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

Mithun Chakraborty\textsuperscript{1}, Kai Yee Phoebe\textsuperscript{2} Chua, Sanmay Das\textsuperscript{1}, and Brendan Juba\textsuperscript{1}

\textsuperscript{1}Dept. of Computer Science and Engineering, Washington University in St. Louis
\texttt{\{mithunchakraborty,sanmay,bjuba\}@wustl.edu}
\textsuperscript{2}School of Information and Computer Science, University of California, Irvine
\texttt{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 multiplicative weights update algorithm for deciding which arm to pull. We then introduce a value-of-information based strategy based on the centralized algorithm for making broadcasts in the decentralized setting, and show experimentally that the algorithm converges rapidly to the performance of the centralized method.

Keywords: multi-agent systems, coordination and cooperation, online learning, multi-armed bandits

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 [21]; a setting in which agents need to cooperate without any pre-coordination and work toward a common goal [20].

Standard approaches to teamwork, e.g. SharedPlans [13], STEAM [22], or GPGP [6], 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 [4].

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 the classical (single-agent) version is the role of information-sharing. Although various solutions to single-agent MAB problems are well known [12, 11, 2, 1, 15] 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. But, in general, information sharing comes at a cost that can take many forms: a fixed penalty term or fractional reduction applied to the immediate reward gathered from a pull, or the preclusion of pulling an arm, by an agent while transmitting 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 will make this point more explicit in the description of our algorithm in Section 4.

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 [25, 7, 8, 16]. 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 [10].

The problem that we actually want to solve involves a decentralized multi-agent system, albeit one that is working towards a common goal; 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 how they utilize publicly available information. We let the team maintain a shared base of all historically broadcast information that acts as 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 her 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 [24]: In the single-agent softmax strategy, 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 variant, individual agents each use a similar weight distribution to choose an arm, the parameters of this distribution being informed by the algorithm we developed for the central benchmark. 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 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 demonstrate experimentally that our decentralized algorithm for the multi-agent multi-armed bandit achieves performance close to that of the centralized algorithm, thus efficiently solving the exploration / exploitation / communication “trilemma”.

1.1 Related Work

Barrett et al. [4] 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 constructing a common communication protocol for every agent in a decentralized team in a situation where broadcasting precludes action (as defined above) in the same round. Our work is also different from the line of literature on distributed multi-agent MAB models for cognitive radio networks [17, 14, 23] 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 information value to each additional pull, but a direct reward value.

We must also mention multi-agent MAB models, where agents access and utilize one another’s historical information, that have been used for the study (both theoretical and experimental) of social learning or imitation in the social and cognitive sciences [19, 5, 18]. The main difference between these models and
ours is that their agents are selfishly motivated (and may follow strategies fundamentally different from those of others) whereas ours work towards the shared goal of maximizing the overall reward of the team. Moreover, in these models, no cost is incurred by an agent for transmitting its information to (an)other agent(s) but may be sustained in acquiring such information; e.g., if an agent is receiving (perhaps noisy) information about another’s action and reward in an epoch, it cannot pull an arm in that epoch.

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, $\mu_{\text{max}} > \mu_{\text{min}} > 0$, 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 $t$, 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 broadcast, 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 up-
date 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 $P_t = (P_{1,t}, P_{2,t}, \ldots, P_{n,t})$ where $P_{i,t} \geq 0 \forall i$, $\sum_{i=1}^{n} P_{i,t} = 1$, and assigns $m P_{i,t}$ agents to arm $i$ in epoch $t$. 1. 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 $P_{t+1}$ at the beginning of the next epoch using the following multiplicative weights update (MWU) approach [16, 9, 10]:

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

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

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

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

where $P = (P_1, P_2, \ldots, P_n)$ 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 $T$.

**Theorem 1.** Suppose, a bandit has $n$ arms producing Gaussian rewards with the same known standard deviation $\sigma$, 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 $\delta$, $0 < \delta < \min\{2nT \Phi(R_\mu/2\sigma), 1\}$, if we use a centralized MWU strategy with a uniform initial weight distribution and the update rule (1) with parameters set as follows:

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

1 In an actual implementation, if $m P_{i,t}$ is fractional, then $\lfloor m P_{i,t} \rfloor$ agents are initially assigned to arm $i$, and then all the remaining $(m - \sum_{i=1}^{n} \lfloor m P_{i,t} \rfloor)$ are optimistically allocated to the arm with the current highest empirical mean.
where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function, then with probability at least $(1 - \delta),$

$$R_{\text{central}}(T) = O\left((R_\mu + \sigma)\sqrt{\frac{\ln(nT)}{T} \ln(n)}\right).$$

We outline the proof of the theorem below, and defer the complete proof to the appendix. But first, we note that the above MWU algorithm is equivalent to a decreasing Softmax strategy [24] over empirical means $\hat{\mu}_{i,t} = \sum_{s=1}^{T} r_{i,s}/t$ with a temperature $\tau_t$ where

$$\tau_t = \tau_0/t, \quad \tau_0 = \left[ R_\mu - 2\sigma \Phi^{-1}(\delta/(2nT)) \right] / \ln(1 + \sqrt{2\ln(n)/T}).$$

(2)

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

$$P_t = \text{Softmax}(\{\hat{\mu}_{i,t}\}_{i=1}^{n}, \tau_t).$$

Proof sketch. We rewrite the multiplicative factor in (1) as $\beta^{L_{i,t}}$ where we define $L_{i,t} = (-r_{i,t} - \lambda) / \kappa$ as the normalized loss. We now recall the amortized analysis that Freund and Schapire [10] 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 $\lambda$ and $\kappa$ so that the following inequality holds:

$$\Pr(\beta^{L_{i,t}} \leq 1 - (1 - \beta)L_{i,t}) \geq 1 - \delta/(nT) \quad \forall i, t.$$  

(3)

Our potential function is $RE(\tilde{P}||P_t)$, the Kullback-Leibler divergence of the current probability distribution $P_t$ from an arbitrary fixed distribution $\tilde{P}$; we further define

$$\Delta_{RE_{t,t'}} = \Delta_{RE(\tilde{P}||P_{t'})} - \Delta_{RE(\tilde{P}||P_t)}.$$

It is easy to see that $\Delta_{RE_{1,T+1}} \geq -\ln(n)$ since $P_1$ is a uniform distribution; combining the analysis of Freund and Schapire [10] with a union bound of probabilities (over arms and epochs) applied to (3), we can further deduce a high-probability upper bound on $\Delta_{RE_{1,T+1}} = \sum_{t=1}^{T} \Delta_{RE_{t,t+1}}$. Using these upper and lower bounds, we can establish that, with probability at least $(1 - \delta),$

$$\frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{n} r_{i,t}\tilde{P}_i - \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{n} r_{i,t}P_{i,t} \leq \Delta,$$

where $\Delta \triangleq \kappa \left(\sqrt{2\ln(n)/T} + \ln(n)/T\right)$.

Finally, using the Gaussian tail inequality $\Phi(-a) \leq 0.5e^{-a^2/2}$ for any $a > 0$, we can show that $\kappa \leq R_\mu + 2\sigma \sqrt{2 \ln(nT)}$, and hence, after some algebra, that

$$\Delta < 2\sqrt{2} \left(\sqrt{2} + 1\right) (R_\mu + \sigma) \sqrt{\frac{\ln(nT) \ln(n)}{T}}$$

for $T \geq \max(2, \ln(n))$. This completes the proof. $\square$
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 two entries for each arm $i$: $\nu_{i,t}$ and $\nu_{i,t}^{\text{cum}}$. The table contains, for each arm $i$, $\nu_{i,t} = \lceil T_{i,t}^{\text{cum}} \rceil$, where $T_{i,t}^{\text{cum}}$ is the set of epochs at which each agent communicated the arm's information 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). The public agent's empirical mean for arm $i$ during epoch $t$ is $\hat{\mu}_{i,t}^{\text{public}} = \nu_{i,t}^{\text{cum}} / \nu_{i,t}$.

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 every agent. Agent \( j \) has a private table with four entries for each arm \( i \); \( \nu_{a,j,t}^a = |T_{i,j,t}^a| \), \( r_{a,j,t}^a = \sum_{t \in T_{i,j,t}^a} r_{i,j,t} \), \( \nu_{i,j,t}^b = |T_{i,j,t}^b| \), \( r_{i,j,t}^b = \sum_{t \in T_{i,j,t}^b} r_{i,j,t} \) where \( T_{i,j,t}^a \) is the set of epochs in which agent \( j \) has pulled arm \( i \), and \( T_{i,j,t}^b \subset T_{i,j,t}^a \) 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 across arms \( \hat{\mu}_{i,j,t}^a \); \( \hat{r}_{i,j,t}^a = \hat{r}_{i,j,t}^a / \hat{\nu}_{i,j,t}^a \) where \( \hat{r}_{i,j,t}^a = r_{i,j,t}^a - r_{i,j,t}^b + r_{i,t}^{cum} \) and \( \hat{\nu}_{i,j,t}^a = \nu_{i,j,t}^a - \nu_{i,j,t}^b + \nu_{i,t}^{cum} \). It then applies a softmax function to these means with the decreasing temperature parameter \( \tau_t = \tau_0 / t \), \( \tau_0 \) defined in (2) in Section 3, 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:

\[
\begin{align*}
\hat{r}_{i,j,t+1}^a &= \hat{r}_{i,j,t}^a + r^*; \\
\nu_{i,j,t+1}^a &= \nu_{i,j,t}^a + 1; \\
\hat{r}_{i,j,t+1}^b &= \hat{r}_{i,j,t}^b; \\
\nu_{i,j,t+1}^b &= \nu_{i,j,t}^b. \\
\end{align*}
\]

**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 to decide whether or not broadcast information on the arm \( i^* \) it just pulled in the next epoch \( (t + 1) \). We depict a flow-chart for this procedure in Figure 1.

First, it checks if the reward \( r^* \) from this pull is greater than the public agent’s empirical mean \( \hat{\mu}_{i,t}^{\text{public}} \); 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 \( \hat{\mu}_{i,j,t} \) 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 \( \sigma \sqrt{2 \ln \left( \frac{n_{i,j,t}}{2\varepsilon_{voi}} \right) / \hat{\nu}_{i,j,t+1}} \), 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 \( A_j^a \), and the other in which it is a broadcast round, say \( A_j^b \). For computing \( A_j^a \), 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 weight

\[ \sigma = \left( \frac{\mu_{\min} + \mu_{\max}}{2} \right). \]

Each agent’s private table is so initialized that the initial empirical mean for each arm is \( (\mu_{\min} + \mu_{\max}) / 2 \).
distribution \( \mathbf{w}_t^a = \{w_{i,t}^a\}_{i=1}^n \), where the temperature \( \tau \) again comes from (2):

\[
\mathbf{w}_t^a = \text{SOFTMAX}(\{\hat{\mu}_{i,t}^\text{public}\}_{i=1}^n, \tau_t). \tag{4}
\]

\( A_{j,t}^a = (T - t)m \sum_{i=1}^n w_{i,t}^a \hat{\mu}_{i,j,t} \).

For evaluating \( A_{j,t}^b \), agent \( j \) assumes that the public agent uses the weight distribution \( \mathbf{w}_t^a \) 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 \( \mathbf{w}_{t+1}^b = \{w_{i,t+1}^b\}_{i=1}^n \), and use this distribution henceforth.

\[
\begin{align*}
\hat{\mu}_{i^*, t+1}^b &= \frac{r_{i^*, t+1}^*}{\rho_{i^*, t+1}^*}, \quad \hat{\mu}_{i,t+1}^b = \mu_{i,t}^\text{public} \quad \forall i \neq i^*; \\
\mathbf{w}_{t+1}^b &= \text{SOFTMAX}(\{\hat{\mu}_{i,t+1}^b\}_{i=1}^n, \tau_t); \\
A_{j,t}^b &= (m - 1) \sum_{i=1}^n w_{i,t}^b \hat{\mu}_{i,j,t} \\
&+ (T - t - 1)m \sum_{i=1}^n w_{i,t+1}^b \hat{\mu}_{i,j,t}.
\end{align*}
\]

Notice that the computation of \( A_{j,t}^b \) 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.

Define the the agent's current Value of Information \( \text{VoI} \) as

\[
\text{VoI} \triangleq A_{j,t}^b - A_{j,t}^a,
\]

which estimates the long-term benefit accrued by the team if the agent under consideration forgoes immediate reward collection to share her latest information. Thus, agent \( j \) decides to not broadcast in epoch \((t + 1)\) if \( \text{VoI} \leq 0 \).

If \( \text{VoI} > 0 \), 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 \( w_{i^*, t}^a \) on the arm; if \( \hat{m}_{i^*} \) is 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 \( 1/\lceil \hat{m}_{i^*} \rceil \). The idea is to keep the expected number of broadcasts per arm per epoch at 1 since no broadcast leads to information loss but more broadcasts provide redundant information and impede reward collection.

If (and only if) agent \( j \) decides that epoch \((t + 1)\) is a broadcast round, the entries \( r_{i^*, j,t}^b \) and \( \nu_{i^*, j,t}^b \) in its private table are incremented by \( r^* \) and 1 respectively.
Fig. 2. Error bars, being small, are omitted. (a) Regret vs. number of arms for fixed “intermediate” time-horizon $T = 500$ epochs: UCB1-Normal is still in its initial exploration phase (see Auer et al. [2] 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. Vol 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).

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.

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 2 (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 ap-
proaches – EXP3 [3] and UCB1-Normal [2] – 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 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 $A_{j,t}^b - A_{j,t}^a$) 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 $\mu_{\text{max}} = \mu_1 = 1$ and ending at $\mu_{\text{min}} = \mu_n = 0.05$, 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 $N_{\text{sim}} = 10^5$ repetitions. We set $\delta = 0.01$, and the VoI error parameter $\varepsilon_{\text{voi}} = 0.05/N_{\text{sim}}$ to ensure that our confidence bounds hold for all experiments.

The graphs in Figure 2 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 formulated 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 introduced an algorithm (decentralized softmax with VoI communication strategy) for achieving an exploration / exploitation / communication trade-off in this model. In order to benchmark the performance of this algorithm, we also designed a centralized algorithm and prove that it achieves no-regret. Finally, we demonstrated empirically that the performance of our decentralized algorithm, measured in terms of regret, is comparable to that of the centralized method. Directions for future work include considering information acquisition costs along the lines of, e.g., Rendell et al. [18], evaluating our strategy under other communication cost structures, and analytically deriving the rate of convergence of our decentralized approach to the centralized benchmark.
References


Appendix: Proof of Theorem 1.

To make our exposition consistent with that of Freund and Schapire [10], we will present our analysis in terms of losses simply defined as negative rewards.

Let $L_{i,t}$ be the raw Gaussian loss generated by arm $i$ at epoch $t$, $i = 1, \ldots, n$, $t = 1, \ldots, T$. In our model, regardless of whether an arm actually gets pulled at epoch $t$, it “sits ready” with this sample which gets added to the cumulative reward of any agent who selects the arm at $t$.

$$\frac{L_{i,t} - \mu_i}{\sigma} \sim \text{i.i.d.} \mathcal{N}(0, 1),$$

where $\mu_i$ is the negative of the time-invariant expected reward generated by arm $i$, and $\sigma$ is the common standard deviation of all reward distributions.

Thus, $\mu_{\min} \leq \mu_i \leq \mu_{\max} \forall i$, $R_{\mu} = \mu_{\max} - \mu_{\min}$.

Thus, we can restate the multiplicative weights update rule used by our central agent as follows:

$$P_{i,t+1} = \frac{P_{i,t}}{Z_t} \beta \hat{L}_{i,t},$$

(6)

where $\hat{L}_{i,t} = (L_{i,t} - \lambda)/\kappa$ is the normalized loss generated by arm $i$ at epoch $t$, for parameters $\beta \in (0, 1)$, $\lambda \in \mathbb{R}$, $\kappa > 0$. We can interpret $\beta$ as $e^{-1/\tau}$ where $\tau > 0$ is the temperature parameter of a traditional softmax function formulation.

Moreover, let us define, for any weight distribution $Q = \{Q_1, Q_2, \ldots, Q_n\}$,

$$L_t(Q) \triangleq \sum_{i=1}^{n} Q_i L_{i,t}, \quad \text{and} \quad \hat{L}_t(Q) \triangleq \sum_{i=1}^{n} Q_i \hat{L}_{i,t} \forall t.$$

This allows us to redefine our regret as

$$\text{regret}(T) = \frac{1}{T} \sum_{t=1}^{T} L_t(P_t) - \min_P \left[\frac{1}{T} \sum_{t=1}^{T} L_t(P)\right].$$

We now restate our theorem in slightly different terms.

Restatement of Theorem 1. Suppose, a bandit has $n$ arms producing Gaussian losses with the same known standard deviation $\sigma$, and unknown means with known upper and lower bounds $\overline{\mu}_i \in [\overline{\mu}_{\min}, \overline{\mu}_{\min} + R_{\mu}] \forall i = 1, \ldots, n$. Then, for any time horizon $T \in \mathbb{Z}^+$ and an arbitrarily small number $\delta$, $0 < \delta < \min(2nT\Phi(R_{\mu}/2\sigma), 1)$, if the MWU update rule (6) has parameters given by

$$\beta = 1/\left(1 + \sqrt{\frac{2\ln(n)}{T}}\right),$$
$$\lambda = \overline{\mu}_{\min} + \sigma \Phi^{-1}(\delta/(2nT)),$$
$$\kappa = R_{\mu} - 2\sigma \Phi^{-1}(\delta/(2nT)),$$
then with probability at least \((1 - \delta)\),

\[
\text{regret}(T) = O\left((R_\mu + \sigma)\sqrt{\ln\left(\frac{n^2 T}{\delta}\right) \ln(\ln n)}\right).
\]

**Proof.** Note that our (normalized) weight distributions \(P_t\) can be viewed as categorical probability distributions. For any positive integers \(t, t',\) and categorical probability distributions \(\tilde{P}, P_t,\) and \(P_{t'}\), each representing a vector of normalized weights assigned to \(n\) arms of a multi-armed bandit, define

\[
\Delta \tilde{R}E_{t, t'} \triangleq \text{RE}(\tilde{P} || P_{t'}) - \text{RE}(\tilde{P} || P_t),
\]

where \(\text{RE}(Q^* || Q)\) is the relative entropy or Kullback-Leibler divergence of \(Q\) from \(Q^*\).

Proceeding as in Freund and Schapire [10], for each epoch \(t\), we can show that

\[
\Delta \tilde{R}E_{t, t+1} = \text{RE}(\tilde{P} || P_{t+1}) - \text{RE}(\tilde{P} || P_t)
\]

\[
= (\ln \frac{1}{\beta}) \sum_{i=1}^{n} \tilde{P}_i \hat{L}_{i,t} + \ln Z_t
\]

\[
= (\ln \frac{1}{\beta}) \hat{L}_t(\tilde{P}) + \ln \left[ \sum_{i=1}^{n} P_i \beta \hat{L}_{i,t} \right].
\]  \hfill (7)

Let us define the following propositions for \(t = 1, \ldots, T:\)

- **TermUB_{i,t} \triangleq \left\{ \beta \hat{L}_{i,t} \leq 1 - (1 - \beta) \hat{L}_{i,t} \right\} \forall i. \hfill (8)**
- **AllTermsUB_t \triangleq \bigcap_{i=1}^{n} \text{TermUB}_{i,t}, \hfill (9)**
- **SumUB_t \triangleq \left\{ \sum_{i=1}^{n} P_{i,t} \beta \hat{L}_{i,t} \leq \sum_{i=1}^{n} P_{i,t} (1 - (1 - \beta) \hat{L}_{i,t}) \right\}, \hfill (10)**
- **LogLinIneq_t \triangleq \left\{ \ln \left( \sum_{i=1}^{n} P_{i,t} (1 - (1 - \beta) \hat{L}_{i,t}) \right) \right\} \leq -(1 - \beta) \hat{L}_t(P_t), \hfill (11)**
- **RELossIneq_t \triangleq \left\{ \Delta \tilde{R}E_{t, t+1} \leq (\ln \frac{1}{\beta}) \hat{L}_t(\tilde{P}) - (1 - \beta) \hat{L}_t(P_t) \right\}. \hfill (12)**

First, note that for any real \(x\) and \(\beta \in (0, 1)\), we can easily show from the convexity of the function \(f(x) = \beta x\):\[
\{ \beta x \leq 1 - (1 - \beta)x \} \equiv \{ 0 \leq x \leq 1 \}.
\]
Hence, for each \(i, t\), from (8),
\[
\text{TermUB}_{i,t} \equiv \{0 \leq \hat{L}_{i,t} \leq 1\}
\]
so that
\[
\Pr(\text{TermUB}_{i,t}) = \Pr(0 \leq \hat{L}_{i,t} \leq 1)
= \Pr(\lambda \leq L_{i,t} \leq \lambda + \kappa)
= \Phi((\lambda + \kappa - \bar{\mu}_i)/\sigma_i)
- \Phi((\lambda - \bar{\mu}_i)/\sigma_i),
\]
where \(\Phi(\cdot)\) is the standard normal CDF. Since \(\Phi((a-x)/s)\) is a strictly decreasing function of \(x \in \mathbb{R}\) and \(\bar{\mu}_\text{min} \leq \bar{\mu}_i \leq \bar{\mu}_\text{max}\ \forall i = 1, 2, \ldots, n\), we have
\[
\Phi((\lambda + \kappa - \bar{\mu}_i)/\sigma_i) \geq \Phi((\lambda + \kappa - \bar{\mu}_\text{max})/\sigma)
= \Phi(-\Phi^{-1}(\delta/(2nT)))
= 1 - \delta/(2nT),
\]
plugging in the expressions for \(\lambda, \kappa, \) and \(R_{\mu}\), and simplifying. Likewise,
\[
\Phi((\lambda - \bar{\mu}_i)/\sigma_i) \leq \Phi((\lambda - \bar{\mu}_\text{min})/\sigma)
= \Phi(\Phi^{-1}(\delta/(2nT)))
= \delta/(2nT),
\]
again using the definition of \(\lambda\). Thus, from (14),
\[
\Pr(\text{TermUB}_{i,t}) \geq (1 - \delta/(2nT)) - \delta/(2nT)
= 1 - \delta/nT.
\]
Moreover, from definitions (8) and (9),
\[
\text{AllTermsUB}_t \equiv \bigcap_{i=1}^n \text{TermUB}_{i,t}.
\]
Hence, by De Morgan’s Laws,
\[
\text{AllTermsUB}_t \equiv \bigcup_{i=1}^n \text{TermUB}_{i,t}.
\]
Applying the union bound,
\[
\Pr(\text{AllTermsUB}_t) \leq \sum_{i=1}^n \Pr(\text{TermUB}_{i,t})
= \sum_{i=1}^n (1 - \Pr(\text{TermUB}_{i,t}))
\leq \sum_{i=1}^n \frac{\delta}{nT} = n \cdot \frac{\delta}{nT}
\text{ from (15)},
\]
i.e.,
\[
\Pr(\text{AllTermsUB}_t) \leq \frac{\delta}{T}.
\]
We shall get back to the above result later. Let us first bound the probability of the required event in terms of the probabilities \( \{\text{Pr}(\text{AllTermsUB}_t)\}_{t=1}^T \).

Since \( \mathbf{P}_t \) is a probability distribution, it is obvious from (8), (9) and (10) that
\[
\text{AllTermsUB}_t \implies \text{SumUB}_t. \tag{18}
\]

Also, note that
\[
\sum_{i=1}^n P_{i,t}(1 - (1 - \beta)\hat{L}_{i,t}) = 1 - (1 - \beta) \sum_{i=1}^n P_{i,t}\hat{L}_{i,t} = 1 - (1 - \beta)\hat{L}_t(\mathbf{P}_t).
\]

Hence, from (11),
\[
\text{LogLinIneq}_t \equiv \left\{ \ln(1 - (1 - \beta)\hat{L}_t(\mathbf{P}_t)) \leq -(1 - \beta)\hat{L}(\mathbf{P}_t) \right\},
\]
which, it can be shown\(^3\), is true whenever
\[
(1 - \beta)\hat{L}_t(\mathbf{P}_t) < 1, \text{ i.e. } \hat{L}_t(\mathbf{P}_t) < 1/(1 - \beta)
\]
for any \( \beta \in (0, 1) \). Thus, from (16) and (13),
\[
\text{AllTermsUB}_t \equiv \bigcap_{i=1}^n \{0 \leq \hat{L}_{i,t} \leq 1\} \implies \{0 \leq \hat{L}_t(\mathbf{P}_t) \leq 1\}
\]
\[
\text{since } \hat{L}(\mathbf{P}_t) = \sum_{i=1}^n P_{i,t}\hat{L}_{i,t} \text{ with } P_{i,t} \geq 0 \ \forall i, \ \sum_{i=1}^n P_{i,t} = 1,
\]
\[
\implies \{\hat{L}(\mathbf{P}_t) < 1/(1 - \beta)\}
\]
\[
\text{since } 1/(1 - \beta) > 1,
\]
\[
\implies \text{LogLinIneq}_t, \tag{19}
\]
as noted above. Thus, combining (18) and (19),
\[
\text{AllTermsUB}_t \implies \text{SumUB}_t \cap \text{LogLinIneq}_t \tag{20}
\]

Now, from (12) and (7),
\[
\text{RELossIneq}_t \equiv \left\{ \ln \left[ \sum_{i=1}^n P_{i,t}\beta\hat{L}_{i,t} \right] \leq -(1 - \beta)\hat{L}(\mathbf{P}_t) \right\}.
\]

\(^3\) Since the function \( g(x) = e^{-x} \) is convex for any real \( x \) and \( g'(x) = -e^{-x} \), we must have \( e^{-x} \geq e^{-0} + (x-0)(-e^{-0}) = 1-x \). For \( x < 1 \), we can take the natural logarithm of both sides to obtain \( -x \geq \ln(1-x) \).
From definitions!(10) and (11), using the increasing monotonicity of ln(·) and the transitivity of \( \leq \), we get
\[
\text{SumUB}_t \cap \text{LogLinIneq}_t \implies \text{RELossIneq}_t.
\]
Hence, from (20),
\[
\text{AllTermsUB}_t \implies \text{RELossIneq}_t \quad \forall \ t \quad \text{so that}
\]
\[
\bigcap_{t=1}^{T} \text{AllTermsUB}_t \implies \bigcap_{t=1}^{T} \text{RELossIneq}_t. \tag{21}
\]
Since \( \Delta \tilde{R}E_{1,T+1} = \sum_{t=1}^{T} \Delta \tilde{R}E_{t,T+1} \), the definition (12) of RELossIneq implies that
\[
\bigcap_{t=1}^{T} \text{RELossIneq}_t \implies \text{Ineq}_1^T, \quad \text{where}
\]
\[
\text{Ineq}_1^T \triangleq \left\{ \Delta \tilde{R}E_{1,T+1} \leq \ln \left( \frac{1}{\beta} \right) \sum_{t=1}^{T} \hat{L}_t(\tilde{P}) - (1-\beta) \sum_{t=1}^{T} \hat{L}_t(P_t) \right\}. \tag{22}
\]
Again, from definition,
\[
\Delta \tilde{R}E_{1,T+1} = \text{RE}(\tilde{P}||P_{T+1}) - \text{RE}(\tilde{P}||P_1) \geq -\text{RE}(\tilde{P}||P_1) \geq -\ln(n). \tag{23}
\]
Inequality (23) follows from the fact that relative entropy is non-negative by definition so that \( \text{RE}(\tilde{P}||P_{T+1}) \geq 0 \) for any choice of \( \tilde{P} \) and any realizations of the random variables \( L_{i,t}, i = 1, 2, \ldots, n, t = 1, 2, \ldots, T \), which (along with \( P_1 \)) determine \( P_{T+1} \). To see the validity of inequality (24), note that for \( P_t, t = 1/n \forall i = 1, 2, \ldots, n \),
\[
\text{RE}(\tilde{P}||P_1) = \sum_{i=1}^{n} \tilde{P}_i \ln \left( \frac{\tilde{P}_i}{1/n} \right) \leq \sum_{i=1}^{n} \tilde{P}_i \ln(n) = \ln(n),
\]
since \( 0 \leq \tilde{P}_i \leq 1 \forall i = 1, 2, \ldots, n \). Hence, from (22),
\[
\text{Ineq}_1^T \implies \text{Ineq}_2^T \quad \text{where}
\]
\[
\text{Ineq}_2^T \triangleq \left\{ \ln \left( \frac{1}{\beta} \right) \sum_{t=1}^{T} \hat{L}_t(\tilde{P}) - (1-\beta) \sum_{t=1}^{T} \hat{L}_t(P_t) \geq -\ln(n) \right\}
\]
\[
= \left\{ \frac{1}{T} \sum_{t=1}^{T} \hat{L}_t(P_t) \leq a_{\beta} \frac{T}{T} \sum_{t=1}^{T} \hat{L}_t(\tilde{P}) + c_{\beta} \frac{\ln(n)}{T} \right\}, \tag{25}
\]
dividing both sides by $T(1 - \beta)$ and rearranging:

$$a_\beta = \frac{\ln(1/\beta)}{1 - \beta} > 0, \quad c_\beta = \frac{1}{1 - \beta} > 1.$$  

Combining (21), (22), and (25), we get

$$\bigcap_{t=1}^{T} \text{AllTermsUB}_t \implies \text{Ineq}_T^2. \quad (26)$$

Observe that, for $\beta \in (0, 1]$, $-\ln \beta \leq (1 - \beta^2)/(2\beta)$ so that $a_\beta \leq (1 + \beta)/(2\beta) = \frac{1}{2} (1 + 1/\beta)$.

Hence, for $0 \leq \hat{L}_t(\hat{P}) \leq 1,$

$$a_\beta \hat{L}_t(\hat{P}) \leq \frac{1}{2} \left(1 + \frac{1}{\beta}\right) \hat{L}_t(\hat{P})$$

$$= \frac{1}{2} \left(2 + \sqrt{\frac{2\ln(n)}{T}}\right) \hat{L}_t(\hat{P})$$

for $\beta = \frac{1}{1 + \sqrt{\frac{2\ln(n)}{T}}}$,

$$= \hat{L}_t(\hat{P}) + \frac{1}{2} \sqrt{\frac{2\ln(n)}{T}} \hat{L}_t(\hat{P})$$

$$\leq \hat{L}_t(\hat{P}) + \frac{1}{2} \sqrt{\frac{2\ln(n)}{T}} \quad \text{since } \hat{L}_t(\hat{P}) \leq 1.$$

Thus, for each epoch $t$,

$$\text{AllTermsUB}_t \equiv \bigcap_{i=1}^{n} \{0 \leq \hat{L}_{i,t} \leq 1\}$$

$$\implies \{0 \leq \hat{L}_t(P_t) \leq 1\}$$

since $\hat{L}_t(P_t) = \sum_{i=1}^{n} P_{i,t} \hat{L}_{i,t}$ with

$$P_{i,t} \geq 0 \quad \forall i, \quad \sum_{i=1}^{n} P_{i,t} = 1$$

$$\implies \left\{ a_\beta \hat{L}_t(\hat{P}) \leq \hat{L}_t(\hat{P}) + \frac{1}{2} \sqrt{\frac{2\ln(n)}{T}} \right\}$$

---

\(^4\) Let $h(x) \triangleq -\ln x - \frac{1-x^2}{2x}$ for $x \in (0, 1]$. Then, for $0 < x \leq 1$, $h'(x) = \frac{1-x^2}{2x^2} \geq 0$, hence $h(x) \leq h(1) = 0.$
Hence,
\[
\bigcap_{t=1}^{T} \text{AllTermsUB}_t \\
\implies \bigcap_{t=1}^{T} \left\{ a_\beta \hat{L}_t(\tilde{P}) \leq \hat{L}_t(\tilde{P}) + \frac{1}{2} \sqrt{\frac{2 \ln(n)}{T}} \right\}
\]
\[
\implies \left\{ a_\beta \sum_{t=1}^{T} \hat{L}_t(\tilde{P}) \leq \sum_{t=1}^{T} \hat{L}_t(\tilde{P}) + \frac{T}{2} \sqrt{\frac{2 \ln(n)}{T}} \right\}
\]
\[
\equiv \text{Ineq}_T^3
\]
where
\[
\text{Ineq}_T^3 \triangleq \left\{ a_\beta \frac{T}{T} \sum_{t=1}^{T} \hat{L}_t(\tilde{P}) \leq \frac{1}{T} \sum_{t=1}^{T} \hat{L}_t(\tilde{P}) + \frac{1}{2} \sqrt{\frac{2 \ln(n)}{T}} \right\}.
\]
Note that, for \( \beta = \frac{1}{1+\sqrt{\frac{2 \ln(n)}{T}}} \),
\[
c_\beta \frac{\ln(n)}{T} = \frac{1}{2} \sqrt{\frac{2 \ln(n)}{T}} + \frac{\ln(n)}{T}.
\]
Thus, combining (26) and (27),
\[
\bigcap_{t=1}^{T} \text{AllTermsUB}_t \implies \text{Ineq}_T^2 \bigcap \text{Ineq}_T^3
\]
\[
\implies \text{Ineq}_T^4,
\]
where
\[
\text{Ineq}_T^4 \triangleq \left\{ \frac{1}{T} \sum_{t=1}^{T} L_t(P_t) \leq \frac{1}{T} \sum_{t=1}^{T} \hat{L}_t(\tilde{P}) \\
+ \sqrt{\frac{2 \ln(n)}{T}} + \frac{\ln(n)}{T} \right\}
\]
\[
\equiv \left\{ \frac{1}{T} \sum_{t=1}^{T} L_t(P_t) \leq \frac{1}{T} \sum_{t=1}^{T} \hat{L}_t(\tilde{P}) \\
+ \kappa \left( \sqrt{\frac{2 \ln(n)}{T}} + \frac{\ln(n)}{T} \right) \right\},
\]
since \( \hat{L}_{t,t} = (L_{t,t} - \lambda)/\kappa \Rightarrow \hat{L}_t(Q) = (L_t(Q) - \lambda)/\kappa \) for any probability distribution \( Q \). Let us define
\[
\Delta \triangleq \kappa \left( \sqrt{\frac{2 \ln(n)}{T}} + \frac{\ln(n)}{T} \right).
\]
Then, plugging in the expression for $\kappa$, we have, for $T \geq \ln(n)$,

\[
\Delta \leq (R_\mu - 2\sigma \Phi^{-1}(\delta/(2nT))) \left( \sqrt{\frac{2\ln(n)}{T}} + \sqrt{\frac{\ln(n)}{T}} \right)
\]

\[
= (R_\mu - 2\sigma \Phi^{-1}(\delta/(2nT))) \left( \sqrt{2} + 1 \right) \sqrt{\frac{\ln(n)}{T}}
\]

\[
\leq \left( \sqrt{2} + 1 \right) \left( R_\mu + 2\sigma \sqrt{2\ln \left( \frac{nT}{\delta} \right)} \right) \sqrt{\frac{\ln(n)}{T}}
\]

from Lemma 1 below. Moreover, for $T \geq 2$, $2\sqrt{2 \ln \left( \frac{nT}{\delta} \right)} > 2\sqrt{2\ln 2} > 1$, since $n/\delta > 1$, so that $R_\mu < 2R_\mu \sqrt{2 \ln \left( \frac{nT}{\delta} \right)}$. Hence,

\[
\Delta < \bar{\Delta} \quad \text{for } T \geq \max(2, \ln(n)),
\]

where $\bar{\Delta} = 2\sqrt{2} \left( \sqrt{2} + 1 \right) (R_\mu + \sigma) \sqrt{\ln \left( \frac{nT}{\delta} \right) \frac{\ln(n)}{T}}$.

This leads us to the conclusion that

\[
\text{Ineq}_4^T \implies \text{ReqdIneq}_T, \quad \text{where}
\]

\[
\text{ReqdIneq}_T \triangleq \left\{ \frac{1}{T} \sum_{t=1}^{T} \hat{L}_t^T(P_t) \leq \frac{1}{T} \sum_{t=1}^{T} \hat{L}_t^T(\tilde{P}) + \bar{\Delta} \right\}.
\]

(29)

Combining (29) and (28),

\[
\Pr(\text{ReqdIneq}_T) \geq \Pr \left( \bigcap_{t=1}^{T} \text{AllTermsUB}_t \right)
\]

\[
\geq 1 - \Pr \left( \bigcup_{t=1}^{T} \text{AllTermsUB}_t \right)
\]

\[
= 1 - \Pr \left( \bigcup_{t=1}^{T} \text{AllTermsUB}_t \right)
\]

by De Morgan’s laws,

\[
\geq 1 - \sum_{t=1}^{T} \Pr(\text{AllTermsUB}_t)
\]

using the union bound,

\[
\geq 1 - \frac{\delta}{T} \quad \text{by (17)},
\]

\[
= 1 - \delta.
\]
Hence, for any choice of $\tilde{P}$, with probability at least $1 - \delta$,

$$\frac{1}{T} \sum_{t=1}^{T} \tilde{L}_t(P_t) \leq \frac{1}{T} \sum_{t=1}^{T} \tilde{L}_t(\tilde{P}) + \tilde{\Delta}.$$ 

So, this result must also hold for $\tilde{P} = \arg\min P \sum_{t=1}^{T} \tilde{L}_t(P)$. This completes the proof.

**Lemma 1.** For positive integers $n$, $T$, and any $\delta \in (0, 1)$,

$$-\Phi^{-1}(\delta/(2nT)) \leq \sqrt{2 \ln \left(\frac{nT}{\delta}\right)}.$$

**Proof.** For any $a > 0$, note that

$$\Phi(-a) = \int_{-\infty}^{-a} \frac{e^{-x^2/2}}{\sqrt{2\pi}} \, dx$$

$$= \int_{0}^{\infty} \frac{e^{-(z+a)^2/2}}{\sqrt{2\pi}} \, dz, \quad z = -(x+a)$$

$$= e^{-a^2/2} \int_{0}^{\infty} \frac{e^{-z^2/2}e^{-az}}{\sqrt{2\pi}} \, dz$$

$$\leq e^{-a^2/2} \int_{0}^{\infty} \frac{e^{-z^2/2}}{\sqrt{2\pi}} \, dz$$

since $e^{-z^2/2} > 0$ for any real $z$, and $e^{-az} < 1$ for $a, z > 0$.

$$e^{-a} < 1$$

$$= e^{-a^2/2}(1 - \Phi(0))$$

$$= e^{-a^2/2}$$

$$= \frac{1}{2}.$$

Hence, plugging in $a = \sqrt{2 \ln \left(\frac{nT}{\delta}\right)}$ on both sides of the above inequality, using the increasing monotonicity of $\Phi(-\cdot)$, and transposing, we obtain the desired result.