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

Mithun Chakraborty, Sanmay Das,*Brendan Juba†
Dept. of Computer Science and Engineering
Washington University in St. Louis
{mithunchakraborty,sanmay,bjuba}@wustl.edu

Kai Yee Phoebe Chua‡
School of Information and Computer Science
University of California, Irvine
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 and hence collect a reward, 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 her 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 standard multiplicative weights update algorithm for deciding which arm to pull. We then introduce a value-of-information based strategy for making broadcasts in the decentralized setting, and show experimentally that the algorithm 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 play 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 in which agents need to work well with possibly unfamiliar teammates 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].

*Chakraborty and Das are supported by NSF IIS awards 1414452 and 1527037.

†Supported by an AFOSR Young Investigator Award.

‡Work performed during an REU program at Washington University in St. Louis while enrolled as an undergraduate student at Messiah College.

In a multi-agent multi-armed bandit problem, a team of agents is playing a MAB. The critical question that makes this problem interesting beyond the intrinsic exploration / exploitation tradeoff feature of classical (single-agent) version is the role of communication and information-sharing. Although various solutions to MAB problems with a single agent are well known [Gittins and Jones, 1979; Gittins, 1979; Auer et al., 2002a, 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, the agents can approach a good strategy much faster.

In this paper, we first define a new multi-agent multi-armed bandit model, different from those that have been developed thus far (see Section 1.1), 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) can indeed obtain performance similar to these single-agent solutions that use full information. Our method is based on the multiplicative weight updates algorithm and we prove a regret bound for this strategy using a technique due to Freund and Schapire [1999]. We then describe a decentralized algorithm for the multi-agent multi-armed bandit, and demonstrate experimentally that it achieves performance close to that of the centralized algorithm, thus efficiently solving the exploration / exploitation / communication "trilemma".

Our decentralized algorithm is a variant of the *softmax* strategy, which is known to perform well empirically for the single-agent problem [Vermorel and Mohri, 2005]. In the single-agent softmax strategy, the agent chooses an arm according to a distribution in which each arm is pulled with probability proportional to the exponential of its empirical reward, weighted appropriately. In our multi-agent variant, individual agents each use a similar weight distribution to choose an arm to pull. To summarize our set of criteria for deciding when to broadcast, which we call the *"Value of Information" (VoI)* communication strategy, an agent optimistically estimates the gain in the total reward that could possibly result from broadcasting its last observation; this value of the broadcast is then compared to the estimated value of pulling an arm. If the estimated reward from pulling an arm is higher than this optimistic estimate of the gain from communicating, then the agent naturally abstains from broadcasting its observation in the subsequent 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).

1.1 Related Work

Barrett et al. [2014] were the first to use a multi-agent MAB (in particular, a two-armed bandit with Bernoulli payoff distributions) to formalize ad hoc teamwork with costly communication: 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), and an action phase (pulling an arm to extract a reward). We, on the other hand, are interested in devising a common communication protocol for every agent in a decentralized team where there is an *opportunity cost* associated with broadcasting – the decision to communicate precludes pulling an arm and gathering reward during that round. Our work is also different from the line of literature on distributed multi-agent MAB models for 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. – thus there is no additional information value to each additional pull, but there is a direct reward value.

2 Formal Problem Description

Our model follows the basic definitions of a classical bandit problem where 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, $\mu_{\min} > \mu_{\max} > 0$, 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, she 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 communicate, which arm should it pull now?

Before presenting our strategy for handling the above issues in Section 4, we describe in Section 3 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.

3 Ideal Centralized Multi-Agent MAB

The public agent maintains a normalized weight (in other words, probability) distribution across the n arms, denoted $\mathbf{P}_t = (P_{1,t}, P_{2,t}, \cdots, P_{n,t})$ 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.[4]The starting distribution is uniform, i.e. $P_{i,1} = 1/n\ \forall i$. 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, 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. $\sum_{t=1}^T \sum_{i=1}^n r_{i,t} m P_{i,t}$, or equivalently the time-averaged per-agent cumulative reward $\frac{1}{T}\sum_{t=1}^T r_{i,t}P_{i,t}$. Hence, we define the (hindsight) regret of the centralized strategy with updates (1) as

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

where $\mathbf{P} = (P_1, P_2, \cdots, P_n)$ is an n-point probability distribution. Theorem 1 shows that the regret of the above centralized MWU method becomes vanishingly small for a large enough T; we provide a proof sketch here, and defer the complete proof to a full version of the paper.

[4]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.

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 set as follows:*

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

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

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

Proof sketch. We consider 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 reward for any action lies in $[-1, 0]$, and extend it to the case of Gaussian rewards. In our version, we use the parameters λ and κ, set to the values specified in the theorem statement, to shift and rescale the Gaussian reward so that the exponent of β in (1) lies within the required bounds with a high probability. Our potential function is the Kullback-Leibler divergence of the current probability distribution $\mathbf{P}_t$ from an arbitrary fixed distribution $\widetilde{\mathbf{P}}$. Using the union bound of probabilities over arms and epochs in addition to the upper and lower bounds on the difference between the values of this potential function at the initial and final epochs within the time-horizon as deduced by Freund and Schapire [1999], we show that $\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{n} r_{i,t}\widetilde{P}_i - \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{n} r_{i,t}P_{i,t} \leq \Delta \triangleq \kappa\left(\sqrt{2\ln(n)/T} + \ln(n)/T\right)$ with probability at least $(1-\delta)$. 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 $T \geq \max(2, \ln(n))$, thus completing the proof. □

Evidently, 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 a temperature $\tau_t = \tau_0/t$ where

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

4 Decentralized MAB With Value-of-Information Communication Strategy

The decentralized version of this problem introduces communication as a costly option for agents. Our idea for a solution is to design a scheme for how each agent decides whether to broadcast or pull an arm in such a way that the overall behavior of the team mimics that of the centralized version described in Section 3 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 that has been publicly communicated by the team until the current epoch: a shared table containing for each arm i two entries, $\nu_{i,t}^{cum} = |T_{i,t}^{cum}|$, where $T_{i,t}^{cum}$ is the set of epochs at each of which information on arm i was broadcast by at least one agent, and the cumulative reward $r_{i,t}^{cum} = \sum_{t \in T_{i,t}^{cum}} r_{i,t}$ (note that by summing over distinct information broadcasts, we ignore excess broadcasts with duplicate contents when an arm's information is communicated separately by two agents in the same round).

For each agent j, any epoch is either an an *action* round, when it pulls an arm, or a *broadcast* round, when it communicates a message to the team, the starting epoch in the time-horizon being necessarily an action round for each agent. Agent j has a private table with four entries for each arm i: $\nu_{i,j,t}^a = |T_{i,j,t}^a|$, $r_{i,j,t}^a = \sum_{t \in T_{i,j,t}^a} r_{i,t}$, $\nu_{i,j,t}^b = |T_{i,j,t}^b|$, $r_{i,j,t}^b = \sum_{t \in T_{i,j,t}^b} r_{i,t}$ where $T_{i,j,t}^a$ is the set of epochs in which agent j has pulled arm i, and $T_{i,j,t}^b \subset T_{i,j,t}^b$ is the subset of these pulls that she 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*: $\hat{\mu}_{i,j,t} = \hat{r}_{i,j,t}/\hat{\nu}_{i,j,t}$ where $\hat{r}_{i,j,t} = r_{i,j,t}^a - r_{i,j,t}^b + r_{i,t}^{cum}$

and $\hat{\nu}_{i,j,t} = \nu^a_{i,j,t} - \nu^b_{i,j,t} + \nu^{cum}_{i,t}$.[5] 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 3, to generate a probability distribution over arms, and draws an arm according to this distribution. Then j's private table is updated, and the reward it collects is added to the team's cumulative reward.

Value-of-Information (VoI) communication criterion. At the end of each action round, say epoch t, agent j follows a three-step procedure to decide whether or not broadcast information on the arm it just pulled, $i^* = i_{j,t}$, in the next epoch. First, it checks if the reward from this pull, $r^* = r_{i^*,t}$, is greater than the public agent's empirical mean $(r^{cum}_{i^*,t}/\nu^{cum}_{i^*,t})$; 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* $\{\hat{\mu}^*_{i,j,t}\}_{i=1}^n$ across arms for further comparison purposes; an upper (resp. lower) bound is obtained by adding to (resp. subtracting from) the empirical mean the quantity $\sigma\sqrt{2\ln(\frac{nT}{2\varepsilon_{voi}})/\hat{\nu}_{i,j,t}}$, where $\varepsilon_{voi} \in (0,1)$ is a free (error) parameter, and $\hat{\nu}_{i,j,t}$ is as defined above .

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 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 $\Lambda^a_{j,t}$, the agent acts as if the public agent will receive no further communication, and will allocate arms to agents for each of the remaining epochs starting at $t+1$ using the (softmax) weight distribution $\{w^a_{i,t}\}_{i=1}^n$ based on the public agent's current empirical means. Thus, $\Lambda^a_{i,t} = (T-t)m\sum_{i=1}^n w^a_{i,t}\hat{\mu}^*_{i,j,t}$. For computing $\Lambda^b_{i,t}$, agent j assumes that the public agent uses the weight distribution $\{w^a_{i,t}\}_{i=1}^n$ 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 to update its weight distribution to $\{w^b_{i,t+1}\}_{i=1}^n$, and use this distribution henceforth. Thus, $\Lambda^b_{j,t} = (m-1)\sum_{i=1}^n w^a_{i,t}\hat{\mu}^*_{i,j,t} + (T-t-1)m\sum_{i=1}^n w^b_{i,t+1}\hat{\mu}^*_{i,j,t}$. The agent decides to not broadcast in epoch $t+1$ if $\Lambda^b_{j,t} - \Lambda^a_{j,t} < 0$.

If this inequality is violated, agent j 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 on the arm; if $\hat{m}_{i^*}$ is one or less, agent j decides to broadcast deterministically at $t+1$, otherwise it broadcasts on the success of a Bernoulli trial with success probability $1/\lfloor\hat{m}_{i^*}\rfloor$.

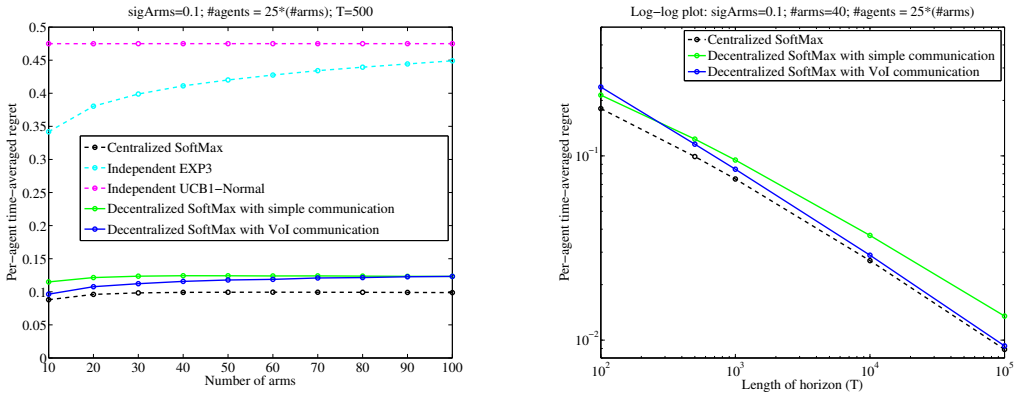
Message broadcasting. If epoch t is a broadcast round for agent j, it publicly sends out the message (i^*, r^*) where $i^* = i_{j,t-1}$ and $r^* = r_{i^*,t-1}$, pulls no arm at t but sets epoch $t+1$ as an action round; before epoch $t+1$ commences, the public agent is augmented with messages broadcast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

5 Experimental Evaluation

In this section, we describe two sets of experiments that we performed in order to compare the performance of the decentralized multi-agent MAB exploration-exploitation algorithm with VoI communication strategy that we proposed in Section 4 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 1 (a) and (b).

Our main benchmark for both sets is the centralized softmax / MWU strategy, detailed in Section 3, 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 4 (skipping the first two stages of

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



(a) (b)

Figure 1: Error bars, being small, are omitted. (a) Regret vs. number of arms for a fixed "intermediate" time-horizon $T = 500$ epochs. For this T, 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 significantly outperforming simple communication strategy for a lower number of arms. (b) Regret vs. length of horizon for a 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 such short horizons; however, it overtakes simple communication for longer horizons, performing on a par with the centralized strategy for a large enough T ($\sim 10^4$ and higher).

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 in the team is set at $m = 25n$ where n is the number of arms in that experiment. The means of the Gaussian reward distributions on the arms of our bandit 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 first arm, which is the "best arm" by design, is used as the standard for computing regrets (as in the classical stochastic setting of a bandit problem).

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 horizon, and $\mu_1 = 1$. This definition of regret is different from, and in fact stronger than, the one we used in Section 3. Each plotted data-point is generated by averaging the per-agent time-averaged regret values over $N_{sim} = 10^5$ repetitions. We set $\delta = 0.01$, and the VoI error parameter $\varepsilon_{voi} = \frac{0.05}{N_{sim}}$ to ensure that our confidence bounds hold for all experiments.

The graphs in Figure 1 provide strong empirical evidence that, for a range of values of the parameters n and T, the VoI communication strategy enables a decentralized team with a sufficient number of members to achieve overall performance close to that effected by a central controller.

6 Conclusion

In this paper, we formulate a novel model for the problem of reward collection by an ad hoc team from multiple (stochastic) sources with costly communication by extending the classic multi-armed bandit model to a new multi-agent setting. We introduce an algorithm (decentralized softmax with VoI communcation strategy) for achieving an exploration / exploitation / communication trade-off in this model. In order to benchmark the performance of this algorithm, we also design a centralized algorithm and prove that it achieves no-regret. Finally, we demonstrate empirically that the performance of our decentralized algorithm, measured in terms of regret, is comparable to that of the centralized method. Our next step will be to establish by an analytical treatment that our method indeed converges to the centralized approach at a reasonable rate.

References

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. 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.

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

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002b.

S. Barrett, N. Agmon, N. Hazon, S. Kraus, and P. Stone. Communicating with unknown teammates. In *Proc. ECAI-14*, 2014.

K. Decker and V. R. Lesser. Designing a family of coordination algorithms. In *Proc. ICMAS-95*, pages 73–80, 1995.

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

D. P. Foster and R. V. Vohra. A randomized rule for selecting forecasts. *Operations Research*, 41: 704–709, 1993.

Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

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

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

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

B. J. Grosz and S. Kraus. Collaborative plans for complex group actions. *AIJ*, 86:269–358, 1996.

D. Kalathil, N. Nayyar, and R. 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.

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

N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, 1994.

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

P. Stone and S. Kraus. To teach or not to teach? decision making under uncertainty in ad hoc teams. In *Proc. AAMAS-2010*, 2010.

P. Stone, G. A. Kaminka, S. Kraus, and J. S. Rosenschein. Ad hoc autonomous agent teams: Collaboration without pre-coordination. In *Proc. AAAI-10*, 2010.

M. Tambe. Towards flexible teamwork. *JAIR*, 7:81–124, 1997.

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

J. Vermorel and M. Mohri. Multi-armed bandit algorithms and empirical evaluation. In *Proc. ECML*, pages 437–446. Springer, 2005.

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