Automatica 125 (2021) 109445

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Distributed cooperative decision making in multi-agent multi-armed bandits[☆]

Peter Landgren[a], Vaibhav Srivastava[b], Naomi Ehrich Leonard[a,*]

[a] Department of Mechanical & Aerospace Engineering, Princeton University, NJ, USA
[b] Department of Electrical and Computer Engineering, Michigan State University, East Lansing, MI, USA

ARTICLE INFO

Article history:
Received 4 March 2020
Received in revised form 3 August 2020
Accepted 27 October 2020
Available online xxxx

Keywords:
Multi-armed bandits
Multi-agent systems
Distributed decision making
Explore–exploit dilemma

ABSTRACT

We study a distributed decision-making problem in which multiple agents face the same multi-armed bandit (MAB), and each agent makes sequential choices among arms to maximize its own individual reward. The agents cooperate by sharing their estimates over a fixed communication graph. We consider an unconstrained reward model in which two or more agents can choose the same arm and collect independent rewards. And we consider a constrained reward model in which agents that choose the same arm at the same time receive no reward. We design a dynamic, consensus-based, distributed estimation algorithm for cooperative estimation of mean rewards at each arm. We leverage the estimates from this algorithm to develop two distributed algorithms: coop-UCB2 and coop-UCB2-selective-learning, for the unconstrained and constrained reward models, respectively. We show that both algorithms achieve group performance close to the performance of a centralized fusion center. Further, we investigate the influence of the communication graph structure on performance. We propose a novel graph explore–exploit index that predicts the relative performance of groups in terms of the communication graph, and we propose a novel nodal explore–exploit centrality index that predicts the relative performance of agents in terms of the agent locations in the communication graph.

1. Introduction

Many engineered and natural systems are faced with the challenge of decision making under uncertainty, in which an agent must make decisions among alternatives while still learning about those options. Decision making under uncertainty inherently features the *explore–exploit tradeoff*, where one must decide between selecting options with a high expected payoff (exploitation) and selecting options with less well-known but potentially better payoff (exploration). Often systems feature multiple networked decision makers, where performance of the system may require *cooperative* decision making, in which disparate and distributed elements of a group act collaboratively.

The explore–exploit tradeoff can be formally investigated within the context of the multi-armed bandit (MAB) problem. In a stochastic MAB problem, an agent is presented with a set of arms (options), and each arm is represented by a stochastic reward with a mean that is unknown to the agent. An agent's goal is to select arms sequentially in order to maximize its own cumulative expected reward over time. Good performance in the MAB problem requires an agent to balance learning the mean reward of each arm (exploration) with choosing the arm with the highest estimated mean (exploitation).

The explore–exploit tradeoff has been widely investigated using the MAB problem across a variety of scientific fields and has found diverse application in control and robotics (Cheung, Leighton, & Hover, 2013; Srivastava, Reverdy, & Leonard, 2014), ecology (Krebs, Kacelnik, & Taylor, 1978; Srivastava, Reverdy, & Leonard, 2013), and communications (Anandkumar, Michael, Tang, & Swami, 2011). The MAB problem, and particularly the classical single-agent variant, has been studied extensively (see Bubeck & Cesa-Bianchi, 2012 for a survey). In Lai and Robbins (1985), Lai and Robbins established a limit on the expected performance of any optimal policy in a frequentist setting by proving a lower bound on the number of times an agent selects a sub-optimal arm.

[☆] This research has been supported in part by ONR, USA grants N00014-14-1-0635, N00014-19-1-2556, ARO, USA grants W911NF-14-1-0431, W911NF-18-1-0325 and by the Department of Defense (DoD), USA through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program, USA. The material in this paper was partially presented at: the 15th annual European Control Conference, June 29–July 1, 2016, Aalborg, Denmark. the 55th IEEE Conference on Decision and Control, December 12–14, 2016, Las Vegas, NV, USA. This paper was recommended for publication in revised form by Associate Editor Luca Schenato under the direction of Editor Christos G. Cassandras.

[*] Corresponding author.
E-mail addresses: peterclandgren@gmail.com (P. Landgren), vaibhav@egr.msu.edu (V. Srivastava), naomi@princeton.edu (N.E. Leonard).

https://doi.org/10.1016/j.automatica.2020.109445
0005-1098/© 2020 Elsevier Ltd. All rights reserved.

To date most research on the MAB problem has focused on single-agent policies, but the rising importance of networked systems and large-scale information networks have motivated the investigation of the MAB problem with multiple agents. In this paper, we study two variants of the multi-agent MAB problem, in which each agent makes choices to maximize its own individual reward but cooperates by communicating its estimates across a network. The first variant assumes an *unconstrained reward* model, in which agents are not penalized if they choose the same arm at the same time. The second variant assumes a *constrained reward*, in which agents that choose the same arm at the same time receive a reduced reward. Consider agents in a remote setting choosing among communication channels to send data back to a base station. The constrained reward model applies to the case in which data cannot be sent if agents choose the same channel. The constrained reward model can also be used to prevent mobile agents from searching for resource in the same patch when there exist multiple resource-rich patches.

When a centralized fusion center that has access to all the information available to every agent decides which arms will be sampled by the agents, the agents are inherently coordinated and no two agents ever sample the same arm at the same time. In this setting, the above two variants become almost the same. Anantharam, Varaiya, and Walrand (1987) extended the classical single-agent MAB problem to the setting of such a fusion center and derived a fundamental lower bound on the performance of the fusion center. In this paper, we design *distributed* algorithms that yield group performance close to that of a centralized fusion center.

Kolla, Jagannathan, and Gopalan (2016) and Landgren, Srivastava, and Leonard (2018) studied the multi-agent MAB problem under the unconstrained reward model. In their setup, each agent can share its actions and the associated rewards at each time with its neighbors in the communication graph. In this setting, group performance improves when each agent acts individually. However, group performance might not be close to the performance of a centralized fusion center, especially for large sparse networks. Madhushani and Leonard (2019, 2020) have extended this setting to examine dynamic interactions among agents governed by a heterogeneous stochastic process and to design strategies that minimize sampling regret as well as communication costs.

Several researchers (Anandkumar et al., 2011; Gai & Krishnamachari, 2014; Kalathil, Nayyar, & Jain, 2014; Liu & Zhao, 2010; Wei & Srivastava, 2018) have studied the distributed multi-agent MAB problem under the constrained reward model. In these works, agents seek to converge on the set of best arms, but they do not explicitly communicate with one another. In Gai and Krishnamachari (2014), Kalathil et al. (2014), agents are ranked and they target the best arm associated with their rank. Anandkumar et al. (2011) also studied distributed policies for agents to learn their ranks while solving the multi-agent MAB problem. Bistritz and Leshem (2018) studied the distributed multi-agent MAB problem under no communication among agents. Assuming no a priori ranking of agents, they developed a game-of-thrones algorithm, inspired by Marden, Young, and Pao (2014), to enable coordination among agents.

Shahrampour, Rakhlin, and Jadbabaie (2017) studied a variant of the multi-agent MAB problem in which the reward associated with each arm may be different for every agent. The best arm is defined as the arm with the maximum average mean reward over all agents. Unlike in other multi-agent MAB setups, in which each agent makes a decision at each time, they consider a single group decision obtained using a majority rule on individual decisions.

In early versions (Landgren, Srivastava, & Leonard, 2016a, 2016b) of the present work, we studied distributed cooperative decision making in the multi-agent MAB problem with the unconstrained reward model and Bayesian as well as frequentist updates. In comparison, this paper considers a broader class of reward distributions and studies both unconstrained and constrained reward models. We present new proofs that improve on the preliminary versions and a much broader exploration of the influence of communication graph structure on individual and group decision-making performance.

Martínez-Rubio, Kanade, and Rebeschini (2019) extended our preliminary versions (Landgren et al., 2016a, 2016b) in the context of the unconstrained reward model. Their work is complementary to the approach discussed here. A key difference between their algorithm and the algorithm discussed in this paper, is that our algorithm requires only the knowledge of total number of agents to tune the decision-making heuristic, while their algorithm requires the knowledge of the spectral gap of the communication graph. They do not investigate the influence of the network graph on performance.

In this paper, we study distributed cooperative decision making in the multi-agent MAB problem under both unconstrained and constrained reward models. We use a set of running consensus algorithms for cooperative estimation of the mean reward at each arm over an undirected graph and develop algorithms for individual decision making based on these estimates for both reward models. We also derive indices of graph structure that are predictive of individual as well as group performance. The major contributions of the paper are as follows.

First, we employ and rigorously analyze running consensus algorithms for distributed cooperative estimation of mean reward at each arm, and we derive bounds on key quantities.

Second, we propose and thoroughly analyze the coop-UCB2 algorithm for the multi-agent MAB problem under the unconstrained reward model and sub-Gaussian reward distributions.

Third, we propose and thoroughly analyze the coop-UCB2-selective-learning algorithm for the multi-agent MAB problem under the constrained reward model and sub-Gaussian rewards distributions.

Fourth, we utilize the derived bounds on the decision-making performance of the group to introduce a novel graph explore–exploit index that predicts the ordering of graphs in terms of group explore–exploit performance and a novel nodal explore–exploit centrality index as a function of an agent's location in a graph that predicts the ordering of agents in terms of individual explore–exploit performance. We illustrate the effectiveness of these indices with simulations.

The remainder of the paper is organized as follows. In Section 2 we describe the multi-agent MAB problem studied in this paper. In Section 3 we present and analyze the cooperative estimation algorithm. We propose and analyze the coop-UCB2 algorithm in Section 4 and the coop-UCB2-selective-learning algorithm in Section 5. We illustrate our analytic results with numerical examples in Section 6. We conclude in Section 7.

2. Problem description

We consider a distributed multi-agent MAB problem in which M agents make sequential choices among the same set of N arms with the goal of maximizing their individual reward. The M agents cooperate by sharing their estimates over a bi-directional communication network. The network is modeled by an undirected graph $\mathcal{G}$ in which each node represents a decision-making agent and edges represent the communication links between them (Bullo, Cortés, & Martínez, 2009). Let $A \in \mathbb{R}^{M\times M}$ be the adjacency matrix associated with $\mathcal{G}$ and $L \in \mathbb{R}^{M\times M}$ the corresponding Laplacian matrix. We assume that the graph $\mathcal{G}$ is connected, i.e., there exists a path between every pair of nodes.

Let the reward associated with arm $i \in \{1,\dots,N\}$ be a stationary random variable with an unknown mean m_i. Using its local information, each agent $k \in \{1, \dots, M\}$ selects arm $i^k(t)$ at time $t \in \{1, \dots, T\}$, where $T \in \mathbb{N}$ is the time horizon.

We study two reward models that determine how the reward associated with arm $i^k(t)$ is received by agent k. In the *unconstrained reward model*, agent k receives a reward equal to the realized value of the reward at arm $i^k(t)$, irrespective of the choices of the other agents. In the *constrained reward model*, agent k receives a reward equal to the realized value of the reward at arm $i^k(t)$, only if it is the only agent to select arm $i^k(t)$ at time t; otherwise it receives no reward.

The objective of the distributed cooperative multi-agent MAB problem is to maximize the expected cumulative group reward. This objective is equivalent to minimizing the *expected cumulative group regret* defined by the difference between the best possible expected cumulative group reward and the achieved expected cumulative group reward.

Let $\{b^i\}_{i\in\{1,\dots,N\}}$ be the permuted sequence of arms such that $m_{b^1} > m_{b^2} > \cdots > m_{b^N}$.[1] Under the unconstrained reward model, the expected cumulative group regret is defined by

$$R_T^{\text{unc}} = MTm_{b^1} - \sum_{t=1}^{T}\sum_{k=1}^{M} m_{i^k(t)} = \sum_{i=1}^{N}\sum_{k=1}^{M} \Delta_i \mathbb{E}[n_i^k(T)], \tag{1}$$

where $n_i^k(T)$ is the total number of times arm i is selected by agent k until time T and $\Delta_i = m_{b^1} - m_i$. In the following, we use b^1 and i^* interchangeably to denote the arm with the highest mean reward. Under the unconstrained reward model, the regret at time t is minimized if every agent chooses arm i^*.

Similarly, under the constrained reward model and assuming $M \le N$, the expected cumulative group regret is defined by

$$R_T^{\text{con}} = T\sum_{k=1}^{M} m_{b^k} - \sum_{t=1}^{T}\sum_{k=1}^{M} m_{i^k(t)} \mathbb{I}_{i^k(t)}^k(t), \tag{2}$$

where $\mathbb{I}_i^k(t) = 1$ if agent k is the only agent to sample arm i at time t, and 0 otherwise. In the following, we denote the set of j best arms by $\mathcal{O}_j^* = \{b^1, \dots, b^j\}$. Under the constrained reward model, the regret at time t is minimized if each agent chooses a different arm in the set $\mathcal{O}_M^*$.

Let p_i be the probability distribution of the reward associated with arm i. For a centralized fusion center that has access to information available to each agent, and under the unconstrained reward model, the lower bound

$$\sum_{k=1}^{M} \mathbb{E}[n_i^k(T)] \ge \left(\frac{1}{\mathcal{D}(p_i \parallel p_{i^*})} + o(1)\right) \ln T \tag{3}$$

holds asymptotically as $T \to +\infty$ for any suboptimal arm $i \neq i^*$ (Anantharam et al., 1987; Lai & Robbins, 1985). Here, $\mathcal{D}(p_i \parallel p_{i^*})$ represents the Kullback–Leibler divergence between p_i and p_{i^*}. Substituting, the lower bound on $\sum_k \mathbb{E}[n_i^k(T)]$ in (3) into the expression (1) for R_T^{unc} yields

$$R_T^{\text{unc}} \ge \sum_{i\neq i^*} \left(\frac{\Delta_i}{\mathcal{D}(p_i \parallel p_{i^*})} + o(1)\right) \ln T. \tag{4}$$

For the constrained reward model, consider a centralized fusion center that has access to the information available to each agent and can assign the arm to be selected by each agent at each time. For such a fusion center, no two agents ever select the same arm at the same time, and, under the constrained reward model, the lower bound

$$\sum_{k=1}^{M} \mathbb{E}[n_i^k(T)] \ge \left(\frac{1}{\mathcal{D}(p_i \parallel p_{b^M})} + o(1)\right) \ln T \tag{5}$$

holds asymptotically as $T \to +\infty$ for any suboptimal arm $i \notin \mathcal{O}_M^*$ (Anantharam et al., 1987). Thus, the asymptotic regret of the fusion center satisfies

$$R_T^{\text{con}} \ge \sum_{i\in\{b^{M+1},\dots,b^N\}} \left(\frac{\Delta_i}{\mathcal{D}(p_i \parallel p_{b^M})} + o(1)\right) \ln T. \tag{6}$$

Note that the expected regret under the constrained reward model is higher if multiple agents select the same arm. Thus, the above lower bound holds even if agents themselves make arm selections instead of being assigned an arm by the fusion center. The situation in which multiple agents select the same arm is referred to as a *collision*.

Our objective in this paper is to design a distributed cooperative algorithm estimating mean reward at each arm and a decision-making algorithm for each agent that yields expected cumulative group regret close to that of a centralized fusion center. We consider rewards drawn from a sub-Gaussian distribution.

Definition 1 (*Sub-Gaussian Random Variable (Boucheron, Lugosi, & Pascal, 2016)*). A real-valued random variable X, with $\mathbb{E}[X] = m \in \mathbb{R}$, is sub-Gaussian if

$$\phi_X(\beta) \le m\beta + \frac{\sigma_g^2\beta^2}{2},$$

where $\sigma_g \in \mathbb{R}_{>0}$, $\beta \in \mathbb{R}$, and $\phi_X : \mathbb{R} \to \mathbb{R}$ is the cumulant generating function of X defined by

$$\phi_X(\beta) = \ln\left(\mathbb{E}[\exp(\beta X)]\right).$$

Sub-Gaussian distributions include Bernoulli, uniform, and Gaussian distributions, and distributions with bounded support.

3. Cooperative estimation of mean rewards

In this section we study cooperative estimation of mean rewards at each arm. We propose two running (dynamic) consensus algorithms (Braca, Marano, & Matta, 2008; Olfati-Saber & Murray, 2004) for each arm and analyze performance.

3.1. Cooperative estimation algorithm

For distributed cooperative estimation of the mean reward at each arm i, we propose two running consensus algorithms: (i) for estimation of total reward provided at arm i, and (ii) for estimation of the total number of times arm i has been sampled.

Let $\hat{s}_i^k(t)$ be agent k's estimate of the total reward provided at arm i until time t per unit agent. Let $\hat{n}_i^k(t)$ be agent k's estimate of the total number of times arm i has been selected until time t per unit agent. Recall that $i^k(t)$ is the arm sampled by agent k at time t and let $\xi_i^k(t) = \mathbb{1}(i^k(t) = i)$. $\mathbb{1}(\cdot)$ is the indicator function, here equal to 1 if $i^k(t) = i$ and 0 otherwise. For all i and k, we define $r_i^k(t)$ as the realized reward at arm i for agent k at time t, which is a random variable sampled from a sub-Gaussian distribution. The corresponding reward received by agent k at time t is $r^k(t) = r_i^k(t) \cdot \mathbb{1}(i^k(t) = i)$.

The estimates $\hat{s}_i^k(t)$ and $\hat{n}_i^k(t)$ are updated using running consensus as follows:

$$\hat{\mathbf{n}}_i(t) = P\left(\hat{\mathbf{n}}_i(t-1) + \boldsymbol{\xi}_i(t)\right), \tag{7}$$

$$\text{and} \quad \hat{\mathbf{s}}_i(t) = P\left(\hat{\mathbf{s}}_i(t-1) + \mathbf{r}_i(t)\right), \tag{8}$$

where $\hat{\mathbf{n}}_i(t)$, $\hat{\mathbf{s}}_i(t)$, $\boldsymbol{\xi}_i(t)$, and $\mathbf{r}_i(t)$ are vectors of $\hat{n}_i^k(t)$, $\hat{s}_i^k(t)$, $\xi_i^k(t)$, and $r_i^k(t) \cdot \mathbb{1}(i^k(t) = i)$, $k \in \{1, \dots, M\}$, respectively; P is a row stochastic matrix given by

$$P = \mathcal{I}_M - \frac{\kappa}{d_{\max}} L. \tag{9}$$

[1] We rely on the assumption that $m_{b^i} \neq m_{b^j}$ for all i, j for the constrained reward model; it can be relaxed for the unconstrained reward model.

$\mathcal{I}_M$ is the identity matrix of order M, $\kappa \in (0, 1]$ is a step size parameter (Olfati-Saber & Murray, 2004), $d_{\max} = \max\{\deg(i) \mid i \in \{1, \ldots, M\}\}$, and $\deg(i)$ is the degree of node i. In the following, we assume without loss of generality that the eigenvalues of P are ordered such that $\lambda_1 = 1 > \lambda_2 \geq \cdots \geq \lambda_M > -1$.

In the running consensus updates (7) and (8), each agent k collects information $\xi_i^k(t)$ and $r^k(t)$ at time t, adds it to its current opinion, and then averages its updated opinion with the updated opinion of its neighbors.

Using $\hat{s}_i^k(t)$ and $\hat{n}_i^k(t)$, agent k can calculate $\hat{\mu}_i^k(t)$, the estimated empirical mean of arm i at time t defined by

$$\hat{\mu}_i^k(t) = \frac{\hat{s}_i^k(t)}{\hat{n}_i^k(t)}. \tag{10}$$

3.2. Analysis of the cooperative estimation algorithm

We now analyze the performance of the estimation algorithm defined by (7), (8) and (10). Let $n_i^{\text{cent}}(t) \equiv \frac{1}{M}\sum_{\tau=1}^{t} \mathbf{1}_M^\top \boldsymbol{\xi}_i(\tau) = \frac{1}{M}\sum_{k=1}^{M} n_i^k(t)$ be the total number of times arm i has been selected per unit agent until time t, and let $s_i^{\text{cent}}(t) \equiv \frac{1}{M}\sum_{\tau=1}^{t} \boldsymbol{\xi}_i^\top(\tau)\mathbf{r}_i(\tau)$ be the total reward provided at arm i per unit agent until time t. Let $\mathbf{u}_i$ be the eigenvector corresponding to λ_i, u_i^d the dth entry of $\mathbf{u}_i$. Note $\lambda_1 = 1$ and $\mathbf{u}_1 = \mathbf{1}_M/\sqrt{M}$. Let

$$\nu_{pj}^{\text{+sum}} = \sum_{d=1}^{M} u_p^d u_j^d \mathbb{1}(u_p^k u_j^k \geq 0), \quad \nu_{pj}^{\text{-sum}} = \sum_{d=1}^{M} u_p^d u_j^d \mathbb{1}(u_p^k u_j^k \leq 0),$$

$$a_{pj}(k) = \begin{cases} \nu_{pj}^{\text{+sum}} u_p^k u_j^k, & \text{if } \lambda_p\lambda_j \geq 0 \ \& \ u_p^k u_j^k \geq 0, \\ \nu_{pj}^{\text{-sum}} u_p^k u_j^k, & \text{if } \lambda_p\lambda_j \geq 0 \ \& \ u_p^k u_j^k \leq 0, \\ \nu_{pj}^{\max} |u_p^k u_j^k|, & \text{if } \lambda_p\lambda_j < 0, \end{cases} \tag{11}$$

where $\nu_{pj}^{\max} = \max\{|\nu_{pj}^{\text{-sum}}|, \nu_{pj}^{\text{+sum}}\}$.

We define the *graph explore–exploit index* ϵ_n as

$$\epsilon_n = \sqrt{M}\sum_{p=2}^{M} \frac{|\lambda_p|}{1-|\lambda_p|}, \tag{12}$$

and the *nodal explore–exploit centrality index* ϵ_c^k for node k as

$$\epsilon_c^k = M\sum_{p=1}^{M}\sum_{j=2}^{M} \frac{|\lambda_p\lambda_j|}{1-|\lambda_p\lambda_j|} a_{pj}(k). \tag{13}$$

Since $|\lambda_p| < 1$, for all $p \geq 2$, definitions (12)–(13) imply that ϵ_n and ϵ_c^k decrease with a decrease in $|\lambda_p|$ for any $p \geq 2$. A small value of ϵ_n reflects a high level of symmetry and connectivity in the graph. This will be shown to predict low error in each agent's estimate of the average number of times a suboptimal arm has been chosen and thus high group explore–exploit performance. Dependence on the kth component of the eigenvectors in (13) makes ϵ_c^k an index for node k, which measures how well agent k estimates the second-order moments of rewards. In an asymmetric graph, $\epsilon_c^k < \epsilon_c^l$ reflects a more favorable location in the graph for node k as compared to node l.

Both ϵ_n and ϵ_c^k depend only on the topology of the communication graph, yet they predict distributed cooperative estimation performance, as we show next, and explore–exploit performance, as we show in subsequent sections.

Proposition 1 (*Performance of Cooperative Estimation*)**.** *For the distributed estimation algorithm defined in* (7), (8) and (10), *and a doubly stochastic matrix P defined in* (9), *the following statements hold:*

(i) *the estimate $\hat{n}_i^k(t)$ satisfies*

$$n_i^{cent}(t) - \epsilon_n \leq \hat{n}_i^k(t) \leq n_i^{cent}(t) + \epsilon_n;$$

(ii) *the following inequality holds for the estimate $\hat{n}_i^k(t)$ and the sequence $\{\xi_i^j(\tau)\}_{\tau\in\{1,\ldots,t\}}$, $j \in \{1, \ldots, M\}$:*

$$\sum_{\tau=1}^{t}\sum_{j=1}^{M}\left(\sum_{p=1}^{M}\lambda_p^{t-\tau+1}u_p^k u_p^j\right)^2 \xi_i^j(\tau) \leq \frac{\hat{n}_i^k(t)+\epsilon_c^k}{M}.$$

Proof. The proof uses some algebraic manipulations on the modal decomposition of (7). See Appendix A. □

We now derive concentration bounds for the estimated mean computed with the cooperative estimation algorithm. Standard concentration inequalities, such as the Chernoff–Hoeffding inequality, rely on the sample size being independent of the realized values of samples. In the context of MABs, the arm selected at time t depends on the rewards accrued at previous times. This makes the number of times an arm is sampled and the total reward accrued at that arm dependent random variables. For the case of a single agent, the specific kind of dependence between these random variables that occurs in MAB problems is leveraged to derive a concentration inequality in Garivier and Moulines (2008). In the following, we extend this concentration inequality to the distributed estimation algorithm studied here.

For $i \in \{1, \ldots, N\}$ and $k \in \{1, \ldots, M\}$, $\{r_i^k(t)\}_{t\in\mathbb{N}}$ is a sequence of i.i.d. sub-Gaussian rewards with mean $m_i \in \mathbb{R}$. Let $\mathcal{F}_t$ be the filtration defined by the sigma-algebra of all the measurements until time t. Let $\{\xi_i^k(t)\}_{t\in\mathbb{N}}$ be a sequence of Bernoulli variables such that $\xi_i^k(t)$ is deterministically known given $\mathcal{F}_{t-1}$, i.e., $\xi_i^k(t)$ is pre-visible with respect to $\mathcal{F}_{t-1}$. Let $\phi_i(\beta) = \ln\left(\mathbb{E}\left[\exp\left(\beta r_i^k(t)\right)\right]\right)$ denote the cumulant generating function of $r_i^k(t)$.

Theorem 1 (*Concentration Bounds for the Mean Estimator*)**.** *For the estimates $\hat{s}_i^k(t)$ and $\hat{n}_i^k(t)$ obtained using* (7) and (8) *given rewards drawn from a sub-Gaussian distribution as defined in* Definition 1, *the following concentration inequality holds:*

$$\mathbb{P}\left(\frac{\hat{s}_i^k(t) - m_i\hat{n}_i^k(t)}{\left(\frac{1}{M}\left(\hat{n}_i^k(t)+\epsilon_c^k\right)\right)^{1/2}} > \delta\right) < \left\lceil\frac{\ln(t+\epsilon_n)}{\ln(1+\eta)}\right\rceil \exp\left(\frac{-\delta^2}{2\sigma_g^2}G(\eta)\right), \tag{14}$$

where $\delta > 0$, $\eta \in (0, 4)$, $G(\eta) = (1 - \frac{\eta^2}{16})$, and ϵ_n and ϵ_c^k are defined in (12) and (13), *respectively.*

Proof. The proof recursively computes a moment generating function of $\hat{s}_i^k(t)$ using modal decomposition of (8) and conditioning on the appropriate filtration. It subsequently uses the Markov inequality and a peeling-argument-based union bound to establish the inequality. See Appendix B. □

4. Cooperative decision making: unconstrained reward

In this section, we extend the UCB algorithm (Auer, Cesa-Bianchi, & Fischer, 2002), for single-agent decision making among arms, to design decision making in the distributed cooperative setting in which a group of M agents communicate with one another over a network with fixed graph. At every time t, each agent k updates its estimates of the mean rewards at each arm i according to the cooperative estimation algorithm of Section 3. Then each agent chooses an arm to maximize its own individual reward. We consider the case of unconstrained sub-Gaussian rewards here and the case of constrained sub-Gaussian rewards in Section 5.

Intuitively, each agent will perform better with communication than without. However, the extent of the performance advantage of each agent and the group as whole, as a result of communication, depends on the network structure. We compute bounds on group performance by computing bounds on the expected group cumulative regret, and we show how the bounds depend on graph explore–exploit index ϵ_n and nodal explore–exploit centrality indices ϵ_c^k, $k = 1, \ldots, M$.

4.1. The coop-UCB2 algorithm

The coop-UCB2 algorithm is initialized by each agent sampling each arm once and proceeds as follows (see Appendix C for pseudocode implementation). At time t each agent k selects the arm with maximum $Q_i^k(t-1) = \hat{\mu}_i^k(t-1) + C_i^k(t-1)$, where

$$C_i^k(t-1) = \sigma_g \sqrt{\frac{2\gamma}{G(\eta)} \cdot \frac{\hat{n}_i^k(t-1) + f(t-1)}{M\hat{n}_i^k(t-1)} \cdot \frac{\ln(t-1)}{\hat{n}_i^k(t-1)}}. \qquad (15)$$

Here, $f(t)$ is an increasing sublogarithmic function of t, $\gamma > 1$, $\eta \in (0, 4)$, and $G(\eta) = 1 - \eta^2/16$.

Then, at each time t, each agent k updates its cooperative estimate of the mean reward at each arm using the distributed cooperative estimation algorithm described in (8)–(10). Note that the heuristic Q_i^k requires agent k to know the total number of agents M but nothing about the graph structure.

Theorem 2 (*Upper Bound on Suboptimal Selections for coop-UCB2 Algorithm*)**.** *For the coop-UCB2 algorithm and the distributed cooperative multi-agent MAB problem under the unconstrained reward model with sub-Gaussian rewards, the number of times a suboptimal arm i is selected by all agents until time T satisfies*

$$\sum_{k=1}^{M} \mathbb{E}[n_i^k(T)] \le \frac{4\sigma_g^2 \gamma \ln T}{\Delta_i^2 G(\eta)} \left(1 + \sqrt{1 + \frac{\Delta_i^2 M G(\eta)}{2\gamma\sigma_s^2} \frac{f(T)}{\ln T}}\right) + L,$$

where

$$L(\epsilon_n, \epsilon_c^1, \ldots, \epsilon_c^M) = \sum_{k=1}^{M} (t_k^\dagger - 1) + M(1 + \epsilon_n) + 1 + \frac{2M}{\ln(1+\eta)} \left(\frac{1}{(\gamma-1)^2} + \frac{\gamma \ln((1+\epsilon_n)(1+\eta))}{\gamma - 1} + 1\right) \qquad (16)$$

is a constant independent of T and $t_k^\dagger = f^{-1}(\epsilon_c^k)$.

Proof. The upper bound is computed as for UCB1 (Auer et al., 2002) and leverages Proposition 1 and Theorem 1. See Appendix D. □

Corollary 1 (*Regret of the coop-UCB2 Algorithm*)**.** *For the coop-UCB2 algorithm and the distributed cooperative multi-agent MAB problem under the unconstrained reward model with sub-Gaussian rewards, the expected cumulative group regret until time T satisfies*

$$R^{\mathrm{unc}}(T) \le \sum_{i=1}^{N} \frac{4\sigma_g^2 \gamma \ln T}{\Delta_i G(\eta)} \left(1 + \sqrt{1 + \frac{\Delta_i^2 M G(\eta)}{2\gamma\sigma_s^2} \frac{f(T)}{\ln T}}\right) + \sum_{i=1}^{N} L\Delta_i.$$

Proof. The corollary follows by substituting the upper bound on $\sum_{k=1}^{M} \mathbb{E}[n_i^k(T)]$ from Theorem 2 into (1). □

From these bounds, we can compare performance for the distributed case relative to the centralized case, and we can draw conclusions about the predictive value of explore–exploit indices ϵ_n and ϵ_c^k as follows.

Remark 1 (*Asymptotic Regret for coop-UCB2*)**.** In the limit $t \to +\infty$, $\frac{f(t)}{\ln(t)} \to 0^+$, $\eta \to 0$, and

$$\sum_{k=1}^{M} \mathbb{E}[n_i^k(T)] \le \left(\frac{8\sigma_g^2\gamma}{\Delta_i^2} + o(1)\right) \ln T.$$

We thus recover the upper bound on regret for a centralized fusion center as given in (3) within a constant factor. □

Remark 2 (*Predicting Relative Performance from Network Graph Topology*)**.** Theorem 2 and Corollary 1 provide bounds on the performance of the group as a function of the graph structure, as measured by the group explore–exploit index ϵ_n and nodal explore–exploit centrality indices ϵ_c^k. While the logarithmic term in the upper bound on group performance is independent of graph structure, the sublogarithmic term L, given in (16), depends on ϵ_n and ϵ_c^k. Our theory predicts that the performance of a group is better for a network with smaller ϵ_n, since a smaller ϵ_n implies a smaller upper bound on expected cumulative group regret. Likewise, our theory predicts that the performance of individual agent j is better than the performance of individual agent l if $\epsilon_c^j < \epsilon_c^l$, since a smaller ϵ_c^k implies a smaller contribution from agent k to the upper bound on expected cumulative group regret. These predictions rely on the bounds being sufficiently tight; we illustrate the usefulness of the predictions with simulations in Section 6. □

5. Cooperative decision making: Constrained reward

In this section we extend our analyses in Section 4 to the case of the constrained reward model.[2] In this setting the optimal solution in terms of group regret is for the M agents to each sample a different arm from among the M-best arms at every time t. Recall that $\mathcal{O}_k^*$ is the set of k-best arms. Let $\Delta_{\min} = \min\{|m_i - m_j| \mid i, j \in \{1, \ldots, N\}, i \neq j\}$. In the following, we assume that each agent k has a preassigned unique rank $\omega^k \in \{1, \ldots, M\}$ and will attempt to sample the arm with the ω^k-th best reward. Without loss of generality, we assume that $\omega^k = k$. We define agent k^i as the index of the agent attempting to sample arm $i \in \mathcal{O}_M^*$. We let $k^i = 0$ if $i \notin \mathcal{O}_M^*$. Therefore, the expected cumulative regret of agent k at time T is

$$R^k(T) = \sum_{t=1}^{T} \left(m_{b^k} - \mathbb{E}\left[\sum_{i=1}^{N} r_i^k(t) \mathbb{1}\{i^k(t) = i\} \mathbb{I}_i^k(t)\right]\right), \qquad (17)$$

where $\mathbb{I}_i^k(t) = 1$ if agent k is the only agent to sample arm i at time t, and 0 otherwise.

In the following, we assume that while agents do not receive any reward if they sample the same arm, they still have access to the value of the reward they did not receive and they can use it in updating their estimates of the mean rewards.

5.1. The coop-UCB2-selective-learning algorithm

In this section, we present the coop-UCB2-selective-learning algorithm in which agent k selectively targets the kth best arm (see Appendix E for pseudocode implementation). The coop-UCB2-selective-learning algorithm for sub-Gaussian rewards is initialized by each agent sampling each arm once in a round-robin fashion with agent k beginning the sampling with the kth arm. At each time t, each agent k updates its cooperative estimate of the mean reward at each arm using the distributed cooperative estimation algorithm described in (8)–(10).

Subsequently, at time t, each agent k estimates $\mathcal{O}_k^*$ by constructing the set $\mathcal{O}_k(t)$ containing k arms associated with the

[2] Some authors (Anandkumar et al., 2011; Kalathil et al., 2014) have considered the case where agents that sample the same arm at the same time receive a split reward. The algorithm presented here is still appropriate for that scenario, and the regret as defined above will upper bound the regret in the case of split rewards.

indices of the k highest values in the set $\{Q_i^k(t-1) = \hat{\mu}_i^k(t-1) + C_i^k(t-1) \mid i \in \{1,\dots,N\}\}$, where

$$C_i^k(t-1) = \sigma_g \sqrt{\frac{2\gamma}{G(\eta)} \cdot \frac{\hat{n}_i^k(t-1) + f(t-1)}{M\hat{n}_i^k(t-1)} \cdot \frac{\ln(t-1)}{\hat{n}_i^k(t-1)}}, \tag{18}$$

$f(t)$ is an increasing sublogarithmic function of t, $\gamma > 1$, $\eta \in (0, 4)$, and $G(\eta) = 1 - \eta^2/16$.

Each agent k then selects the arm associated with the minimum value in the set $\{W_i^k(t-1) = \hat{\mu}_i^k(t-1) - C_i^k(t-1) \mid i \in \mathcal{O}_k\}$.

Our algorithm generalizes the selective-learning algorithm for multi-agent MABs with no communication among agents proposed in Gai and Krishnamachari (2014) to the case of communicating agents.

5.2. Analysis of the coop-UCB2-selective-learning algorithm

We first bound the number of times an arm i is incorrectly selected. We call the selection of arm $i \in \mathcal{O}_M^*$ incorrect if it is selected by an agent $k \neq k^i$. Any selection of arm $i \notin \mathcal{O}_M^*$ is incorrect. Let $\bar{n}_i^k(t)$ be the number of incorrect selections of arm i until time t.

Theorem 3 (*Upper Bound on Incorrect Selections for coop-UCB2-selective-learning Algorithm*)**.** *For the coop-UCB2-selective-learning algorithm and the distributed cooperative multi-agent MAB problem under the constrained reward model with sub-Gaussian rewards, the number of times an arm i is incorrectly selected by all agent until time T satisfies*

$$\sum_{k \neq k^i} \mathbb{E}\left[\bar{n}_i^k(T)\right] \leq \frac{4\sigma_g^2 \gamma}{\Delta_{\min}^2 G(\eta)} \left(1 + \sqrt{1 + \frac{\Delta_{\min}^2 M G(\eta)}{2\sigma_g^2 \gamma} \frac{f(T)}{\ln T}}\right) \ln T + \bar{L},$$

where

$$\bar{L}(\epsilon_n, \epsilon_c^1, \dots, \epsilon_c^M) = \sum_{k=1}^{M} (t_k^\dagger - 1) + M(1+\epsilon_n) + 1$$
$$+ \frac{2M(N+1)}{\ln(1+\eta)} \left(\frac{1}{(\gamma-1)^2} + \frac{\gamma \ln((1+\epsilon_n)(1+\eta))}{\gamma - 1} + 1\right) \tag{19}$$

is a constant independent of T and $t_k^\dagger = f^{-1}(\epsilon_c^k)$*.*

Proof. The upper bound is computed similarly to SL(K) (Gai & Krishnamachari, 2014), leveraging Proposition 1 and Theorem 1. See Appendix F. □

Corollary 2 (*Regret of the coop-UCB2-selective-learning Algorithm*)**.** *For the coop-UCB2-selective-learning algorithm and the distributed cooperative multi-agent MAB problem under the constrained reward model with sub-Gaussian rewards, the expected cumulative regret of the group satisfies*

$$R^{\mathrm{con}}(T) \leq \sum_{k=1}^{M} R^k(T) \leq m_{i^*} N B + \sum_{k=1}^{M} m_{b^k} B,$$

where

$$B = \frac{4\sigma_g^2 \gamma}{\Delta_{\min}^2 G(\eta)} \left(1 + \sqrt{1 + \frac{\Delta_{\min}^2 M G(\eta)}{2\sigma_g^2 \gamma} \frac{f(T)}{\ln T}}\right) \ln T + \bar{L}. \tag{20}$$

Proof. As in Gai and Krishnamachari (2014), agent k incurs regret either by selecting an arm $i \neq b^k$ or when another user $j \neq k$ selects arm b^k. Therefore,

$$\sum_{k=1}^{M} R^k(T) \leq \sum_{k=1}^{M} \sum_{i \neq b^k} \mathbb{E}\left[\bar{n}_i^k(T)\right] m_{b^k} + \sum_{k=1}^{M} \sum_{j \neq k} \mathbb{E}\left[\bar{n}_{b^k}^j(T)\right] m_{b^k}$$
$$\leq m_{i^*} \sum_{i=1}^{N} \sum_{k \neq k^i} \mathbb{E}\left[\bar{n}_i^k(T)\right] + \sum_{k=1}^{M} \sum_{j \neq k} \mathbb{E}\left[\bar{n}_{b^k}^j(T)\right] m_{b^k}$$
$$\leq m_{i^*} \sum_{i=1}^{N} B + \sum_{k=1}^{M} m_{b^k} B,$$

completing the proof. □

From these bounds, we can compare performance in the case of communication between agents relative to the case of no communication between agents, and we can draw conclusions about the predictive value of explore–exploit indices ϵ_n and ϵ_c^k for the unconstrained reward model, as follows.

Remark 3 (*Concise Upper Bound on Regret*)**.** The upper bound on expected cumulative group regret in Corollary 2 can be expressed concisely, at the expense of some tightness, as

$$\sum_{k=1}^{M} R^k(T) \leq m_{i^*} B(M+N).$$

In the limit $\eta \to 0^+$ and $\gamma \to 1^+$, this is a factor of $4M$ tighter than the bounds in Gai and Krishnamachari (2014), demonstrating the benefits of communication between agents for the constrained reward model.

Remark 4 (*Predicting Relative Performance from Network Graph Topology for Constrained Reward Model*)**.** Theorem 3 and Corollary 2 predict the performance of the group as a function of the graph structure for the constrained reward model just as described for the unconstrained reward model in Remark 2, since $\bar{L}$ given in (19) has the same form as L given in (16). □

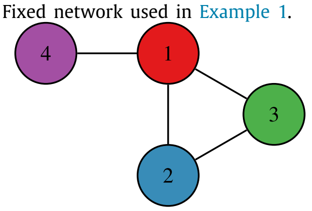
6. Numerical illustrations

In this section, we illustrate our theoretical analyses from the previous sections with numerical examples. We first provide examples in which the ordering of the performance of nodes obtained through numerical simulations is as predicted by the ordering of the nodal explore–exploit centrality indices, as discussed in Remarks 2 and 4. That is, a smaller ϵ_c^k predicts better performance for agent k. We then provide examples in which the ordering over networks of the performance of a group of agents is as predicted by the ordering over networks of the graph explore–exploit index, as discussed in Remarks 2 and 4. That is, a smaller ϵ_n predicts better performance for the group with the corresponding network graph. Our final example illustrates how performance improves with connectivity.

Unless otherwise noted in the simulations, we consider a 10-armed bandit problem with mean rewards drawn from a normal random distribution for each Monte-Carlo run with mean 0 and standard deviation 10. The sampling standard deviation is $\sigma_s = 30$ and the results displayed are the average of 10^6 Monte-Carlo runs. These parameters were selected to give illustrative results within the displayed time horizon, but the relevant conclusions hold across a wide range of parameter values. In the simulations $f(t) = \sqrt{\ln t}$, and consensus matrix P is as in (9) with $\kappa = \frac{d_{\max}}{d_{\max}-1}$.

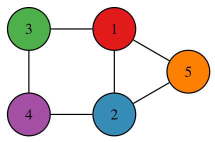
Example 1. Fig. 1 demonstrates the ordering of performance among agents using coop-UCB2 with the underlying graph structure in Table 1. The values of ϵ_c^k for each node are also given in Table 1. As predicted by Theorem 2 (Remark 2), agent 1 should have the lowest regret, agents 2 and 3 should have equal and intermediate regret, and agent 4 should have the highest regret as this is their ordering with respect to ϵ_c^k. These predictions are validated in our simulations shown in Fig. 1.

Table 1
Fixed network used in Example 1.



Agent	Degree	ϵ_c^k
1	3	0
2	2	2.31
3	2	2.31
4	1	5.41

Table 2
Fixed network used in Example 2 and several centrality indices.



Agent	Degree	Info. Cent.	ϵ_c^k
1	3	.35	1.4
2	3	.35	1.4
3	2	.28	3.4
4	2	.28	3.4
5	2	.27	2.9

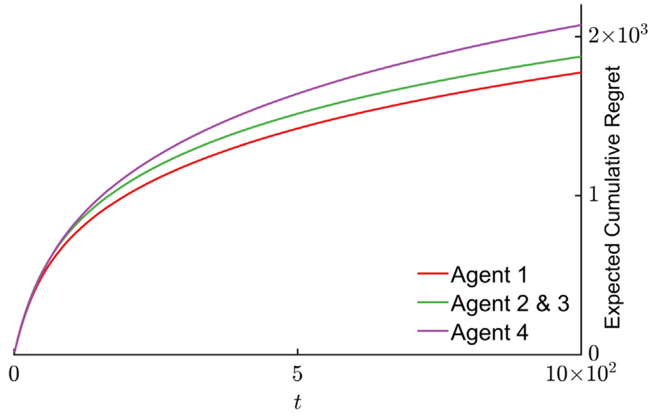
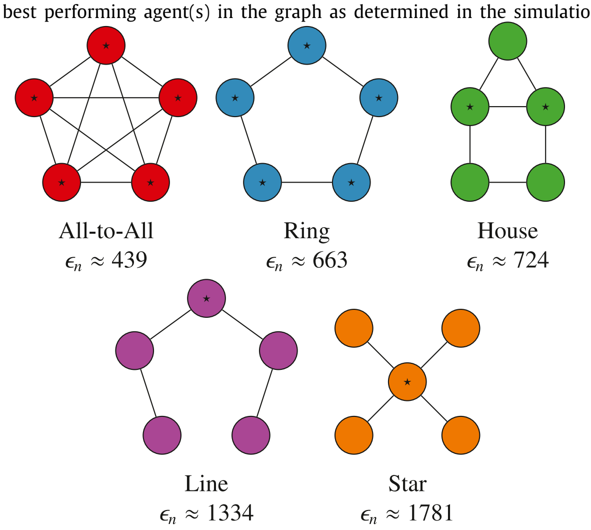


Fig. 1. Simulation results comparing expected cumulative regret for agents in the fixed network shown in Table 1. Agents 2 and 3, with the same centrality index, have nearly identical expected regret. Agent 1, with lowest centrality index, performs best and agent 4, with highest centrality index, performs worst.

Table 3
Fixed networks used in Example 3 arranged in order of increasing value of ϵ_n. Values of ϵ_n are calculated using P as in Eq. (9) and $\kappa = 0.02$. A $\star$ indicates best performing agent(s) in the graph as determined in the simulations.



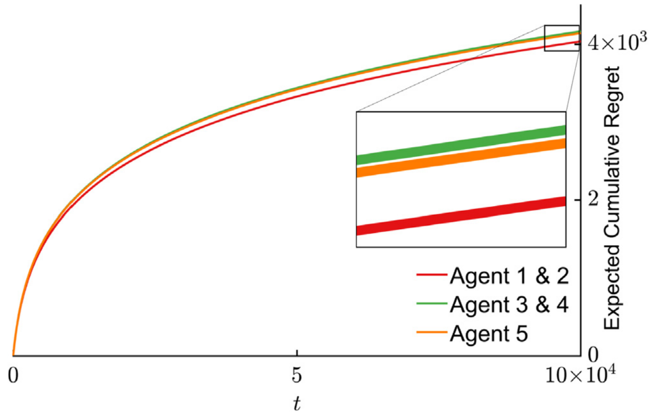


Fig. 2. Simulation results comparing expected cumulative regret for agents in the fixed graph shown in Table 2.

Example 2. Fig. 2 demonstrates the ordering of performance among agents using coop-UCB2 with the underlying graph structure in Table 2. Rewards are drawn from a normal distribution with mean 0 and standard deviation 5. The values of ϵ_c^k for each node are also given in Table 2, along with the values of degree and information centrality for each node (Poulakakis, Young, Scardovi, & Leonard, 2015), for comparison. Degree centrality for a node is defined as the number of neighbors. Information centrality, defined in Stephenson and Zelen (1989), is a nodal measure of the "effective resistance" between the node and every other node in the network.

For this example, degree centrality does not distinguish agent 5 from agents 3 and 4, whereas ϵ_c^k (and information centrality) does. Further, according to information centrality, which is larger the more central the node, node 5 is less information central than nodes 3 and 4. In contrast, according to ϵ_c^k, which is smaller the more central the node, node 5 is more explore–exploit central than nodes 3 and 4.

As in the prior example, the simulation results of Fig. 2 validate the prediction of Theorem 2 (Remark 2) that the ordering of agents by performance, as measured by expected cumulative regret, is the same as the ordering of agents by nodal explore–exploit centrality index ϵ_c^k, with smaller ϵ_c^k corresponding to lower regret. In contrast, for this example, the ordering of agents by degree or information centrality do not predict the ordering of agents by performance.

We have found some parameter regimes, specifically for rewards that are far apart in mean value, where information centrality does give the correct ordering of performance, rather than ϵ_c^k. This is likely due to sensitivity of performance to the Δ_i. However, we have observed that ϵ_c^k is broadly predictive of performance for a variety of regimes and network graphs.

6.1. Validation of relative performance of networks as predicted by graph explore–exploit index ϵ_n

Example 3. Fig. 3 compares the expected cumulative regret averaged over all agents in each of the five graphs in Table 3, where agents use coop-UCB2. The value of ϵ_n is shown in Table 3 for each graph. Theorem 2 predicts that graphs with lower ϵ_n will have lower average expected cumulative regret. Here we use two arms and $\kappa = 0.02$. Fig. 3 verifies this prediction, showing the ordering of graphs by performance is equal to the ordering of graphs by the graph explore–exploit index ϵ_n.

Fig. 4 compares expected cumulative regret for best performing agent(s) in each of the five graphs in Table 3. The central agent in the star graph outperforms the best agent in the all-to-all graph

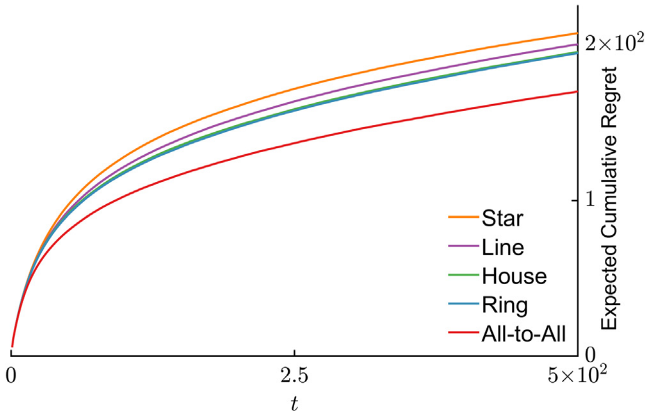


Fig. 3. Simulation results of expected cumulative regret of the group for each of the fixed graphs shown in Table 3.

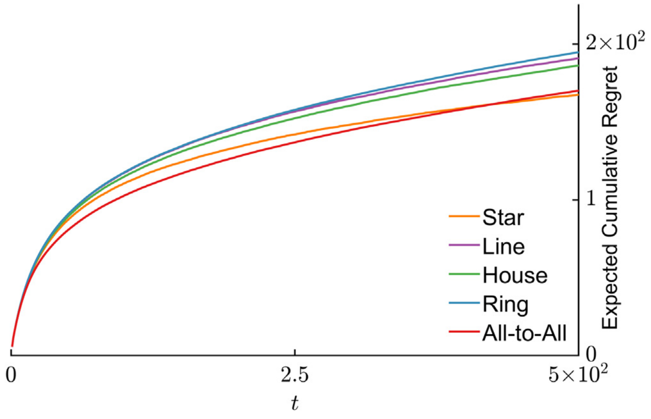


Fig. 4. Simulation results of expected cumulative regret of the agent with lowest regret in each of the fixed graphs shown in Table 3.

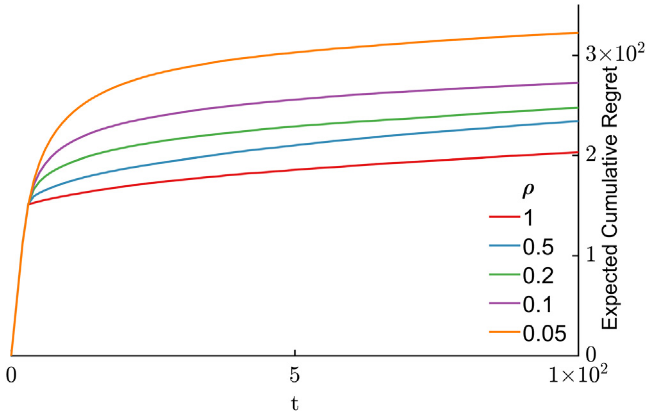


Fig. 5. Simulation results of expected cumulative regret of 100 agents on an Erdös–Rényi random graph for five different values of edge probability ρ.

despite the star graph's poor group performance. This indicates that the four peripheral agents are doing most of the exploration. The stark difference in the propensity to explore between the central and peripheral agents in the star graph demonstrates that regret accumulation for different agents could be controlled by design of the communication graph structure.

Example 4. Fig. 5 compares the average expected cumulative regret of 100 agents using coop-UCB2 (two arms and $\kappa = \frac{d_{\max}}{d_{\max}-1}$) for a range of Erdös-Rényi (ER) random graphs (Bollobás, 1998). We simulate five values of the probability ρ of a connection between any two agents, from $\rho = 0.05$ (weakly connected) to $\rho = 1.0$ (fusion center). For each ρ we randomly generated 15 ER graphs. We show the results of 2×10^4 simulations per graph, or 3×10^5 simulations per ρ. The plot shows how performance improves as the connection between agents increases.

7. Final remarks

We have used a distributed multi-agent MAB problem to explore cooperative decision making under uncertainty for networks of agents. Each agent makes choices among arms to maximize its own individual reward but cooperates with others in the group by communicating its estimates across the network. We considered both an unconstrained reward model, in which agents are not penalized if they choose the same arm at the same time, and a constrained reward model, in which agents that choose the same arm at the same time receive no reward.

We designed an algorithm for distributed cooperative estimation of mean reward at each arm. Building on this, we designed the coop-UCB2 and coop-UCB2-selective-learning algorithms for the unconstrained and constrained reward models, respectively. These are distributed algorithms that enable agents to leverage the information shared by neighbors in their decision making, without requiring that agents know the network graph structure. We proved bounds on performance, showing logarithmic expected cumulative group regret close to that of a centralized fusion center, for both reward models.

From the bounds on regret, we defined a novel graph explore–exploit index and nodal explore–exploit centrality index, which depend only on the network graph topology. The group index predicts the ordering by performance of network graphs and the nodal index predicts the ordering by performance of the nodes.

Future research directions include rigorously exploring other communications schemes, which may offer better performance or be better suited to modeling classes of networked systems. The tradeoff between communication frequency and performance (Madhushani & Leonard, 2020) as well as the presence of noisy communications (Savas, Srivastava, & Leonard, 2017) will be important considerations.

Acknowledgment

The authors thank Tor Lattimore for pointing out an error in a previous version of one of the proofs.

Appendix A. Proof of Proposition 1

We begin with statement (i). From (7) it follows that

$$\begin{aligned}\hat{\mathbf{n}}_i(t) &= P^t\hat{\mathbf{n}}_i(0) + \sum_{\tau=1}^{t} P^{t-\tau+1}\boldsymbol{\xi}_i(\tau)\\ &= \sum_{\tau=1}^{t}\Big(\frac{1}{M}\mathbf{1}_M\mathbf{1}_M^\top\boldsymbol{\xi}_i(\tau) + \sum_{p=2}^{M}\lambda_p^{t-\tau+1}\mathbf{u}_p\mathbf{u}_p{}^\top\boldsymbol{\xi}_i(\tau)\Big)\\ &= n_i^{\text{cent}}(t)\mathbf{1}_M + \sum_{\tau=1}^{t}\sum_{p=2}^{M}\lambda_p^{t-\tau+1}\mathbf{u}_p\mathbf{u}_p{}^\top\boldsymbol{\xi}_i(\tau). \qquad \text{(A.1)}\end{aligned}$$

For (i), we bound the kth entry of the second term of (A.1):

$$\sum_{\tau=1}^{t}\sum_{p=2}^{M}\lambda_p^{t-\tau+1}\left(\mathbf{u}_p\mathbf{u}_p{}^{\top}\boldsymbol{\xi}_i(\tau)\right)_k \le \sum_{\tau=1}^{t}\sum_{p=2}^{M}|\lambda_p^{t-\tau+1}|\|\mathbf{u}_p\|_2^2\|\boldsymbol{\xi}_i(\tau)\|_2$$
$$\le \sqrt{M}\sum_{\tau=1}^{t}\sum_{p=2}^{M}|\lambda_p^{t-\tau+1}| \le \epsilon_n.$$

To prove statement (ii), let $\nu_{pwi}(\tau)=\sum_{j=1}^{M}u_p^ju_w^j\xi_i^j(\tau)$ and then

$$\sum_{\tau=1}^{t}\sum_{j=1}^{M}\left(\sum_{p=1}^{M}\lambda_p^{t-\tau+1}u_p^ku_p^j\right)^2\xi_i^j(\tau)$$
$$=\sum_{\tau=1}^{t}\sum_{p=1}^{M}\sum_{w=1}^{M}(\lambda_p\lambda_w)^{t-\tau+1}u_p^ku_w^k\sum_{j=1}^{M}u_p^ju_w^j\xi_i^j(\tau)$$
$$=\sum_{\tau=1}^{t}\sum_{p=1}^{M}\sum_{w=2}^{M}(\lambda_p\lambda_w)^{t-\tau+1}u_p^ku_w^k\nu_{pwi}(\tau)$$
$$+\frac{1}{M}\sum_{\tau=1}^{t}\sum_{p=1}^{M}\sum_{j=1}^{M}\lambda_p^{t-\tau+1}u_p^ku_p^j\xi_i^j(\tau)$$
$$=\sum_{\tau=1}^{t}\sum_{p=1}^{M}\sum_{w=2}^{M}(\lambda_p\lambda_w)^{t-\tau+1}u_p^ku_w^k\nu_{pwi}(\tau)+\frac{1}{M}\hat{n}_i^k(t). \tag{A.2}$$

This establishes (ii) since for the first term of (A.2):

$$\sum_{\tau=1}^{t}(\lambda_p\lambda_w)^{t-\tau+1}u_p^ku_w^k\nu_{pwi}(\tau)\le\sum_{\tau=1}^{t}|(\lambda_p\lambda_w)^{t-\tau+1}\parallel u_p^ku_w^k\nu_{pwi}(\tau)|$$
$$\le\sum_{\tau=0}^{t-1}|\lambda_p\lambda_w|^{t-\tau+1}a_{pw}(k)\le\frac{|\lambda_p\lambda_w|}{1-|\lambda_p\lambda_w|}a_{pw}(k).$$

Appendix B. Proof of Theorem 1

We begin by noting that $\hat{s}_i^k(t)$ can be decomposed as

$$\hat{s}_i^k(t)=\sum_{\tau=1}^{t}\sum_{p=1}^{M}\lambda_p^{t-\tau+1}\sum_{j=1}^{M}u_p^ku_p^jr_i^j(\tau)\xi_i^j(\tau). \tag{B.1}$$

Let $\hat{s}_i^{kp}(t)=\sum_{\tau=1}^{t}\lambda_p^{t-\tau+1}\sum_{j=1}^{M}u_p^ku_p^jr_i^j(\tau)\xi_i^j(\tau)$. Then,

$$\sum_{p=1}^{M}\hat{s}_i^{kp}(t)=\sum_{p=1}^{M}\sum_{j=1}^{M}\lambda_pu_p^ku_p^jr_i^j(t)\xi_i^j(t)+\sum_{p=1}^{M}\lambda_p\hat{s}_i^{kp}(t-1). \tag{B.2}$$

It follows from (B.1) and (B.2) that for any $\Theta>0$

$$\mathbb{E}\left[\exp\left(\Theta\hat{s}_i^k(t)\right)\Big|\mathcal{F}_{t-1}\right]=\mathbb{E}\left[\exp\left(\Theta\sum_{p=1}^{M}\hat{s}_i^{kp}(t)\right)\Big|\mathcal{F}_{t-1}\right]$$
$$=\mathbb{E}\left[\exp\left(\Theta\sum_{p=1}^{M}\lambda_p\sum_{j=1}^{M}u_p^ku_p^jr_i^j(t)\xi_i^j(t)\right)\Big|\mathcal{F}_{t-1}\right]K_{(t-1)}$$
$$=\prod_{j=1}^{M}\mathbb{E}\left[\exp\left(\Theta\sum_{p=1}^{M}\lambda_pu_p^ku_p^jr_i^j(t)\xi_i^j(t)\right)\Big|\mathcal{F}_{t-1}\right]K_{(t-1)}$$
$$=\exp\left(\sum_{j=1}^{M}\phi_i\left(\Theta\sum_{p=1}^{M}\lambda_pu_p^ku_p^j\xi_i^j(t)\right)\right)K_{(t-1)}$$
$$=\exp\left(\sum_{j=1}^{M}\phi_i\left(\Theta\sum_{p=1}^{M}\lambda_pu_p^ku_p^j\right)\xi_i^j(t)\right)K_{(t-1)},$$

$$K_{(t-1)}=\exp\left(\Theta\sum_{p=1}^{M}\lambda_p\hat{s}_i^{kp}(t-1)\right),$$

and the second-to-last equality follows since, conditioned on $\mathcal{F}_{t-1}$, $\xi_i^j(t)$ is deterministic and $r_i^j(t)$ are i.i.d. for each $j\in\{1,\ldots,M\}$. The last equality follows since $\xi_i^j(t)$ is binary and the two expressions are the same for $\xi_i^j(t)\in\{0,1\}$. Therefore,

$$\mathbb{E}\left[\exp\left(\Theta\sum_{p=1}^{M}\hat{s}_i^{kp}(t)-\sum_{j=1}^{M}\phi_i\left(\Theta\sum_{p=1}^{M}\lambda_pu_p^ku_p^j\right)\xi_i^j(t)\right)\Big|\mathcal{F}_{t-1}\right]=K_{(t-1)}.$$

Using the above argument recursively with $s_i^k(0)=0$, we obtain

$$\mathbb{E}\left[\exp\left(\Theta\hat{s}_i^k(t)-\sum_{\tau=1}^{t}\sum_{j=1}^{M}\phi_i\left(\Theta\sum_{p=1}^{M}\lambda_p^{t-\tau+1}u_p^ku_p^j\right)\xi_i^j(\tau)\right)\right]=1.$$

For sub-Gaussian random variables $\phi_i(\beta)\le\beta m_i+\frac{1}{2}\sigma_g^2\beta^2$, thus

$$1=\mathbb{E}\left[\exp\left(\Theta\left(\hat{s}_i^k(t)-m_i\hat{n}_i^k(t)\right)\right.\right. \tag{B.3}$$
$$\left.\left.-\frac{\sigma_g^2}{2}\sum_{\tau=1}^{t}\sum_{j=1}^{M}\left(\Theta\sum_{p=1}^{M}\lambda_p^{t-\tau+1}u_p^ku_p^j\right)^2\xi_i^j(\tau)\right)\right]$$
$$\ge\mathbb{E}\left[\exp\left(\Theta\left(\hat{s}_i^k(t)-m_i\hat{n}_i^k(t)\right)-\frac{\sigma_g^2\Theta^2}{2M}\left(\hat{n}_i^k(t)+\epsilon_c^k\right)\right)\right],$$

where the last inequality follows from the second statement of Proposition 1. Now using the Markov inequality, we obtain

$$e^{-a}\ge\mathbb{P}\left(\exp\left(\Theta\left(\hat{s}_i^k(t)-m_i\hat{n}_i^k(t)\right)-\frac{\sigma_g^2\Theta^2}{2M}\left(\hat{n}_i^k(t)+\epsilon_c^k\right)\right)\ge e^a\right)$$
$$=\mathbb{P}\left(\frac{\hat{s}_i^k(t)-m_i\hat{n}_i^k(t)}{\left(\frac{1}{M}\left(\hat{n}_i^k(t)+\epsilon_c^k\right)\right)^{\frac{1}{2}}}\ge\frac{a}{\Theta}\left(\frac{1}{M}\left(\hat{n}_i^k(t)+\epsilon_c^k\right)\right)^{-\frac{1}{2}}\right.$$
$$\left.+\frac{\sigma_g^2\Theta}{2}\left(\frac{1}{M}\left(\hat{n}_i^k(t)+\epsilon_c^k\right)\right)^{\frac{1}{2}}\right). \tag{B.4}$$

Random variable $\hat{n}_i^k(t)$ on the right of (B.4) depends on the random variable on the left. So, we use union bounds on $\hat{n}_i^k(t)$ to obtain the concentration inequality. Consider an exponentially increasing sequence of time indices $\{(1+\eta)^{h-1}\mid h\in\{1,\ldots,D\}\}$, where $D=\left\lceil\frac{\ln(t+\epsilon_n)}{\ln(1+\eta)}\right\rceil$ and $\eta>0$. For every $h\in\{1,\ldots,D\}$, define

$$\Theta_h=\frac{1}{\sigma_g}\sqrt{\frac{2aM}{(1+\eta)^{h-\frac{1}{2}}+\epsilon_c^k}}. \tag{B.5}$$

Thus, if $(1+\eta)^{h-1}\le\hat{n}_i^k(t)\le(1+\eta)^h$, then

$$\frac{a}{\Theta_h}\left(\frac{1}{M}\left(\hat{n}_i^k(t)+\epsilon_c^k\right)\right)^{-\frac{1}{2}}+\frac{\sigma_g^2\Theta_h}{2}\left(\frac{1}{M}\left(\hat{n}_i^k(t)+\epsilon_c^k\right)\right)^{\frac{1}{2}}$$
$$=\sigma_g\sqrt{\frac{a}{2}}\left(\left(\frac{(1+\eta)^{h-\frac{1}{2}}+\epsilon_c^k}{\hat{n}_i^k(t)+\epsilon_c^k}\right)^{\frac{1}{2}}+\left(\frac{\hat{n}_i^k(t)+\epsilon_c^k}{(1+\eta)^{h-\frac{1}{2}}+\epsilon_c^k}\right)^{\frac{1}{2}}\right)$$

$$\le \sigma_g\sqrt{\frac{a}{2}}\left(\left(\frac{(1+\eta)^{h-\frac{1}{2}}}{\hat{n}_i^k(t)}\right)^{\frac{1}{2}}+\left(\frac{\hat{n}_i^k(t)}{(1+\eta)^{h-\frac{1}{2}}}\right)^{\frac{1}{2}}\right)$$

$$\le \sigma_g\sqrt{\frac{a}{2}}\left((1+\eta)^{\frac{1}{4}}+(1+\eta)^{-\frac{1}{4}}\right), \tag{B.6}$$

where the second-to-last inequality follows from the fact that for $a, b > 0$, the function $\epsilon \mapsto \sqrt{\frac{a+\epsilon}{b+\epsilon}} + \sqrt{\frac{b+\epsilon}{a+\epsilon}}$ with domain $\mathbb{R}_{\ge 0}$ is monotonically non-increasing, and
the last inequality follows from the fact that for $\eta > 0$, the function $x \mapsto \sqrt{\frac{(1+\eta)^{h-\frac{1}{2}}}{x}} + \sqrt{\frac{x}{(1+\eta)^{h-\frac{1}{2}}}}$ with domain $[(1+\eta)^{h-1}, (1+\eta)^h]$ achieves its maximum at either of the boundaries. Applying union bounds on D possible values of h and using (B.6) for $(1+\eta)^{h-1} \le \hat{n}_i^k(t) \le (1+\eta)^h$, from (B.4) we get

$$\mathbb{P}\left(\frac{\hat{s}_i^k(t)-m_i\hat{n}_i^k(t)}{\left(\frac{1}{M}\left(\hat{n}_i^k(t)+\epsilon_c^k\right)\right)^{\frac{1}{2}}} > \sigma_g\sqrt{\frac{a}{2}}\left((1+\eta)^{\frac{1}{4}}+(1+\eta)^{-\frac{1}{4}}\right)\right)$$

$$\le \sum_{h=1}^{D}\mathbb{P}\left(\frac{\hat{s}_i^k(t)-m_i\hat{n}_i^k(t)}{\left(\frac{1}{M}\left(\hat{n}_i^k(t)+\epsilon_c^k\right)\right)^{\frac{1}{2}}} > \frac{a}{\Theta_h}\left(\frac{1}{M}\left(\hat{n}_i^k(t)+\epsilon_c^k\right)\right)^{-\frac{1}{2}} + \frac{\sigma_g^2\Theta_h}{2}\left(\frac{1}{M}\left(\hat{n}_i^k(t)+\epsilon_c^k\right)\right)^{\frac{1}{2}}\right.$$

$$\left.\& \ (1+\eta)^{h-1} \le \hat{n}_i^k(t)+\epsilon_c^k < (1+\eta)^h\right) \le De^{-a}.$$

Setting $\sigma_g\sqrt{\frac{a}{2}}\left((1+\eta)^{\frac{1}{4}}+(1+\eta)^{-\frac{1}{4}}\right) = \delta$ yields

$$\mathbb{P}\left(\frac{\hat{s}_i^k(t)-m_i\hat{n}_i^k(t)}{\left(\frac{1}{M}\left(\hat{n}_i^k(t)+\epsilon_c^k\right)\right)^{\frac{1}{2}}} > \delta\right)$$

$$\le D\exp\left(\frac{-2\delta^2}{\sigma_g^2\left((1+\eta)^{\frac{1}{4}}+(1+\eta)^{-\frac{1}{4}}\right)^2}\right).$$

It can be verified using Taylor series expansion that

$$\frac{4}{\left((1+\eta)^{\frac{1}{4}}+(1+\eta)^{-\frac{1}{4}}\right)^2} \ge 1-\frac{\eta^2}{16}.$$

Therefore, it holds that

$$\mathbb{P}\left(\frac{\hat{s}_i^k(t)-m_i\hat{n}_i^k(t)}{\left(\frac{1}{M}\left(\hat{n}_i^k(t)+\epsilon_c^k\right)\right)^{\frac{1}{2}}} > \delta\right) \le D\exp\left(\frac{-\delta^2}{2\sigma_g^2}\left(1-\frac{\eta^2}{16}\right)\right)$$

$$= \left\lceil\frac{\ln(t+\epsilon_n)}{\ln(1+\eta)}\right\rceil\exp\left(\frac{-\delta^2}{2\sigma_g^2}\left(1-\frac{\eta^2}{16}\right)\right).$$

Appendix C. Pseudocode for coop-UCB2

See Algorithm 1.

Algorithm 1: *coop-UCB2*

```
Input  : arms {1, ..., N}, agents {1, ..., M};
Input  : parameters $\sigma_g > 0$, $\eta > 0$, $\gamma > 1$, function $f(t)$;
Output : allocation sequence $i^k(t), t \in \{1,\dots,T\}, k \in \{1,\dots,M\}$;

set $\hat{n}_i^k \leftarrow 0, \hat{s}_i^k \leftarrow 0, i \in \{1,\dots,N\}, k \in \{1,\dots,M\}$;
for $t \in \{1,\dots,T\}$ do
    if $t \le N$ then
        % Initialization
        for each agent $k \in \{1,\dots,M\}$ do
            $i^k(t) \leftarrow t$ ;
            collect reward $r^k(t)$ ;
    else
        for each agent $k \in \{1,\dots,M\}$ do
            % select arm with maximum $Q_i^k$
            for each arm $i \in \{1,\dots,N\}$ do
                $Q_i^k \leftarrow \frac{\hat{s}_i^k}{\hat{n}_i^k} + \sigma_g\sqrt{\frac{2\gamma}{G(\eta)} \cdot \frac{\hat{n}_i^k+f(t-1)}{M\hat{n}_i^k} \cdot \frac{\ln(t-1)}{\hat{n}_i^k}}$;
            $i^k(t) \leftarrow \mathrm{argmax}\{Q_i^k \mid i \in \{1,\dots,N\}\}$ ;
            collect reward $r^k(t)$ ;
    for $i \in \{1,\dots,N\}$ do
        update $\hat{\boldsymbol{n}}_i$ and $\hat{\boldsymbol{s}}_i$ using (7) and (8);
```

Appendix D. Proof of Theorem 2

We proceed similarly to Auer et al. (2002). The number of selections of a suboptimal arm i by all agents until time T is

$$\sum_{k=1}^{M} n_i^k(T) \le \sum_{k=1}^{M}(t_k^{\dagger}-1) + \sum_{k=1}^{M}\sum_{t=t_k^{\dagger}}^{T}\mathbb{1}(Q_i^k(t-1) \ge Q_{i^*}^k(t-1))$$

$$\le A + \sum_{k=1}^{M}\left((t_k^{\dagger}-1) + \sum_{t=t_k^{\dagger}}^{T}\mathbb{1}(Q_i^k(t-1) \ge Q_{i^*}^k(t-1), Mn_i^{\text{cent}} \ge A)\right) \tag{D.1}$$

where $A > 0$ is a constant that will be chosen later.

At a given time $t+1$ an individual agent k will choose a suboptimal arm only if $Q_i^k(t) \ge Q_{i^*}^k(t)$. For this condition to be true at least one of the following three conditions must hold:

$$\hat{\mu}_{i^*}(t) \le m_{i^*} - C_{i^*}^k(t) \tag{D.2}$$

$$\hat{\mu}_i(t) \ge m_i + C_i^k(t) \tag{D.3}$$

$$m_{i^*} < m_i + 2C_i^k(t). \tag{D.4}$$

We bound the probability that (D.2) and (D.3) hold using Theorem 1:

$$\mathbb{P}\left(\text{(D.2) holds} \mid t \ge t_k^{\dagger}\right)$$

$$= \mathbb{P}\left(\frac{\hat{s}_i^k - m_i\hat{n}_i^k}{\sqrt{\frac{1}{M}\left(\hat{n}_i^k(t)+f(t)\right)}} \ge \sigma_g\sqrt{\frac{2\gamma\ln(t)}{G(\eta)}}\ \middle|\ t \ge t_k^{\dagger}\right)$$

$$\le \mathbb{P}\left(\frac{\hat{s}_i^k - m_i\hat{n}_i^k}{\sqrt{\frac{1}{M}\left(\hat{n}_i^k(t)+\epsilon_c^k\right)}} \ge \sigma_g\sqrt{\frac{2\gamma\ln(t)}{G(\eta)}}\ \middle|\ t \ge t_k^{\dagger}\right)$$

$$\le \left(\frac{\ln(t)}{\ln(1+\eta)} + \frac{\ln(1+\epsilon_n)}{\ln(1+\eta)} + 1\right)\frac{1}{t^{\gamma}},$$

$$\mathbb{P}\left(\text{(D.3) holds} \mid t \ge t_k^{\dagger}\right) \le \left(\frac{\ln(t)}{\ln(1+\eta)} + \frac{\ln(1+\epsilon_n)}{\ln(1+\eta)} + 1\right)\frac{1}{t^{\gamma}}.$$

We now examine the event (D.4).

$$m_{i^*} < m_i + 2C_i^k(t)$$

$$\implies \hat{n}_i^k(t)^2\frac{\Delta_i^2 MG(\eta)}{8\sigma_g^2} - \gamma\hat{n}_i^k(t)\ln(t) - \gamma f(t)\ln(t) < 0. \tag{D.5}$$

The quadratic equation (D.5) can be solved to find its roots, and if $\hat{n}_i(t)$ is greater than the larger root the inequality will never hold. Solving the quadratic equation (D.5), we obtain that event (D.4) does not hold if

$$\hat{n}_i^k(t) \geq \frac{4\sigma_g^2\gamma\ln(t)}{\Delta_i^2 MG(\eta)} + \sqrt{\left(\frac{4\gamma\sigma_g^2\ln(t)}{\Delta_i^2 MG(\eta)}\right)^2 + \frac{8\sigma_g^2 f(t)\gamma\ln(t)}{\Delta_i^2 MG(\eta)}}$$
$$= \frac{4\sigma_g^2\gamma\ln t}{\Delta_i^2 MG(\eta)}\left(1+\sqrt{1+\frac{\Delta_i^2 MG(\eta)}{2\sigma_g^2\gamma}\frac{f(t)}{\ln t}}\right).$$

Now, we set $A = \left\lceil M\epsilon_n + \frac{4\sigma_g^2\gamma\ln T}{\Delta_i^2 G(\eta)}\left(1+\sqrt{1+\frac{\Delta_i^2 MG(\eta)}{2\gamma\sigma_g^2}\frac{f(T)}{\ln T}}\right)\right\rceil$. It follows from monotonicity of $f(t)$ and $\ln(t)$ and statement (i) of Proposition 1 that event (D.4) does not hold if $Mn_i^{\text{cent}}(t) > A$.

Therefore, from (D.1) we see that

$$\sum_{k=1}^M \mathbb{E}\left[n_i^k(T)\right] \leq \bar{A} + \sum_{k=1}^M (t_k^\dagger - 1)$$
$$+\frac{2}{\ln(1+\eta)}\sum_{k=1}^M\sum_{t=t_k^\dagger}^T\left(\frac{\ln(t)}{t^\gamma}+\frac{\ln((1+\epsilon_n)(1+\eta))}{t^\gamma}\right)$$
$$\leq \bar{A} + \sum_{k=1}^M (t_k^\dagger - 1) + \frac{2M}{\ln(1+\eta)}\sum_{t=1}^T\left(\frac{\ln(t)}{t^\gamma}+\frac{\ln((1+\epsilon_n)(1+\eta))}{t^\gamma}\right)$$
$$\leq \bar{A} + \sum_{k=1}^M (t_k^\dagger - 1) + \frac{2M}{\ln(1+\eta)}\left(\frac{1}{(\gamma-1)^2}+\frac{\gamma\ln((1+\epsilon_n)(1+\eta))}{\gamma-1}+1\right),$$

where $\bar{A} = \max\{M, A\}$ is chosen to account for the M selections of the ith arm during the initialization phase.

Appendix E. Pseudocode for coop-UCB2-selective-learning

See Algorithm 2.

Algorithm 2: *coop-UCB2-selective-learning*

```
Input  : arms $\{1,\ldots,N\}$, agents $\{1,\ldots,M\}$;
Input  : parameters $\sigma_g > 0$, $\eta > 0$, $\gamma > 1$, function $f(t)$;
Output : allocation sequence $i^k(t), t \in \{1,\ldots,T\}, k \in \{1,\ldots,M\}$;
set $\hat{n}_i^k \leftarrow 0, \hat{s}_i^k \leftarrow 0, i \in \{1,\ldots,N\}, k \in \{1,\ldots,M\}$;
for $t \in \{1,\ldots,T\}$ do
    if $t \leq N$ then
        % Initialization
        for each agent $k \in \{1,\ldots,M\}$ do
            $i^k(t) \leftarrow (t-1+k) \mod N$ ;
            collect reward $r^k(t)$ ;
    else
        for each agent $k \in \{1,\ldots,M\}$ do
            for each arm $i \in \{1,\ldots,N\}$ do
                $Q_i^k \leftarrow \frac{\hat{s}_i^k}{\hat{n}_i^k} + \sigma_g\sqrt{\frac{2\gamma}{G(\eta)}\cdot\frac{\hat{n}_i^k+f(t-1)}{M\hat{n}_i^k}\cdot\frac{\ln(t-1)}{\hat{n}_i^k}}$;
            % Compute descending sort indices for $Q_i^k$
            $I_i^k \leftarrow$ sort_index($\{Q_i^k \mid i \in \{1,\ldots,N\}\}$, 'descend');
            % Estimate k-best arms
            $\mathcal{O}_k \leftarrow \{I_1^k,\ldots,I_k^k\}$ ;
            % select the worst arm from k-best arms
            for each arm $i \in \mathcal{O}_k$ do
                $W_i^k \leftarrow \frac{\hat{s}_i^k}{\hat{n}_i^k} - \sigma_g\sqrt{\frac{2\gamma}{G(\eta)}\cdot\frac{\hat{n}_i^k+f(t-1)}{M\hat{n}_i^k}\cdot\frac{\ln(t-1)}{\hat{n}_i^k}}$;
            $i^k(t) \leftarrow \operatorname{argmin}\{W_i^k \mid i \in \mathcal{O}_k\}$ ;
            collect reward $r^k(t)$ ;
    for $i \in \{1,\ldots,N\}$ do
        update $\hat{\boldsymbol{n}}_i$ and $\hat{\boldsymbol{s}}_i$ using (7) and (8);
```

Appendix F. Proof of Theorem 3

We begin by noting that

$$\sum_{k\neq k^i} n_i^k(T) = \sum_{k\neq k^i}\sum_{t=1}^T \mathbb{1}\left\{i^k(t)=i\right\}$$
$$=\sum_{k\neq k^i}\sum_{t=1}^T\left(\mathbb{1}\left\{i^k(t)=i, m_i < m_{b^k}\right\}+\mathbb{1}\left\{i^k(t)=i, m_i \geq m_{b^k}\right\}\right)$$
$$\leq A + \sum_{k=1}^M (t_k^\dagger - 1) + \sum_{k\neq k^i}\sum_{t=t_k^\dagger}^T \mathbb{1}\left\{i^k(t)=i, m_i < m_{b^k}, Mn_i^{\text{cent}}(t) \geq A\right\}$$
$$+\sum_{k\neq k^i}\sum_{t=t_k^\dagger}^T \mathbb{1}\left\{i^k(t)=i, m_i \geq m_{b^k}, Mn_i^{\text{cent}}(t) \geq A\right\}, \tag{F.1}$$

where A is a constant that will be chosen later. In the case where $m_i < m_{b^k}$, agent k picking arm i implies that there exists an arm $j \in \mathcal{O}_k^*$ such that $j \notin \mathcal{O}_k(t)$. Therefore, the following holds:

$$\sum_{k\neq k^i}\sum_{t=t_k^\dagger}^T \mathbb{1}\left\{i^k(t)=i, m_i < m_{b^k}, Mn_i^{\text{cent}}(t) \geq A\right\}$$
$$\leq \sum_{k\neq k^i}\sum_{t_k^\dagger-1}^{T-1}\mathbb{1}\Big\{Q_i^k(t) \geq Q_j^k(t), \text{ for some } j \in \mathcal{O}_k^* \setminus \mathcal{O}_k(t),$$
$$m_i < m_{b^k}, Mn_i^{\text{cent}}(t) \geq A\Big\}$$
$$\leq \sum_{k\neq k^i}\sum_{t=t_k^\dagger-1}^T\sum_{j\in\mathcal{O}_k^*}\mathbb{1}\Big\{Q_i^k(t) \geq Q_j^k(t), m_i < m_{b^k}, Mn_i^{\text{cent}}(t) \geq A\Big\}$$
$$\leq \sum_{k\neq k^i}\sum_{j\in\mathcal{O}_k^*}\sum_{t=t_k^\dagger}^T\mathbb{1}\Big\{Q_i^k(t) \geq Q_j^k(t), m_i < m_{b^k}, Mn_i^{\text{cent}}(t) \geq A\Big\}.$$

As in Theorem 2, $Q_i^k(t-1) \geq Q_j^k(t-1)$ implies that at least one of the following three conditions must hold for any $j \in O_k^*$:

$$\hat{\mu}_j(t) \leq m_j - C_j^k(t) \tag{F.2}$$
$$\hat{\mu}_i(t) \geq m_i + C_i^k(t) \tag{F.3}$$
$$m_j < m_i + 2C_i^k(t). \tag{F.4}$$

The first two equations are bounded using Theorem 1 as in the proof of Theorem 2. The third equation is equivalent to

$$2C_i^k(t) > m_j - m_i > \Delta_{\min},$$

which, as in the proof of Theorem 2, does not hold if

$$n_i^k(t) > \frac{4\sigma_g^2\gamma}{\Delta_{\min}^2 G(\eta)}\left(1+\sqrt{1+\frac{\Delta_{\min}^2 MG(\eta)}{2\sigma_g^2\gamma}\frac{f(T)}{\ln T}}\right)\ln T.$$

Therefore, for

$$A = \left\lceil M\epsilon_n + \frac{4\sigma_g^2\gamma}{\Delta_{\min}^2 G(\eta)}\left(1+\sqrt{1+\frac{\Delta_{\min}^2 MG(\eta)}{2\sigma_g^2\gamma}\frac{f(T)}{\ln T}}\right)\ln T\right\rceil,$$

(F.4) does not hold. This results in

$$\sum_{k\neq k^i}\sum_{j\in\mathcal{O}_k^*}\sum_{t=t_k^\dagger-1}^{T}\mathbb{1}\left\{Q_i^k(t-1)\geq Q_j^k(t-1),\, m_i<m_{b^k},\, Mn_i^{\text{cent}}(t)\geq A\right\}$$
$$\leq\sum_{k\neq k^i}\sum_{j\in\mathcal{O}_k^*}\frac{2}{\ln(1+\eta)}\left(\frac{1}{(\gamma-1)^2}+\frac{\gamma\ln((1+\epsilon_n)(1+\eta))}{\gamma-1}+1\right)$$
$$\leq\frac{M(M+1)}{\ln(1+\eta)}\left(\frac{1}{(\gamma-1)^2}+\frac{\gamma\ln((1+\epsilon_n)(1+\eta))}{\gamma-1}+1\right). \tag{F.5}$$

We now examine the second part of (F.1) when $m_i\geq m_{b^k}$ and split the conditional as

$$\mathbb{1}\left\{i^k(t)=i,\, m_i\geq m_{b^k},\, Mn_i^{\text{cent}}(t)\geq A\right\}$$
$$=\mathbb{1}\left\{i^k(t)=i,\, m_i\geq m_{b^k},\, Mn_i^{\text{cent}}(t)\geq A,\, \mathcal{O}_{\omega^k}(t)=\mathcal{O}_{\omega^k}^*\right\}$$
$$+\mathbb{1}\left\{i^k(t)=i,\, m_i\geq m_{b^k},\, Mn_i^{\text{cent}}(t)\geq A,\, \mathcal{O}_{\omega^k}(t)\neq\mathcal{O}_{\omega^k}^*\right\}$$
$$\leq\mathbb{1}\left\{m_i\geq m_{b^k},\, Mn_i^{\text{cent}}(t)\geq A,\, W_i^k(t-1)\leq W_{b^k}^k(t-1)\right\}$$
$$+\mathbb{1}\left\{m_i\geq m_{b^k},\, Mn_i^{\text{cent}}(t)\geq A,\, W_i^k(t-1)\leq W_h^k(t-1)\right\} \tag{F.6}$$

for any arm $h\notin\mathcal{O}_k^*$. The two indicator functions in (F.6) can be combined as follows:

$$\text{(F.6)}=\mathbb{1}\left\{m_i\geq m_{b^k},\, Mn_i^{\text{cent}}(t)\geq A,\, W_i^k(t-1)\leq W_j^k(t-1)\right\},$$

for any $j\notin\mathcal{O}_k^*\setminus\{b^k\}$. This results in

$$\sum_{k\neq k^i}\sum_{t=t_k^\dagger}^{T}\mathbb{1}\left\{i^k(t)=i,\, m_i\geq m_{b^k},\, Mn_i^{\text{cent}}(t)\geq A\right\}$$
$$\leq\sum_{k\neq k^i}\sum_{j\notin\mathcal{O}_k^*\setminus\{b^k\}}\sum_{t=t_k^\dagger}^{T}\mathbb{1}\Big\{m_i\geq m_{b^k},\, Mn_i^{\text{cent}}(t)\geq A,$$
$$W_i^k(t-1)\leq W_j^k(t-1)\Big\}. \tag{F.7}$$

For $W_i^k(t)\leq W_j^k(t)$ to be true, at least one of the following must hold:

$$\hat{\mu}_i(t)\leq m_i-C_i^k(t) \tag{F.8}$$
$$\hat{\mu}_j(t)\geq m_j+C_j^k(t) \tag{F.9}$$
$$m_i<m_j+2C_j^k(t). \tag{F.10}$$

(F.8) and (F.9) can be bounded using Theorem 1. As before, (F.10) never holds due to our choice of A. Similarly to (F.5)

$$\sum_{k\neq k^i}\sum_{t=1}^{T}\mathbb{P}\left(i^k(t)=i,\, m_i\geq m_{b^k},\, Mn_i^{\text{cent}}(t)\geq A\right)$$
$$\leq\sum_{k\neq k^i}\sum_{j\notin\mathcal{O}_k^*\setminus\{b^k\}}\frac{2}{\ln(1+\eta)}\left(\frac{1}{(\gamma-1)^2}+\frac{\gamma\ln((1+\epsilon_n)(1+\eta))}{\gamma-1}+1\right)$$
$$\leq\frac{2NM-M(M-1)}{\ln(1+\eta)}\left(\frac{1}{(\gamma-1)^2}+\frac{\gamma\ln((1+\epsilon_n)(1+\eta))}{\gamma-1}+1\right). \tag{F.11}$$

Using (F.1), (F.5), and (F.11) and accounting for the selections of arm i during the initialization as in the proof of Theorem 2, we obtain the bound in the theorem statement.

References

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

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

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

Bistritz, I., & Leshem, A. (2018). Distributed multi-player bandits-a game of thrones approach. In *Advances in neural information processing systems* (pp. 7222–7232).

Bollobás, B. (1998). *Random graphs*. Springer.

Boucheron, S., Lugosi, G., & Pascal, M. (2016). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press.

Braca, P., Marano, S., & Matta, V. (2008). Enforcing consensus while monitoring the environment in wireless sensor networks. *IEEE Transactions on Signal Processing*, *56*(7), 3375–3380.

Bubeck, S., & Cesa-Bianchi, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Machine Learning*, *5*(1), 1–122.

Bullo, F., Cortés, J., & Martínez, S. (2009). *Distributed control of robotic networks*. Princeton University Press.

Cheung, M. Y., Leighton, J., & Hover, F. S. (2013). Autonomous mobile acoustic relay positioning as a multi-armed bandit with switching costs. In *IEEE/RSJ int. conf. intelligent robots & systems* (pp. 3368–3373).

Gai, Y., & Krishnamachari, B. (2014). Distributed stochastic online learning policies for opportunistic spectrum access. *IEEE Transactions on Signal Processing*, *62*(23), 6184–6193.

Garivier, A., & Moulines, E. (2008). On upper-confidence bound policies for non-stationary bandit problems. arXiv preprint arXiv:0805.3415.

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

Kolla, R. K., Jagannathan, K., & Gopalan, A. (2016). Stochastic bandits on a social network: Collaborative learning with local information sharing. CoRR, abs/1602.08886.

Krebs, J. R., Kacelnik, A., & Taylor, P. (1978). Test of optimal sampling by foraging great tits. *Nature*, *275*(5675), 27–31.

Lai, T. L., & Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, *6*(1), 4–22.

Landgren, P., Srivastava, V., & Leonard, N. E. (2016). On distributed cooperative decision-making in multiarmed bandits. In *European control conference* (pp. 243–248). Correction in arXiv:1512.06888v3 [cs.SY].

Landgren, P., Srivastava, V., & Leonard, N. E. (2016). Distributed cooperative decision-making in multiarmed bandits: Frequentist and bayesian algorithms. In *IEEE conf. decision and control* (pp. 167–172). Correction in arXiv:1606.00911v3 [cs.SY].

Landgren, P., Srivastava, V., & Leonard, N. E. (2018). Social imitation in cooperative multiarmed bandits: Partition-based algorithms with strictly local information. In *IEEE conf. decision and control* (pp. 5239–5244).

Liu, K., & Zhao, Q. (2010). Distributed learning in multi-armed bandit with multiple players. *IEEE Transactions on Signal Processing*, *58*(11), 5667–5681.

Madhushani, U., & Leonard, N. E. (2019). Heterogeneous stochastic interactions for multiple agents in a multi-armed bandit problem. In *European control conference* (pp. 3502–3507).

Madhushani, U., & Leonard, N. E. (2020). A dynamic observation strategy for multi-agent multi-armed bandit problem. In *European control conference* (pp. 1677–1683).

Marden, J. R., Young, H. P., & Pao, L. Y. (2014). Achieving pareto optimality through distributed learning. *SIAM Journal on Control and Optimization*, *52*(5), 2753–2770.

Martínez-Rubio, D., Kanade, V., & Rebeschini, P. (2019). Decentralized cooperative stochastic bandits. In *Advances in neural information processing systems* (pp. 4531–4542).

Olfati-Saber, R., & Murray, R. M. (2004). Consensus problems in networks of agents with switching topology and time-delays. *IEEE Transactions on Automatic Control*, *49*(9), 1520–1533.

Poulakakis, I., Young, G. F., Scardovi, L., & Leonard, N. E. (2015). Information centrality and ordering of nodes for accuracy in noisy decision-making networks. *IEEE Transactions on Automatic Control*, *61*(4), 1040–1045.

Savas, A., Srivastava, V., & Leonard, N. E. (2017). On distributed linear filtering with noisy communication. In *Am. control conf.* (pp. 2699–2704).

Shahrampour, S., Rakhlin, A., & Jadbabaie, A. (2017). Multi-armed bandits in multi-agent networks. In *2017 IEEE international conference on acoustics, speech and signal processing*.

Srivastava, V., Reverdy, P., & Leonard, N. E. (2013). On optimal foraging and multi-armed bandits. In *Allerton conference on communication, control, and computing* (pp. 494–499), Oct.

Srivastava, V., Reverdy, P., & Leonard, N. E. (2014). Surveillance in an abruptly changing world via multiarmed bandits. In *IEEE conf. decision and control* (pp. 692–697).

Stephenson, K., & Zelen, M. (1989). Rethinking centrality: Methods and examples. *Social Networks*, *11*(1), 1–37.

Wei, L., & Srivastava, V. (2018). On distributed multi-player multiarmed bandit problems in abruptly changing environment. In *IEEE conf. decision and control* (pp. 5783–5788).

Peter Landgren received the B.S. degree in Physics from Whitworth University in Spokane, WA, USA, in 2013, and the M.A. and Ph.D. degrees in Mechanical and Aerospace Engineering from Princeton University, Princeton, NJ, USA, in 2016 and 2018, respectively.

He is currently a Software Development Engineer at Amazon Prime Air in Seattle, WA. His research interests include modeling and analysis of collective behavior in biological and engineered systems, control for multiagent systems, and robotic systems.

Vaibhav Srivastava received the B.Tech. degree (2007) in mechanical engineering from the Indian Institute of Technology Bombay, Mumbai, India; the M.S. degree in mechanical engineering (2011), the M.A. degree in statistics (2012), and the Ph.D. degree in mechanical engineering (2012) from the University of California at Santa Barbara, Santa Barbara, CA. He served as a Lecturer and Associate Research Scholar with the Mechanical and Aerospace Engineering Department, Princeton University, Princeton, NJ from 2013–2016.

He is currently an Assistant Professor with the Electrical and Computer Engineering at Michigan State University. He is also affiliated with Mechanical Engineering, Cognitive Science Program, and Connected and Autonomous Networked Vehicles for Active Safety (CANVAS). His research focuses on Cyber Physical Human Systems with emphasis on mixed human–robot systems, networked multi-agent systems, aerial robotics, and connected and autonomous vehicles.

Naomi Ehrich Leonard received the B.S.E. degree in mechanical engineering from Princeton University, Princeton, NJ, USA, in 1985, and the M.S. and Ph.D. degrees in electrical engineering from the University of Maryland, College Park, MD, USA, in 1991 and 1994, respectively. From 1985 to 1989, she was an Engineer in the electric power industry.

She is currently the Edwin S. Wilsey Professor with the Department of Mechanical and Aerospace Engineering and the Director of the Council on Science and Technology at Princeton University. She is also an Associated Faculty of Princeton's Program in Applied and Computational Mathematics. Her research and teaching are in control and dynamical systems with current interests in networked multiagent systems, robotics, collective animal behavior, and social decision making.