Coordinated Versus Decentralized Exploration In Multi-Agent Multi-Armed Bandits

Mithun Chakraborty[1], **Kai Yee Phoebe Chua**[2], **Sanmay Das**[1], **Brendan Juba**[1]
[1]Washington University in St. Louis
[2]University of California, Irvine
[1]{mithunchakraborty,sanmay,bjuba}@wustl.edu; [2]kychua@uci.edu

Abstract

In this paper, we introduce a multi-agent multi-armed bandit-based model for ad hoc teamwork with expensive communication. The goal of the team is to maximize the total reward gained from pulling arms of a bandit over a number of epochs. In each epoch, each agent decides whether to pull an arm, or to broadcast the reward it obtained in the previous epoch to the team and forgo pulling an arm. These decisions must be made only on the basis of the agent's private information and the public information broadcast prior to that epoch. We first benchmark the achievable utility by analyzing an idealized version of this problem where a central authority has complete knowledge of rewards acquired from all arms in all epochs and uses a multiplicative weights update algorithm for allocating arms to agents. We then introduce an algorithm for the decentralized setting that uses a value-of-information based communication strategy and an exploration-exploitation strategy based on the centralized algorithm, and show experimentally that it converges rapidly to the performance of the centralized method.

1 Introduction

The past decade has seen an increased use of robotic and software agents; more companies and labs are creating their own agents that have different operating strategies and, in many cases, may need to work together as a team in order to achieve certain objectives. This world of increasing interdependence has motivated the research area of ad hoc teamwork [Stone and Kraus, 2010]: a setting in which agents need to cooperate without any pre-coordination and work toward a common goal [Stone *et al.*, 2010]. Standard approaches to teamwork, e.g. SharedPlans [Grosz and Kraus, 1996], STEAM [Tambe, 1997], or GPGP [Decker and Lesser, 1995], rely on common agreements about strategies and communication standards, or other shared assumptions. However, in *ad hoc* teamwork, teammates should be able to leverage each others' knowledge without explicitly relying on the strategy used to generate that knowledge, or assumptions about how others will operate in the future. This is a grand challenge for the state of the art in multi-agent systems, but the multi-armed bandit (MAB) domain has emerged in the last few years as the standard approach to start thinking about it [Barrett *et al.*, 2014].

In a multi-agent multi-armed bandit problem, a team of agents (e.g. a swarm of nanorobots performing a complex set of tasks or of drones patrolling a large area, etc.) is playing a MAB. The question that makes this problem interesting beyond the intrinsic exploration / exploitation tradeoff feature of the single-agent version is the role of information sharing. Although various solutions to classical MAB problems are well known [Gittins and Jones, 1979; Gittins, 1979; Auer *et al.*, 2002a; Auer *et al.*, 1995; Lai and Robbins, 1985] and the agents may individually play such a solution to converge on a good strategy, it is intuitively clear that by *sharing information*, e.g. about their observed payoffs in past rounds, they can approach a good strategy much faster. But, in general, information sharing comes at a cost that can take many forms: for example, a fixed penalty term or fractional reduction applied to the immediate reward gathered from a pull, or the preclusion of pulling an arm while while an agent transmits a message to its teammates. In this paper, we devise and evaluate a solution scheme for the last scenario in the above list: if an agent communicates, it cannot simultaneously access the bandit, thus incurring an *opportunity cost*. However, the scheme generalizes to other definitions of cost.

We first define a new multi-agent MAB model, in order to capture the three-way tradeoff between exploration, exploitation, and communication. We benchmark the performance that can be achieved in a centralized version of this problem (with Gaussian rewards) in which a controller with knowledge of each agent's choices and rewards can decide which agents to allocate to pulling which arm without suffering any communication cost. Intuitively, one would expect that under such circumstances, it should be possible to achieve a total reward that is close to what is attainable in the full-information *experts learning* or *forecasting* variant of the problem [Vovk, 1990; Foster, 1991; Foster and Vohra, 1993; Littlestone and Warmuth, 1994]. We establish that a multi-agent strategy with a centralized coordinator (with costless communication), which we call the *public agent*, can indeed obtain performance similar to these single-agent solutions that use full information. Our method is based on multiplicative weight updates, and we prove a regret bound for this strategy using a technique due to Freund and Schapire [1999].

The problem that we actually want to solve involves a *decentralized* (albeit cooperative) multi-agent system; but the insights we gain from the analysis of the idealized version with a central controller turn out to be useful in designing effective protocols for how these agents communicate and utilize publicly available information. We let the team maintain a shared base of all historically broadcast information that acts as a proxy for the public agent in the following sense: Whenever an agent decides which arm to pull, it does so based on a combination of its private information with that in this shared base; whenever an agent decides whether or not to broadcast and hence has to reason about the long-term impact of its immediate future action on the performance of the team, it approximates the team's future behavior by that of an imaginary controller having access to the shared information base and allocating arms as in the centralized version.

More precisely, the exploration-exploitation strategy in our decentralized algorithm is a variant of the *softmax* approach, which is known to perform well empirically for the single-agent problem [Vermorel and Mohri, 2005]: The agent chooses an arm according to a distribution of probabilities proportional to the exponentials of the empirical rewards obtained from the arms, weighted appropriately. In our multi-agent extension, each agent chooses an arm using a similar weight distribution whose parameters are informed by the algorithm we developed for the central benchmark. To summarize our procedure for deciding when to broadcast, which we call the *"Value of Information" (VoI)* communication strategy, an agent optimistically estimates the total reward that could possibly result from broadcasting its latest observation, which is then compared to the estimated reward from pulling an arm. If the latter is higher, then the agent naturally abstains from broadcasting in the next round; otherwise, it communicates with a probability chosen so that in expectation approximately one of the agents that pulled that arm is communicating in a given round (specifically, a probability inversely proportional to the expected number of agents in the population that pulled the arm, given the current empirical estimates). Finally, we show experimentally that our decentralized algorithm for the multi-agent MAB achieves performance close to that of the centralized algorithm, thus efficiently solving the exploration / exploitation / communication "trilemma".

2 Related Work

Stone and Kraus [2010] were the first to use a MAB model for ad hoc teamwork (a simple two-agent cooperative setting), but Barrett *et al.* [2014] used a multi-agent MAB (in particular, a two-armed bandit with Bernoulli payoff distributions) to formalize ad hoc teamwork with *costly communication* for the first time: They focused on designing a *single* ad hoc agent that can learn optimal strategies when playing with teammates who have specified strategies, in a setting where each round consists of a *communication* phase (broadcasting a message with an associated cost function), and an *action* phase (pulling an arm to extract a reward). More recent work has focused on designing adaptive ad hoc agents that can observe (previously unseen) teammates during operation but cannot exchange messages with them (e.g. [Albrecht and Ramamoorthy, 2013; Barrett and Stone, 2015; Hernandez-Leal *et al.*, 2016] and references therein). We, on the other hand, construct a common communication protocol for every agent in a decentralized team where information sharing is feasible but optional and, if executed, precludes action. This also differentiates our contribution from the distributed stochastic bandit algorithm [Szörényi *et al.*, 2013] for a P2P network where each pull of an arm is necessarily followed by message passing between peers.

Some previous work has modeled communication cost in a collaborative multi-agent system as an extraneous quantity, equal or proportional to the number "of messages sent in the system", that is traded off against regret (in the reward collected from the bandit), in both (distributed) experts [Kanade *et al.*, 2012] and MAB [Hillel *et al.*, 2013; Buccapatnam *et al.*, 2015] settings; in our model, information sharing interferes more intimately with reward collection. Distributed multi-agent MABs have also been employed to model cognitive radio networks [Liu and Zhao, 2010; Kalathil *et al.*, 2012; Tossou and Dimitrakakis, 2015] where collisions (multiple agents pulling the same arm in the same round) are costly. In our model, multiple agents pulling the same arm all receive the same reward for that round. A fundamentally different application of the MAB model to multi-agent systems is the use of MAB policies by agents to select *other agents* for specific services within a service sharing system [Vallée *et al.*, 2014]. Both theoretical and experimental studies of *social learning* or *imitation* in the social and cognitive sciences have involved multi-agent MAB models, where agents access and utilize one another's historical information [Schlag, 1998; Biele *et al.*, 2009; Rendell *et al.*, 2010]. The main difference with our model is that their agents are selfishly motivated whereas ours work towards the shared goal of maximizing the collective reward. Moreover, in these models, no cost is incurred by an agent for *transmitting* information but may be sustained in *acquiring* it; e.g., if an agent is receiving (perhaps noisy) information about another's action and reward, it cannot pull an arm in that epoch.

3 Formal Problem Description

Our model follows the basic definitions of a classical bandit problem: We have a set of n arms such that, in any epoch t over a pre-specified time-horizon of length T, arm i generates a random reward $r_{i,t}$ independently (across arms and epochs) from a time-invariant Gaussian distribution: $r_{i,t} \sim \mathcal{N}(\mu_i, \sigma^2)\ \forall i, t$. We assume that all arms have the same known standard deviation $\sigma > 0$ but unknown means $\{\mu_i\}_{i=1}^n$, where $\mu_i \neq \mu_j$ for at least one pair (i, j), and that the maximum and minimum possible values, $\mu_{\max} > \mu_{\min} > 0$, of these mean rewards are also known *a priori*.

There are $m > n$ agents in our team: In epoch t, each agent j must decide without any knowledge of the others' simultaneous decisions whether to broadcast a message consisting of the index of the arm it pulled and the reward it thus gained in epoch $(t-1)$. If an agent chooses to broadcast at t, then it loses the chance to pull any arm and hence collect any reward during t – this can be viewed as the cost of communication – but its message becomes available to the entire team for use

in decision-making from epoch $(t+1)$ onwards. However if an agent decides not to broadcast at t, it pulls an arm and gets a reward. If multiple agents pull the same arm i in epoch t, each receives the same reward $r_{i,t}$; the fact that an arm generates the same reward regardless of how many times it is pulled in an epoch removes any *learning* benefit from an arm being pulled by more than one agent at any t.

Thus, if $m_{i,t}$ agents pull arm i in epoch t, then $\sum_{i=1}^{n} m_{i,t} \leq m$ in general, and the total reward amassed by the team in this epoch is $\sum_{i=1}^{n} m_{i,t} r_{i,t}$. Every agent's goal is to maximize the team's cumulative total reward over T epochs, i.e. $\sum_{t=1}^{T}\sum_{i=1}^{n} m_{i,t} r_{i,t}$. This is why broadcasting can be beneficial in the long run: By sacrificing immediate gain, an agent enriches the shared pool of knowledge about the unknown parameters, leading to savings in exploration time for the team as a whole. However, each agent now has to resolve a two-stage dilemma: **[Stage 1** (*Communication vs Reward Collection*)**]** Should it broadcast its observation from the previous epoch? **[Stage 2** (*Exploration vs Exploitation*)**]** If it decides not to broadcast, which arm should it pull now?

Before presenting our strategy for handling the above issues in Section 5, we describe in Section 4 an idealized version of our problem in which a central authority that we call the *public agent* always has complete knowledge of rewards generated by all arms, and uses that to allocate arms to agents that do not make individual decisions. We then propose and analyze a multiplicative weights update algorithm to solve this exploration-exploitation problem with instantaneous costless communication. This framework serves a dual purpose: It offers insights that we utilize in the design of our solution scheme for the decentralized problem, and also provides a gold standard for evaluating that scheme.

4 Ideal Centralized Multi-Agent MAB

The public agent maintains a normalized weight (in other words, probability) distribution across the n arms, denoted by $\mathbf{P}_t = (P_{1,t}, P_{2,t}, \cdots, P_{n,t})$ where $P_{i,t} \geq 0\ \forall i$, $\sum_{i=1}^{n} P_{i,t} = 1$, and assigns $mP_{i,t}$ agents to arm i in epoch t.[1] The starting distribution is uniform: $P_{i,1} = 1/n\ \forall i$. During t, the public agent observes the sample reward $r_{i,t}$ generated by each arm i, and hence updates the weight distribution to $\mathbf{P}_{t+1}$ at the beginning of the next epoch using the following *multiplicative weights update* (MWU) approach [Littlestone and Warmuth, 1994; Freund and Schapire, 1997; Freund and Schapire, 1999]:

$$P_{i,t+1} = P_{i,t}\beta^{-\frac{r_{i,t}+\lambda}{\kappa}}/Z_t, \tag{1}$$

where $\beta \in (0,1)$, $\lambda \in \mathbb{R}$, $\kappa > 0$, $Z_t = \sum_{\iota=1}^{n} P_{\iota,t}\beta^{-\frac{r_{\iota,t}+\lambda}{\kappa}}$.

Ideally, the public agent would like to maximize the cumulative reward of the team over a given time-horizon T, i.e. $\sum_{t=1}^{T}\sum_{i=1}^{n} r_{i,t} mP_{i,t}$, or equivalently the time-averaged per-agent cumulative reward $\frac{1}{T}\sum_{t=1}^{T} r_{i,t}P_{i,t}$. We define the *regret* of the centralized strategy with updates (1) as

$$\mathcal{R}_{\text{central}}(T) = \max_{\mathbf{P}}\left[\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{n} r_{i,t}P_i\right] - \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{n} r_{i,t}P_{i,t},$$

where $\mathbf{P} = (P_1, P_2, \cdots, P_n)$ is a probability distribution. Theorem 1 shows that the regret of this centralized MWU method becomes vanishingly small for a large enough T.

Theorem 1 *Suppose, a bandit has n arms producing Gaussian rewards with the same known standard deviation σ, and unknown means $\{\mu_i\}_{i=1}^{n}$ with a known range $R_\mu \triangleq \mu_{\max} - \mu_{\min} > 0$. For any horizon $T \in \mathbb{Z}^+$ and an arbitrarily small number δ, $0 < \delta < \min\{2nT\Phi(R_\mu/2\sigma), 1\}$, if we use a centralized MWU strategy with a uniform initial weight distribution and the update rule* (1) *with parameters*

$$\beta = 1/\Big(1+\sqrt{\tfrac{2\ln(n)}{T}}\Big), \quad \lambda = \sigma\Phi^{-1}(\delta/(2nT)) - \mu_{\max},$$
$$\kappa = R_\mu - 2\sigma\Phi^{-1}(\delta/(2nT)),$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function, then with probability at least $(1-\delta)$,

$$\mathcal{R}_{\textit{central}}(T) = O\Big((R_\mu+\sigma)\sqrt{\ln(\tfrac{nT}{\delta})\tfrac{\ln(n)}{T}}\Big).$$

We outline the proof of the theorem below, and defer the complete proof to a full version of the paper. But first, we note that the above MWU algorithm is equivalent to a **decreasing SOFTMAX** strategy [Vermorel and Mohri, 2005] over empirical means $\hat{\mu}_{i,t} = \sum_{s=1}^{t} r_{i,s}/t$ with temperature $\tau_t = \tau_0/t$,

$$\tau_0 = [R_\mu - 2\sigma\Phi^{-1}(\delta/(2nT))]/\ln(1+\sqrt{2\ln(n)/T}). \tag{2}$$

Thus, we can rewrite the weight distribution in (1) as

$$\mathbf{P}_t = \textsc{Softmax}(\{\hat{\mu}_{i,t}\}_{i=1}^{n}, \tau_t).$$

Proof sketch. We rewrite the multiplicative factor in (1) as $\beta^{\hat{L}_{i,t}}$ where we define $\hat{L}_{i,t} \triangleq (-r_{i,t}-\lambda)/\kappa$ as the normalized loss. We now recall the amortized analysis that Freund and Schapire [1999] used to prove that their MWU algorithm for repeated game playing is no-regret when the player's loss for any action lies in $[0,1]$, and adapt it to the case of Gaussian losses. We have chosen the shifting and rescaling parameters λ and κ so that the following inequality holds:

$$\Pr(\beta^{\hat{L}_{i,t}} \leq 1-(1-\beta)\hat{L}_{i,t}) \geq 1-\delta/(nT) \quad \forall i,t. \tag{3}$$

Our potential function is $\mathrm{RE}(\tilde{\mathbf{P}}||\mathbf{P}_t)$, the Kullback-Leibler divergence of the current probability distribution $\mathbf{P}_t$ from an arbitrary fixed distribution $\widetilde{\mathbf{P}}$; we further define $\Delta\widetilde{\mathrm{RE}}_{t,t'} \triangleq \mathrm{RE}(\tilde{\mathbf{P}}||\mathbf{P}_{t'}) - \mathrm{RE}(\tilde{\mathbf{P}}||\mathbf{P}_t)$. It is easy to see that $\Delta\widetilde{\mathrm{RE}}_{1,T+1} \geq -\ln(n)$ since $\mathbf{P}_1$ is a uniform distribution; combining the analysis of Freund and Schapire [1999] with a union bound of probabilities (over arms and epochs) applied to (3), we can further deduce a high-probability upper bound on $\Delta\widetilde{\mathrm{RE}}_{1,T+1} = \sum_{t=1}^{T}\Delta\widetilde{\mathrm{RE}}_{t,t+1}$. Using these upper and lower bounds, we can establish that, with probability at least $(1-\delta)$, $\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{n} r_{i,t}\widetilde{P}_i - \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{n} r_{i,t}P_{i,t} \leq \Delta$, where

[1] In an actual implementation, if $mP_{i,t}$ is fractional, then $\lfloor mP_{i,t}\rfloor$ agents are initially assigned to arm i, and then all the remaining $(m - \sum_{i=1}^{n}\lfloor mP_{i,t}\rfloor)$ are optimistically allocated to the arm with the current highest empirical mean.

$\Delta \triangleq \kappa\left(\sqrt{2\ln(n)/T} + \ln(n)/T\right)$. Finally, using the Gaussian tail inequality $\Phi(-a) \leq 0.5e^{-a^2/2}$ for any $a > 0$, we can show that $\kappa \leq R_\mu + 2\sigma\sqrt{2\ln(\frac{nT}{\delta})}$, and hence, after some algebra, that $\Delta < 2\sqrt{2}\left(\sqrt{2}+1\right)(R_\mu+\sigma)\sqrt{\ln(\frac{nT}{\delta})\frac{\ln(n)}{T}}$ for $T \geq \max(2, \ln(n))$. This completes the proof. □

5 Decentralized MAB With Value-of-Information Communication Strategy

The decentralized problem takes away the coordinator, introducing communication as a costly option; we design a scheme for how each agent decides whether to broadcast or pull an arm such that the team's overall behavior mimics that of the centralized version (Section 4) as closely as possible. To this end, we employ a device which we will call the *public agent* for a decentralized MAB but which is, in fact, an identical representation held by each agent of all the information publicly communicated until the current epoch: a shared table containing two entries for each arm i, $\nu^{cum}_{i,t} = |T^{cum}_{i,t}|$, where $T^{cum}_{i,t}$ is the set of epochs at each of which information on arm i was broadcast by at least one agent until (but excluding) epoch t, and the cumulative reward $r^{cum}_{i,t} = \sum_{t \in T^{cum}_{i,t}} r_{i,t}$. The public agent's empirical mean for arm i during epoch t is $\hat{\mu}^{\text{public}}_{i,t} = r^{cum}_{i,t}/\nu^{cum}_{i,t}$. For any agent j, any epoch is either an *action* round, when it pulls an arm, or a *broadcast* round, when it sends a message to the team, the starting epoch being necessarily an action round for every agent. Agent j has a private table with four entries for each arm i: $\nu^a_{i,j,t} = |T^a_{i,j,t}|$, $r^a_{i,j,t} = \sum_{t \in T^a_{i,j,t}} r_{i,t}$, $\nu^b_{i,j,t} = |T^b_{i,j,t}|$, $r^b_{i,j,t} = \sum_{t \in T^b_{i,j,t}} r_{i,t}$ where $T^a_{i,j,t}$ is the set of epochs in which agent j has pulled arm i, and $T^b_{i,j,t} \subset T^a_{i,j,t}$ is the subset of these pulls that it has communicated until (but excluding) epoch t.

Exploration-exploitation with softmax strategy. If epoch t is an action round for agent j, then at the beginning of this epoch, this agent combines its private table with the public agent's information set to produce its own vector of *empirical means* across arms $\{\hat{\mu}_{i,j,t}\}_{i=1}^n$: $\hat{\mu}_{i,j,t} = \hat{r}_{i,j,t}/\hat{\nu}_{i,j,t}$ where $\hat{r}_{i,j,t} = r^a_{i,j,t} - r^b_{i,j,t} + r^{cum}_{i,t}$ and $\hat{\nu}_{i,j,t} = \nu^a_{i,j,t} - \nu^b_{i,j,t} + \nu^{cum}_{i,t}$.[2] It then applies a softmax function to these means with the decreasing temperature parameter $\tau_t = \tau_0/t$, τ_0 defined in (2) in Section 4, to generate a probability distribution over arms, and draws an arm according to this distribution, say $i^* = i_{j,t}$. The reward $r^* = r_{i^*,t}$ thus collected is added to the team's cumulative reward; agent j updates its private table entries: $r^a_{i^*,j,t+1} = r^a_{i^*,j,t} + r^*$, $\nu^a_{i^*,j,t+1} = \nu^a_{i^*,j,t} + 1$; $r^a_{i,j,t+1} = r^a_{i,j,t}$, $\nu^a_{i,j,t+1} = \nu^a_{i,j,t}$ $\forall i \neq i^*$.

Value-of-Information (VoI) communication criterion. At the end of each action round, say epoch t, agent j follows a three-step (in general) procedure, depicted as flow-chart in Figure 1, to decide whether or not broadcast information on the arm i^* it just pulled in the next epoch $(t+1)$.

First, it checks if the reward r^* from this pull is greater than the public agent's empirical mean $\hat{\mu}^{\text{public}}_{i^*,t}$; if yes (resp. no), then it uses an upper (resp. a lower) confidence bound on the mean reward of the arm i^* under consideration and a lower (resp. an upper) bound on that of every other arm to generate a vector of *working estimates* $\{\widetilde{\mu}_{i,j,t}\}_{i=1}^n$ across arms for further comparison purposes: An upper (resp. lower) bound is obtained by adding to (resp. subtracting from) the updated empirical mean $\hat{\mu}_{i,j,t+1}$ the quantity $\sigma\sqrt{2\ln(\frac{nT}{2\varepsilon_{voi}})/\hat{\nu}_{i,j,t+1}}$, where $\varepsilon_{voi} \in (0,1)$ is a free (error) parameter.

In the second step, agent j uses the public agent as a proxy for the team's collective behavior to compare the team's estimated expected cumulative reward over the remainder of the horizon in two mutually exclusive and exhaustive scenarios: one in which epoch $(t+1)$ is an action round, say $\Lambda^a_{j,t}$, and the other in which it is a broadcast round, say $\Lambda^b_{j,t}$. For computing $\Lambda^a_{j,t}$, agent j acts as if the public agent will receive no further communication, and will allocate arms to agents in all epochs starting at $(t+1)$ using the weight distribution $\mathbf{w^a_t} = \{w^a_{i,t}\}_{i=1}^n$, where the temperature τ_t comes from (2):

$$\mathbf{w^a_t} = \textsc{Softmax}(\{\hat{\mu}^{\text{public}}_{i,t}\}_{i=1}^n, \tau_t). \tag{4}$$

For evaluating $\Lambda^b_{j,t}$, agent j assumes that the public agent uses the weight distribution $\mathbf{w^a_t}$ to allocate arms to the remaining $(m-1)$ agents during epoch $(t+1)$, after which it will augment its information set with *only* agent j's broadcast message (i^*, r^*) to update its weight distribution to $\mathbf{w^b_{t+1}} = \{w^b_{i,t+1}\}_{i=1}^n$, and use this distribution henceforth.

$$\hat{\mu}^b_{i^*,t+1} = \tfrac{r^{cum}_{i^*,t} + r^*}{\nu^{cum}_{i^*,t}+1}, \qquad \hat{\mu}^b_{i,t+1} = \hat{\mu}^{\text{public}}_{i,t} \quad \forall i \neq i^*;$$
$$\mathbf{w^b_{t+1}} = \textsc{Softmax}(\{\hat{\mu}^b_{i,t+1}\}_{i=1}^n, \tau_t). \tag{5}$$

The algebraic expressions for $\Lambda^a_{j,t}$ and $\Lambda^b_{j,t}$ are provided in Figure 1. Notice that the computation of $\Lambda^b_{j,t}$ is the only time in our scheme when the exact nature of the communication cost comes into play. For other cost types mentioned in our introduction, we will just have (an) additional term(s) to account for it in the above summation; the rest of the scheme remains the same as delineated in this paper.

Let us define the agent's current *Value of Information* $\mathbf{VoI} \triangleq \Lambda^b_{j,t} - \Lambda^a_{j,t}$, which estimates the long-term benefit accrued by the team if the agent under consideration forgoes immediate reward collection to share its latest information. The idea of using the value of information for decision making implicitly goes back a long way; the explicit modern formulation goes back at least to Howard [1966]. Recently, it has been used in related AI contexts by Chajewska *et al.* [2000] and Boutilier [2002], among others.

Agent j decides to not broadcast in epoch $(t+1)$ if $\mathbf{VoI} \leq 0$; if $\mathbf{VoI} > 0$, it uses what we will call the *simple communication criterion* in the final step of its decision-making procedure: It estimates the number of agents $\hat{m}_{i^*}$ that have pulled the arm i^* in epoch t as the product of m and the public agent's current weight $w^a_{i^*,t}$ on the arm; if $\hat{m}_{i^*}$ is

[2] Each agent's private table is so initialized that the initial empirical mean for each arm is $(\mu_{\min} + \mu_{\max})/2$.

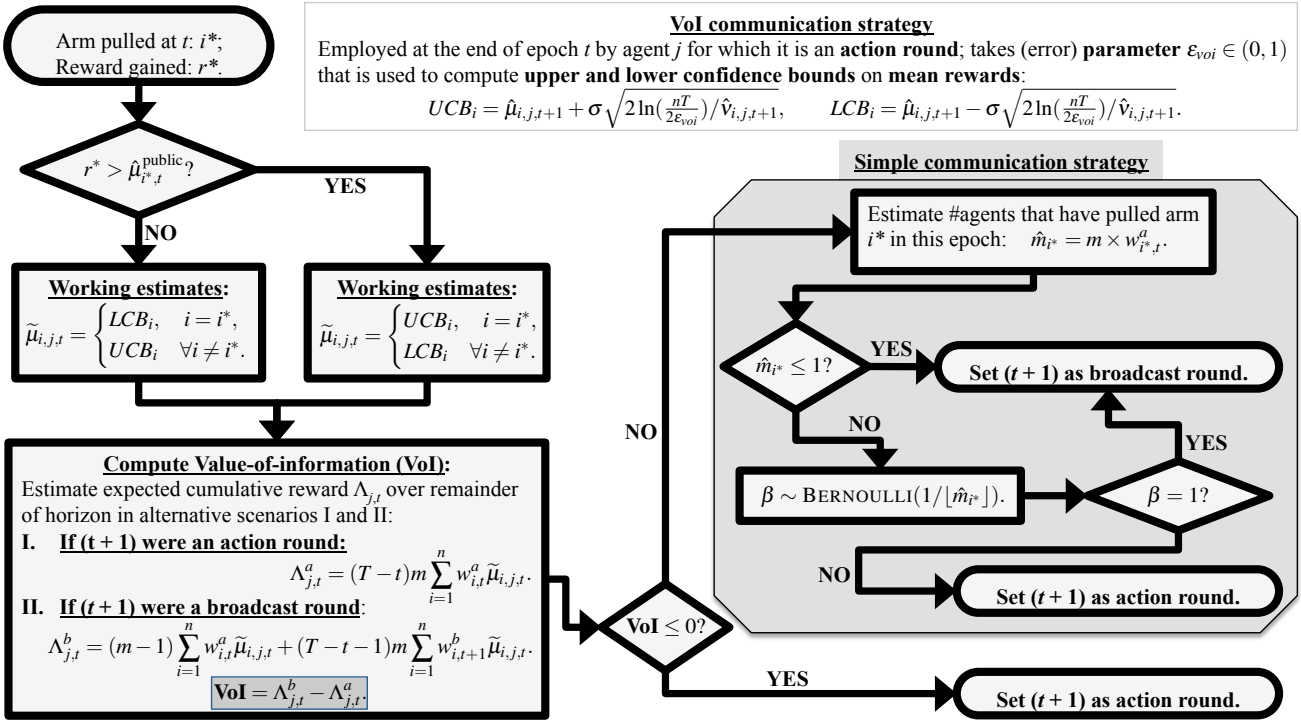

Figure 1: Flow-chart showing the steps of the *VoI* and *simple communication strategies* for an arbitrary agent j at the end of an action round, as described in Section 5. $\hat{\mu}^{\text{public}}_{i,t}$ and $\hat{\mu}_{i,j,t+1}$ denote the public agent and agent j's empirical mean reward values for arm i after agent j has pulled an arm in epoch t; $\{w^a_{i,t}\}_{i=1}^n$ and $\{w^a_{i,t+1}b\}_{i=1}^n$ are defined in (4) and (5) respectively.

one or less, agent j decides to broadcast deterministically at $(t+1)$, otherwise it broadcasts at $(t+1)$ on the success of a Bernoulli trial with success probability $1/\lfloor \hat{m}_{i^*} \rfloor$. The idea is to keep the expected number of broadcasts per arm per epoch at 1 since excess simultaneous broadcasts corresponding to the same arm convey redundant information and only impede reward collection. If (and only if) agent j decides that epoch $(t+1)$ is a broadcast round, the entries $r^b_{i^*,j,t}$ and $\nu^b_{i^*,j,t}$ in its private table are incremented by r^* and 1 respectively.

Message broadcasting. If epoch t is a broadcast round for agent j, it publicly sends out the message (i^*, r^*) where $i^* = i_{j,t-1}$ and $r^* = r_{i^*,t-1}$, pulls no arm at t but sets epoch $(t+1)$ as an action round; before epoch $(t+1)$ commences, the public agent is augmented with messages transmitted by all broadcasting team members in epoch t, discarding duplicates.

6 Experimental Evaluation

In this section, we describe two sets of experiments we ran to compare the performance of the decentralized multi-agent MAB exploration-exploitation algorithm with VoI communication strategy that we proposed in Section 5 with several benchmarks described below. In these two sets, we studied the variation of the regret of each algorithm over the number of arms and over different lengths of the time-horizon (keeping the other variable fixed) respectively – we report the corresponding results in Figures 2 (a) and (b).

Our main benchmark for both sets is the centralized softmax / MWU strategy, detailed in Section 4, which gives us a lower bound on the regret achievable by any decentralized scheme. Additionally, for the first set of experiments, we used two other benchmarks – agents exploring-and-exploiting the bandit arms independently (i.e. with no communication) all using one of two standard approaches – EXP3 [Auer *et al.*, 2002b] and UCB1-Normal [Auer *et al.*, 2002a] – to demonstrate that regret can be lowered drastically by allowing agents to engage in broadcasting, even if the latter is expensive. Finally, for both sets, we also ran experiments where agents made their broadcasting decisions using only the simple communication criterion described in Section 5 and demarcated in Figure 1 (skipping the first two stages of VoI) in order to show the improvement, if any, that can be achieved by incorporating the value of information (i.e. the difference $\Lambda^b_{j,t} - \Lambda^a_{j,t}$) in one's decision-making process.

For each experiment, the number of agents is set at $m = 25n$, n being the number of arms The means of the Gaussian reward distributions on the bandit arms form a decreasing arithmetic sequence starting at $\mu_{\max} = \mu_1 = 1$ and ending at $\mu_{\min} = \mu_n = 0.05$, so that the magnitude of the common difference is $\Theta(\frac{1}{n})$; the shared standard deviation $\sigma = 0.1$ is independent of the number of arms.

The per-agent time-averaged regret, plotted on the vertical axis, is defined as the difference between the total reward accumulated by the team over the time-horizon, divided by the number of agents and the number of epochs in the hori-

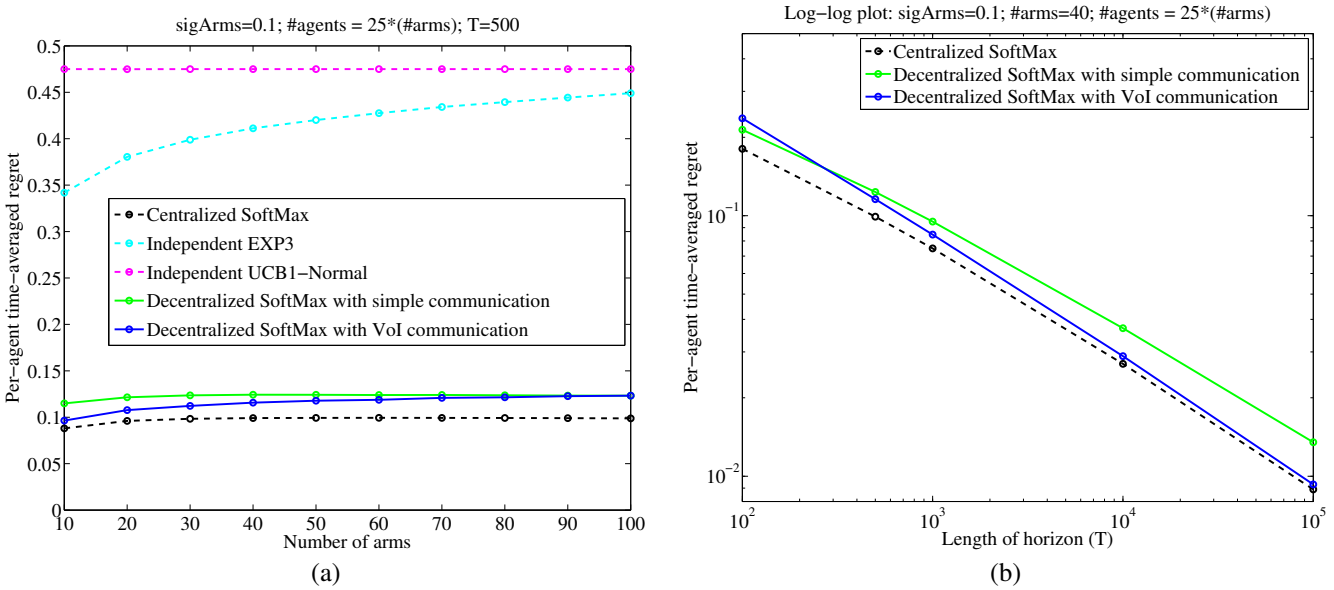

Figure 2: Error bars, being small, are omitted. (a) Regret vs. number of arms for fixed "intermediate" time-horizon $T = 500$: UCB1-Normal is still in its initial exploration phase (see Auer *et al.* [2002a] for details). Both communication strategies offer significant improvements over independent reward-collection schemes, VoI outperforming simple for a lower number of arms. (b) Regret vs. length of horizon for fixed number of arms $n = 40$, both axes using logarithmic scales. VoI is slightly worse than simple for smaller $T \sim 10^2$ presumably because the former results in relatively fewer broadcasts preventing agents from utilizing others' information over short horizons; however, it overtakes simple for longer horizons, performing on a par with the centralized strategy for a large enough T ($\sim 10^4$ and higher).

zon, and $\mu_1 = 1$ (this regret concept is stronger than that in Section 4). Each data-point is generated by averaging the regret values over $N_{sim} = 10^5$ repetitions. We set $\delta = 0.01$, $\varepsilon_{voi} = \frac{0.05}{N_{sim}}$ to ensure that our confidence bounds hold for all experiments. Figure 2 provides strong empirical evidence that, for a range of values of n and T, the VoI strategy enables a decentralized team with a sufficient number of members to achieve performance close to that with a central controller.

7 Conclusion

We formulated a novel model for the problem of reward collection by a team from multiple (stochastic) sources with costly communication, by extending the classic multi-armed bandit model to a multi-agent setting. We introduced an algorithm (decentralized softmax with VoI communication strategy) for achieving an exploration / exploitation / communication trade-off in this model. To benchmark the performance of this algorithm, we designed a centralized algorithm and proved its no-regret property. Finally, we demonstrated empirically that the performance of our decentralized algorithm, measured in terms of regret, approaches that of the centralized method. Directions of future work include evaluating our strategy under other communication cost structures, considering information acquisition costs (see, e.g., Rendell *et al.* [2010]), and analytically deriving the rate of convergence of our decentralized approach to the centralized benchmark.

Acknowledgments

Chakraborty and Das are grateful for support from NSF award 1560191. Juba is supported by an AFOSR Young Investigator Award.

References

[Albrecht and Ramamoorthy, 2013] Stefano V. Albrecht and Subramanian Ramamoorthy. A game-theoretic model and best-response learning method for ad hoc coordination in multiagent systems. In *Proc. AAMAS*, pages 1155–1156, 2013.

[Auer *et al.*, 1995] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proc. FOCS*, volume 36, pages 322–331. IEEE Computer Society Press, 1995.

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

[Auer *et al.*, 2002b] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

[Barrett and Stone, 2015] Samuel Barrett and Peter Stone. Cooperating with unknown teammates in complex domains: A robot soccer case study of ad hoc teamwork. In *Proc. AAAI*, pages 2010–2016, 2015.

[Barrett *et al.*, 2014] Samuel Barrett, Noa Agmon, Noam Hazon, Sarit Kraus, and Peter Stone. Communicating with unknown teammates. In *Proc. ECAI*, 2014.

[Biele *et al.*, 2009] Guido Biele, Jörg Rieskamp, and Richard Gonzalez. Computational models for the com-

bination of advice and individual learning. *Cognitive science*, 33(2):206–242, 2009.

[Boutilier, 2002] Craig Boutilier. A POMDP formulation of preference elicitation problems. In *Proc. AAAI/IAAI*, pages 239–246, 2002.

[Buccapatnam *et al.*, 2015] Swapna Buccapatnam, Jian Tan, and Li Zhang. Information sharing in distributed stochastic bandits. In *Proc. INFOCOM*, pages 2605–2613. IEEE, 2015.

[Chajewska *et al.*, 2000] Urszula Chajewska, Daphne Koller, and Ronald Parr. Making rational decisions using adaptive utility elicitation. In *Proc. AAAI/IAAI*, pages 363–369, 2000.

[Decker and Lesser, 1995] Keith Decker and Victor R. Lesser. Designing a family of coordination algorithms. In *Proc. ICMAS-95*, pages 73–80, 1995.

[Foster and Vohra, 1993] Dean P. Foster and Rakesh V. Vohra. A randomized rule for selecting forecasts. *Operations Research*, 41:704–709, 1993.

[Foster, 1991] Dean P. Foster. Prediction in the worst case. *Ann. Statistics*, 19:1084–1090, 1991.

[Freund and Schapire, 1997] Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comp. and Sys. Sciences*, 55(1):119–139, 1997.

[Freund and Schapire, 1999] Yoav Freund and Robert E. Schapire. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29(1):79–103, 1999.

[Gittins and Jones, 1979] John C. Gittins and David M. Jones. A dynamic allocation index for the discounted multiarmed bandit problem. *Biometrika*, 66(3):561–565, 1979.

[Gittins, 1979] John C. Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2):148–177, 1979.

[Grosz and Kraus, 1996] Barbara J. Grosz and Sarit Kraus. Collaborative plans for complex group actions. *AIJ*, 86:269–358, 1996.

[Hernandez-Leal *et al.*, 2016] Pablo Hernandez-Leal, Matthew E. Taylor, Benjamin Rosman, Luis Enrique Sucar, and Enrique Munoz de Cote. Identifying and tracking switching, non-stationary opponents: A Bayesian approach. In *Proc. AAMAS*, pages 1315–1316, 2016.

[Hillel *et al.*, 2013] Eshcar Hillel, Zohar S. Karnin, Tomer Koren, Ronny Lempel, and Oren Somekh. Distributed exploration in multi-armed bandits. In *Proc. NIPS*, pages 854–862, 2013.

[Howard, 1966] Ronald A. Howard. Information value theory. *IEEE Trans. Sys. Sci. and Cybernetics*, 2(1):22–26, 1966.

[Kalathil *et al.*, 2012] Dileep Kalathil, Naumaan Nayyar, and Rahul Jain. Multi-player multi-armed bandits: Decentralized learning with IID rewards. In *Proc. Annual Allerton Conference on Communication, Control, and Computing*, pages 853–860. IEEE, 2012.

[Kanade *et al.*, 2012] Varun Kanade, Zhenming Liu, and Bozidar Radunovic. Distributed non-stochastic experts. In *Proc. NIPS*, pages 260–268, 2012.

[Lai and Robbins, 1985] T. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

[Littlestone and Warmuth, 1994] Nick Littlestone and Manfred K Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, 1994.

[Liu and Zhao, 2010] Keqin Liu and Qing Zhao. Distributed learning in multi-armed bandit with multiple players. *IEEE Transactions on Signal Processing*, 58(11):5667–5681, 2010.

[Rendell *et al.*, 2010] Luke Rendell, Robert Boyd, Daniel Cownden, Marquist Enquist, Kimmo Eriksson, Marc W Feldman, Laurel Fogarty, Stefano Ghirlanda, Timothy Lillicrap, and Kevin N Laland. Why copy others? insights from the social learning strategies tournament. *Science*, 328(5975):208–213, 2010.

[Schlag, 1998] Karl H Schlag. Why imitate, and if so, how?: A boundedly rational approach to multi-armed bandits. *J. Econ. Theory*, 78(1):130–156, 1998.

[Stone and Kraus, 2010] Peter Stone and Sarit Kraus. To teach or not to teach? decision making under uncertainty in ad hoc teams. In *Proc. AAMAS*, 2010.

[Stone *et al.*, 2010] Peter Stone, Gal. A. Kaminka, Sarit Kraus, and Jeffrey S. Rosenschein. Ad hoc autonomous agent teams: Collaboration without pre-coordination. In *Proc. AAAI*, 2010.

[Szörényi *et al.*, 2013] Balázs Szörényi, Róbert Busa-Fekete, István Hegedüs, Róbert Ormándi, Márk Jelasity, and Balázs Kégl. Gossip-based distributed stochastic bandit algorithms. In *Proc. ICML*, pages 19–27, 2013.

[Tambe, 1997] Milind Tambe. Towards flexible teamwork. *JAIR*, 7:81–124, 1997.

[Tossou and Dimitrakakis, 2015] Aristide C. Y. Tossou and Christos Dimitrakakis. Differentially private, multi-agent multi-armed bandits. In *European Workshop on Reinforcement Learning (EWRL)*, 2015. (presentation only).

[Vallée *et al.*, 2014] Thibaut Vallée, Grégory Bonnet, and François Bourdon. Multi-armed bandit policies for reputation systems. In *The PAAMS Collection*, pages 279–290. Springer, 2014.

[Vermorel and Mohri, 2005] Joannès Vermorel and Mehryar Mohri. Multi-armed bandit algorithms and empirical evaluation. In *Proc. ECML*, pages 437–446. Springer, 2005.

[Vovk, 1990] Vladimir Vovk. Aggregating strategies. In *Proc. 3rd COLT*, pages 371–383, 1990.