\mu}_N^k(t)\}\} \subseteq \{\widehat{\mu}_{i^*}^k(t) \leq \mu_{i^*} - C_{i^*}^k(t)\}\\ \cup\{\widehat{\mu}_{i^*}^k(t) \geq \mu_{i^*} - C_{i^*}^k(t), \exists i\ s.t.\ \widehat{\mu}_i^k(t) \geq \widehat{\mu}_{i^*}^k(t), \varphi_k^t = i^*\}.\end{aligned}$$

Thus, we have

$$\begin{aligned}\sum_{i=1}^{T}\mathbb{P}\left(\varphi_k^t = i^*, \widehat{\mu}_{i^*} \neq \max\{\widehat{\mu}_i^k(t), \cdots, \widehat{\mu}_N^k(t)\}\right) \leq \sum_{i=1}^{T}\mathbb{P}\left(\widehat{\mu}_{i^*}^k(t) \leq \mu_{i^*} - C_{i^*}^k(t)\right)\\ +\sum_{i=1}^{T}\mathbb{P}\left(\widehat{\mu}_{i^*}^k(t) \geq \mu_{i^*} - C_{i^*}^k(t), \exists i\ s.t.\ (\widehat{\mu}_i^k(t) \geq \widehat{\mu}_{i^*}^k(t), \varphi_k^t = i^*\right).\end{aligned} \tag{3}$$

From Lemma 1 in Madhushani & Leonard (2020) we get

$$\sum_{i=1}^{T}\mathbb{P}\left(\widehat{\mu}_{i^*}^k(t) \leq \mu_{i^*} - C_{i^*}^k(t)\right) = O(\log T), \tag{4}$$

Now we proceed to upper bound the second summation term of (3). Note that for some $\beta_i^k(t) > 0$ we have

$$\begin{aligned}
&\sum_{i=1}^{T} \mathbb{P}\left(\widehat{\mu}_{i^*}^k(t) \geq \mu_{i^*} - C_{i^*}^k(t), \widehat{\mu}_i^k(t) \geq \widehat{\mu}_{i^*}^k(t), \varphi_k^t = i^*\right) \\
&\qquad \leq \sum_{i=1}^{T} \mathbb{P}\left(\widehat{\mu}_{i^*}^k(t) \geq \mu_{i^*} - C_{i^*}^k(t), \widehat{\mu}_i^k(t) \geq \widehat{\mu}_{i^*}^k(t), N_{i^*}^k(t) \leq \beta_i^k(t), \varphi_k^t = i^*\right) \\
&\qquad + \sum_{i=1}^{T} \mathbb{P}\left(\widehat{\mu}_{i^*}^k(t) \geq \mu_{i^*} - C_{i^*}^k(t), \widehat{\mu}_i^k(t) \geq \widehat{\mu}_{i^*}^k(t), N_{i^*}^k(t) > \beta_i^k(t)\right) \\
&\qquad \leq \beta_i^k(T) + \sum_{i=1}^{T} \mathbb{P}\left(\widehat{\mu}_{i^*}^k(t) \geq \mu_{i^*} - C_{i^*}^k(t), \widehat{\mu}_i^k(t) \geq \widehat{\mu}_{i^*}^k(t), N_{i^*}^k(t) > \beta_i^k(t)\right).
\end{aligned}$$

Let i be the sub-optimal option with highest estimated expected reward for agents k at time t. Then we have $i = \arg\max\{\widehat{\mu}_1^k(t), \cdots, \widehat{\mu}_N^k(t)\}$ and $i \neq i^*$. If agent k chooses option i^* at time $t+1$ we have $Q_{i^*}^k(t) > Q_i^k(t)$. Thus we have $\widehat{\mu}_i^k(t) > \widehat{\mu}_{i^*}^k(t)$ and $C_i^k(t) < C_{i^*}^k(t)$. Then for $\beta_i^k(t) = \frac{8\sigma_{i^*}(\xi+1)}{\Delta_i^2}\log t$ we have

$$\begin{aligned}
&\sum_{i=1}^{T} \mathbb{P}\left(\widehat{\mu}_{i^*}^k(t) \geq \mu_{i^*} - C_{i^*}^k(t), \widehat{\mu}_i^k(t) \geq \widehat{\mu}_{i^*}^k(t), N_{i^*}^k(t) > \beta_i^k(t)\right) \\
&\qquad \leq \sum_{i=1}^{T} \mathbb{P}\left(\widehat{\mu}_i^k(t) \geq \mu_{i^*} - C_{i^*}^k(t), \mu_{i^*} > \mu_i + 2C_{i^*}^k(t)\right) \\
&\qquad \leq \sum_{t=1}^{T} \mathbb{P}\left(\widehat{\mu}_i^k(t) \geq \mu_i + C_i^k(t)\right) = O(\log T).
\end{aligned}$$

The last equality follows from Lemma 1 by Madhushani & Leonard (2020).

Then we have

$$\sum_{i=1}^{T} \mathbb{P}\left(\widehat{\mu}_{i^*}^k(t) \geq \mu_{i^*} - C_{i^*}^k(t), \exists i \; s.t. \; (\widehat{\mu}_i^k(t) \geq \widehat{\mu}_{i^*}^k(t), \varphi_k^t = i^*\right) = O(\log T). \tag{5}$$

The proof of Lemma 1 follows from Equations (2)-(5).