DISTRIBUTED LEARNING: SEQUENTIAL DECISION MAKING IN RESOURCE-CONSTRAINED ENVIRONMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study cost-effective communication strategies that can be utilized to improve the performance of distributed learning systems in resource-constrained environments. First, we propose a new cost-effective partial communication protocol for distributed learning in sequential decision making. We illustrate that with this protocol the group obtains the same order of performance they obtain with full communication. Next, we prove that under the proposed partial communication protocol communication cost is $O(\log T)$ as opposed to the full communication that obtains a communication cost of $O(T)$, where T is the time horizon of the decision making process. Finally, we validate the theoretical results using numerical simulations.

1 INTRODUCTION

In resource constrained environments, the difficulty of constructing and maintaining large scale infrastructures limits the possibility of developing a centralized learning system with access to acquire all information, resources to effectively process obtained information and capacity to make all decisions. Consequently, developing distributed learning systems, i.e., groups of units who collectively process information and make decisions, that utilize a minimum amount of resources 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), 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.

In particular, we focus on developing cost-effective distributed learning techniques for sequential decision making under stochastic outcomes. This work is motivated by the growing number of real-world applications such as clinical trials, recommender systems, and user-targeted online advertising. Consider a set of organizations networked to recommend educational programs to online users under high demand. Each company makes a series of sequential decisions about which programs to recommend according to the user feedback (Warlop et al., 2018; Féraud et al., 2018). Similarly to it, 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 cases received feedback is stochastic i.e., 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 external and internal factors such as environmental conditions and state of their mind. Establishing a communication network that facilitates *full communication:* each company shares all feedback immediately with others, can significantly improve the performance of these systems. However, oftentimes communication can 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, this communication protocol becomes impractical. We aim to address this problem by proposing a *partial communication* strategy that obtains the *same order of performance* as full communication protocol while using a significantly small amount of data communication.

To solve this problem, we consider 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 a given fixed set of options (Kalathil et al., 2014; Landgren et al., 2016a;b; Martínez-Rubio et al., 2019). After every choice, each agent receives a numerical reward drawn from a probability distribution associated with the chosen option. The objective of each agent is maximizing the individual cumulative reward while contributing to maximizing the group cumulative reward. The best strategy in this situation is to repeatedly choose the optimal option, i.e., the option that provides the maximum average reward. However, agents are unaware of 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 the options that might potentially provide higher rewards. This process is sped up through collective learning by sharing reward values and actions with their neighbors. We consider sharing information only when agents execute exploiting actions as *exploit based communication*. Similarly, we consider sharing information only when agents execute exploring actions as *explore based communication*. Note that

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

We propose a new partial communication protocol that shares information only when agents execute an exploring action. We illustrate that explore based communication obtains the same order of performance as full communication while incurring a significantly small communication cost.

Key Contributions In this work, we study effective 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.
- We illustrate that with this protocol the group obtains the same order of performance they obtain with full communication.
- We prove that under the proposed partial communication protocol communication cost is $O(\log T)$ as opposed to the full communication that obtains a communication cost of $O(T)$, where T is the number of decision making steps.

Related Work A large number of previous works (Kalathil et al., 2014; Landgren et al., 2016a;b; 2018; Martínez-Rubio et al., 2019) have considered distributed bandit problem without a communication cost. They analyze how communication structure affects individual and group performance. A decentralized multi-agent setting is considered in Kalathil et al. (2014); Landgren et al. (2016a;b); Martínez-Rubio et al. (2019). Landgren et al. (2016a;b) use a running consensus algorithm to update estimates and provide a graph structure-dependent performance measure that predicts the relative performance of agents. Martínez-Rubio et al. 2019 provides an improved bound for this problem by proposing an accelerated consensus procedure. Szörényi et al. 2013 considers a P2P communication where an agent is only allowed to communicate with two other agents at each time step. 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) considers the pure exploration bandit problem with a communication cost equivalent to the number of times agents communicate. Wang et al. (2020) proposes an algorithm that achieves near-optimal performance with a communication cost equivalent to the amount of data transmitted. Authors (2020) proposes 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 $i \in \{1, 2, \dots, N\}$. Define $\mathbb{E}\left(X_i\right) = \mu_i$ as the expected reward of option i. We define the option with maximum expected reward as the optimal option $i^* = \arg\max\{\mu_1, \dots, \mu_N\}$. 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 chose the option i at time t and 0 otherwise.

We define the communication network as follows. Let $G(\mathcal{V}, \mathcal{E})$ be a fixed non trivial graph that defines neighbors, where $\mathcal{V}$ denotes the set of agents and $e(k, j) \in \mathcal{E}$ denotes the communication link between agent k and j. Let $\mathbb{I}_{\{\cdot,k\}}^t$ 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 we see that $\mathbb{I}_{\{j,k\}}^t = 0, \forall k, j, t$ 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 agent k obtained from its neighbors until time t. So, by definition

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

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

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

The goal of each agent is to maximize its individual cumulative reward while contributing to maximizing the group cumulative reward. In this work, we consider the case with known variance proxy. We formally state this assumption as follows.

Assumption 2 We assume that 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 $\{\varphi_t^k\}_1^T$ of the agent k at time $t \in \{1, \dots, T\}$ 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

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

Here the term $C_i^k(t)$ is associated with the uncertainty of the estimated mean of the option i. Note that when the number of observations taken from the option i is high, the uncertainty associated with the estimated mean of the option i is low and vice-versa. Since $C_i^k(t)$ and number of observations $N_i^k(t)$ are inversely related, where high $C_i^k(t)$ corresponds to high uncertainty and low $C_i^k(t)$ corresponds to low uncertainty.

Exploiting actions correspond to choosing the options with high estimated mean values, i.e., an option with maximum objective function value is same as the option with maximum estimated mean value, and exploring actions correspond to choosing options with high uncertainties, i.e., 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. Oftentimes executing exploring actions leads to taking samples from sub-optimal options. Thus we define a partial communication protocol such that agents share their reward values with their neighbors only when they execute an exploring action.

Followed by this intuition we propose a partial communication rule as follows:

Definition 2 (Communication Rule) *The communication rule of the agent k at time $t \in \{1, \dots, T\}$ is defined as*

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

3 Results

The goal of maximizing the cumulative reward is equivalent to minimizing the cumulative regret, which is the loss incurred by the agent through 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 N-armed bandit problem for T time steps, expected group cumulative regret can be given 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, the expected group cumulative regret can be minimized by minimizing the expected number of samples taken from suboptimal options.

Communication Cost Recall that through communication agents share their reward values and actions with their neighbors. Consequently, each communicated message has the same length. Thus we define the 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,

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

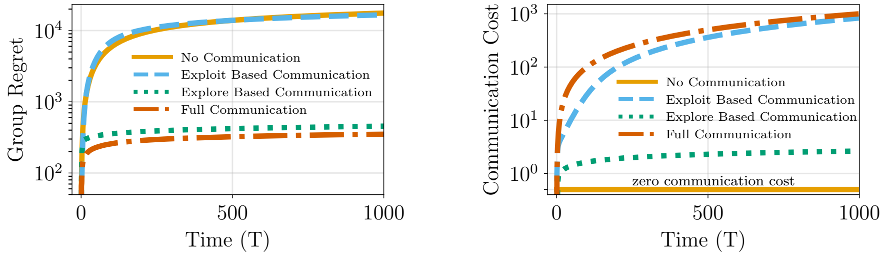
Note that under full communication expected communication cost is $O(T)$. Now we processed to analyze the expected communication cost under the proposed partial communication protocol.

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

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

The proof of Lemma 1 follows from Lemma 3 in the paper Authors (2020).

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 consider the communication network (Graph G) as a complete graph. 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 using the sampling rule given in Definitions 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 strategies.

Figure 1(a) presents expected cumulative group regret for 1000 time steps. This illustrates that both full communication and explore based communication significantly improves the performance of the group compared to no communication. Performance improvement obtained by utilizing exploit based communication is insignificant compared to the performance under no communication. Further, this shows that the performance of explore based communication is of the same order as the group performance under full communication. This illustrates that sharing reward values obtained through executing exploiting actions do not contribute to significant performance improvement. However, it incurs a significant communication cost. Figure 1(b) presents the results for expected cumulative communication cost per agent for 1000 time steps. This illustrates that communication cost incurred by explore based communication is significantly smaller than the cost incurred by full communication. Communication cost incurred by exploit based communication is close to the cost incurred by full communication. These results illustrate that explore based communication protocol incurs only a small communication cost while significantly improving the group performance.

4 Discussion and Conclusion

We studied the development of cost-effective communication protocols that are desirable in resource constrained environments. In particular, we proposed a new partial communication protocol for distributed multi-armed bandit problem. We illustrated that the proposed communication protocol has a significantly small communication cost as opposed to full communication while obtaining the same order of performance. Another aspect of this problem is developing effective communication protocols for the networks that are prone to communication errors. A future extension for this problem can be analyzing and improving the performance of the proposed communication protocol under random communication failures.

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.

Anonymous Authors. Suppressed for anonymity. 2020.

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. IEEE, 2016a.

Peter Landgren, Vaibhav Srivastava, and Naomi Ehrich Leonard. Distributed cooperative decision-making in multiarmed bandits: Frequentist and bayesian algorithms. In *IEEE 55th Conference on Decision and Control (CDC)*, pp. 167–172. IEEE, 2016b.

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

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 Societ, 2013.

Chao Tao, Qin Zhang, and Yuan Zhou. Collaborative learning with limited interaction: Tight bounds for distributed exploration in multi-armed bandits. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 126–146. IEEE, 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.