

# Heterogeneous Stochastic Interactions for Multiple Agents in a Multi-armed Bandit Problem

Udari Madhushani and Naomi Ehrich Leonard

**Abstract**— We define and analyze a multi-agent multi-armed bandit problem in which decision-making agents can observe the choices and rewards of their neighbors. Neighbors are defined by a network graph with heterogeneous and stochastic interconnections. These interactions are determined by the sociability of each agent, which corresponds to the probability that the agent observes its neighbors. We design an algorithm for each agent to maximize its own expected cumulative reward and prove performance bounds that depend on the sociability of the agents and the network structure. We use the bounds to predict the rank ordering of agents according to their performance and verify the accuracy analytically and computationally.

## I. INTRODUCTION

Animal and robotic foraging problems that involve searching over spatially distributed patches with uncertain distribution of resource (reward) result in the explore-exploit dilemma. Each animal or robotic forager chooses which patch to sample, sequentially in time. The dilemma at every time in the sequence is whether to choose a well sampled patch that is expected to reap the highest reward (exploit) or to choose a poorly sampled patch that is expected to reduce uncertainty and possibly identify an option with an even higher reward (explore). Choices to explore can be costly or risky since they may not return much reward, but successful exploiting requires sufficient information that comes from exploration. The challenge is sorting out protocols for when and where to explore versus when and where to exploit.

In decision theory, multi-armed bandit (MAB) problems serve as models that capture the salient features of the explore-exploit trade-off [1], [2]. Thus, advances in addressing MAB problems directly benefit understanding and solving foraging and search problems. The MAB problem is analogous to a scenario in which an agent is repeatedly faced with several different options, each returning uncertain reward, and aims to make a sequence of choices to maximize the cumulative reward [3]. This is equivalent to minimizing the cumulative regret [4].

In their seminal work, Lai and Robbins [4] established a lower bound for the expected cumulative regret in the finite time horizon case. Specifically, they derived a logarithmic lower bound for the expected number of times a sub-optimal option needs to be sampled by an optimal sampling rule. They also established a confidence bound and a sampling rule to achieve logarithmic cumulative regret. These results were further simplified in [5] by introducing a confidence bound

using a sample mean based method. Improving on these results, a family of Upper Confidence Bound (UCB) algorithms for achieving asymptotic and uniform logarithmic cumulative regret were proposed in [6]. These UCB algorithms are based on the notion that minimizing the expected cumulative regret is realized by choosing an appropriate uncertainty model, which results in an optimal trade-off between reward gain and information gain through uncertainty.

For the standard MAB problem the reward associated with an option is considered to be an iid stochastic process. Therefore, in the frequentist setting, the natural way of estimating the expectation of the reward is to consider the sample average [4], [5], [6]. The papers [7], [8] present how to incorporate prior knowledge about reward expectation in the estimation step by leveraging the theory of conditional expectation in the Bayesian setting.

The papers [9], [10] extend this work to the multi-agent setting where the explore-exploit problem is defined for a group of agents. The objective is to understand how the performance of individuals or the group can benefit from inter-agent observations. A centralized multi-agent setting is considered in [9] and a decentralized setting is considered in [11]. The papers [12], [10] use a running consensus algorithm in which agents observe the estimates of their neighbors. A multi-agent multi-armed bandit (MAMAB) problem, where agents observe instantaneous rewards and actions in a leader-follower setting, is considered in [13], [14].

In the above works, agents observe their neighbors, defined according to a static network graph. In this paper we consider a MAMAB problem where agents observe instantaneous rewards and actions of their neighbors through stochastic interactions. Observing rewards and actions rather than estimates is motivated by the foraging success of social groups where agents can observe neighbors even when their neighbors don't necessarily want to share what they know [15]. Further, we assume that each agent  $k$  only observes its neighbors with some probability  $p_k$ . This provides a framework for evaluating efficiency and robustness to changes in communication in terms of the  $p_k$ .

The setting is formulated by defining an underlying undirected network graph and imposing directed observation probabilities on each edge. The underlying graph models the inherent observation constraints present in the network. Imposed observation probabilities capture the heterogeneous social effort of agents in observing neighbors. We introduce the notion of *sociability* of each agent  $k$  as the likelihood  $p_k$  that the agent observes its neighbors. We derive analytical upper bounds for expected cumulative regret and propose

This research has been supported in part by ARO grant W911NF-18-1-0325 and ONR grant N00014-14-1-0635. Department of Mechanical and Aerospace Engineering, Princeton University, NJ 08544, USA. {udarim, naomi}@princeton.edu

a measure to rank agents according to their relative performance as a function of each agent's sociability and the sociability of its neighbors. We show that our model predicts how high performance requires an agent to have *both* high sociability *and* neighbors with low sociability. This is an important result: it implies that making an investment to observe neighbors may be worthwhile only if those neighbors are sufficiently explorative.

In Section II we introduce the MAMAB problem with time-varying (stochastic) observation structure. We propose an efficient sampling rule for an agent to maximize cumulative expected reward. We analyze the performance of the proposed sampling rule in Section III. In Section III-A we analytically upper bound the expected cumulative regret and in Section III-B we propose a measure to predict the ranks of agents according to their relative performance. In Section IV we provide numerical simulation results and computationally validate the analytical results. We conclude in Section V and provide additional mathematical details in the Appendix.

## II. MULTI-AGENT MULTI-ARMED BANDIT PROBLEM

Consider a MAMAB with  $K$  agents and  $N$  arms, which represent  $N$  options (patches). Let  $X_i$  be a sub-Gaussian random variable with variance proxy  $(\sigma'_i)^2$  that denotes the reward associated with option  $i \in \mathcal{I} \triangleq \{1, 2, \dots, N\}$ . Define  $\mu_i$  as the expected reward  $\mathbb{E}(X_i)$  of option  $i$ . Let  $i^*$  be the optimal option such that  $\mu_{i^*} = \max\{\mu_1, \dots, \mu_N\}$ . Each agent  $k \in \{1, \dots, K\}$  chooses one option at each time step with the goal of maximizing its cumulative reward, which is equivalent to minimizing its cumulative regret. We assume that the  $\mu_i$  are unknown and the  $\sigma'_i$  are known to the agents (e.g., if  $\sigma'_i$  is the magnitude of sensor noise).

Let  $\mathcal{G}(\mathcal{V}, \mathcal{E})$  be an undirected graph that encodes the observation structure of the system. If there exists an observation link  $e_{kj}$  from agent  $k$  to agent  $j$ ,  $j, k \in \mathcal{V}$ , then  $\{k, j\}, \{j, k\} \in \mathcal{E}$ ,  $e_{kj} = e_{jk} = 1$ , and we say agents  $k$  and  $j$  are neighbors. For social or robotic groups searching over physical space, observation links may exist between pairs of agents when they are sufficiently close to be visible (or otherwise observable) to each other. Let  $d_k$  be the number of neighbors and  $\mathcal{V}_{d_k}$  the set of neighbors of agent  $k$ .

Let  $\varphi_k^t \in \mathcal{I}$  and  $X_i^t$  be random variables that denote the option chosen and reward received by agent  $k$  at time  $t$ , respectively. Let  $X_i^t, \forall t$ , be i.i.d copies of  $X_i$ . Define  $\sigma_i = \sqrt{K}\sigma'_i$ . Consider the probability space  $(\Omega, \mathcal{U}, \mathbb{P})$  and the increasing sequence of subalgebras  $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \dots \subset \mathcal{F}_t \subset \dots \subset \mathcal{F}_{T-1} \subset \mathcal{U}$  for  $t = 0, 1, \dots, T$ , where  $\mathbb{P}$  is the probability measure on the sigma algebra  $\mathcal{U}$  of  $\Omega$ . Here  $\mathcal{F}_t$  is the sigma algebra generated by information available at time  $t$ . Let  $\mathbb{I}_{\{\varphi_k^t=i\}}$  be a  $\mathcal{F}_{t-1}$  measurable indicator random variable that takes the value one if option  $i$  is chosen by agent  $k$  at time  $t$  and is zero otherwise. Define  $\mathbb{I}_{\{k,j\}}^t$  to be an  $\mathcal{F}_{t-1}$  measurable indicator random variable that takes the value one if agent  $k$  can observe agent  $j$  at time  $t$  and is zero otherwise.

Let  $p_k$  be the probability that agent  $k$  observes the instantaneous actions and rewards of its neighbors. Then we

have  $\mathbb{E}\left(\mathbb{I}_{\{k,j\}}^t\right) = p_k, \forall j \neq k$  such that  $\{k, j\} \in \mathcal{E}$ . We let  $\mathbb{I}_{\{k,k\}}^t = 1$  and  $\mathbb{I}_{\{k,j\}}^t = 0, \forall j \neq k$  such that  $\{k, j\} \notin \mathcal{E}$ . An agent  $k$  that has high probability  $p_k$  is more likely to obtain observations from its neighbors. We introduce the notion of an agent's *sociability* to refer to its value of  $p_k$ : we interpret agents with high  $p_k$  values to be more sociable and agents with low  $p_k$  values to be less sociable.

In order to maximize the cumulative reward in the long run, agents need to both identify the best options through exploring and sample the best options through exploiting. Since agents with high sociability values are more likely to obtain a greater number of observations, they can identify the best options with less exploring. However the usefulness of their observations is affected by the sociability values of their neighbors. Better performance can be obtained by an agent when it observes neighbors that do a lot of exploring (because they are less sociable), as compared to when it observes agents that do a lot of exploiting (because they are more sociable). This is, the agent will be able to exploit more, without compromising performance, when it has neighbors that are less sociable.

Let the number of times agent  $k$  samples option  $i$  in  $T$  trials be given by the  $\mathcal{F}_{T-1}$  measurable random variable  $n_i^k(T) \triangleq \sum_{t=1}^T \mathbb{I}_{\{\varphi_k^t=i\}}$ . And let the total number of times that agent  $k$  observes rewards from option  $i$  be the  $\mathcal{F}_{T-1}$  measurable random variable  $N_i^k(T)$ , given as

$$N_i^k(T) = \sum_{t=1}^T \sum_{j=1}^K \mathbb{I}_{\{\varphi_j^t=i\}} \mathbb{I}_{\{k,j\}}^t.$$

We define a sampling rule for agent  $k$  as follows. Let the  $\mathcal{F}_t$  measurable random variable  $\hat{\mu}_i^k(T)$  be the estimate of option  $i$  by agent  $k$  at time  $t$ . We define

$$\hat{\mu}_i^k(T) = \frac{S_i^k(T)}{N_i^k(T)},$$

where  $S_i^k(T) = \sum_{t=1}^T \sum_{j=1}^K X_i^t \mathbb{I}_{\{\varphi_j^t=i\}} \mathbb{I}_{\{k,j\}}^t$  is the total reward observed by agent  $k$  from option  $i$  in  $T$  trials.

*Definition 1:* The *sampling rule*  $\{\varphi_k^t\}_{t=1}^T$  for agent  $k$  at time  $t \in \{1, \dots, T\}$  is defined as

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

with

$$Q_i^k(t) \triangleq \hat{\mu}_i^k(t) + C_i^k(t) \quad (2)$$

$$C_i^k(t) \triangleq \sigma_i \sqrt{\frac{2(\xi + 1)(N_i^k(t) + f(t)) \log t}{N_i^k(t) N_i^k(t)}} \quad (3)$$

where  $\xi > 1$  and  $f(t)$  a sublogarithmic nondecreasing nonnegative function.

## III. PERFORMANCE ANALYSIS

In this section we proceed to analyze the performance of the proposed sampling rule by analyzing the expected cumulative regret of the agents. Recall that the expected

cumulative regret depends on the expected number of times the agents sample suboptimal options and the goal of each agent is to maximize its individual cumulative reward.

### A. Regret Analysis

Let  $i$  be a suboptimal option. The total number of times agent  $k$  samples from option  $i$  can be upper bounded as follows:

$$n_i^k(T) = \sum_{t=1}^T \mathbb{I}_{\{\varphi_i^k = i\}} \leq \sum_{t=1}^T \mathbb{I}_{\{Q_i^k(t) \geq Q_{i^*}^k(t)\}}.$$

Here  $\mathbb{I}_{\{Q_i^k(t) > Q_{i^*}^k(t)\}}$  is an indicator function that takes value one if the objective function of option  $i$  is greater than the objective function of the optimal option  $i^*$  and zero otherwise, i.e.,

$$\mathbb{I}_{\{Q_i^k(t) > Q_{i^*}^k(t)\}} = \begin{cases} 1 & , \quad Q_i^k(t) \geq Q_{i^*}^k(t) \\ 0 & , \quad \text{o.w.} \end{cases}$$

Thus we have

$$\mathbb{E}(n_i^k(T)) \leq \sum_{t=1}^T \mathbb{P}(Q_i^k(t) \geq Q_{i^*}^k(t)).$$

Let  $R_i^k(T)$  be the cumulative regret of agent  $k$  from option  $i$ . Define  $\Delta_i \triangleq \mu_{i^*} - \mu_i$ . Then we have, from [16],

$$\mathbb{E}(R_i^k(T)) = \Delta_i \mathbb{E}(n_i^k(T)). \quad (4)$$

We proceed to analyze the expected number of samples from suboptimal options as follows.

First we note that  $\forall i, k, t$  we have

$$\begin{aligned} \{Q_i^k(t) \geq Q_{i^*}^k(t)\} &\subseteq \{\mu_{i^*} < \mu_i + 2C_i^k(t)\} \\ &\cup \{\hat{\mu}_{i^*}^k \leq \mu_{i^*} - C_{i^*}^k(t)\} \cup \{\hat{\mu}_i^k \geq \mu_i + C_i^k(t)\} \end{aligned} \quad (5)$$

$$\begin{aligned} \mathbb{E}(n_i^k(T)) &\leq \sum_{t=1}^T \mathbb{P}(\mu_{i^*} < \mu_i + 2C_i^k(t)) \\ &+ \sum_{t=1}^T \mathbb{P}(\hat{\mu}_{i^*}^k \leq \mu_{i^*} - C_{i^*}^k(t)) + \sum_{t=1}^T \mathbb{P}(\hat{\mu}_i^k \geq \mu_i + C_i^k(t)). \end{aligned} \quad (6)$$

Next we proceed to analyze concentration probability bounds on the estimates of options.

*Theorem 1:* For any  $\zeta > 1$  and for  $\sigma_i > 0$  there exists a  $\vartheta > 0$  such that

$$\mathbb{P}\left(\left|\hat{\mu}_i^k(t) - \mu_i\right| > \sqrt{\frac{\vartheta}{N_i^k(t)}}\right) \leq \frac{\nu \log Kt}{\exp(2\kappa\vartheta)}$$

where

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

*Proof of Theorem 1:* See Appendix.

Using symmetry we conclude that

$$\mathbb{P}\left(\left|\hat{\mu}_i^k(t) - \mu_i\right| > \sqrt{\frac{\vartheta}{N_i^k(t)}}\right) \leq \frac{\nu \log Kt}{\exp(2\kappa\vartheta)}.$$

*Lemma 1:* For  $\vartheta = 2\sigma_i^2(\delta(\xi) + \delta'(\epsilon)) \log t$ , with  $\epsilon > 0$ ,  $\delta(\xi) > 0$ ,  $\delta'(\epsilon) = \delta(\xi)\epsilon/4$ , there exists a  $\zeta > 1$  such that

$$\mathbb{P}\left(\left|\hat{\mu}_i^k(t) - \mu_i\right| > \sigma_i \sqrt{\frac{2(\delta(\xi) + \delta'(\epsilon)) \log t}{N_i^k(t)}}\right) \leq \frac{\nu \log Kt}{t^{\delta(\xi)}}.$$

*Proof of Lemma 1:* Define  $\Gamma(\zeta) = \frac{1}{(\zeta^{\frac{1}{4}} + \zeta^{-\frac{1}{4}})^2}$  where  $\zeta > 1$ .  $\Gamma(\zeta)$  is a monotonically decreasing function and  $\lim_{\zeta \rightarrow 1} \Gamma(\zeta) = \frac{1}{4}$ . This implies that  $\kappa < \frac{1}{4\sigma_i^2}$  since  $\kappa = \Gamma(\zeta)/\sigma_i^2$ . Choose  $\zeta > 1$  such that  $\kappa = \frac{1}{(4+\epsilon)\sigma_i^2}$ , where  $\epsilon > 0$ . Then  $\exists \delta'(\epsilon) > 0$  such that  $4\kappa(\delta(\xi) + \delta'(\epsilon))\sigma_i^2 = \delta(\xi)$ . This proves that  $\exists \zeta > 1$  such that  $2\kappa\vartheta = \delta(\xi) \log t$ . The lemma follows from Theorem 1.

We proceed to upper bound the summation of the probabilities of the events  $\{\mu_{i^*} < \mu_i + 2C_i^k(t)\}$  for  $t \in \{1, 2, \dots, T\}$  as follows. Using the equation (3) we have that the inequality  $\mu_{i^*} < \mu_i + 2C_i^k(t)$  implies

$$\frac{\Delta_i^2}{4\sigma_i^2} (N_i^k(t))^2 - 2(\xi + 1) \log t (N_i^k(t) + f(t)) < 0.$$

This inequality does not hold for  $N_i^k(t) > \eta_i(t)$ , where

$$\eta_i(t) = \frac{4\sigma_i^2(\xi + 1)}{\Delta_i^2} \left(1 + \sqrt{1 + \frac{\Delta_i^2}{2\sigma_i^2(\xi + 1)} \frac{f(t)}{\log t}}\right) \log t.$$

Thus we have

$$\sum_{t=1}^T \mathbb{P}(Q_i^k(t) \geq Q_{i^*}^k(t), N_i^k(t) < \eta_i(t)) \leq \eta_i(T). \quad (7)$$

Using probability bounds given in Lemma 1 and equation (7) we prove our main result on performance bounds.

*Theorem 2:* The total expected number of times agent  $k$  samples suboptimal option  $i$  until time  $T$  is upper bounded as

$$\begin{aligned} \mathbb{E}(n_i^k(T)) &\leq \Gamma(\zeta, \xi, K) + \frac{1}{T^{\xi-1} \log \zeta} \left(\frac{\log K}{T\xi} + \frac{1}{\xi-1}\right) \\ &+ \frac{4\sigma_i^2(\xi + 1)}{\Delta_i^2} \left(1 + \sqrt{1 + \frac{\Delta_i^2}{2\sigma_i^2(\xi + 1)} \frac{f(T)}{\log T}}\right) \log T \end{aligned}$$

where  $\Gamma(\zeta, \xi, K) = \frac{1}{\log \zeta} (1 + \log K) + \frac{1}{2\xi \log \zeta} \left(\frac{\log K}{\xi} + \frac{2}{\xi-1}\right)$  and  $\zeta, \xi > 1$ .

*Proof of Theorem 2:* See Appendix.

Since  $f(T)$  is a sublogarithmic function, the expected number of suboptimal samples are logarithmically bounded.

### B. Performance Measure

In this section we provide a measure to rank the agents according to their relative performance. We motivate with the following two cases for a set of four agents where there is an underlying all-to-all observation structure:

	$k$	1	2	3	4
Case 1	$p_k$	0.5	0	0	0
Case 2	$p_k$	0.5	1	1	1

In Case 1, neighbors of agent 1 are not at all sociable and in Case 2 they are maximally sociable. Therefore neighbors

of agent 1 explore more in Case 1 than in Case 2. As a result, in Case 1, agent 1 tends to obtain observations from neighbors about lesser known options and this allows agent 1 to exploit more. In Case 2, agent 1 tends to obtain observations from neighbors about well sampled options, and this forces agent 1 to explore more. As a result agent 1 performs better in Case 1 as illustrated in Figure 1.

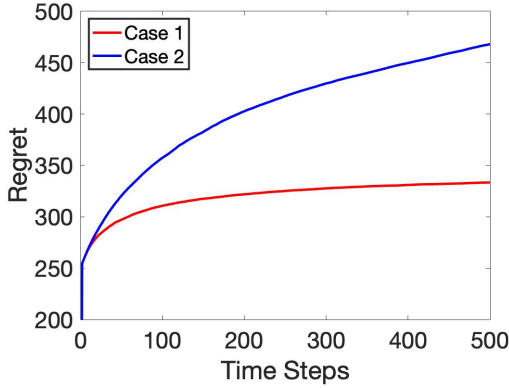


Fig. 1. Expected cumulative regret of agent 1 in Cases 1 and 2.

With this intuition we propose a measure as follows. First we restrict our attention to a class of problems where the underlying observation structure of the agents is a symmetric  $d$ -regular graph. This means that  $d_k = d, \forall k$ , i.e., every agent has the same number of neighbors. The all-to-all graph is a special case of the class of  $d$ -regular graphs with  $d = K - 1$ . We also assume  $p_k \in \{.05, .10, .15, \dots, 1\}, \forall k$ .

We define performance measure  $\epsilon_p^k \in (0, 1]$  for agent  $k$  as

$$\epsilon_p^k \triangleq \frac{1}{p_k + 1} \sqrt{\frac{1}{d_k} \sum_{j=1, j \in \mathcal{V}_{d_k}}^K p_j}. \quad (8)$$

Our goal is to show that a lower  $\epsilon_p^k$  implies a lower cumulative regret and therefore higher performance for agent  $k$ . The measure  $\epsilon_p^k$  is inversely related to agent  $k$ 's sociability  $p_k$  and directly related to the sociability  $p_j$  of the neighbors  $j$  of agent  $k$ . It then makes sense intuitively that lower  $\epsilon_p^k$  implies higher performance for agent  $k$ , since the higher the  $p_k$  the more agent  $k$  observes and the lower the  $p_j$  the more its neighbors explore and the more valuable is their information.

We next design a protocol for each agent  $k$  that depends on  $\epsilon_p^k$  and show in Corollary 1 that the bound on agent  $k$ 's cumulative regret is directly related to  $\epsilon_p^k$ , which suggests that the ordering of agents by  $\epsilon_p^k$  predicts the ordering of agents by performance. We plan to prove in a future publication that the same ordering is predicted even when the protocol does not depend on  $\epsilon_p^k$ . Simulations in Section IV provide validation of these assertions.

Let  $f(t) = \epsilon_p^k$ . Then the objective function becomes

$$Q_i^k(t) \triangleq \hat{\mu}_i^k(t) + \sigma_i \sqrt{\frac{2(\xi + 1)(N_i^k(t) + \epsilon_p^k) \log t}{N_i^k(t)}}.$$

This assumes that each agent  $k$  knows the  $p_j$  of each of its neighbors  $j$ . From Theorem 2, the expected number of times agent  $k$  samples the suboptimal option  $i$  is bounded as

$$\mathbb{E}(n_i^k(T)) \leq \Gamma(\zeta, \xi, K) + \frac{1}{T^{\xi-1} \log \zeta} \left( \frac{\log K}{T\xi} + \frac{1}{\xi - 1} \right) + \frac{4\sigma_i^2(\xi + 1)}{\Delta_i^2} \left( 1 + \sqrt{1 + \frac{\Delta_i^2}{2\sigma_i^2(\xi + 1)} \frac{\epsilon_p^k}{\log T}} \right) \log T.$$

This suggests that lower  $\epsilon_p^k$  values correspond to lower cumulative expected regret and hence better performance. Using the bound on  $\mathbb{E}(n_i^k(T))$  and equation (4) we upper bound the cumulative expected regret of agent  $k$  as follows.

*Corollary 1:* Let  $R^k(T)$  be the regret of agent  $k$  up to time  $T$ . Then we have

$$\mathbb{E}(R^k(T)) \leq \sum_{i=1}^N \Delta_i \Gamma(\zeta, \xi, K) + \sum_{i=1}^N \left( \frac{\Delta_i}{T^{\xi-1} \log \zeta} \left( \frac{\log K}{T\xi} + \frac{1}{\xi - 1} \right) \right) + \sum_{i=1}^N \frac{4\sigma_i^2(\xi + 1)}{\Delta_i} \left( 1 + \sqrt{1 + \frac{\Delta_i^2}{2\sigma_i^2(\xi + 1)} \frac{\epsilon_p^k}{\log T}} \right) \log T.$$

Expected cumulative regret is then logarithmically bounded.

The performance measure  $\epsilon_p^k$  can be bounded as

$$\epsilon_p^k \leq \frac{1}{p_k + 1}, \quad \forall k.$$

Accounting for the sociability of neighbors, as we have done, provides a more accurate and tighter bound than only considering individual sociability values. However, for an all-to-all observation structure the rank order can be predicted using only individual sociability values as shown next.

*Lemma 2:* Let  $\mathcal{G}$  be an all-to-all graph. Let  $p_k, p_{k'}$  be the sociability of agents  $k, k'$  such that  $p_k > p_{k'}$ . Then  $\epsilon_p^k < \epsilon_p^{k'}$ .

*Proof of Lemma 2:* Since  $p_k > p_{k'}$  and  $d_k = d_{k'} = K - 1$  we have

$$\frac{1}{p_k + 1} < \frac{1}{p_{k'} + 1}; \quad \sqrt{\frac{1}{d_k} \sum_{j=1, j \in \mathcal{V}_{d_k}}^K p_j} < \sqrt{\frac{1}{d_{k'}} \sum_{j=1, j \in \mathcal{V}_{d_{k'}}}^K p_j}$$

By equation (8), this proves that  $\epsilon_p^k < \epsilon_p^{k'}$ .

#### IV. NUMERICAL SIMULATIONS

We ran numerical simulations to evaluate the performance of sampling rule (1)–(2) with  $f(t) = \epsilon_p^k$  for agent  $k$  and with  $f(t) = \log \log t$ . The results in both cases verify the accuracy of agent ranks predicted by the performance measure  $\epsilon_p^k$ : lower  $\epsilon_p^k$  corresponds to lower cumulative regret, hence higher performance. We show plots in the case  $f(t) = \epsilon_p^k$ .

We consider 6 agents playing 10-armed bandit problems with two distinct observation structures: A) an all-to-all graph and B) a cyclic graph. In all simulations we let the reward distributions be Gaussian with variance  $(\sigma'_i)^2 = 25$ ,  $i = 1, \dots, 10$ , mean values given by

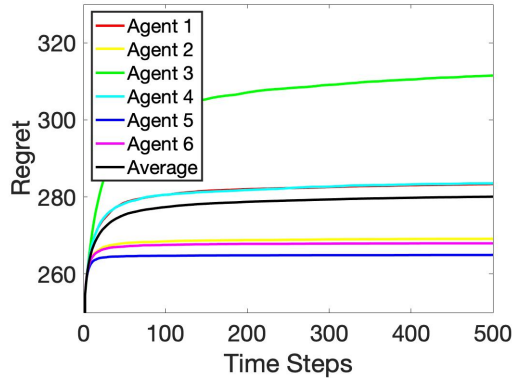


Fig. 2. Expected cumulative regret of the 6 agents using the sampling rule given in (1)–(2) and performance measure defined in (8) with distinct observation probabilities  $p_k$  and underlying all-to-all observation structure.

$i$	1	2	3	4	5	6	7	8	9	10
$\mu_i$	40	50	50	60	70	70	80	90	92	95

and sociability values given by

$k$	1	2	3	4	5	6
$p_k$	0.50	0.85	0.05	0.50	1.00	0.90

We provide results for 500 time steps with 1000 Monte Carlo simulations. We set the sampling rule parameter  $\xi = 1.1$ .

#### A. All-to-all observation

The underlying observation graph structure is all-to-all, equivalently a 5-regular graph. We calculate the performance measure  $\epsilon_p^k$  for each agent  $k$  using equation (8):

$k$	1	2	3	4	5	6
$\epsilon_p^k$	0.542	0.415	0.825	0.542	0.374	0.401

The best predicted performer is agent 5 with lowest performance measure  $\epsilon_p^5 = 0.374$  and highest sociability value  $p_5 = 1.0$ . Second and third best predicted performers are agent 6 and agent 2, respectively, with second and third highest sociability. Agents 1 and 4 are ranked next with equal sociability and  $\epsilon_p^1 = \epsilon_p^4$ . The worst predicted performer is agent 3, with lowest sociability and highest performance measure  $\epsilon_p^3$ . These predictions on performance ranking are verified in the simulation results of Figure 2. The results also verify that for an underlying all-to-all graph, the performance rank ordering is predicted by the sociability rank ordering.

#### B. Cyclic observation

The underlying observation graph structure is a cycle, equivalently a 2-regular graph, defined as  $\{k, j\} \in \mathcal{E}$ , where  $|(k - j) \bmod 6| = 1$ . We calculate the  $\epsilon_p^k$  using (8):

$k$	1	2	3	4	5	6
$\epsilon_p^k$	0.624	0.284	0.783	0.483	0.418	0.456

The best predicted performer is agent 2 with the lowest performance measure  $\epsilon_p^2 = 0.284$ , but not the highest sociability. In fact, agents 5 and 6 have higher sociability than agent 2. However, while all three agents 2, 5, and 6 have one

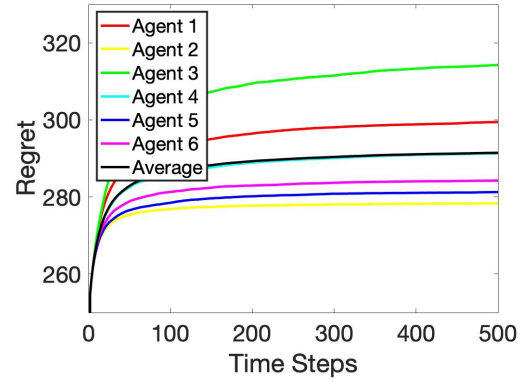


Fig. 3. Expected cumulative regret of the 6 agents using the sampling rule given in (1)–(2) and performance measure defined in (8) with distinct observation probabilities  $p_k$  and underlying cyclic observation structure.

neighbor with sociability 0.5, the other neighbor of agents 5 and 6 has sociability 0.9 and 1, respectively, whereas the other neighbor of agent 2 has sociability 0.05. The very low sociability of one of agent 2's neighbors improves agent 2's performance significantly enough that it outperforms agents 5 and 6. This result illustrates the important role of sociability of neighbors in an agent's performance. Further, while agents 1 and 4 are indistinguishable by their own sociability, agent 4 has a neighbor with sociability 0.05 whereas agent 1 has neighbors with relatively high sociability. Therefore as predicted by the performance measure, agent 4 outperforms agent 1. Figure 3 validates the predicted rankings.

## V. CONCLUSIONS

We studied a MAMAB problem where agents observe instantaneous actions and rewards of their neighbors according to a stochastic network graph in which agents are distinguished by their sociability, defined as the probability of observing their neighbors. We derived an upper bound for expected cumulative regret of agents. We proposed a measure to predict relative performance ranking of the agents as a function of sociability of agents and their neighbors. We verified that having less sociable neighbors improves the performance of agents. Accuracy of the measure has been verified analytically through expected cumulative regret bounds and computationally through numerical simulations.

## ACKNOWLEDGMENT

The first author wishes to thank Peter Landgren for helpful comments during the preparation of this paper.

## APPENDIX

*Proof of Theorem 1:* Since  $X_i$  is a sub-Gaussian random variable with variance proxy  $(\sigma'_i)^2$ , we have

$$E(\exp(\lambda(X_i - \mu_i))) \leq \exp\left(\frac{\lambda^2(\sigma'_i)^2}{2}\right).$$

Define a new random variable such that

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

Note that  $E(Y_i^k(\tau)) = E(Y_i^k(\tau)|\mathcal{F}_{\tau-1}) = 0$ . Let  $Z_i^k(t) = \sum_{\tau=1}^t Y_i^k(\tau)$ . Let  $T_i^k(\tau) = \sum_{j=1}^K \mathbb{I}_{\{\varphi_j^\tau=i\}} \mathbb{I}_{\{k,j\}}^\tau$ . For any  $\lambda > 0$

$$E \left( \exp(\lambda Y_i^k(\tau)) | \mathcal{F}_{\tau-1} \right) \leq \exp \left( \frac{\lambda^2 \sigma_i^2}{2} T_i^k(\tau) \right).$$

$\mathbb{I}_{\{\varphi_j^\tau=i\}}$  is an  $\mathcal{F}_{\tau-1}$  measurable random variable, and so

$$E \left( \exp \left( \lambda Y_i^k(\tau) - \frac{\lambda^2 \sigma_i^2}{2} T_i^k(\tau) \right) | \mathcal{F}_{\tau-1} \right) \leq 1.$$

Further, using the properties of conditional expectations

$$\begin{aligned} E \left( \exp \left( \sum_{1 \leq \tau \leq t} \lambda Y_i^k(\tau) - \frac{\lambda^2 \sigma_i^2}{2} T_i^k(\tau) \right) | \mathcal{F}_{t-1} \right) \\ \leq \exp \left( \sum_{1 \leq \tau \leq t-1} \lambda Y_i^k(\tau) - \frac{\lambda^2 \sigma_i^2}{2} T_i^k(\tau) \right). \end{aligned}$$

Thus we see that

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

The rest of the proof closely follows the papers [17], [18]. For clarity and completeness we include the main steps. Let  $\zeta > 1$ . Then  $1 \leq N_i^k(t) \leq \zeta^{D_t}$  where  $D_t = \frac{\log K t}{\log \zeta}$ . 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

$$\frac{2\kappa\vartheta}{\lambda_j \sqrt{N_i^k(t)}} + \frac{\sigma_i^2}{2} \lambda_j \sqrt{N_i^k(t)} \leq \sqrt{\vartheta},$$

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

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

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

*Proof of Theorem 2:* From equations (6)–(7) we have

$$\begin{aligned} \mathbb{E} (n_i^k(T)) &\leq \sum_{t=1}^T \mathbb{P} (\hat{\mu}_{i^*}^k \leq \mu_{i^*} - C_{i^*}^k(t), N_i^k(t) > \eta_i(t)) \\ &+ \eta_i(T) + \sum_{t=1}^T \mathbb{P} (\hat{\mu}_i^k \geq \mu_i + C_i^k(t), N_i^k(t) > \eta_i(t)). \end{aligned}$$

Note that  $\frac{N_i^k(t)+f(t)}{N_i^k(t)} \geq 1, \forall t \geq 1$ . Using Lemma 1 and  $\delta(\xi) = \xi + 1 - \delta'(\epsilon)$  we have

$$\mathbb{E} (n_i^k(T)) \leq \eta_i(T) + 2\nu \sum_{t=1}^T \frac{\log K t}{t^{\delta(\xi)}}.$$

Since  $\delta'(\epsilon)$  can be made arbitrarily small, the summation term can be evaluated as follows:

$$\begin{aligned} \sum_{t=1}^T \frac{\log K t}{t^{\delta(\xi)}} &\leq \sum_{t=1}^T \frac{\log K}{t^{\xi+1}} + \sum_{t=1}^T \frac{1}{t^\xi} \\ &\leq \log K \left( 1 + \int_2^T \frac{1}{t^{\xi+1}} dt \right) + 1 + \int_2^T \frac{1}{t^\xi} dt \end{aligned}$$

$$\begin{aligned} &= 1 + \log K + \frac{1}{2^\xi} \left( \frac{\log K}{\xi} + \frac{2}{\xi-1} \right) \\ &+ \frac{1}{T^\xi} \left( \frac{\log K}{\xi} + \frac{T}{\xi-1} \right). \end{aligned}$$

Thus we have

$$\begin{aligned} \mathbb{E} (n_i^k(T)) &\leq \frac{1}{\log \zeta} (1 + \log K) + \frac{1}{2^\xi \log \zeta} \left( \frac{\log K}{\xi} + \frac{2}{\xi-1} \right) \\ &+ \frac{1}{T^{\xi-1} \log \zeta} \left( \frac{\log K}{T^\xi} + \frac{1}{\xi-1} \right) \\ &+ \frac{4\sigma_i^2(\xi+1)}{\Delta_i^2} \left( 1 + \sqrt{1 + \frac{\Delta_i^2}{2\sigma_i^2(\xi+1)} \frac{f(T)}{\log T}} \right) \log T. \end{aligned}$$

## REFERENCES

- [1] R. S. Sutton and A. G. Barto, *Introduction to Reinforcement Learning*. MIT Press Cambridge, MA, USA, 1998.
- [2] H. Robbins, *Some Aspects of the Sequential Design of Experiments*. Springer New York, 1985.
- [3] J. C. Gittins, "Bandit processes and dynamic allocation indices," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 41, pp. 148–177, 1979.
- [4] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in Applied Mathematics*, vol. 6, no. 1, pp. 4–22, 1985.
- [5] R. Agrawal, "Sample mean based index policies with  $o(\log n)$  regret for the multi-armed bandit problem," *Advances in Applied Probability*, vol. 27, pp. 1054–1078, 1995.
- [6] P. Auer, N. Cesa-Bianchi, and P. Fisher, "Finite-time analysis of the multi-armed bandit problem," *Machine Learning*, vol. 47, pp. 235–256, 2002.
- [7] E. Kauffman, O. Cappe, and A. Garivier, "On Bayesian upper confidence bounds for bandit problem," in *International Conference on Artificial Intelligence and Statistics*, April 2012, pp. 592–600.
- [8] P. Reverdy, V. Srivastava, and N. E. Leonard, "Modeling human decision-making in generalized Gaussian multi-armed bandits," in *Proceedings of the IEEE*, vol. 102, no. 4, 2014, pp. 544–571.
- [9] V. Anantharam, P. Varaiya, and J. Walrand, "Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part i: I.i.d. rewards," *IEEE Transactions on Automatic Control*, vol. 32, no. 11, pp. 968–976, 1987.
- [10] P. Landgren, V. Srivastava, and N. E. Leonard, "Distributed cooperative decision-making in multiarmed bandits: Frequentist and Bayesian algorithms," in *IEEE Conference on Decision and Control*, December 2016, pp. 167–172.
- [11] D. Kalathil, N. Nayyar, and R. Jain, "Decentralized learning for multiplayer multiarmed bandits," *IEEE Transactions on Information Theory*, vol. 60, no. 4, pp. 2331–2345, 2014.
- [12] P. Landgren, V. Srivastava, and N. E. Leonard, "On distributed cooperative decision-making in multiarmed bandits," in *European Control Conference*, June 2016, pp. 243–248.
- [13] R. K. Kolla, K. Jagannathan, and A. Gopalan, "Collaborative learning of stochastic bandits over a social network," *arXiv:1602.0886v2*, 2016.
- [14] P. Landgren, V. Srivastava, and N. E. Leonard, "Social imitation in cooperative multiarmed bandits: partition-based algorithms with strictly local information," in *IEEE Conference on Decision and Control*, December 2018, pp. 5239–5244.
- [15] A. R. Tilman, J. R. Watson, and S. Levin, "Maintaining cooperation in social-ecological systems," *Theoretical Ecology*, vol. 10, pp. 155–165, 2016.
- [16] T. L. Lai, "Adaptive treatment allocation and the multi-armed bandit problem," *Ann. Statist.*, vol. 15, no. 3, pp. 1091–1114, 09 1987. [Online]. Available: <https://doi.org/10.1214/aos/1176350495>
- [17] A. Garivier and E. Moulines, *On Upper-Confidence Bound Policies for Switching Bandit Problems*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 174–188.
- [18] —, "On upper-confidence bound policies for non-stationary bandit problems," *arXiv:0805.3415v1*, 2008.