Optimal Algorithms for Multiplayer Multi-Armed Bandits

Po-An Wang **Alexandre Proutiere** **Kaito Ariu*** **Yassir Jedra*** **Alessio Russo***

KTH, Royal Institute of Technology
Stockholm, Sweden

Abstract

The paper addresses various Multiplayer Multi-Armed Bandit (MMAB) problems, where M decision-makers, or players, collaborate to maximize their cumulative reward. We first investigate the MMAB problem where players selecting the same arms experience a collision (and are aware of it) and do not collect any reward. For this problem, we present DPE1 (Decentralized Parsimonious Exploration), a decentralized algorithm that achieves the same asymptotic regret as that obtained by an optimal centralized algorithm. DPE1 is simpler than the state-of-the-art algorithm SIC-MMAB Boursier and Perchet (2019), and yet offers better performance guarantees. We then study the MMAB problem without collision, where players may select the same arm. Players sit on vertices of a graph, and in each round, they are able to send a message to their neighbours in the graph. We present DPE2, a simple and asymptotically optimal algorithm that outperforms the state-of-the-art algorithm DD-UCB Martínez-Rubio et al. (2019). Besides, under DPE2, the expected number of bits transmitted by the players in the graph is finite.

1 Introduction

In Multiplayer MAB, there are M independent decision-makers, or players. At every round: (i) each decision-maker selects an arm in the set $\mathcal{K} = \{1, \dots, K\}$, (ii) receives some feedback about this arm, and (iii) possibly communicates with neighbouring players. To simplify the presentation, we assume that in round t, when arm k is selected, the potential collected reward is a random variable (independent of the rewards of the other arms) $X_k(t)$ with Bernoulli distribution with mean μ_k. We further assume that the average rewards $\mu = (\mu_1, \dots, \mu_K)$ are such that $\mu_1 > \mu_2 > \dots > \mu_K$. MMAB problems have received a lot of attention recently. In this paper, we investigate the two most studied MMAB problems: MMAB *with* collisions, motivated by radio channel assignment problems in cognitive radios Jouini et al. (2009), and MMAB *without* collisions, motivated by sequential decisions in social networks Landgren et al. (2016).

1.1 Multiplayer MAB with collisions

In this model, when a player selects an arm, she collects the corresponding reward only if no other player has selected this same arm. More precisely, when in round t, the player selects k, she observes (1) whether her decision collides with those of other players, and (2) $X_k(t)$ in the absence of collision. This feedback scenario is referred to as *collision sensing* in Boursier and Perchet (2019). The different players are not communicating, and they only sense the presence of other players through experienced collisions. A policy π determines in each round which arm every decision-maker will select. We are interested in distributed policies where each decision-maker decides which arm to select independently. This choice depends on the available information to the decision-maker: the past observed collisions and rewards. We denote by $k_i^\pi(t)$ the arm selected by the decision-maker i in round t under the policy π.

Regret lower bound. The optimal expected reward that can be collected in each round is $\sum_{k=1}^M \mu_k$ (when the M best arms are played). Hence the regret up to round T of a policy π is defined as:

$$R^\pi(T) = T\sum_{k=1}^{M}\mu_k - \sum_{t=1}^{T}\sum_{i=1}^{M}\mathbb{E}[\mu_{k_i^\pi(t)}\mathbf{1}_{\{k_i^\pi(t)\neq k_j^\pi(t),\forall j\neq i\}}].$$

*Equal contribution. Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

As in the classical bandit literature Lai and Robbins (1985), we say that a policy π is *uniformly* good if it regret satisfies $R^\pi(T) = o(T^\alpha)$ for all $\alpha > 0$ for any possible μ. We know from Anantharam et al. (1987) that any uniformly good policy π, centralized or not, satisfies:

$$\liminf_{T\to\infty} \frac{R^\pi(T)}{\log(T)} \geq C(\mu) := \sum_{k>M} \frac{\mu_M - \mu_k}{\mathrm{kl}(\mu_k, \mu_M)}, \quad (1)$$

where $\mathrm{kl}(a,b)$ denotes the KL divergence between two Bernoulli distributions of respective means a and b. This result is a simple extension of the classical result derived by Lai and Robbins (1985). Anantharam et al. (1987) also presents a centralized policy achieving the above asymptotic regret lower bound.

State-of-the-art algorithm. In a recent paper Boursier and Perchet (2019), the authors develop SIC-MMAB, an algorithm that uses collisions as a communication tool, and whose regret satisfies:

$$\begin{aligned} R^{\mathrm{SIC}}(T) \leq & c_1 \sum_{k>M} \min\left\{\frac{\log T}{\mu_M - \mu_k}, \sqrt{T\log T}\right\} \\ & + c_2 KM \log T \\ & + c_3 KM^3 \log^2\left(\min\left\{\frac{\log T}{(\mu_M - \mu_{M+1})^2}, T\right\}\right), \end{aligned}$$

for some constants $c_1, c_2, c_3 > 0$. The regret of SIC-MMAB is logarithmically increasing with the time horizon, but does not match the regret lower bound (1). In addition, SIC-MMAB needs to know the time horizon in advance. More importantly, it requires involved communication phases (players need to exchange their estimates of the arms' mean rewards), and in turn, the number of collisions used to communicate grows large with T.

Our contributions. We present DPE1 (Decentralized Parsimonious Exploration), a simple policy that achieves the asymptotic fundamental regret limit (1). The policy relies on the observation that in a MAB problem where the decision-maker selects M arms in each round (a model referred to as MAB with multiple plays Anantharam et al. (1987)), an optimal algorithm consists in playing the $(M-1)$ best empirical arms and exploring using the remaining arm according to an optimal index policy, such as KL-UCB Garivier and Cappé (2011). This observation that such *parsimonious* exploration suffices was already made and exploited in Combes et al. (2015) for the design of learning-to-rank algorithms. It is powerful in the design of a decentralized MMAB algorithm: indeed, it implies that the exploration can be only performed by a single player, the so-called *leader*. The leader maintains the set of the M best empirical arms based on the rewards she received so far for the various arms. The other players, referred to as the *followers*, just need to play these best empirical arms greedily. To this aim, the leader just needs to inform the followers when the set of the M best empirical arms changes – and it can be done using collisions as proposed in Boursier and Perchet (2019).

Our finite-time analysis of the regret of DPE1 reveals that: for all $T \geq 3$ and any $0 < \delta < \delta_0 = \min_{1\leq k\leq K-1} \frac{\mu_k - \mu_{k+1}}{2}$:

$$\begin{aligned} R^{\mathrm{DPE1}}(T) \leq & \sum_{k>M} \frac{\mu_M - \mu_k}{\mathrm{kl}(\mu_k + \delta, \mu_M - \delta)} f(T) \\ & + K^2M^2\left[\frac{1}{K-M} + 284K^{1/2}M(7+\delta^{-2})\right], \end{aligned}$$

where $f(T) = \log(T) + 4\log\log(T)$. In particular, by letting first T tend to ∞ and then δ tend to 0, the above result implies that DPE1 is asymptotically optimal: its regret matches the regret lower bound (1). DPE1 achieves the regret of the best possible centralized algorithm. In addition, DPE1 is simpler than SIC-MMAB, since the leader just needs to communicate the indexes of the best empirical arms, when the latter change. In fact, the expected number of collisions used for communication – equivalently the number of communication bits (one may see a collision as communicating one bit) is finite (it is upper bounded by $K^2M^2\left[\frac{1}{(K-M)} + 242K^{1/2}(7+\delta^{-2})\right]$).

1.2 Multiplayer MAB without collision

In the absence of collisions, different players can select the same arm. When a player selects arm k in round t, she collects the reward $X_k(t)$[1]. In this model, the players are the vertices of a *communication* graph $G = (V, E)$. At the end of each round, a player can communicate to her neighbors in G. In recent papers Landgren et al. (2016, 2018); Martínez-Rubio et al. (2019), players are assumed to be able to communicate real numbers to their neighbors in each round. We study a more realistic setting where players can only send a finite number of bits per round. As for the model with collision, we are interested in distributed arm selection policies. Under such a policy π, a player i selects in round t arm $k_i^\pi(t)$ and design messages to be sent to neighbors depending on her past observations (the collected rewards and the messages received from her neighbors).

Regret lower bound. The maximum expected reward that can be collected in one round is $M\mu_1$. Hence

[1] When two arms select the same arm, we assume that they receive the same random rewards for simplicity. However, they could well collect stochastically independent rewards; the analysis of the average regret would not be affected.

the regret up to round T of a policy π is defined as:

$$R^{\pi}(T) = MT\mu_1 - \sum_{t=1}^{T}\sum_{i=1}^{M}\mathbb{E}[\mu_{k_i^{\pi}(t)}].$$

The regret of any uniformly good policy (centralized or not) should satisfy:

$$\liminf_{T\to\infty}\frac{R^{\pi}(T)}{\log(T)} \geq C_1(\mu) := \sum_{k>1}\frac{\mu_1-\mu_k}{\mathrm{kl}(\mu_k,\mu_1)}. \quad (2)$$

Indeed, the asymptotic regret above corresponds to the best possible regret of a single player.

State-of-the-art algorithm. In Martínez-Rubio et al. (2019), the authors present Distributed Delayed UCB (DD-UCB), an algorithm that combines UCB and a consensus algorithm. The latter is meant so that all players share similar estimates of the mean rewards of the arms, and requires that each player sends a few real numbers to her neighbours in each round. DD-UCB enjoys the following finite-time regret guarantee[2]:

$$R^{\mathrm{DDUCB}}(T) \leq c_1\sum_{k>1}\frac{\log(MT)}{\mu_1-\mu_k} + c_2M\log(M)\sum_{k>1}(\mu_1-\mu_k),$$

for some constants $c_1, c_2 > 0$. DD-UCB suffers from the same issues as SIC-MMAB for the MMAB with collisions: its regret does not match the regret lower bound (2), and it requires players to communicate a lot.

Our contributions. We present DPE2, an algorithm based on the same *parsimonious exploration* principle as for DPE1. The algorithm starts by electing a leader among the players. After this election, the leader is the only player exploring arms, again using KL-UCB indexes. The other players, the followers, just play the best empirical arm announced by the leader. We show that the regret of DPE2 satisfies: for all $T \geq 3$ and any $0 < \delta < \delta_0$:

$$R^{\mathrm{DPE2}}(T) \leq \sum_{k>1}\frac{\mu_1-\mu_k}{\mathrm{kl}(\mu_k+\delta,\mu_1-\delta)}f(T) + 9DKM(29+K\delta^{-2}),$$

where D is the diameter of the graph G. Hence, DPE2 achieves the regret of the best possible centralized algorithm. In addition, under DPE2, the expected number of bits used for communication is finite (it is upper bounded by $4DM^2\log_2(M) + 8KM^2D\log_2(K)(29 + K\delta^{-2})$).

[2]This upper bound is derived for subgaussian rewards, so it is valid for Bernoulli rewards as well.

2 Multiplayer MAB with collisions

This section is devoted to problems where players experience a collision when they select the same arm. We present the DPE1 algorithm, and analyze its performance.

2.1 The DPE1 algorithm

We provide a detailed description of DPE1 and explain its advantages over SIC-MMAB. DPE1 starts with an initialization phase whose objective is to assign different ranks in $\{1,\dots,M\}$ to players. This rank assignment will be used to avoid collisions. After the initialization phase, DPE1 alternates between exploitation and exploration as usual. However, exploration is conducted by one player only (the leader).

2.1.1 Initialization phase

The first phase consists of coordinating the players. After this phase, a single player becomes the *leader*; this player is ranked first and is aware of this rank. The other players are *followers* and get to know their respective ranks $2,\dots,M$. All players learn, in passing, the number of players M. After this phase, they can coordinate and avoid collisions except if they need collisions to communicate. SIC-MMAB also starts with such an initialization phase; this phase has by design a fixed duration $T_0 = \lceil K\log(T)\rceil$, which implies in particular that its cost in terms of expected regret is $KM\log(T)$. In contrast, the initialization phase in DPE1 has a random duration: it lasts until all its objectives are achieved. The expected duration of the DPE1 initialization phase is finite, and hence just generates a constant expected regret.

DPE1 initialization phase consists of two sub-phases:

A. Orthogonalization. This first sub-phase aims at assigning in a distributed manner M different arms within $\{1,\dots,K-1\}$ to the various players. In this sub-phase, the players maintain an internal state with values in $\{0,1\dots,K-1\}$: when the state is '0', it means that the player is not satisfied, and still needs to find a free arm. When the state is 'k' with $k \neq 0$, it means that the player manages to select arm k without collision, and she will keep this state until the end of the sub-phase. The sub-phase consists of a sequence of blocks of $K+1$ rounds: in the first round of a block, players with state different than '0' select the arm corresponding to their state, and players with '0' state randomly select an arm in $\{0,1\dots,K-1\}$. The K remaining rounds of the block are used to communicate the outcomes of the first round. This communication is done by selecting arm K and by observing collisions. More precisely, if a player is in state $k \neq 0$, then she

selects arm k except in the k-th round where she selects K. If a player is in state '0', she selects arm K in the K rounds. Note that as long as there is a player in state '0', collisions are experienced by all players in the K last rounds of the block. Hence, all the players know that all players are satisfied when no collision is experienced in a block. When such a block occurs for the first time, the sub-phase terminates, and all players are aware of this termination. We prove (see Appendix) that the expected duration of the orthogonalization phase does not exceed $\frac{M(K-1)(K+1)}{K-M}$ rounds.

B. Rank assignment. After the orthogonalization sub-phase, all the players have different states in $\{1,\dots,K-1\}$. The rank assignment sub-phase consists of $2K-2$ consecutive rounds, denoted by $t_1,\dots,t_{2K-2}$. Should a player be in state 'k', she selects arms in the following manner: (i) in a round $t_s \in \{t_1,\dots,t_{2k}\} \cup \{t_{K+k},\dots,t_{2K-2}\}$, she selects arm k; (ii) otherwise, when $t_s \in \{t_{2k+1},\dots,t_{K+k-1}\}$ she selects arm $s-k$, which corresponds to selecting sequentially the arms $k+1,\dots,K-1$.

It is easy to observe that with the above procedure, two players only collide once (see Appendix). Furthermore, to determine her rank, a player initially in state k just needs to count the number i_k of collisions experienced in the first $2k$ rounds. i_k+1 becomes her rank. The rank-1 player is the leader.

It should be noted that the leader-follower structure is also adopted in Tibrewal et al. (2019). There, however, the orthogonalization phase could well stop before players have distinct ranks.

2.1.2 Exploration-exploitation phase

In DPE1, the leader is responsible for exploring and maintaining the set of the M best empirical arms. Exploration is conducted using the following KL-UCB indexes. The index of arm k in round t is

$$b_k(t) = \sup\{q \geq 0 : N_k(t)\mathrm{kl}(\hat{\mu}_k(t), q) \leq f(t)\},$$

where $f(t) = \log(t) + 4\log\log(t)$, $N_k(t)$ denotes the number of times the leader has played arm k up to round t, and $\hat{\mu}_k(t)$ is the empirical average of arm k based on the rewards obtained before round t. The leader is also responsible for communicating to the followers when the set $\mathcal{M}(t)$ of the M best empirical arms changes. To this aim, she leverages collisions in the same manner as in SIC-MMAB. Each time $\mathcal{M}(t)$ changes, a communication phase is initiated by the leader, and this phase lasts a finite number of rounds. The algorithm is designed so that the expected number of times $\mathcal{M}(t)$ changes is finite. The followers just play different arms from $\mathcal{M}(t)$. Note that the followers do not need to communicate anything to the leader; in particular, the rewards they collect are not taken into account by the leader. Each communication phase has a fixed and finite duration and is known to all players – see Subsection 2.1.3 for detail. Hence without loss of generality, we ignore these periods of communication and we can assume that the leader communicates the new $\mathcal{M}(t)$ *instantaneously* whenever required. Communication phases will be, however, accounted for when deriving the regret of the algorithm.

The set of rounds is divided into blocks of MJ rounds where $J = \lceil K^{1/2} \rceil$. In rounds belonging to the same block, the empirical means of the arms, the KL-UCB indexes, and the set of best empirical arms are kept constant. More precisely, the decisions made in one block are based on: For each $k \in \{1,\dots,K\}$,

$$\hat{\nu}_k(t) = \hat{\mu}_k\Big(\Big\lfloor \frac{t}{MJ} \Big\rfloor MJ\Big), \quad d_k(t) = b_k\Big(\Big\lfloor \frac{t}{MJ} \Big\rfloor MJ\Big),$$
$$\text{and } \mathcal{N}(t) = \mathcal{M}\Big(\Big\lfloor \frac{t}{MJ} \Big\rfloor MJ\Big).$$

At the beginning of a block, the leader updates the above variables. The block structure is designed so that (i) the leader gathers J samples of each of the $(M-1)$ best empirical arms, and in expectation $J/2$ samples from the M-th best empirical arm; (ii) each follower selects each arm in $\mathcal{N}(t)$ J times, and (iii) the leader explores only when the followers play the $(M-1)$ best empirical arms. In particular, a block can be split into J sub-block: each sub-block consists of M consecutive rounds and the follower plays each of the M empirical best arms once in a sub-block. Now $\mathcal{K}$ is decomposed into J subsets $\mathcal{K}_j = \{k \in \mathcal{K} : (j-1)J < k \leq jJ\}$. In the j-th sub-block, we impose that the leader may explore an arm from $\mathcal{K}_j$ only. We introduce in DPE1 the J sub-blocks so as to optimize the constant term (w.r.t. T) of our regret upper bound. We could well run DPE1 without sub-blocks ($J=1$), it would be still asymptotically optimal, but the constant term in the regret upper bound would be multiplied by $\sqrt{K}$.

Next, we describe in detail the decisions taken by the leader and the followers under DPE1.

Leader. At the beginning of round t, if $t = 0 (\mathrm{mod}\ MJ)$, the leader updates the vectors $\hat{\nu}(t)$, $d(t)$, and $\mathcal{N}(t)$. The set $\mathcal{N}(t)$ is ordered: $\mathcal{N}(t) = \{\ell_1(t),\dots,\ell_M(t)\}$. This order is arbitrary, but independent of the empirical means of the arms. In particular, the order is kept fixed even if the relative empirical means of the arms in $\mathcal{N}(t)$ evolve, so that the leader only needs to communicate to the followers when $\mathcal{N}(t)$ changes. Ordering $\mathcal{N}(t)$ is important to avoid collisions. In the following, we denote by $\hat{M}(t)$ the arm in $\mathcal{N}(t)$ with the smallest empirical mean.

If $\mathcal{N}(t) \neq \mathcal{N}(t-1)$, the leader communicates to the followers the identity of the arm leaving the set and that of the new arm that replaces it in $\mathcal{N}(t)$ (the rank of the new arm inherits that of the arm that left).

The sequential arm selections made by the leader are as follows. In round t, define $m = [(t-1)(\text{mod } M)]+1$ and $j = \lfloor t/M \rfloor (\text{mod } J) + 1$. If $\ell_m(t) \neq \hat{M}(t)$, then the leader selects $\rho(t) = \ell_m(t)$. If $\ell_m(t) = \hat{M}(t)$, then with probability $1/2$, the leader selects arm $\hat{M}(t)$, and with probability $1/2$, the leader plays an arm $k \notin \mathcal{N}(t)$ such that $d_k(t) > \hat{\nu}_{\hat{M}(t)}$, should such an arm exists in $\mathcal{K}_j \backslash \mathcal{N}(t)$, and plays $\hat{M}(t)$ otherwise.

Followers. The followers just exploit the knowledge of the leader: they play greedily different arms of $\mathcal{N}(t)$. More precisely, the follower with rank $i \in \{1, \dots, M-1\}$ plays in round t the arm $\ell_{m_i}(t)$ where $m_i = [(t-1+i)(\text{mod } M)] + 1$.

The pseudo-code of the exploration-exploitation phase of the DPE1 algorithm is presented in Algorithm 1.

Algorithm 1: The DPE1 algorithm: Exploration-exploitation phase

Leader.
Initialization: Set $\hat{\nu}_k(1) = d_k(1) = 0$ for all k
Initialize the set of best empirical arms $\mathcal{N}(1)$ and $\hat{M}(1)$ arbitrarily
Communicate $\mathcal{N}(1) = \{\ell_1(1), \ell_2(1), \dots, \ell_M(1)\}$ to the followers
For round $t \geq 1$:
1. $m \leftarrow [(t-1)(\text{mod } M)] + 1$, $j \leftarrow \lfloor t/M \rfloor (\text{mod } J) + 1$
2. If $t > 1$ and $j = m = 1$, update $\hat{\nu}_k(t)$, $d_k(t)$ for each arm
   k, $\hat{M}(t)$, and update the ordered set
   $\mathcal{N}(t) \leftarrow \{\ell_1(t), \ell_2(t), \dots, \ell_M(t)\}$
   (the set of the M best empirical arms)
3. If $t > 1$ and $\mathcal{N}(t) \neq \mathcal{N}(t-1)$, communicate $\mathcal{N}(t)$ to the followers
4. For each $s = 1, 2, \dots, J$,
$\mathcal{D}_s(t) \leftarrow \left\{ k \notin \mathcal{N}(t) : d_k(t) \geq \hat{\nu}_{\hat{M}(t)}(t), \lceil \frac{k}{J} \rceil = s \right\}$,
If $\mathcal{D}_j(t) = \emptyset$ or $\ell_m(t) \neq \hat{M}(t)$, $\rho(t) \leftarrow \ell_m(t)$
Else
   w.p. $1/2$, $\rho(t) \leftarrow \hat{M}(t)$
   w.p. $1/2$, $\rho(t) \leftarrow k$ where $k \sim \mathcal{D}_j(t)$ uniformly
Select arm $\rho(t)$
Follower with rank $i \in \{1, 2, \dots, M-1\}$.
In round $t \geq 1$: $m_i \leftarrow [(t-1+i)(\text{mod } M)] + 1$, select arm $\ell_{m_i}(t)$

2.1.3 Communication phases

When $\mathcal{N}(t) \neq \mathcal{N}(t-1)$, the leader starts a communication of the new ordered set $\mathcal{M}(t)$ as follows. She uses 3 blocks: (i) a block of $M-1$ rounds to initiate a communication with the $M-1$ followers, (ii) a block of M rounds to inform the followers of the arm a^- to be removed from the list, and finally (iii) a block of K rounds to inform the followers of the arm a^+ to be added to the list. Hence a communication phase requires $K + 2M - 1$ rounds.

Initial block. In the first block of $(M-1)$ rounds, the leader sequentially selects the arms selected by the followers ranked $2, \dots, M-1$. The rank-i follower experiences a collision indicating the communication phase, and she knows the round when the communication phase started (she is aware of her rank).

Second block: Removing a^- from $\mathcal{N}(t-1)$. In the next M rounds, the leader selects arm a^-. Followers continue the arm selection as in the exploration-exploitation phase: thus, at some round in this block, they select a^- and collide with the leader, which indicates the arm to be removed.

Third block: Adding a^+ to $\mathcal{N}(t-1)$. During the final K rounds, the leader selects arm a^+. Followers change their arm selection, and select all arms during this block without colliding with each other (except with the leader). More precisely, in the m-th round of this block, the follower with rank i selects the arm $[(m+i)(\text{mod } K)]+1$. Each follower will experience a collision when selecting a^+, and they can add it to $\mathcal{N}(t-1)$.

2.2 Regret and communication complexity analysis

The next theorem provides a finite-time analysis of the regret of DPE1. It also gives an upper bound on the expected number of collisions involved in the algorithm, including those of the initialization and communication phases.

Theorem 1. *For any μ, $T \geq 3$, and $0 < \delta < \min_{1 \leq k \leq K-1}(\mu_k - \mu_{k+1})/2$, the regret of* DPE1 *satisfies:*

$$R^{\text{DPE1}}(T) \leq K^2 M^2 \left[\frac{1}{(K-M)} + 284 K^{1/2} M (7 + \delta^{-2}) \right] + \sum_{k > M} \frac{(\mu_M - \mu_k) f(T)}{\text{kl}\,(\mu_k + \delta, \mu_M - \delta)}.$$

Irrespective of the time horizon T, the expected number of collisions under DPE1 *is upper bounded by:*

$$K^2 M^2 \left[\frac{1}{(K-M)} + 242 K^{1/2} (7 + \delta^{-2}) \right].$$

By letting T tend to ∞, and then δ tend to 0 in the above regret upper bound, we simply deduce that DPE1 is asymptotically optimal:

$$\limsup_{T\to\infty} \frac{R^{\mathrm{DPE1}}(T)}{\log T} \leq \sum_{k>M} \frac{\mu_M - \mu_k}{\mathrm{kl}\left(\mu_k, \mu_M\right)}.$$

To establish Theorem 1, we prove that the expected number of rounds where $\mathcal{M}(t)$ does not correspond to the actual M best arms is finite. This is the key ingredient whose proof actually exploits the arguments used in Combes et al. (2015) to establish a regret upper bound of a centralized algorithm for some MAB problems with multiple plays. From this result, we know that the number of communication phases is finite in expectation, as well as the average regret experienced by the followers. The last term of the regret upper bound just corresponds to the regret paid by the leader.

3 Multiplayer MAB without collisions

This section is devoted to MMAB problems without collisions: players can select the same arm and collect the corresponding rewards. Players sit on the vertices of a graph G, and are assumed to be able to send messages to their neighbors in G in each round. We present DPE2, an algorithm similar to DPE1 and adapted to this new setting.

3.1 The DPE2 algorithm

DPE2 starts with a leader election phase. After the election, DPE2 alternates between exploration and exploitation. As in DPE1, under DPE2, the exploration is conducted by the leader only. The followers just play the best empirical arm seen by the leader. The latter needs to communicate to the followers only when her best empirical arm changes.

3.1.1 Leader election

The leader election problem is a well studied problem in the field of distributed computing Fokkink (2013); Tel (1994); Casteigts et al. (2019). To simplify, we assume here that each player has initially a unique id in $\{1, \ldots, P\}$, where P could be potentially very large[3]. For the analysis of the communication complexity of the algorithm, we will take $P = M$ for simplicity. With this assumption, leader election can be performed in $O(D + \log M)$ rounds using $\mathcal{O}(1)$ bits per messages Casteigts et al. (2019), where D denotes the diameter of the graph. We propose a simpler alternative election process whose duration is enough for our purposes.

Every player initializes her state to her id. For $(D+1)$ consecutive rounds, every player sends her state to all her neighbours. When a player receives states from other players, she updates her state to the minimal value of the states received and her state. After $(D+1)$ rounds, the player whose state corresponds to her initial id is the leader. Note that the total number of bits sent during this procedure is at most $2|E|(D+1)\lceil \log_2 M \rceil$ (where $|E|$ is the number of edges in G).

3.1.2 Exploration-exploitation phase

In round t, the leader maintains the set $\mathcal{D}(t)$, which consists of those arms with a larger KL-UCB index than that of the best empirical arm $\ell_1(t)$. When this set is empty, she plays $\ell_1(t)$ and updates the empirical means $\hat{\nu}_k(t)$ of the arms and KL-UCB $d_k(t)$ indexes of the arms as well as $\mathcal{D}(t+1)$. She also communicates to her neighbours if $\ell_1(t)$ has changed.
If the set $\mathcal{D}(t)$ is not empty, a block of rounds starts. In the first round of this block, the leader play $\ell_1(t)$; in the subsequent rounds, she plays the arms in $\mathcal{D}(t)$ until this set is exhausted, which then ends the block.

Algorithm 2: The DPE2 algorithm: Exploration-exploitation phase

Initialization: Set $\hat{\nu}_k(1) = d_k(1) = 0$ for all k, $\mathcal{D}(1) = \emptyset$, and $s = 0$. Initialize best empirical arm $\ell_1(1)$ arbitrarily.
For round $t \geq 1$:
Leader.
If $\mathcal{D}(t) = \emptyset$,
  If $t > 1$, update $\hat{\nu}_k(t)$, $d_k(t)$ for all k, and $\ell_1(t)$
    $\mathcal{D}(t) \leftarrow \left\{k \neq \ell_1(t) : d_k(t) \geq \hat{\nu}_{\ell_1(t)}(t)\right\}$,
    $s \leftarrow \mathbf{1}_{\{|\mathcal{D}(t)|>0\}}$
    If $\ell_1(t) \neq \ell_1(t-1)$, communicate $\ell_1(t)$ to the followers
  $\rho(t) \leftarrow \ell_1(t)$, select arm $\rho(t)$
Else
  If $s = 1$, $\rho(t) \leftarrow \ell_1(t)$, $s \leftarrow 0$, select arm $\rho(t)$
  Else, $\rho(t) \leftarrow \arg\max_{k\in\mathcal{D}(t)} d_k(t)$, select arm $\rho(t)$
    $\mathcal{D}(t+1) \leftarrow \mathcal{D}(t) \setminus \{\rho(t)\}$
Follower.
  Select arm $\ell_1(t)$

3.1.3 Communication phases

When $\ell_1(t-1) \neq \ell_1(t)$, the leader sends the new best empirical arm to all her neighbours. In every round, each time a follower receives the id of an arm that does not correspond to the arm she is playing: (i) she starts

[3] This assumption is mild compared to those made in recent papers, e.g. Martínez-Rubio et al. (2019), addressing the MMAB problem without collision. There, players can pass real numbers to their neighbours, in which case the leader election becomes trivial. Note also that anonymous leader election can be also performed in finite expected time Fokkink (2013).

playing the new arm from the next round, and (ii) forward the id of the new arm to her neighbours except to the player from whom she received the information.

Each communication phase takes at most D rounds, and the total number of bits transmitted during such a phase does not exceed $2|E|\lceil \log_2(K) \rceil$.

3.2 Regret and communication complexity analysis

The next theorem provides a finite-time analysis of the regret of DPE2, and an upper bound on the expected amount of communication involved in the algorithm.

Theorem 2. *For any μ, $T \geq 3$, and $0 < \delta < \min_{1 \leq k \leq K-1}(\mu_k - \mu_{k+1})/2$, the regret of DPE2 satisfies:*

$$R^{\mathrm{DPE2}}(T) \leq 9MDK(29 + K\delta^{-2}) + \sum_{k>1} \frac{\mu_1 - \mu_k}{\mathrm{kl}(\mu_k + \delta, \mu_1 - \delta)} f(T).$$

The total expected number of bits sent under DPE2 is lower than:

$$4DM^2 \log_2(M) + 8KM^2 D \log_2(K)(29 + K\delta^{-2}).$$

The above theorem implies that DPE2 is asymptotically optimal. Its proof relies on similar arguments as those used to establish Theorem 1.

4 Numerical experiments

In this section, we provide initial experiments to illustrate the performance of DPE1 and DPE2. Further experiments are presented in the appendix.

4.1 DPE1

The experiments we run follow the same setting as in Boursier and Perchet (2019). We consider $K = 9$ arms with Bernoulli rewards, and a fixed number of players $M = 6$. We compare the regret of DPE1 with those of SIC-MMAB Boursier and Perchet (2019) and MCTopM Besson and Kaufmann (2018). All the regret and communication complexity values are averaged over 208 runs.

Figure 1 compares DPE1, MCTopM, and SIC-MMAB, over a time horizon of $T = 5 \cdot 10^5$ rounds. The means of the arms are linearly distributed between 0.9 and 0.89, so the gap is approximately $\Delta = 1.1 \cdot 10^{-3}$. DPE1 significantly outperforms both SIC-MMAB and MCTopM.

In Figure 2, we plot the communication complexity vs. time, i.e., the average number of communication

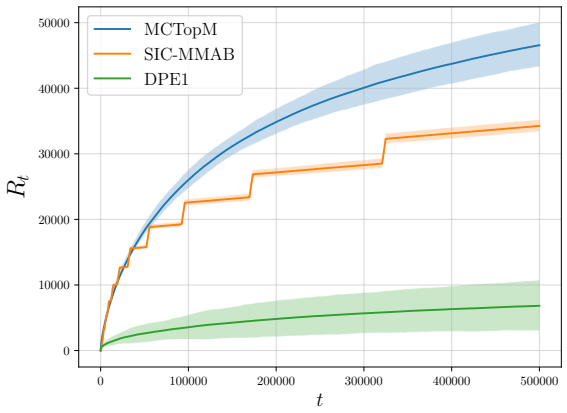

Figure 1: Regret evolution over a horizon of $5 \cdot 10^5$ rounds. The continuous curves represent the empirical average value, and the shadowed area 3 times the standard deviation.

phases under DPE1 over time. Figure 2 confirms our theoretical insights: the expected number of times the set $\mathcal{M}(t)$ gets updated is finite. We show the communication complexity for several values of the gap Δ between two consecutive arms, keeping the average reward of the best arm equal to 0.9. Note that the communication complexity increases as Δ approaches 0. In the appendix we also show the average number of communication phases as a function of the gap Δ.

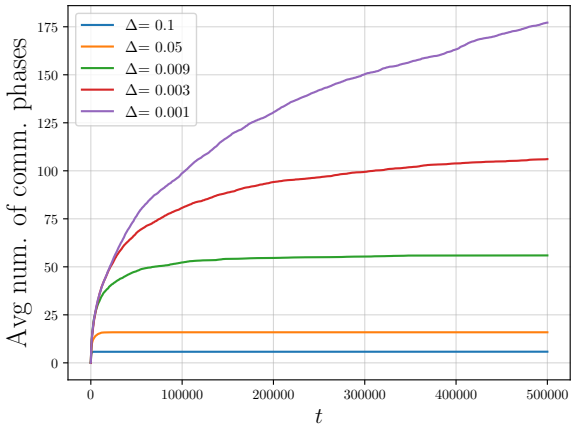

Figure 2: Communication complexity, i.e, the expected number of communication phases. For smaller value of Δ, it is harder for the leader to identify the best arm, and the number of communication phases increases.

4.2 DPE2

To evaluate DPE2, we used the same communicating graph as in Landgren et al. (2016). The graph has 4

players – 3 players are connected to each other, and a fourth player is only connected to one player only. We consider 10 arms with Bernoulli rewards. The range of the rewards is $[0.1, 0.9]$, with a gap of $\Delta = 8/9$ between two consecutive arms. In our experiments, we noticed that the identity of the leader (for this scenario) does not impact regret significantly, so we choose one player arbitrarily as the leader. We use a time horizon of $T = 5 \cdot 10^4$ rounds, and average regrets over 1024 runs.

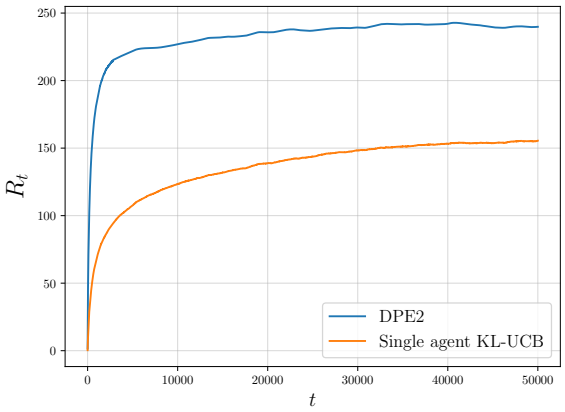

Figure 3: Regret evolution of a time horizon of $5 \cdot 10^4$ rounds. In *blue* is shown DPE2, whilst in *orange* is shown the regret of a single agent running KL-UCB.

In Figure 3, we compare the regret obtained under DPE2 to the regret that would be obtained by a single player implementing KL-UCB. The regret of DPE2 is not really worse than the regret experienced by a single player. The communication complexity of DPE2 is numerically illustrated in the appendix, and as predicted by our analysis, it is rather small.

5 Related Work

MMAB problems with collisions have attracted a lot of attention over the last decade due to their applicability to the design of decentralized channel selection schemes in cognitive radio systems. Early works include Liu and Zhao (2010) and Anandkumar et al. (2011), proposing various algorithms with regret guarantees far from the best possible regret achieved by DPE1. Even in more recent papers such as Rosenski et al. (2016) and Besson and Kaufmann (2018), the regret upper bounds were actually bigger than expressions of the form: $M \sum_{k>M} \frac{\log(T)}{\mu_M - \mu_k}$. The multiplicative factor M was considered unavoidable, until Boursier and Perchet Boursier and Perchet (2019) proposed to actually use collisions as a way to communicate. In turn, SIC-MMAB exploits collisions to share the estimated mean rewards of arms. The idea of exploiting collisions had also been suggested in Lugosi and Mehrabian (2018).

As already mentioned in the introduction, SIC-MMAB was, until now, the algorithm with the best regret guarantees. Our algorithm, DPE1, is simpler, has better regret guarantees, and as our numerical experiments suggest, outperforms other algorithms. Note that the leader-follower framework used by DPE1 had been also proposed in the algorithms presented in Hanawal and Darak (2018), but the latter exhibit worse regret guarantees than DPE1.

The literature on MMAB problems without collisions is not as abundant. For the case of subgaussian rewards with the same and known variance, Landgren et al. (2016) propose COOP-UCB, an algorithm where each player maintains an estimate of the mean reward of each arm using information obtained by their neighbours via a running consensus scheme Braca et al. (2008). Martínez-Rubio et al. (2019) recently managed to develop a similar algorithm, DD-UCB, with better regret guarantees than that of COOP-UCB. Under both COOP-UCB and DD-UCB, the players need to communicate real numbers to their neighbours (to run a consensus scheme). DPE2, our algorithm, has better regret guarantees, and keeps the total expected number of bits communicated finite.

6 Conclusion

In this paper, we have studied two Multiplayer MAB problems, where we were able to devise decentralized algorithms that achieve the same regret as the one we would obtain by using an optimal centralized algorithm. The design of these algorithms leveraged the critical observation that minimal regret can be achieved by letting only one player exploring arms, and by allowing other players to select the best empirical arms greedily. This observation also implies that the player exploring arms just needs to inform other players regarding the change of the best empirical arms. Moreover, our algorithms do require very little communication: the expected number of bits used for communication does not depend on the time horizon. Further research includes the case of heterogeneous rewards, i.e., the case where the average reward of an arm depends on the player selecting that arm. On the other hand, it would be interesting to further investigate scenarios where players sit on the vertices of a graph. For instance, we could consider that players who are neighbors in the graph collide (collision graphs are instrumental in modelling radio communication systems). In this case, several players would need to explore.

References

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

Anantharam, V., Varaiya, P., and Walrand, J. (1987). Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part i: I.i.d. rewards. *IEEE Transactions on Automatic Control*, 32(11):968–976.

Besson, L. and Kaufmann, E. (2018). Multi-player bandits revisited. *Algorithmic Learning Theory*.

Boursier, E. and Perchet, V. (2019). SIC-MMAB: synchronisation involves communication in multiplayer multi-armed bandits. In *Advances in Neural Information Processing Systems 32*, pages 12048–12057.

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

Casteigts, A., Métivier, Y., Robson, J. M., and Zemmari, A. (2019). Deterministic leader election takes $\Theta(D + \log(n))$ bit rounds. *Algorithmica*, 81(5):1901–1920.

Combes, R., Magureanu, S., Proutiere, A., and Laroche, C. (2015). Learning to rank: Regret lower bounds and efficient algorithms. *SIGMETRICS Perform. Eval. Rev.*, 43(1):231–244.

Fokkink, W. (2013). *Distributed algorithms: an intuitive approach*. MIT Press.

Garivier, A. and Cappé, O. (2011). The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory*, pages 359–376.

Hanawal, M. K. and Darak, S. J. (2018). Multi-player bandits: A trekking approach. *arXiv preprint arXiv:1809.06040*.

Jouini, W., Ernst, D., Moy, C., and Palicot, J. (2009). Multi-armed bandit based policies for cognitive radio's decision making issues. In *2009 3rd International Conference on Signals, Circuits and Systems (SCS)*, pages 1–6. IEEE.

Lai, T. L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22.

Landgren, P., Srivastava, V., and Leonard, N. E. (2016). Distributed cooperative decision-making in multiarmed bandits: Frequentist and bayesian algorithms. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 167–172. IEEE.

Landgren, P., Srivastava, V., and Leonard, N. E. (2018). Social imitation in cooperative multiarmed bandits: Partition-based algorithms with strictly local information. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 5239–5244. IEEE.

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

Lugosi, G. and Mehrabian, A. (2018). Multiplayer bandits without observing collision information. *arXiv preprint arXiv:1808.08416*.

Martínez-Rubio, D., Kanade, V., and Rebeschini, P. (2019). Decentralized cooperative stochastic bandits. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32*, pages 4529–4540. Curran Associates, Inc.

Rosenski, J., Shamir, O., and Szlak, L. (2016). Multiplayer bandits–a musical chairs approach. In *International Conference on Machine Learning*, pages 155–163.

Tel, G. (1994). Network orientation. *International Journal of Foundations of Computer Science*, 5(01):23–57.

Tibrewal, H., Patchala, S., Hanawal, M. K., and Darak, S. J. (2019). Multi-player bandits for optimal assignment with heterogeneous rewards.