Game of Thrones: Fully Distributed Learning for Multi-Player Bandits

Ilai Bistritz and Amir Leshem

Abstract

We consider an N-player multi-armed bandit game where each player chooses one out of M arms for T turns. Each player has different expected rewards for the arms, and the instantaneous rewards are independent and identically distributed or Markovian. When two or more players choose the same arm, they all receive zero reward. Performance is measured using the expected sum of regrets, compared to optimal assignment of arms to players that maximizes the sum of expected rewards. We assume that each player only knows her actions and the reward she received each turn. Players cannot observe the actions of other players, and no communication between players is possible. We present a distributed algorithm and prove that it achieves an expected sum of regrets of near-$O(\log T)$. This is the first algorithm to achieve a near order optimal regret in this fully distributed scenario. All other works have assumed that either all players have the same vector of expected rewards or that communication between players is possible.

I. Introduction

In online learning problems, an agent needs to learn on the run how to behave optimally. The crux of these problems is the tradeoff between exploration and exploitation. This tradeoff is well captured by the multi-armed bandit problem, which has attracted enormous attention from the research community [2]–[8]. In the multi-armed bandit problem, on each turn, for a total of T turns, an agent has to choose to pull one of the arms of several slot machines (bandits). Each arm provides a stochastic reward with a distribution which is unknown to the agent. The agent's performance is measured by the expected difference between the sum of rewards and the sum of rewards she could have achieved if she knew the statistics of the machines. In their seminal work, [9] proved that the best policy achieves a regret of $O(\log T)$.

Recently, there has been growing interest in the case of the multi-player multi-armed bandit. In the multi-player scenario, the nature of the interaction between the players can take many forms. Players may want to solve the problem of finding the best arms as a team [10]–[16], or may compete over the arms as resources they all individually require [17]–[26]. The idea of regret in the competitive multi-player multi-armed bandit problem is the expected sum of regrets and is defined as the performance loss compared to the optimal assignment of arms to players. The rationale for this notion of regret is formulated from the designer's perspective, who wants the distributed system of individuals to converge to a globally good solution.

Many works have considered a scenario where all the players have the same expectations for the rewards of all arms. Some of these works assume that communication between players is possible [20]–[22], [24], whereas others consider a fully distributed scenario [17], [23], [25].

One of the main reasons for studying resource allocation bandits has to do with their applications in wireless networks. In these scenarios, the channels are interpreted as the arms and the channel gain (or signal to noise ratio) as the arm's reward. However, since users are scattered in space, the physical reality dictates that different arms have different expected channel gains for different players. Different users having different preferences is of course the typical case in many other resource allocation scenarios as well.

This essential generalization for a matrix of expected rewards introduces the famous assignment problem [27]. Achieving a sublinear expected total regret in a distributed manner requires a distributed solution to the assignment problem, which has been explored in [28]–[30]. This generalization was first considered in [19], and later enhanced in [18], where an algorithm that achieves an expected sum of regrets of near-$O(\log T)$ was presented. However, this algorithm requires communication between players. It is based on the distributed auction algorithm in [28], which is not fully distributed. It requires that players can observe the bids of other players. This was possible in [18], [19] since it was assumed that the players could observe the actions of other players, which allowed them to communicate by using arm choices as a signaling method. The work in [31] suggested an algorithm that only assumes users can sense all channels without knowing which channels were chosen by whom. This algorithm requires less communication than [18], but has no regret guarantees. If the multi-armed bandit problem is relaxed such that the expected rewards are multiplies of some constant and players can choose not to transmit but instead only to sense a chosen channel, then it is shown in [32] that a regret of $O(\log T)$ can be achieved.

This research was supported by the Israeli Ministry of Science and Technology under grant 3-13038 and by a joint ISF-NRF research grant number 2277/16. A paper with a preliminary version of this study ([1]) was accepted to the thirty-second Conference on Neural Information Processing Systems (NIPS 2018).

In wireless networks, assuming that each user can hear all the other transmissions (fully connected network) is very demanding in practice. It requires a large sensing overhead or might simply be impossible due to the geometry of the network (e.g., exposed and hidden terminals). In a fully distributed scenario, players only have access to their previous actions and rewards. However, to date there is no completely distributed algorithm that converges to the exact optimal solution of the assignment problem. The fully distributed multi-armed bandit problem remains unresolved.

Our work generalizes [17] for different expectations for different players and [18], [19], [31] for a fully distributed scenario with no communication between players.

Recently, powerful payoff-based dynamics were introduced by [33], [34]. These dynamics only require each player to know her own action and the reward she received for that action. The dynamics in [33] guarantee that the Nash equilibrium (NE) with the best sum of utilities strategy profile will be played a sufficiently large portion of the time. The dynamics in [34] guarantee that the optimal sum of utilities strategy profile will be played a sufficiently large portion of the time, even if it is not a NE. In [35], equipping the dynamics of [34] with a decreasing exploration rate sequence was shown to provide a convergence in probability guarantee to the optimal sum of utilities solution. However, no explicit probability of convergence in a specific finite time was provided, which is essential for regret computation. Nevertheless, the crucial issue of applying these results to our case is that they all assume interdependent games. Interdependent games are games where each group of players can always influence at least one player from outside this group. In the multiplayer multi-armed bandit collision model, this does not hold. A player who shares an arm with another receives zero reward. Nothing that other players (who chose other arms) can do will change this.

In this paper, we suggest novel modified dynamics that behave similarly to [34], but in our non-interdependent game. These dynamics guarantee that the optimal solution to the assignment problem is played a considerable amount of the time. We present a fully distributed multi-player multi-armed bandit algorithm for the resource allocation and collision scenario, based on these modified dynamics. By fully distributed we mean that players only have access to their own actions and rewards. To the best of our knowledge, this is the first algorithm that achieves a near-optimal ($O(\log T)$) expected sum of regrets with a matrix of expected rewards and no communication at all between players, or equivalently, a game where players cannot observe the actions of other players.

A preliminary version of this paper appeared in [1]. The algorithm and analysis in this paper substantially extend the results as follows:

1) The total regret is improved to near-$O(\log T)$ instead of near-$O\left(\log^2 T\right)$ of [1], (almost) closing the gap with the lower bound of $O(\log T)$.
2) A tighter stationary distribution analysis of the GoT phase (in Appendix A) allows for choosing a much better parameter c for the algorithm, improving the convergence time significantly.
3) The stochastic arm rewards can be unbounded or Markovian, and not only bounded i.i.d. variables as in [1].

A. *Outline*

Section 2 formalizes the multi-player multi-armed bandit resource allocation game. Section 3 describes our fully distributed Game of Thrones (GoT) algorithm. In the first phase of every epoch, players explore in order to estimate the expectations of the arm rewards. This phase is analyzed in Section 4. In the second phase of every epoch, players use our modified dynamics and play the optimal solution most of the time. This is analyzed in Section 5. In the third and final phase of every epoch, players play the action they played most of the time in the recent GoT phases. Section 6 generalizes our main result to the case of Markovian rewards. Section 7 demonstrates our algorithm's performance using numerical experiments and Section 8 concludes the paper.

II. Problem Formulation

We consider a stochastic game with the set of players $\mathcal{N} = \{1, ..., N\}$ and a finite time horizon T. The horizon T is not known in advance by any of the players. The discrete turn index is denoted by t. The strategy space of each player is a set of M arms denoted by $\mathcal{A}_n = \{1, ..., M\}$ for each n. We assume that $M \geq N$, such that an allocation without collisions is possible. At each turn t, all players simultaneously pick one arm each. The arm that player n chooses at time t is denoted by $a_n(t)$ and the strategy profile at time t is $\boldsymbol{a}(t)$. Players do not know which arms the other players chose, and need not even know how many other players there are.

Define the set of players that chose arm i in strategy profile $\boldsymbol{a}$ as $\mathcal{N}_i(\boldsymbol{a}) = \{n \mid a_n = i\}$. The no-collision indicator of arm i in strategy profile $\boldsymbol{a}$ is defined as

$$\eta_i(\boldsymbol{a}) = \begin{cases} 0 & \left|\mathcal{N}_i(\boldsymbol{a})\right| > 1 \\ 1 & o.w. \end{cases} . \tag{1}$$

The assumptions on the stochastic rewards are summarized as follows. We also study the case of Markovian rewards in Section VI.

Definition 1. The sequence of rewards $\{r_{n,i}(t)\}_{t=1}^{T}$ of arm i for player n is i.i.d. with expectation $\mu_{n,i} > 0$ and variance $\sigma_{n,i}^2$ such that:

1) The distribution of $r_{n,i}(t)$ is continuous for each n, i.
2) For some positive parameter $b_{n,i}$ we have $E\left\{|r_{n,i}(t) - \mu_{n,i}|^k\right\} \leq \frac{1}{2}k!\sigma_{n,i}^2 b_{n,i}^{k-2}$ for all integers $k \geq 3$.
3) The sequences $\{r_{n,i}(t)\}_t$ are independent for different n or different i.

The family of distributions that satisfies the second condition (known as Bernstein's condition, see [36]) includes, among many others, the normal and Laplace distributions and, trivially, any bounded distribution. The instantaneous utility of player n in strategy profile $\boldsymbol{a}(t)$ at time t is

$$\upsilon_n(\boldsymbol{a}(t)) = r_{n,a_n(t)}(t)\,\eta_{a_n(t)}(\boldsymbol{a}(t)). \tag{2}$$

Our goal is to design a distributed algorithm that minimizes the expected total regret, defined next.

Definition 2. Denote the expected utility of player n in the strategy profile $\boldsymbol{a}$ by $g_n(\boldsymbol{a}) = E\{\upsilon_n(\boldsymbol{a})\}$. The total regret is defined as the random variable

$$R = \sum_{t=1}^{T}\sum_{n=1}^{N}\upsilon_n(\boldsymbol{a}^*) - \sum_{t=1}^{T}\sum_{n=1}^{N} r_{n,a_n(t)}(t)\,\eta_{a_n(t)}(\boldsymbol{a}(t)) \tag{3}$$

where

$$\boldsymbol{a}^* \in \arg\max_{\boldsymbol{a}} \sum_{n=1}^{N} g_n(\boldsymbol{a}). \tag{4}$$

The expected total regret $\bar{R} \triangleq E\{R\}$ is the average of (3) over the randomness of the rewards $\{r_{n,i}(t)\}_t$, that dictate the random actions $\{a_n(t)\}$.

The problem in (4) is no other than the famous assignment problem [27] on the $N \times M$ matrix of expectations $\{\mu_{n,i}\}$. In this sense, our problem is a generalization of the distributed assignment problem to an online learning framework.

Assuming continuously distributed rewards is well justified in wireless networks. Given no collision, the quality of an arm (channel) always has a continuous measure like SNR or a channel gain. Since the probability of zero reward in a non-collision is zero, players can safely deduce their non-collision indicator and rule out collisions in their estimation of the expected rewards. In Section VI, we extend our results to discrete Markovian rewards, that include i.i.d. discrete rewards as a special case. This reflects a case where each user can operate in one out of a finite number of qualities of service.

In the case where the probability for a zero reward is not zero, we can assume instead that each player can observe her collision indicator in addition to her reward. Knowing whether other players chose the same arm is a very modest requirement compared to assuming that players can observe the actions of other players.

According to the seminal work [9], the optimal regret of the single-player case is logarithmic; i.e., $O(\log T)$. The multiple players do not help each other; hence, we expect the expected total regret lower bound to be logarithmic at best, as shown by the following proposition.

Proposition 3. *The expected total regret is at least* $\Omega(\log T)$.

Proof: Assume that for $N > 1$ there is a policy that results in a total expected regret lower than $\Omega(\log T)$. Some player with given expected rewards, denoted player n, can simulate $N-1$ other players and set their expected rewards such that she gets her best arm at the optimal matching, and all other players have the same expected reward for all arms. In this case, her personal term in the sum of regrets in (3) is her single player regret, and is lower than the total sum which is lower than $\Omega(\log T)$. This player can generate the rewards of other players at random, all of which are independent of the actual rewards she receives. This player also simulates the policies for other players, and even knows when a collision occurred for herself and can assign zero reward in that case. Hence, simulating $N-1$ fictitious players is a valid single player multi-armed bandit policy that violates the $\Omega(\log T)$ bound, which is a contradiction. We conclude that this bound is also valid for $N > 1$. ■

III. Game of Thrones Algorithm

When all players have the same arm expectations, the exploration phase is used to identify the N best arms. Once the best arms are identified, players need to coordinate to be sure that each of them will sit on a different "chair" (see the Musical Chairs algorithm in [17]). When players have different arm expectations, a non-cooperative **game** is induced where the estimated expected rewards serve as utilities. In this game, players cannot sit on an ordinary chair without causing a linear regret, and must strive for a single **throne**. This throne is the arm they must play in the allocation that maximizes the sum of the expected rewards of all players. Any other solution will result in linear (in T)

expected total regret. Note that our assignment problem has a unique optimal allocation with probability 1 (as shown in Lemma 11).

The total time needed for exploration increases with T since the cost of being wrong becomes higher. When T is known by the players, a long enough exploration can be accomplished at the beginning of the game. In order to maintain the right balance between exploration and exploitation when T is not known in advance to the players, we divide the T turns into epochs, one starting immediately after the other. Each epoch is further divided into three phases. In the k-th epoch:

1) **Exploration Phase -** this phase has a length of $c_1 k^\delta$ turns for an arbitrary positive integer c_1 and a positive δ. The goal of this phase is to estimate the expectation of the arms. This phase is described in detail and analyzed in Section IV. It adds a $O\left(\log^{1+\delta} T\right)$ to the expected total regret.
2) **Game of Thrones (GoT) Phase -** this phase has a length of $c_2 k^\delta$ turns for an arbitrary positive integer c_2 and a positive δ. In this phase, players play a non-cooperative game with the estimated expectations from the exploration phase as their deterministic utilities. They choose their action at random according to payoff-based dynamics that always assign a positive probability for exploration. These dynamics induce a perturbed (ergodic) Markov chain that tends to visit the optimal sum of utilities (in (4)) state more often than other states. When the exploration rate is low enough, the optimal state has a probability greater than $\frac{1}{2}$ in the stationary distribution of the chain. Hence, all players play the optimal action most of the time and can agree on the optimal state distributedly. These arguments are described in detail and analyzed in Section V. This phase adds a $O\left(\log^{1+\delta} T\right)$ to the expected total regret. The GoT dynamics are described in the next subsection.
3) **Exploitation Phase -** this phase has a length of $c_3 2^k$ turns for an arbitrary positive integer c_3. During this phase, each player plays the most frequent action she has played during the last $\left\lfloor \frac{k}{2} \right\rfloor + 1$ GoT phases combined. Since the error probability of the exploration phase is kept small, it is very likely that all the last $\left\lfloor \frac{k}{2} \right\rfloor + 1$ exploration phases agree on the same optimal state. Hence, by playing their most frequent action, players are highly likely to play the optimal strategy profile that has a stationary distribution of more than $\frac{1}{2}$ in all the last $\left\lfloor \frac{k}{2} \right\rfloor + 1$ GoT Markov chains. This phase adds a vanishing term (with T) to the expected total regret.

This division into epochs is depicted in Fig. 1. The GoT algorithm is described in Algorithm 1. We note that the constants c_1, c_2, c_3 can be arbitrarily chosen and do not need to satisfy any condition. Their only role is to adjust the length of the initial three phases, which is only of interest for numerical or practical reasons (see Section VII).

A. Game of Thrones Dynamics

The core of our algorithm is the GoT dynamics of the second phase, where players play the game of thrones with the utility function $u_n(\boldsymbol{a})$ for player n. Denote the optimal objective by J_1 and $J_0 = \sum_n u_{n,\max}$ for $u_{n,\max} = \max_{\boldsymbol{a}} u_n(\boldsymbol{a})$. Let $c \geq J_0 - J_1$. Each player n has a personal state $S_n = \{C, D\}$ where C is content and D is discontent. Each player keeps a baseline action $\overline{a}_n$. In each turn during the GoT phase:

- A content player has a very small probability of deviating from her current baseline action:

$$p_n^{a_n} = \begin{cases} \frac{\varepsilon^c}{\left|\mathcal{A}_n\right| - 1} & a_n \neq \overline{a}_n \\ 1 - \varepsilon^c & a_n = \overline{a}_n \end{cases} . \tag{5}$$

- A discontent player chooses an action uniformly at random; i.e.,

$$p_n^{a_n} = \frac{1}{\left|\mathcal{A}_n\right|}, \forall a_n \in \mathcal{A}_n. \tag{6}$$

The transitions between C and D are determined as follows:

- If $\overline{a}_n = a_n$ and $u_n > 0$, then a content player remains content with probability 1:

$$[\overline{a}_n, C] \to [\overline{a}_n, C] . \tag{7}$$

- If $\overline{a}_n \neq a_n$ or $u_n = 0$ or $S_n = D$, then the state transition is (C/D denoting either C or D):

$$[\overline{a}_n, C/D] \to \begin{cases} [a_n, C] & \frac{u_n}{u_{n,\max}} \varepsilon^{u_{n,\max} - u_n} \\ [a_n, D] & 1 - \frac{u_n}{u_{n,\max}} \varepsilon^{u_{n,\max} - u_n} \end{cases} . \tag{8}$$

We now formulate our main result on the regret of our algorithm, which is generalized to Markovian rewards in Section VI. The proof of Theorem 14 provides an explicit requirement on the parameter ε.

Theorem 4 (Main Theorem)**.** *Assume i.i.d. rewards $\{r_{n,i}(t)\}_t$ with positive expectations $\{\mu_{n,i}\}$, as in Definition 1. Let the game have a finite horizon T, unknown to the players. Let each player play according to Algorithm 1, with any*

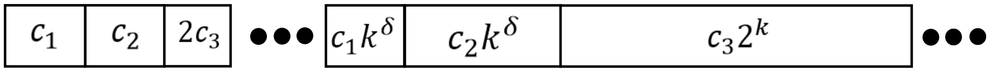


Fig. 1: Epoch structure. Depicted are the first and the k-th epochs.

integers $c_1, c_2, c_3 > 0$ *and* $\delta > 0$. *Then there exists a small enough* ε *such that for large enough* T, *the expected total regret is bounded by*

$$\bar{R} \leq 4\left(\max_{n,i}\mu_{n,i}\right)(c_1+c_2)\,N\log_2^{1+\delta}\left(\frac{T}{c_3}+2\right) = O\left(\log^{1+\delta}T\right). \tag{9}$$

Proof: Let $\delta > 0$. Let k_0 be the index of a sufficiently large epoch. We now bound the expected total regret of epoch $k > k_0$, denoted by $\bar{R}_k$. Define E_k as the event where the k-th exploitation phase does not consist of playing the optimal assignment $\boldsymbol{a}^*$. We have

$$\Pr(E_k) \leq \Pr\left(\bigcup_{r=0}^{\left\lfloor\frac{k}{2}\right\rfloor} P_{e,k-r}\right) + P_{c,k} \tag{10}$$

where $P_{e,k}$ is the probability that the k-th exploration phase results in an estimation for which the optimal assignment is different from $\boldsymbol{a}^*$, and $P_{c,k}$ is the probability that $\boldsymbol{a}^*$ was not the most frequent state in the last $\left\lfloor\frac{k}{2}\right\rfloor+1$ GoT phases combined. If neither of these failures occurred, then $\boldsymbol{a}^*$ is played in the k-th exploitation phase, which establishes (10). Only under E_k will this exploitation phase contribute to the total regret. Let $\mu_{\max} = \max_{n,i}\mu_{n,i}$. For $k > k_0$ we obtain, for any $0 < \eta < \frac{1}{2}$, that

$$\bar{R}_k \leq \mu_{\max}(c_1+c_2)\,k^\delta N + \left(\Pr\left(\bigcup_{r=0}^{\left\lfloor\frac{k}{2}\right\rfloor} P_{e,k-r}\right) + P_{c,k}\right)\mu_{\max}c_32^kN \underset{(a)}{\leq} \mu_{\max}(c_1+c_2)\,k^\delta N +$$
$$\left(\frac{2NM}{1-e^{-wc_1\left(\frac{k}{4}\right)^\delta}}e^{-\frac{w}{2}c_1\left(\frac{k}{4}\right)^\delta k} + \frac{NM}{1-e^{-\frac{1}{36M^2}c_1\left(\frac{k}{4}\right)^\delta}}e^{-\frac{1}{72M^2}c_1\left(\frac{k}{4}\right)^\delta k}\right)\mu_{\max}c_32^kN +$$
$$+\left(C_0e^{-\frac{c_2\eta^2}{144T_m\left(\frac{1}{8}\right)}\left(\pi_{z^*}-\frac{1}{2(1-\eta)}\right)\left(\frac{k}{2}\right)^\delta}\right)^k\mu_{\max}c_32^kN \underset{(b)}{\leq} 2\mu_{\max}(c_1+c_2)\,k^\delta N \tag{11}$$

where (a) is due to the upper bounds on $\Pr\left(\bigcup_{r=0}^{\left\lfloor\frac{k}{2}\right\rfloor} P_{e,k-r}\right)$, $P_{c,k}$ from Lemma 8 and Lemma 16, respectively, where w is a positive constant defined in (19). Lemma 16 holds for a small enough ε, for which $\pi_{z^*} > \frac{1}{2(1-\eta)}$, which exists due to Theorem 14. Inequality (b) follows since for $k > k_0$ we have

$$\max\left\{e^{-\frac{w}{2}c_1\left(\frac{k}{4}\right)^\delta}, e^{-\frac{1}{72M^2}c_1\left(\frac{k}{4}\right)^\delta}, C_0e^{-\frac{c_2\eta^2}{144T_m\left(\frac{1}{8}\right)}\left(\pi_{z^*}-\frac{1}{2(1-\eta)}\right)\left(\frac{k}{2}\right)^\delta}\right\} < \frac{1}{2} \tag{12}$$

where π_{z^*} is the component of the optimal state z^* (of $\boldsymbol{a}^*$) in the stationary distribution of the GoT phase Markov chain. Let E be the number of epochs that start within T turns. We conclude that

$$\bar{R} \underset{(a)}{\leq} \sum_{k=1}^{E}\bar{R}_k \underset{(b)}{\leq} N\mu_{\max}\sum_{k=1}^{k_0}c_32^k + 2N\mu_{\max}\sum_{k=1}^{E}(c_1+c_2)\,k^\delta \underset{(c)}{\leq}$$
$$N\mu_{\max}c_32^{k_0+1} + 2\mu_{\max}(c_1+c_2)\,N\log_2^{1+\delta}\left(\frac{T}{c_3}+2\right) \tag{13}$$

where (a) follows since completing the last epoch to be a full epoch only increases $\bar{R}_k$. In (b) we used (11) for $k > k_0$ and $\bar{R}_k \leq 2\mu_{\max}N(c_1+c_2)\,k^\delta + \mu_{\max}Nc_32^k$ for $k \leq k_0$. Inequality (c) follows from $\sum_{k=k_0+1}^{E}k^\delta \leq E^{1+\delta}$ and bounding E using $T \geq \sum_{k=1}^{E-1}c_32^k \geq c_3\left(2^E-2\right)$. ■

The goal of the parameter $\delta > 0$ is to guarantee that Algorithm 1 can work without the knowledge of the problem parameters $w, T_m\left(\frac{1}{8}\right)$ and π_{z^*} (see (12)). Since the k-th exploration and GoT phases both have an increasing length of k^δ, they are guaranteed to eventually be long enough compared to any of these parameters. The cost is achieving only a near-optimal regret of $O\left(\log^{1+\delta}T\right)$ instead of $O(\log T)$. If bounds on these parameters are available, it is possible to choose c_1 and c_2 such that (12) holds even for $\delta = 0$.

Algorithm 1 Game of Thrones Algorithm

Initialization -Set $\delta > 0$, $\varepsilon > 0$ and integers $c_1, c_2, c_3 > 0$. Set $V_{n,i}(0) = 0$, $s_{n,i}(0) = 0$ for each arm $i = 1, .., M$.

For $k = 1, ..., E$

1) **Exploration Phase** - for the next $c_1 k^\delta$ turns
   a) Sample an arm i uniformly at random from all M arms.
   b) Receive the reward $r_{n,i}(t)$ and set $\eta_i(t) = 0$ if $r_{n,i}(t) = 0$ and $\eta_i(t) = 1$ otherwise.
   c) If $\eta_i(t) = 1$ then update $V_{n,i}(t) = V_{n,i}(t-1) + 1$ and $s_{n,i}(t) = s_{n,i}(t-1) + r_{n,i}(t)$.
   d) Estimate the expectation of arm i by $\mu^k_{n,i} = \frac{s_{n,i}(t)}{V_{n,i}(t)}$, for each $i = 1, ..., M$.
2) **GoT Phase -** for the next $c_2 k^\delta$ turns, play according to the GoT Dynamics with ε. Set $S_n = C$ and set $\overline{a}_n$ to the last action played in the $k - \left\lfloor \frac{k}{2} \right\rfloor - 1$ GoT phase, or a random action if $k = 1, 2$.
   a) If $S_n = C$ then play according to (5) and if $S_n = D$ then choose an arm at random (6).
   b) If $\overline{a}_n \neq a_n$ or $u_n = 0$ or $S_n = D$ then set $S_n = C$ w.p. $\frac{u_n}{u_{n,\max}} \varepsilon^{u_{n,\max} - u_n}$ (otherwise $S_n = D$).
   c) Keep track of the number of times each action was played and resulted in being content:
$$F^n_k(i) \triangleq \sum_{t \in \mathcal{G}_k} I(a_n(t) = i, S_n(t) = C) \tag{14}$$
   where I is the indicator function and $\mathcal{G}_k$ is the set of turns of the k-th GoT phase.
3) **Exploitation Phase** - for the next $c_3 2^k$ turns, play
$$a^k_n = \arg \max_{i=1,...,M} \sum_{r=0}^{\left\lfloor \frac{k}{2} \right\rfloor} F^n_{k-r}(i) \tag{15}$$

End

If either the exploration or the GoT phases fail, the regret becomes linear with T. Like many other bandit algorithms, we avoid a linear expected regret by ensuring that the error probabilities vanish exponentially with T. By using a single epoch with a constant duration for the first two phases, we get an alternative formulation of our result, as in [17]. In this case, with high probability (in T) our algorithm achieves a constant regret.

Corollary 5. *For any $\eta > 0$, there exist c_1, c_2 such that a GoT algorithm with a single epoch (of length T) has a constant regret in T with probability of at least $1 - \eta$.*

Proof: This corollary follows since $P_{e,k}$ in (20) vanishes with c_1 and $P_{c,1}$ in (31) vanishes with c_2. ■

IV. The Exploration Phase - Estimation of the Expected Rewards

In this section, we describe the exploration phase, and analyze its addition to the expected total regret. At the beginning of the game, players still do not have any evaluation of the M different arms. They estimate these values on the run, based on the rewards they get. We propose a pure exploration phase where each player picks an arm uniformly at random, similar to the one suggested in [17]. Note that in contrast to [17], we do not assume that T is known to the players. Hence, the exploration phase is repeated in each epoch. However, the estimation uses all previous exploration phases, so that the number of samples for estimation grows like $O\left(k^{1+\delta}\right)$ over time.

We use the following notion of the second best objective of the assignment problem. Note that $J_2 < J_1$ is in general different from the objective of second best allocation, which might be also optimal. If all allocations have the same objective then there is no need for any estimation and all the results of this section trivially hold with $c_1 \geq 1$.

Definition 6. Let $J_1 = \max_{\boldsymbol{a}} \sum_{n=1}^N g_n(\boldsymbol{a})$. We define the second best objective J_2 as the maximal value of $\sum_{n=1}^N g_n(\boldsymbol{a})$ over all $\boldsymbol{a}$ such that $\sum_{n=1}^N g_n(\boldsymbol{a}) < J_1$.

The estimation of the expected rewards is never perfect. Hence, the optimal solution to the assignment problem given the estimated expectations might be different from the optimal solution with the correct expectations. However, if the uncertainty of the true value of each expectation is small enough, we expect both these optimal assignments to coincide, as formulated in the following lemma.

Lemma 7. *Assume that $\{\mu_{n,i}\}$ are known up to an uncertainty of Δ, i.e., $|\hat{\mu}_{n,i} - \mu_{n,i}| < \Delta$ for each n and i for some $\{\hat{\mu}_{n,i}\}$. Let the optimal assignment be $\boldsymbol{a}_1 = \arg\max_{\boldsymbol{a}} \sum_{n=1}^N g_n(\boldsymbol{a})$ and its objective be $J_1 = \sum_{n=1}^N g_n(\boldsymbol{a}_1)$. Let the second* *best objective and the corresponding assignment be* J_2 *and* $\boldsymbol{a}_2$*, respectively. If* $\Delta \leq \frac{J_1 - J_2}{2N}$ *then*

$$\arg\max_{\boldsymbol{a}} \sum_{n=1}^{N} g_n(\boldsymbol{a}) = \arg\max_{\boldsymbol{a}} \sum_{n=1}^{N} \hat{\mu}_{n,a_n} \eta_{a_n}(\boldsymbol{a}) \tag{16}$$

so that the optimal assignment does not change due to the uncertainty.

Proof: First note that an optimal solution must not have any collisions, otherwise it can be improved since $M \geq N$ and the expected rewards are positive. Hence $J_1 = \sum_{n=1}^{N} \mu_{n,a_{1,n}}$. For all n and i we have $\hat{\mu}_{n,i} = \mu_{n,i} + z_{n,i}$ such that $|z_{n,i}| < \Delta$. In the perturbed assignment problem, $\boldsymbol{a}_1$ performs at least as well as

$$\sum_{n=1}^{N} \hat{\mu}_{n,a_{1,n}} = \sum_{n=1}^{N} \left(\mu_{n,a_{1,n}} + z_{n,a_{1,n}}\right) > \sum_{n=1}^{N} \mu_{n,a_{1,n}} - \Delta N \tag{17}$$

and any assignment $\boldsymbol{a} \neq \boldsymbol{a}_1$ such that $J(\boldsymbol{a}) < J(\boldsymbol{a}_1)$ performs at most as well as

$$\sum_{n=1}^{N} \hat{\mu}_{n,a_n} \eta_{a_n}(\boldsymbol{a}) = \sum_{n=1}^{N} \left(\mu_{n,a_n} + z_{n,a_n}\right) \eta_{a_n}(\boldsymbol{a}) < \sum_{n=1}^{N} \mu_{n,a_{2,n}} \eta_{a_{2,n}}(\boldsymbol{a}_2) + \Delta N. \tag{18}$$

Therefore, if $\Delta \leq \frac{J_1 - J_2}{2N}$ then $\sum_{n=1}^{N} \hat{\mu}_{n,a_{1,n}} > \sum_{n=1}^{N} \hat{\mu}_{n,a_n} \eta_{a_n}(\boldsymbol{a})$ for every $\boldsymbol{a} \neq \boldsymbol{a}_1$, which gives (16). ■

There need not be a perturbation $\{z_{n,i}\}$ such that $|z_{n,i}| = \frac{J_1 - J_2}{2N}$ for which the optimal allocation of the perturbed problem is different from $\boldsymbol{a}_1$. It is possible only if the allocation a_2 that yields J_2 satisfies $a_{2,n} \neq a_{1,n}$ for all n. Therefore, by taking into account the linear programming constraints of the assignment problem, it is possible to achieve a looser requirement than $\Delta \leq \frac{J_1 - J_2}{2N}$. Moreover, in practice much larger random perturbations are not likely to change the optimal assignment.

The following lemma concludes this section by providing an upper bound for the probability that the k-th exploration phase failed, as well as for the probability that at least one of the last $\left\lfloor \frac{k}{2} \right\rfloor + 1$ exploration phases failed.

Lemma 8 (Exploration Error Probability)**.** *Assume i.i.d. rewards* $\{r_{n,i}(t)\}_t$ *as in Definition 1 where* $b_{\max} = \max_{n,i} b_{n,i}$ *and* $\sigma_{\max} = \max_{n,i} \sigma_{n,i}$*. Let* $\left\{\mu_{n,i}^k\right\}$ *be the estimated reward expectations using all the exploration phases up to epoch* k*. Let* $\boldsymbol{a}^* = \arg\max_{\boldsymbol{a}} \sum_{n=1}^{N} g_n(\boldsymbol{a})$ *and* $\boldsymbol{a}^{k*} = \arg\max_{\boldsymbol{a}} \sum_{n=1}^{N} \mu_{n,a_n}^k \eta_{a_n}(\boldsymbol{a})$*. Also define* $J_1 = \sum_{n=1}^{N} g_n(\boldsymbol{a}^*)$ *and the second best objective* J_2*. Define*

$$w \triangleq \frac{(J_1 - J_2)^2}{MN^2 \left(80\sigma_{\max}^2 + \frac{40 b_{\max}}{N}(J_1 - J_2)\right)}. \tag{19}$$

If the length of the k*-th exploration phase is* $c_1 k^\delta$ *then after the* k*-th epoch we have*

$$P_{e,k} \triangleq \Pr\left(\boldsymbol{a}^* \neq \boldsymbol{a}^{k*}\right) \leq 2NMe^{-wc_1\left(\frac{k}{2}\right)^\delta k} + NMe^{-\frac{c_1\left(\frac{k}{2}\right)^\delta}{36M^2}k}. \tag{20}$$

Furthermore,

$$\Pr\left(\bigcup_{r=0}^{\left\lfloor \frac{k}{2} \right\rfloor} P_{e,k-r}\right) \leq \frac{2NM}{1 - e^{-wc_1\left(\frac{k}{4}\right)^\delta}} e^{-\frac{w}{2}c_1\left(\frac{k}{4}\right)^\delta k} + \frac{NM}{1 - e^{-\frac{1}{36M^2}c_1\left(\frac{k}{4}\right)^\delta}} e^{-\frac{1}{72M^2}c_1\left(\frac{k}{4}\right)^\delta k}. \tag{21}$$

Proof: After the k-th exploration phase, the number of samples that are used for estimating the expected rewards is

$$T_e(k) \triangleq c_1 \sum_{i=1}^{k} i^\delta \geq c_1 \left(\frac{k}{2}\right)^{\delta+1}. \tag{22}$$

Define $A_{n,i}(t)$ as the indicator that is equal to one if only player n chose arm i at time t. Define $\mathcal{G}_{n,i}^A$ as set of times for which $A_{n,i}(t) = 1$. Define $V_{n,i}(t) \triangleq \sum_{\tau \in \mathcal{G}_{n,i}^A} A_{n,i}(\tau)$, which is the number of visits of player n to arm i with no collision, up to time t and $V_{\min} = \min_{n,i} V_{n,i}(t)$. From Definition 1, for each n, i and for some positive parameter $b_{n,i}$ we have $E\left\{\left|r_{n,i}(t)-\mu_{n,i}\right|^k\right\} \leq \frac{1}{2}k!\sigma_{n,i}^2 b_{n,i}^{k-2}$ for all integers $k \geq 3$. Define E as the event in which there exists a player n that has an estimate of some arm i with an accuracy worse than Δ. We have

$$\Pr\left(E|V_{\min}\geq v\right)=\Pr\left(\bigcup_{i=1}^{M}\bigcup_{n=1}^{N}\left\{\left|\frac{1}{V_{n,i}(t)}\sum_{\tau\in\mathcal{G}_{n,i}^{A}}r_{n,i}(\tau)-\mu_{n,i}\right|>\Delta\,|\,V_{\min}\geq v\right\}\right)$$
$$\underset{(a)}{\leq} NM\max_{n,i}\Pr\left(\left|\frac{1}{V_{n,i}(t)}\sum_{\tau\in\mathcal{G}_{n,i}^{A}}r_{n,i}(\tau)-\mu_{n,i}\right|>\Delta\,|\,V_{\min}\geq v\right)\underset{(b)}{\leq} 2NMe^{-\frac{\Delta^2}{2\sigma_{\max}^2+2b_{\max}\Delta}v}. \quad (23)$$

where (a) follows by taking the union bound over all players and arms and (b) from using Bernstein's inequality (see [36, Page 205]). Since the exploration phase consists of uniform and independent arm choices we have

$$\Pr\left(A_{n,i}(t)=1\right)=\frac{1}{M}\left(1-\frac{1}{M}\right)^{N-1}. \quad (24)$$

Therefore

$$\Pr\left(V_{\min}<\frac{T_e(k)}{5M}\right)=\Pr\left(\bigcup_{i=1}^{M}\bigcup_{n=1}^{N}\left\{V_{n,i}(t)<\frac{T_e(k)}{5M}\right\}\right)\underset{(a)}{\leq} NM\Pr\left(V_{1,1}(t)<\frac{T_e(k)}{5M}\right)\underset{(b)}{\leq}$$
$$NMe^{-2\frac{1}{M^2}\left(\left(1-\frac{1}{M}\right)^{N-1}-\frac{1}{5}\right)^2 T_e(k)}\underset{(c)}{\leq} NMe^{-\frac{1}{18M^2}T_e(k)} \quad (25)$$

where (a) follows from the union bound, (b) from Hoeffding's inequality for Bernoulli random variables and (c) since $M \geq N$ and $\left(1-\frac{1}{M}\right)^{M-1}-\frac{1}{5}\geq e^{-1}-\frac{1}{5}>\frac{1}{6}$. We conclude that

$$\Pr(E)=\sum_{v=0}^{T_e(k)}\Pr\left(E|V_{\min}=v\right)\Pr\left(V_{\min}=v\right)\leq\sum_{v=0}^{\left\lfloor\frac{T_e(k)}{5M}\right\rfloor}\Pr\left(V_{\min}=v\right)+$$
$$\sum_{v=\left\lfloor\frac{T_e(k)}{5M}\right\rfloor+1}^{T_e(k)}\Pr\left(E|V_{\min}=v\right)\Pr\left(V_{\min}=v\right)\leq\Pr\left(V_{\min}<\frac{T_e(k)}{5M}\right)+\Pr\left(E|\,V_{\min}\geq\frac{T_e(k)}{5M}\right)\underset{(a)}{\leq}$$
$$2NMe^{-\frac{\Delta^2 T_e(k)}{M\left(10\sigma_{\max}^2+10b_{\max}\Delta\right)}}+NMe^{-\frac{T_e(k)}{18M^2}}\underset{(b)}{\leq}2NMe^{-\frac{\Delta^2 c_1\left(\frac{k}{2}\right)^\delta}{M\left(20\sigma_{\max}^2+20b_{\max}\Delta\right)}k}+NMe^{-\frac{c_1\left(\frac{k}{2}\right)^\delta}{36M^2}k} \quad (26)$$

where (a) follows from (23) and (25), and (b) from (22). By requiring $\Delta=\frac{J_1-J_2}{2N}$ we know from Lemma 7 that $\Pr\left(\boldsymbol{a}^*\neq\boldsymbol{a}^{k*}\right)\leq\Pr(E)$, which together with (26) establishes (20). Now define $w=\frac{(J_1-J_2)^2}{MN^2\left(80\sigma_{\max}^2+\frac{40b_{\max}}{N}(J_1-J_2)\right)}$. We obtain

$$\Pr\left(\bigcup_{r=0}^{\left\lfloor\frac{k}{2}\right\rfloor}P_{e,k-r}\right)\underset{(a)}{\leq}2NM\sum_{r=0}^{\left\lfloor\frac{k}{2}\right\rfloor}e^{-wc_1\left(\frac{k-r}{2}\right)^\delta(k-r)}+NM\sum_{r=0}^{\left\lfloor\frac{k}{2}\right\rfloor}e^{-\frac{c_1}{36M^2}\left(\frac{k-r}{2}\right)^\delta(k-r)}\underset{(b)}{\leq}$$
$$2NMe^{-wc_1\left(\frac{k}{4}\right)^\delta k}\sum_{r=0}^{\left\lfloor\frac{k}{2}\right\rfloor}e^{wc_1\left(\frac{k}{4}\right)^\delta r}+NMe^{-\frac{c_1}{36M^2}\left(\frac{k}{4}\right)^\delta k}\sum_{r=0}^{\left\lfloor\frac{k}{2}\right\rfloor}e^{\frac{c_1}{36M^2}\left(\frac{k}{4}\right)^\delta r}\underset{(c)}{\leq}$$
$$2NMe^{-wc_1\left(\frac{k}{4}\right)^\delta k}\frac{e^{wc_1\left(\frac{k}{4}\right)^\delta\left(\frac{k}{2}+1\right)}-1}{e^{wc_1\left(\frac{k}{4}\right)^\delta}-1}+NMe^{-\frac{c_1}{36M^2}\left(\frac{k}{4}\right)^\delta k}\frac{e^{\frac{1}{36M^2}c_1\left(\frac{k}{4}\right)^\delta\left(\frac{k}{2}+1\right)}-1}{e^{\frac{1}{36M^2}c_1\left(\frac{k}{4}\right)^\delta}-1}\leq$$
$$\frac{2NM}{1-e^{-wc_1\left(\frac{k}{4}\right)^\delta}}e^{-\frac{w}{2}c_1\left(\frac{k}{4}\right)^\delta k}+\frac{NM}{1-e^{-\frac{1}{36M^2}c_1\left(\frac{k}{4}\right)^\delta}}e^{-\frac{1}{72M^2}c_1\left(\frac{k}{4}\right)^\delta k} \quad (27)$$

where (a) follows by using the union bound on (26) , (b) follows since $e^{-wc_1\left(\frac{k-r}{2}\right)^\delta(k-r)}\leq e^{-wc_1\left(\frac{k}{4}\right)^\delta(k-r)}$ for $r\leq\left\lfloor\frac{k}{2}\right\rfloor$, and (c) is the geometric series formula. ■

V. Game of Thrones Dynamics Phase

In this section, we analyze the game of thrones (GoT) dynamics between players. These dynamics guarantee that the optimal state will be played a significant amount of time, and only require the players to know their own action and their received payoff on each turn. Note that these dynamics assume deterministic utilities. We use the estimated expected reward of each arm as the utility for this step, and zero if a collision occurred. This means that players ignore the numerical reward they receive by choosing the arm, as long as it is non-zero.

Definition 9. The game of thrones G of epoch k has the N players of the original multi-armed bandit game. Each player can choose from among the M arms, so $\mathcal{A}_n = \{1, ..., M\}$ for each n. The utility of player n in the strategy profile $\boldsymbol{a} = (a_1, ..., a_N)$ is

$$u_n(\boldsymbol{a}) = \mu^k_{n,a_n} \eta_{a_n}(\boldsymbol{a}) \tag{28}$$

where μ^k_{n,a_n} is the estimation of the expected reward of arm a_n, from all the exploration phases that have ended, up to epoch k. Also define $u_{n,\max} = \max\limits_{\boldsymbol{a}} u_n(\boldsymbol{a})$.

Our dynamics belong to the family introduced in [33]–[35]. These are very powerful dynamics that guarantee that the optimal strategy profile (in terms of the sum of utilities) will be played a sufficiently large portion of the turns. However, [33]–[35] all rely on the following structural property of the game, called interdependence.

Definition 10. A game G with finite action spaces $\mathcal{A}_1, ..., \mathcal{A}_N$ is interdependent if for every strategy profile $\boldsymbol{a} \in \mathcal{A}_1 \times ... \times \mathcal{A}_N$ and every set of players $J \subset N$, there exists a player $n \notin J$ and a choice of actions $\boldsymbol{a}'_J \in \prod_{m \in J} \mathcal{A}_m$ such that $u_n(\boldsymbol{a}'_J, \boldsymbol{a}_{-J}) \neq u_n(\boldsymbol{a}_J, \boldsymbol{a}_{-J})$.

Our GoT is not interdependent. To see this, pick any strategy profile $\boldsymbol{a}$ such that some players are in a collision while others are not. Choose J as the set of all players that are not in a collision. All players outside this set are in a collision, and there does not exist any colliding player such that the actions of the non-colliding players can make her utility non-zero.

Our GoT Dynamics, described in Subsection III-A, modify [34] such that interdependency is no longer needed. In comparison to [34], our dynamics assign zero probability that a player with $u_n = 0$ (in a collision) will be content. Additionally, we do not need to keep the benchmark utility as part of the state. A player knows with probability 1 whether there was a collision, and if there was not, she gets the same utility for the same arm.

Unlike the exploration during the exploration phase, the exploration of the GoT dynamics is meant to allow the players to reach the optimal strategy profile. Since the GoT must have well-defined and time-invariant utility functions, the reward samples from the GoT phase cannot be used to change the estimation for the arm expectations already in the current epoch. However, players can use these samples in the next epoch and improve the accuracy of their estimation by doing so. While it does not improve our regret analysis, it can enhance the algorithm's performance in practice.

The GoT dynamics (see Subsection 3.1) induce a Markov chain over the state space $Z = \prod_{n=1}^{N} (\mathcal{A}_n \times \mathcal{M})$, where $\mathcal{M} = \{C, D\}$. We denote the transition matrix of this Markov chain by P^ε. We are interested in the invariant distribution of P^ε, which, for small ε, is concentrated only on the states with an optimal sum of utilities. However, it is only guaranteed that the dynamics will visit these optimal states very often for a small enough ε. There could be multiple optimal states and the dynamics might fluctuate between them. This could prevent players on distributedly agreeing which optimal state to play. Fortunately, as shown in the following lemma, in our case there is a unique optimal state with probability one. This result arises from the continuous distribution of the rewards that makes the distribution of the empirical estimation for the expectations continuous as well.

Lemma 11. *The optimal solution to* $\max\limits_{\boldsymbol{a}} \sum_{n=1}^{N} u_n(\boldsymbol{a})$ *is unique with probability 1.*

Proof: First note that an optimal solution must not have any collisions, otherwise it can be improved since $M \geq N$. Let $\left\{\mu^k_{n,i}\right\}$ be the estimated reward expectations in epoch k. For two different solutions $\tilde{\boldsymbol{a}} \neq \boldsymbol{a}^*$ to be optimal, we must have $\sum_{n=1}^{N} \mu^k_{n,\tilde{a}_n} = \sum_{n=1}^{N} \mu^k_{n,a^*_n}$. However, $\tilde{\boldsymbol{a}}$ and $\boldsymbol{a}^*$ must differ in at least one assignment. Since the distributions of the rewards r_{n,a_n} are continuous, so are the distributions of $\sum_{n=1}^{N} \mu^k_{n,a_n}$ (as a sum of the average of the rewards). Hence $\Pr\left(\sum_{n=1}^{N} \mu^k_{n,\tilde{a}_n} = \sum_{n=1}^{N} \mu^k_{n,a^*_n}\right) = 0$, which is a contradiction, so the result follows. ■

Next we prove a lower bound on the probability for the unique optimal state z^* in the stationary distribution of the GoT dynamics. This lower bound is a function of ε, and hence allows for a choice of ε that guarantees that $\pi_{z^*} > \frac{1}{2}$. Note that the analysis in [34] cannot be applied here since our game is not interdependent. Moreover, our proof only requires that $c > \sum_n u_{n,\max} - J_1$ where J_1 is the optimal objective. This significantly improves the $c > N$ requirement in [34], and has a crucial effect on the mixing time of the GoT dynamics and therefore on the convergence time to the optimal state. Note that the analysis in [34] assumed utilities that are in $[0, 1]$ so $\sum_n u_{n,\max} \leq N$.

The following notion is often useful for the analysis of perturbed Markov chains.

Definition 12. Let Z be a finite set and let $z_r \in Z$. A graph consisting of edges $z' \to z$ such that $z' \in Z \setminus \{z_r\}$ and $z \in Z$ is called a $\{z_r\}$-graph if it satisfies the following conditions:

1) Every point $z' \in Z \setminus \{z_r\}$ is the initial point of exactly one edge.
2) For any point $z' \in Z \setminus \{z_r\}$ there exists a sequence of edges leading from it to z_r.

We denote by $G(z_r)$ the set of $\{z_r\}$-graphs and by g a specific graph. A graph in $G(z_r)$ is a tree rooted in z_r such that from every $z' \neq z_r$ there is a path to z_r. This follows since the two conditions above imply that there are no closed cycles in a $\{z_r\}$-graph.

The following lemma provides an explicit expression for the stationary distribution of a Markov chain.

Lemma 13 ([37, Lemma 3.1, Chapter 6])**.** *Consider a Markov chain with a set of states Z and transition probabilities $P_{z'z}$. Assume that every state can be reached from any other state in a finite number of steps. Then the stationary distribution of the chain is*

$$\pi(z) = \frac{Q(z)}{\sum_{z' \in Z} Q(z')} \tag{29}$$

where

$$Q(z) = \sum_{g \in G(z)} \prod_{(z' \to z) \in g} P_{z'z}. \tag{30}$$

Using the lemma above, we can prove the following lower bound on π_{z^*} as a function of ε.

Theorem 14. *Let $\varepsilon > 0$ and let π be the stationary distribution of Z. Let $\boldsymbol{a}^{k*} = \arg\max_{\boldsymbol{a}} \sum_{n=1}^{N} u_n(\boldsymbol{a})$ and let the optimal state be $z^* = \left[\boldsymbol{a}^{k*}, C^N\right]$. Let $J_1 = \sum_{n=1}^{N} u_n\left(\boldsymbol{a}^{k*}\right)$. If $c > \sum_n u_{n,\max} - J_1$ then for any $0 < \eta < \frac{1}{2}$ there exists a small enough ε such that $\pi_{z^*} > \frac{1}{2(1-\eta)}$.*

Proof: See Appendix A. ■

We need that $\pi_{z^*} > \frac{1}{2}$ in order for the GoT phase to succeed. The main role of the theorem above is to show that a small enough ε, such that $\pi_{z^*} > \frac{1}{2}$, exists. However, the proof of Theorem 14 also tells which ε values are small enough to guarantee that $\pi_{z^*} > \frac{1}{2}$, as a function of the problem parameters N, M and the matrix of utilities $\{u_n\}$. The designer that has to choose ε does not typically know the parameters of the problem. Nevertheless, even having a coarse bound on these parameters is enough since we only need to choose a small enough ε instead of tuning ε as a function of these parameters. In practice, the designer is likely to have either uncertainty bounds for the problem parameters or a random model that is either known or can be simulated. Using this random model, the designer can always take ε small enough to make the probability that $\pi_{z^*} > \frac{1}{2}$ arbitrarily close to one.

Next we prove a probabilistic lower bound on the number of times the optimal state has been played during the k-th GoT phase.

Lemma 15. *Let $\boldsymbol{a}^{k*} = \arg\max_{\boldsymbol{a}} \sum_{n=1}^{N} u_n(\boldsymbol{a})$ and let the optimal state be $z^* = \left[\boldsymbol{a}^{k*}, C^N\right]$. Denote the stationary distribution of Z by π. Let $\mathcal{G}_k$ be the set of turns of the k-th GoT phase. Then for any $0 < \eta < 1$ we have, for a sufficiently large k, that*

$$P_g \triangleq \Pr\left(\sum_{t \in \mathcal{G}_k} I(z(t) = z^*) \leq (1-\eta)\pi_{z^*} c_2 k^{\delta}\right) \leq \underbrace{B_0 \left\|\varphi_k\right\|_{\pi}}_{A_k} e^{-\frac{\eta^2 \pi_{z^*} c_2 k^{\delta}}{72 T_m\left(\frac{1}{8}\right)}} \tag{31}$$

where I is the indicator function, B_0 is a constant independent of η and π_{z^*}, φ_k is the probability distribution of the state played in the $k - \left\lfloor \frac{k}{2} \right\rfloor - 1$-th exploitation phase, and $\left\|\varphi_k\right\|_{\pi} \triangleq \sqrt{\sum_{i=1}^{|Z|} \frac{\varphi_{k,i}^2}{\pi_i}}$.

Proof: The events $I(z(t) = z^*)$ are not independent but rather form a Markov chain. Hence, Markovian concentration inequalities are required. The result follows by a simple application of the concentration bound in [38, Theorem 3]. We use $f(z) = I(z = z^*)$, which counts the number of visits to the optimal state. We denote by $T_m\left(\frac{1}{8}\right)$ the mixing time of Z with an accuracy of $\frac{1}{8}$. Our initial state is the state $\hat{z}_k$ played in the $k - \left\lfloor \frac{k}{2} \right\rfloor - 1$ exploitation phase. ■

The next lemma concludes this section by providing an upper bound for the probability that the k-th exploitation phase failed due to the last $\left\lfloor \frac{k}{2} \right\rfloor + 1$ GoT phases (including the k-th GoT phase).

Lemma 16 (GoT Error Probability)**.** *Let $\boldsymbol{a}^{k*} = \arg\max_{\boldsymbol{a}} \sum_{n=1}^{N} u_n(\boldsymbol{a})$. Let the optimal state of the k-th GoT phase be $z^{k*} = \left[\boldsymbol{a}^{k*}, C^N\right]$. Let $\mathcal{G}_k$ be the set of turns of the k-th GoT phase. Define the number of times the optimal state has been played in the last $\left\lfloor \frac{k}{2} \right\rfloor + 1$ GoT phases combined by*

$$F_k(z^*) \triangleq \sum_{i=k-\left\lfloor \frac{k}{2} \right\rfloor}^{k} \sum_{t \in \mathcal{G}_i} I\left(z(t) = z^{i*}\right). \tag{32}$$

Let $\pi_{z^*}=\min_{k-\left\lfloor\frac{k}{2}\right\rfloor\leq i\leq k}\pi_{z^{i*}}$. *If* $\pi_{z^*}>\frac{1}{2(1-\eta)}$ *for some* $0<\eta<\frac{1}{2}$ *then for a sufficiently large* k

$$P_{c,k}\triangleq\Pr\left(F_k\left(z^*\right)\leq\frac{1}{2}\sum_{i=k-\left\lfloor\frac{k}{2}\right\rfloor}^{k}c_2 i^{\delta}\right)\leq\left(C_0 e^{-\frac{c_2\eta^2}{144T_m\left(\frac{1}{8}\right)}\left(\pi_{z^*}-\frac{1}{2(1-\eta)}\right)\left(\frac{k}{2}\right)^{\delta}}\right)^{k}. \tag{33}$$

where C_0 *is a constant that is independent of* k, π_{z^*} *and* η.

Proof: Define the total length of the last $\left\lfloor\frac{k}{2}\right\rfloor+1$ GoT phases by $L_k=c_2\sum_{i=k-\left\lfloor\frac{k}{2}\right\rfloor}^{k}i^{\delta}$. Define the independent random variables $S_i=\sum_{t\in\mathcal{G}_i}I\left(z\left(t\right)=z^{i*}\right)$ for each i. For any $t>0$, Chernoff bound on (32) yields

$$\Pr\left(\sum_{i=k-\left\lfloor\frac{k}{2}\right\rfloor}^{k}S_i\leq\frac{L_k}{2}\mid\left\{z^i\left(0\right)\right\}_{i=k-\left\lfloor\frac{k}{2}\right\rfloor}^{k}\right)\leq e^{t\frac{L_k}{2}}\prod_{i=k-\left\lfloor\frac{k}{2}\right\rfloor}^{k}E\left\{e^{-tS_i}\mid z^i\left(0\right)\right\} \tag{34}$$

where $z^i\left(0\right)$ is the initial state at the beginning of the i-th GoT phase. For $t=\frac{\frac{\eta^2}{1-\eta}}{72T_m\left(\frac{1}{8}\right)}$ we have the following bound

$$E\left\{e^{-tS_i}\mid z^i\left(0\right)\right\}\underset{(a)}{\leq}P_g\cdot 1+\left(1-P_g\right)\cdot e^{-t(1-\eta)\pi_{z^*}c_2 i^{\delta}}\underset{(b)}{\leq}\left(1+A_i\right)e^{-\frac{\eta^2\pi_{z^*}c_2 i^{\delta}}{72T_m\left(\frac{1}{8}\right)}} \tag{35}$$

where (a) is since $e^{-tS_i}\leq 1$ and (b) follows from Lemma 15 on P_g. Plugging (35) back into (34) and using the law of total expectation over $\left\{z^i\left(0\right)\right\}_{i=k-\left\lfloor\frac{k}{2}\right\rfloor}^{k}$ gives

$$\Pr\left(\sum_{i=k-\left\lfloor\frac{k}{2}\right\rfloor}^{k}S_i\leq\frac{L_k}{2}\right)\underset{(a)}{\leq}\left[\prod_{i=k-\left\lfloor\frac{k}{2}\right\rfloor}^{k}\left(1+E^{z^i(0)}\left\{A_i\right\}\right)\right]e^{\frac{\frac{\eta^2}{1-\eta}}{72T_m\left(\frac{1}{8}\right)}\frac{L_k}{2}}e^{-\frac{\eta^2\pi_{z^*}}{72T_m\left(\frac{1}{8}\right)}L_k}\underset{(b)}{\leq}$$
$$C_0^k e^{-\frac{\eta^2}{72T_m\left(\frac{1}{8}\right)}\left(\pi_{z^*}-\frac{1}{2(1-\eta)}\right)L_k}\underset{(c)}{\leq}\left(C_0 e^{-\frac{c_2\eta^2}{144T_m\left(\frac{1}{8}\right)}\left(\pi_{z^*}-\frac{1}{2(1-\eta)}\right)\left(\frac{k}{2}\right)^{\delta}}\right)^{k} \tag{36}$$

where (a) uses $\prod_{i=k-\left\lfloor\frac{k}{2}\right\rfloor}^{k}e^{-ac_2 i^{\delta}}=e^{-a\sum_{i=k-\left\lfloor\frac{k}{2}\right\rfloor}^{k}c_2 i^{\delta}}$ for any a. In (b), note that for every $z^i\left(0\right)$ we have $A_i=c\left\|\varphi_i\right\|_{\pi}$ for some constant c, so $E^{z^i(0)}\left\{A_i\right\}$ is bounded for any i so the above bound vanishes with k. However, by choosing $z^i\left(0\right)$ to be the state played in the $i-\left\lfloor\frac{i}{2}\right\rfloor-1$ exploitation phase, the same bound guarantees that $\Pr\left(z^i\left(0\right)=z^{i*}\right)\to 1$ as $i\to\infty$, so $E\left\{\left\|\varphi_i\right\|_{\pi}\right\}\to\frac{1}{\sqrt{\pi_{z^{i*}}}}<\sqrt{2}$. Inequality (c) uses $L_k=\sum_{i=k-\left\lfloor\frac{k}{2}\right\rfloor}^{k}c_2 i^{\delta}\geq\frac{k}{2}\left(\frac{k}{2}\right)^{\delta}$. Note that the counting in (32) is on the same optimal state $z^{i*}=z^*$ for $k-\left\lfloor\frac{k}{2}\right\rfloor\leq i\leq k$ only if the previous $\left\lfloor\frac{k}{2}\right\rfloor+1$ exploration phases succeeded. ■

VI. Markovian Rewards

In this section, we generalize our main result to the case of (rested) Markovian rewards. This generalization is valid for the same GoT algorithm (Algorithm 1) without any modifications. In this section we assume that each player can observe her collision indicator in addition to her reward. The reason for this additional assumption is that with discrete rewards, the collision indicator cannot be deduced from the rewards with probability 1. Nevertheless, knowing whether other players chose the same arm is a very modest requirement compared to assuming that players can observe the actions of other players.

We use the standard model for Markovian bandits (see for example [39]). As in the i.i.d. case, we assume that the reward processes (now Markov chains) of players are independent.

Definition 17. Let $V_{n,i}$ be the number of visits without collision of player n to arm i. The sequence of rewards $\left\{r_{n,i}\left(V_{n,i}\right)\right\}$ of arm i for player n is an ergodic Markov chain such that:

1) $r_{n,i}\left(V_{n,i}\right)$ has a finite state space $\mathcal{R}_{n,i}$ for each n,i, consisting of positive numbers.
2) The transition matrix of $r_{n,i}\left(V_{n,i}\right)$ is $P_{n,i}$ and the stationary distribution is $\pi^{n,i}$.
3) The Markov chains $\left\{r_{n,i}\left(V_{n,i}\right)\right\}$ are independent for different n or different i.

The total regret is now defined as

Definition 18. The expectation of arm i for player n is defined as:

$$\mu_{n,i} = \sum_{r \in \mathcal{R}_{n,i}} r\pi^{n,i}(r). \tag{37}$$

Define $g_n(\boldsymbol{a}) = \mu_{n,a_n}\eta_{a_n}(\boldsymbol{a})$ as the expected utility of player n in strategy profile $\boldsymbol{a}$. The total regret is defined as the random variable

$$R = \sum_{t=1}^{T}\sum_{n=1}^{N} g_n(\boldsymbol{a}^*) - \sum_{t=1}^{T}\sum_{n=1}^{N} r_{n,a_n(t)}(V_{n,i}(t))\,\eta_{a_n(t)}(\boldsymbol{a}(t)) \tag{38}$$

where

$$\boldsymbol{a}^* \in \arg\max_{\boldsymbol{a}} \sum_{n=1}^{N} g_n(\boldsymbol{a}). \tag{39}$$

The expected total regret $\bar{R} \triangleq E\{R\}$ is the average of (38) over the randomness of the rewards $\{r_{n,i}(V_{n,i})\}$, that dictate the random actions $\{a_n(t)\}$.

The division into epochs and phases makes the generalization of our result to other reward models convenient. Only the exploration phase requires a different analysis. This is formalized in the following theorem. As before, an explicit requirement on ε is given in the proof of Theorem 14.

Theorem 19 (Generalization to Markovian Rewards)**.** *Assume that the rewards $\{r_{n,i}(V_{n,i})\}$ are Markovian as in Definition 17. Assume that for every n, i, the stationary distribution of $r_{n,i}(V_{n,i})$, denoted $\pi^{n,i}$, is generated at random using a continuous distribution on the M dimensional simplex. Let the game have a finite horizon T, unknown to the players. Let each player play according to Algorithm 1, with any integers $c_1, c_2, c_3 > 0$ and $\delta > 0$. Then there exists a small enough ε such that for large enough T, the expected total regret is bounded by*

$$\bar{R} \leq 4\left(\max_{n,i}\mu_{n,i}\right)(c_1 + c_2)\,N\log_2^{1+\delta}\left(\frac{T}{c_3} + 2\right) = O\left(\log^{1+\delta} T\right). \tag{40}$$

Proof: First note that an optimal solution must not have any collisions, otherwise it can be improved since $M \geq N$ and the expected rewards are positive. Since the stationary distributions are continuous, the probability for any specific value of the expected rewards $\{\mu_{n,i}\}$ defined in (37), or any specific value of $\sum_{n=1}^{N}\mu_{n,\tilde{a}_n}$, is zero. For two different solutions $\tilde{\boldsymbol{a}} \neq \boldsymbol{a}^*$ to be optimal, they must have $\sum_{n=1}^{N}\mu_{n,\tilde{a}_n} = \sum_{n=1}^{N}\mu_{n,a_n^*}$. However, they must differ in at least one assignment. Hence $\Pr\left(\sum_{n=1}^{N}\mu_{n,\tilde{a}_n} = \sum_{n=1}^{N}\mu_{n,a_n^*}\right) = 0$, so there is a unique solution to (39) with probability 1. If the exploration phase of the k-th epoch did not fail, then the solution to $\max_{a}\sum_{n=1}^{N}\mu_{n,a_n}^{k}$ is identical to this unique solution.

The rest of the proof replaces the exploration error bound in Lemma 8. The estimated expected rewards $\hat{\mu}_{n,i}$ are now defined with respect to the stationary distribution of the reward chain. However, there is still a single estimated value for each player and arm and the role and analysis of the GoT phase remain the same. Hence, we need to show that the exploration phase results in an accurate enough estimation of $\mu_{n,i}$.

As in (22), after the k-th exploration phase, the number of samples that are used for estimating the expected rewards is $T_e(k) \geq c_1\left(\frac{k}{2}\right)^{\delta+1}$. Let $T_{n,i}\left(\frac{1}{8}\right)$ be the mixing time of $r_{n,i}(t)$ with an accuracy of $\frac{1}{8}$. We define for the initial distribution φ on $\mathcal{R}_{n,i}$

$$\|\varphi\|_{\pi_{n,i}} \triangleq \sqrt{\sum_{j=1}^{|\mathcal{R}_{n,i}|}\frac{\varphi_j^2}{\pi_{n,i}(j)}}. \tag{41}$$

Define $A_{n,i}(t)$ as the indicator that is equal to one if only player n chose arm i at time t. Let $\mathcal{G}_{n,i}^{A}$ be the set of times for which $A_{n,i}(t) = 1$. Define $V_{n,i}(t) \triangleq \sum_{\tau \in \mathcal{G}_{n,i}^{A}} A_{n,i}(\tau)$, which is the number of visits of player n to arm i with no collision, up to time t and define $V_{\min} = \min_{n,i} V_{n,i}(t)$. Define E as the event in which there exists a player n that has an estimate of some arm i with an accuracy worse than Δ. We have

$$\Pr\left(E|V_{\min} \geq v\right) = \Pr\left(\bigcup_{i=1}^{M}\bigcup_{n=1}^{N}\left\{\left|\frac{1}{V_{n,i}(t)}\sum_{\tau\in\mathcal{G}_{n,i}^{A}} r_{n,i}(\tau) - \mu_{n,i}\right| \geq \Delta \,|\, V_{\min} \geq v\right\}\right) \underset{(a)}{\leq}$$
$$NM\max_{n,i}\left(\Pr\left(\sum_{\tau\in\mathcal{G}_{n,i}^{A}} r_{n,i}(\tau) \geq \left(1+\frac{\Delta}{\mu_{n,i}}\right)\mu_{n,i}v\right) + \Pr\left(\sum_{\tau\in\mathcal{G}_{n,i}^{A}} r_{n,i}(\tau) \leq \left(1-\frac{\Delta}{\mu_{n,i}}\right)\mu_{n,i}v\right)\right) \underset{(b)}{\leq}$$
$$2NMc\max_{n,i}\left(\|\varphi\|_{\pi_{n,i}}\right)e^{-\min\limits_{n,i}\frac{\Delta\min\left\{\frac{\Delta}{\mu_{n,i}},1\right\}}{72T_{n,i}\left(\frac{1}{8}\right)}v} \tag{42}$$

where (a) follows by taking the union bound over all players and arms and (b) from using the bound in [38] with $\eta = \frac{\Delta}{\mu_{n,i}}$, where c is a constant independent of η and π_{z^*}. We conclude that

$$\Pr(E) = \sum_{v=0}^{T_e(k)} \Pr\left(E|V_{\min} = v\right)\Pr\left(V_{\min} = v\right) \leq$$
$$\sum_{v=0}^{\left\lfloor\frac{T_e(k)}{5M}\right\rfloor} \Pr\left(V_{\min} = v\right) + \sum_{v=\left\lfloor\frac{T_e(k)}{5M}\right\rfloor+1}^{T_e(k)} \Pr\left(E|V_{\min} = v\right)\Pr\left(V_{\min} = v\right) \leq \Pr\left(V_{\min} < \frac{T_e(k)}{5M}\right) +$$
$$\Pr\left(E|\ V_{\min} \geq \frac{T_e(k)}{5M}\right) \underset{(a)}{\leq} 2NMc\max_{n,i}\left(\|\varphi\|_{\pi_{n,i}}\right)e^{-\min\limits_{n,i}\frac{\Delta\min\left\{\frac{\Delta}{\mu_{n,i}},1\right\}c_1\left(\frac{k}{2}\right)^{\delta}}{720T_{n,i}\left(\frac{1}{8}\right)M}k} + NMe^{-\frac{c_1\left(\frac{k}{2}\right)^{\delta}}{36M^2}k}. \tag{43}$$

where (a) follows from (42) and (25), that still holds since arm choices are independent and uniform in the exploration phase. Finally, by requiring $\Delta = \frac{J_1-J_2}{2N}$ we know from Lemma 7 that $\Pr\left(\boldsymbol{a}^* \neq \boldsymbol{a}^{k*}\right) \leq \Pr(E)$, which together with (43) establishes the same bound (up to a constant factor) as in (20) but with $\tilde{w} = \min_{n,i}\frac{(J_1-J_2)\min\left\{\frac{J_1-J_2}{2N\mu_{n,i}},1\right\}}{1440T_{n,i}\left(\frac{1}{8}\right)NM}$ replacing w. Using this new bound, the proof of Theorem 4 remains the same using $\tilde{w}$ instead of w. ■

VII. Simulation Results

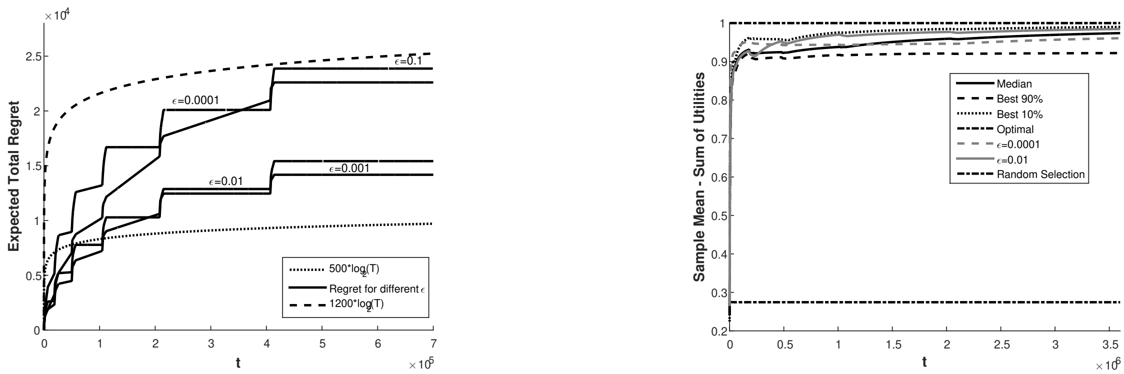
In this section, we illustrate the behavior of Algorithm 1 using numerical simulations. We use $\delta = 0$ since it yields good results in practice. We conjecture that the bound (31) is not tight for our particular Markov chain and indicator function, since it applies to all Markov chains with the same mixing time and all functions on the states. This explains why modest choices of c_2 are large enough to satisfy (12) even for $\delta = 0$, making the k^{δ} factor in the exponent unnecessary in practice. In order not to waste the initial phases, their lengths should be chosen so that the exploitation phase already occupies most of the turns in early epochs, while allowing for a considerable GoT phase. In the implementation of our algorithm we also use the GoT and exploration phases for estimating the expected rewards, which improves the estimation significantly in practice, making the GoT phase the main issue. We chose the parameter $c = \frac{\log\left(\frac{2}{c_2N}\right)}{\log\varepsilon} \approx 1.4$ (for $\varepsilon = 0.01$), in order for the escape probability from a content state to be $\varepsilon^c = \frac{2}{c_2N}$, with an expectation of approximately 2 escapes in each GoT phase. Note that this choice is a significant relaxation from $c = N$ (the lowest possible value in [34]), and has a dramatic positive effect on the mixing time of the GoT Dynamics and the convergence time. The rewards are generated as $r_{n,i}(t) = \mu_{n,i} + z_{n,i}(t)$ where $\{\mu_{n,i}\}$ are taken from a matrix U and $\{z_{n,i}(t)\}$ are independent Gaussian variables with zero mean and a variance of $\sigma^2 = 0.05$ for each n, i.

First we demonstrate our theoretical result in the following scenario with $N = M = 3$:

$$U = \begin{pmatrix} 0.1 & 0.05 & 0.9 \\ 0.1 & 0.25 & 0.3 \\ 0.4 & 0.2 & 0.8 \end{pmatrix}$$

for which the optimal allocation is $a_1 = 3, a_2 = 2, a_1 = 1$. Here we used $c_1 = 500, c_2 = c_3 = 6000$. The expected total regret as a function of time, averaged over 100 realizations of the algorithm, is depicted in part (a) of Fig. 2. After the second epoch, all epochs lead to the optimal solution in their exploitation phase. It can be seen that the regret increases like $O(\log_2 T)$, and more specifically, is between $400\log_2 T$ and $700\log_2 T$. This demonstrates the theoretical result of Theorem 4. Furthermore, we see that the regret behaves very similarly across 4 different orders of magnitude of ε. This suggests that choosing a small enough ε is easy in practice.

The total regret compares the sum of utilities to the ideal one that could have been achieved in a centralized scenario. With no communication between players and with a matrix of expected rewards, the gap from this ideal naturally increases. In this scenario, converging to the exact optimal solution might take a long time, even for the (unknown) optimal algorithm. A nice property of our algorithm that makes it appealing in practice is that the GoT dynamics are not specifically oriented to converging to the optimal solution, but they probabilistically prefer states with a higher sum of utilities. This is simply because these states have incoming paths with high probabilities (i.e., "low resistance"). To demonstrate this property, we generated 100 independent realizations of U with elements that were chosen uniformly at random on $[0.05, 0.95]$. In part (b) of Fig. 2, we present the sample mean of the accumulated sum of utilities $\sum_{n=1}^{N} \frac{1}{t} \sum_{\tau=1}^{t} u_n\left(\boldsymbol{a}\left(\tau\right)\right)$ as a function of time t and averaged over 100 experiments for $M = N = 5$. Here we used $c_1 = 500, c_2 = c_3 = 60000$ and $\varepsilon = 0.001$. The performance was normalized by the optimal solution to the assignment problem (for each U) and compared to the performance of a random choice of arms. Clearly the sum of utilities becomes close to optimal (more than 90%) fast, with only a small variation between different realizations of U. Additionally, our algorithm behaved very similarly for a wide range of ε values (3 orders of magnitude). This supports the intuition that there is no threshold phenomenon on ε (becoming "small enough"), since the dynamics prefer states with higher sum of utilities for all $\varepsilon < 1$. Furthermore, it indicates that there is no penalty for choosing smaller ε than necessary just to have a safety margin.



(a) (b)

Fig. 2: In (a): Total Regret for $N = M = 3$ as a function of time. In (b): Sample mean of the sum of utilities for $N = M = 5$ as a function of time. Both figures are averaged over 100 simulations.

VIII. Conclusion and Open Questions

In this paper, we considered a multi-player multi-armed bandit game where players distributedly learn how to allocate the arms (i.e., resources) between them. In contrast to all other multi-player bandit problems, we both allow for different expected rewards between users **and** assume users only know their actions and rewards. We proposed a novel fully distributed algorithm that achieves an expected total regret of near-$O\left(\log T\right)$, when the horizon T is unknown by the players.

Our simulations suggest that tuning the parameters for our algorithm is a relatively easy task in practice. The designer can do so by simulating a random model for the unknown environment and varying the parameters, knowing that only very slack accuracy is needed for the tuning.

In the single-player case, if the algorithm knows enough about the problem parameters, it is possible to even achieve a bounded regret (see [40]). It is an interesting question whether a bounded regret can be achieved in the multi-player scenario with enough knowledge of the problem parameters.

Analytically, a single epoch in our algorithm converges to the optimal solution of the assignment problem with high probability. This might be valuable even outside the context of bandit learning algorithms, since it is the first fully distributed algorithm that solves the assignment problem exactly. Our dynamics may take a long time to converge in terms of N. While this does not effect the regret for large T, it might be significant if our algorithm were to be used for this purpose. Our game is not a general one but has a structure that allowed us to modify the dynamics in [34] such that the interdependence assumption can be dropped. We conjecture that this structure can also be exploited to accelerate the converge rate of this kind of dynamics. Studying the converge time of perhaps further refined dynamics is another promising future research direction.

IX. Appendix A - Proof of Theorem 14

In this section we prove Theorem 14, which bounds from below the probability of z^* in the stationary distribution of Z. From Lemma 13 we know that

$$\pi_{z^*} = \frac{1}{1 + \underbrace{\frac{\sum_{z_r \in Z, z_r \neq z^*} \sum_{g \in G(z_r)} \prod_{(z' \to z) \in g} P_{z'z}}{\sum_{g \in G(z^*)} \prod_{(z' \to z) \in g} P_{z'z}}}_{A}}. \tag{44}$$

In Subsection IX-A, we bound the denominator of A from below by identifying the maximal term in the sum (the trees in $G(z^*)$ with the maximal probability). In Subsection IX-B we upper bound the numerator of A.

A. The Maximal Probability Tree

For each z_r, the dominant term in (30) is

$$\max_{g \in G(z_r)} \prod_{(z' \to z) \in g} P_{z'z}. \tag{45}$$

Define S_z as the personal state vector in z. We decompose the transition probabilities as follows

$$P_{z'z} = \Pr(z \mid z') = \Pr(S_z, \bar{a}_z \mid S_{z'}, \bar{a}_{z'}) = \Pr(\bar{a}_z \mid S_{z'}, \bar{a}_{z'}) \Pr(S_z \mid S_{z'}, \bar{a}_{z'}, \bar{a}_z). \tag{46}$$

So the objective becomes

$$\prod_{(z' \to z) \in g} P_{z'z} = \prod_{(z' \to z) \in g} \Pr(\bar{a}_z \mid S_{z'}, \bar{a}_{z'}) \prod_{(z' \to z) \in g} \Pr(S_z \mid S_{z'}, \bar{a}_{z'}, \bar{a}_z). \tag{47}$$

Define $\boldsymbol{a}^{z_r}$ as the strategy profile associated with z_r and define its objective by $J_{z_r} = \max_{a} \sum_{n=1}^{N} u_n(\boldsymbol{a}^{z_r})$. Throughout the proof we use the following definition.

Definition 20. Define Z_d as the set of states for which exactly d players are discontent (so $N-d$ players are content).

We first provide an upper bound for (45) by upper bounding each factor in (47) separately. We then construct a specific tree $g \in G(z_r)$ that achieves the upper bound.

1) Upper bounding the probability of personal state transition - $\prod_{(z' \to z) \in g} \Pr(S_z \mid S_{z'}, \bar{a}_{z'}, \bar{a}_z)$ *:* If $z_r \in Z_0$ then in the path from $z_N \in Z_N$ to z_r, the personal state of each player needs to change from discontent to content with $\boldsymbol{a}^{z_r}$. This occurs with a probability of no more than $\left(\prod_{n=1}^{N} \frac{u_n(\boldsymbol{a}^{z_r})}{u_{n,\max}}\right) \varepsilon^{\sum_n u_{n,\max} - J_{z_r}}$ since each transition to content with u_n is only possible with probability $\frac{u_n}{u_{n,\max}} \varepsilon^{u_{n,\max} - u_n}$ (see (8)). From the second constraint on $G(z_r)$ (see Definition 5), this path exists in any tree in $G(z_r)$ such that $z_r \in Z_0$. We conclude that for any $z_r \in Z_0$

$$\prod_{(z' \to z) \in g} \Pr(S_z \mid S_{z'}, \bar{a}_{z'}, \bar{a}_z) \leq \varepsilon^{\sum_n u_{n,\max} - J_z} \prod_{n=1}^{N} \frac{u_n(\boldsymbol{a}^{z_r})}{u_{n,\max}}. \tag{48}$$

If $z_r \in Z_d$ for $d \geq 1$ then trivially

$$\prod_{(z' \to z) \in g} \Pr(S_z \mid S_{z'}, \bar{a}_{z'}, \bar{a}_z) \leq 1. \tag{49}$$

2) Upper bounding the probability of choosing a different arm - $\prod_{(z' \to z) \in g} \Pr(\bar{a}_z \mid S_{z'}, \bar{a}_{z'})$*:* Assume that $\varepsilon < \left(1 - \frac{1}{M}\right)^{\frac{1}{c}}$ so $\frac{\varepsilon^c}{M-1} < 1 - \varepsilon^c$ and $1 - \varepsilon^c > \frac{1}{M}$. We use the first constraint on $G(z_r)$ (see Definition 5) and count the most likely outgoing edges from each non-root state.

1) Any outgoing transition from $z' \in Z_0$ is bounded by the maximal probability transition (in (5))

$$\Pr(\bar{a}_z \mid S_{z'}, \bar{a}_{z'}, d = 0) \leq (1 - \varepsilon^c)^{N-1} \frac{\varepsilon^c}{M-1} \tag{50}$$

even if the total state transition is infeasible.

2) Any outgoing transition from $z' \in Z_d$ for $d \geq 1$ is bounded by the maximal probability transition (in (5) and (6))

$$\Pr(\bar{a}_z \mid S_{z'}, \bar{a}_{z'}, d \geq 1) \leq (1 - \varepsilon^c)^{N-d} \frac{1}{M^d} \tag{51}$$

even if the total state transition is infeasible.

Since all $z \in Z \setminus \{z_r\}$ must have a single outgoing edge in g, going over all paths in $(z' \to z) \in g$ must at least include all these outgoing edges. By sorting these outgoing edges according to the number of discontent players in their source state, we conclude that for $z_r \in Z_0$

$$\prod_{(z' \to z) \in g} \Pr\left(\bar{a}_z \mid S_{z'}, \bar{a}_{z'}\right) \leq \left(\frac{\left(1-\varepsilon^c\right)^{N-1} \varepsilon^c}{M-1}\right)^{|Z_0|-1} \prod_{d'=1}^{N} \left(\left(1-\varepsilon^c\right)^{N-d'} \frac{1}{M^{d'}}\right)^{|Z_{d'}|} \tag{52}$$

and for $z_r \in Z_d$ for $d \geq 1$ we have

$$\prod_{(z' \to z) \in g} \Pr\left(\bar{a}_z \mid S_{z'}, \bar{a}_{z'}\right) \leq \frac{M^d}{\left(1-\varepsilon^c\right)^{N-d}} \left(\frac{\left(1-\varepsilon^c\right)^{N-1} \varepsilon^c}{M-1}\right)^{|Z_0|} \prod_{d'=1}^{N} \left(\left(1-\varepsilon^c\right)^{N-d'} \frac{1}{M^{d'}}\right)^{|Z_{d'}|} \tag{53}$$

where the factor $\frac{M^d}{(1-\varepsilon^c)^{N-d}}$ compensates for counting one Z_d state too many (the root).

3) Upper bounding $\prod_{(z' \to z) \in g} P_{z'z}$*:* Define

$$A\left(\varepsilon, N, M\right) = \left(\frac{\left(1-\varepsilon^c\right)^{N-1} \varepsilon^c}{M-1}\right)^{|Z_0|-1} \prod_{d'=1}^{N} \left(\left(1-\varepsilon^c\right)^{N-d'} \frac{1}{M^{d'}}\right)^{|Z_{d'}|}. \tag{54}$$

Gathering the terms together, we conclude that

$$\max_{g \in G(z_r)} \prod_{(z' \to z) \in g} P_{z'z} \leq \begin{cases} \left(\prod_{n=1}^{N} \frac{u_n(\boldsymbol{a}^{z_r})}{u_{n,\max}}\right) \varepsilon^{\sum_n u_{n,\max} - J_{z_r}} A\left(\varepsilon, N, M\right) & z_r \in Z_0 \\ \left(1-\varepsilon^c\right)^{d-1} \frac{M^d}{M-1} \varepsilon^c A\left(\varepsilon, N, M\right) & z_r \in Z_d,\ d \geq 1 \end{cases}. \tag{55}$$

4) Constructing a tree that achieves the bound in (55) *:* Now we construct a specific tree in $G(z_r)$ that achieves the upper bound of (55), and hence must maximize (45). We call this tree the maximal tree of z_r. This tree consists solely of the maximal probability outgoing edges of each state. Fig. 3 illustrates the maximal tree of a content state $z_0 \in Z_0$ for $N = M = 2$. Note that the maximal probability path between two states is not necessarily the shortest one. For instance, there is a single edge in Z from the lower content state to the root, with a probability of ε^{2c}. However, for $c > u_{\max,1} - u_1 + u_{\max,2} - u_2$ and a small enough ε, the probability of the most likely path is $\frac{\mu_1 \mu_2}{u_{\max,1} u_{\max,2}} \left(1-\varepsilon^c\right) \varepsilon^c \varepsilon^{u_{\max,1} - u_1 + u_{\max,2} - u_2}$.

The maximal tree is constructed as follows:

1) Connect all $z_0 \in Z_0$ to some state $z_2 \in Z_2$ with a probability of $\left(1-\varepsilon^c\right)^{N-1} \frac{\varepsilon^c}{M-1}$. This is possible since the player that changed her arm (with a probability of $\frac{\varepsilon^c}{M-1}$) chooses the arm of one of the other players, making both of them discontent with probability 1 since they both receive $u_n = 0$.
2) Connect all $z_d \in Z_d$ with $1 \leq d < \frac{N}{2}$ to some state $z_{d'} \in Z_{d'}$ with $d' > d$ with a probability of $\left(1-\varepsilon^c\right)^{N-d} \frac{1}{M^d}$. This is possible when all the discontent players choose, with a probability of $\frac{1}{M^d}$, one or more of the arms of the content players (all kept their arms with a probability of $\left(1-\varepsilon^c\right)^{N-d}$), thus making all the colliding players discontent with probability 1.
3) Repeat Step 2 for $z_{d'} \in Z_{d'}$ until $d' \geq \frac{N}{2}$.
4) Connect all $z_d \in Z_d$ with $\frac{N}{2} \leq d < N$ to some state $z_N \in Z_N$ with a probability of $\left(1-\varepsilon^c\right)^{N-d} \frac{1}{M^d}$. This is possible since the discontent players together can choose all the arms of the content players. Hence, after this transition, all players receive $u_n = 0$ and become discontent with probability one.
5) Choose $\tilde{z}_N \in Z_N$ such that all players are in a collision. Connect all other $z_N \in Z_N$ to $\tilde{z}_N$ with a probability of $\frac{1}{M^N}$, recalling that the collision makes all players discontent with probability 1.
6) If $z_r \in Z_N$ then pick $\tilde{z}_N$ as the root. Otherwise if $z_r \in Z_0$ then disconnect the outgoing edge of z_r from step 1 and connect $\tilde{z}_N \in Z_N$ to z_r with a probability of $\frac{1}{M^N} \left(\prod_{n=1}^{N} \frac{u_n(\boldsymbol{a}^{z_r})}{u_{n,\max}}\right) \varepsilon^{\sum_n u_{n,\max} - J_{z_r}}$.

If the root is $z_r \in Z_N$ then we have constructed $g_N \in G(z_r)$ with

$$\prod_{(z' \to z) \in g} P_{z'z} = \left(1-\varepsilon^c\right)^{N-1} \frac{M^N}{M-1} \varepsilon^c A\left(\varepsilon, N, M\right). \tag{56}$$

If the root is $z_r \in Z_0$, then we have constructed $\widetilde{g}_N \in G(z_r)$ with

$$\prod_{(z' \to z) \in g} P_{z'z} = \left(\prod_{n=1}^{N} \frac{u_n\left(a^{z_r}\right)}{u_{n,\max}}\right) \varepsilon^{\sum_n u_{n,\max} - J_{z_r}} A\left(\varepsilon, N, M\right). \tag{57}$$

The maximal tree of z_r is depicted in Fig. 3.

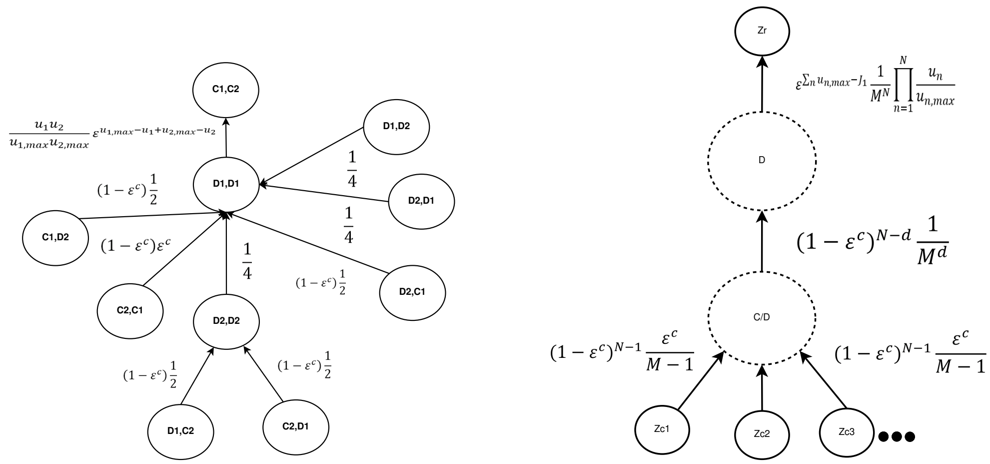


Fig. 3: The maximal tree of $z_r \in Z_0$, for $N = M = 2$ and for the general case. C1,D2 is the state where player 1 is content with arm 1 and player 2 is discontent with arm 2.

B. Upper bounding the probability of all non z^* trees

We are left with upper bounding $\sum_{z_r\in Z, z_r\neq z^*}\sum_{g\in G(z_r)}\prod_{(z'\to z)\in g} P_{z'z}$. First we identify a probability factor that appears in $\prod_{(z'\to z)\in g} P_{z'z}$ for every tree g. In any possible outgoing edge from $z_d \in Z_d$ there is a probability of $\frac{1}{M^d}$ for the action choices of the d discontent players. In a similar manner, there is a factor of $\frac{\varepsilon^c}{M-1}$ in the probability of any possible outgoing edge from any $z_0 \in Z_0$. Hence, any tree has

$$B(\varepsilon, N, M) = \left(\frac{\varepsilon^c}{M-1}\right)^{|Z_0|-1}\prod_{d'=1}^{N}\left(\frac{1}{M^{d'}}\right)^{|Z_{d'}|} \tag{58}$$

as a factor of $\prod_{(z'\to z)\in g} P_{z'z}$. Moreover:

1) All trees $g \in G(z_r)$ with $z_r \in Z_0$ have a factor of $\prod_{n=1}^{N}\frac{u_n(\boldsymbol{a}^{z_r})}{u_{n,\max}}\varepsilon^{\sum_n u_{n,\max}-J_{z_r}}$ in $\prod_{(z'\to z)\in g} P_{z'z}$ since there must exist a path from any $z_N \in Z_N$ to the root, and all players change from being discontent to being content with $\boldsymbol{a}_{z_r}$ at least once on this path.
2) All trees $g \in G(z_r)$ with $z_r \notin Z_0$ must have a factor of $\left(\frac{\varepsilon^c}{M-1}\right)^{|Z_0|}$ (instead of $\left(\frac{\varepsilon^c}{M-1}\right)^{|Z_0|-1}$) in $\prod_{(z'\to z)\in g} P_{z'z}$ since there are $|Z_0|$ nodes from Z_0 to be connected.

Now we divide all the trees in $G(z)$ into different categories. Define $a^z_{l,q}$ as the number of trees in $G(z)$ that have an l extra $\varepsilon^{u_{n,\max}-u_n} < 1$ factors and q extra ε^c factors, where extra means more than the factors that must exist in any tree as argued above. Note that since there are N players and $|Z|-1$ edges in every tree, l and q cannot exceed $N|Z|$. Let $\mathcal{A}^n_{\max} = \arg\max_{\boldsymbol{a}} u_n(\boldsymbol{a})$ and define

$$\alpha = \min_{n}\min_{\boldsymbol{a}\notin\mathcal{A}^n_{\max}}\left(u_{n,\max} - u_n(\boldsymbol{a})\right). \tag{59}$$

Also define $J_0 \triangleq \sum_n u_{n,\max}$ and recall Definition 3 of the second best objective J_2. Using the above definitions, we can write

$$\sum_{z_r\in Z, z\neq z^*} \sum_{g\in G(z_r)} \prod_{(z'\to z)\in g} P_{z'z} = B(\varepsilon,N,M) \frac{\varepsilon^c}{M-1} \sum_{d=1}^{N} \sum_{z_r\in Z_d} \sum_{g\in G(z_r)} \frac{\prod_{(z'\to z)\in g} P_{z'z}}{\frac{\varepsilon^c}{M-1} B(\varepsilon,N,M)} +$$
$$B(\varepsilon,N,M) \sum_{z_r\in Z_0, z\neq z^*} \varepsilon^{J_0-J_{z_r}} \prod_{n=1}^{N} \frac{u_n(\boldsymbol{a}^{z_r})}{u_{n,\max}} \sum_{g\in G(z_r)} \frac{\prod_{(z'\to z)\in g} P_{z'z}}{\varepsilon^{J_0-J_{z_r}} \prod_{n=1}^{N} \frac{u_n(\boldsymbol{a}^{z_r})}{u_{n,\max}} B(\varepsilon,N,M)} \underset{(a)}{\leq}$$
$$B(\varepsilon,N,M) \left(\frac{\varepsilon^c}{M-1} \sum_{d=1}^{N} M^d \sum_{z_r\in Z_d} \sum_{l=0}^{N|Z|} \sum_{q=0}^{N|Z|} a_{l,q}^{z_r} \varepsilon^{l\alpha} \varepsilon^{qc} + \sum_{z_r\in Z_0, z\neq z^*} \varepsilon^{J_0-J_{z_r}} \prod_{n=1}^{N} \frac{u_n(\boldsymbol{a}^{z_r})}{u_{n,\max}} \sum_{l=0}^{N|Z|} \sum_{q=0}^{N|Z|} a_{l,q}^{z_r} \varepsilon^{l\alpha} \varepsilon^{qc} \right)$$
$$\underset{(b)}{\leq} \frac{B(\varepsilon,N,M)}{(1-\varepsilon^\alpha)(1-\varepsilon^c)} \left(\frac{\varepsilon^c}{M-1} \sum_{d=1}^{N} M^d \sum_{z_r\in Z_d} \max_{l,q} a_{l,q}^{z_r} + \sum_{z_r\in Z_0, z\neq z^*} \max_{l,q} a_{l,q}^{z_r} \prod_{n=1}^{N} \frac{u_n(\boldsymbol{a}^{z_r})}{u_{n,\max}} \varepsilon^{J_0-J_{z_r}} \right) \leq$$
$$\frac{B(\varepsilon,N,M)}{(1-\varepsilon^\alpha)(1-\varepsilon^c)} \left(|Z| \max_{z_r,l,q} a_{l,q}^{z_r} \right) \left(\frac{M\left(M^N-1\right)}{(M-1)^2} \varepsilon^c + \varepsilon^{J_0-J_2} \right) \tag{60}$$

where (a) follows since $B(\varepsilon,N,M)$ and $\frac{\varepsilon^c}{M-1}$ or $\varepsilon^{J_0-J_{z_r}} \prod_{n=1}^{N} \frac{u_n(\boldsymbol{a}^{z_r})}{u_{n,\max}}$ are factors of $\prod_{(z'\to z)\in g} P_{z'z}$, and so are $\varepsilon^{ld}\varepsilon^{qc}$ if g is counted by $a_{l,q}^{z}$. The factor M^d compensates for counting, in (58), the (non-existing) outgoing edge of $z_r \in Z_d$. Inequality (b) follows from the two geometric series that appear by bounding as follows

$$\sum_{l=0}^{N|Z|} \sum_{q=0}^{N|Z|} a_{l,q}^{z_r} \varepsilon^{l\alpha} \varepsilon^{qc} \leq \max_{l,q} a_{l,q}^{z_r} \sum_{l=0}^{\infty} \sum_{q=0}^{\infty} \varepsilon^{l\alpha} \varepsilon^{qc} = \frac{1}{1-\varepsilon^\alpha} \frac{1}{1-\varepsilon^c} \max_{l,q} a_{l,q}^{z_r}. \tag{61}$$

C. *Lower bounding π_{z^*}*

Using (60) we obtain

$$\frac{\sum_{z_r\in Z, z\neq z^*} \sum_{g\in G(z_r)} \prod_{(z'\to z)\in g} P_{z'z}}{\sum_{g\in G(z^*)} \prod_{(z'\to z)\in g} P_{z'z}} \underset{(a)}{\leq} \frac{\frac{B(\varepsilon,N,M)}{(1-\varepsilon^\alpha)(1-\varepsilon^c)} \left(|Z| \max_{z_r,l,q} a_{l,q}^{z_r} \right) \left(\frac{M\left(M^N-1\right)}{(M-1)^2} \varepsilon^c + \varepsilon^{J_0-J_2} \right)}{G^* \left(\prod_{n=1}^{N} \frac{u_n(\boldsymbol{a}^*)}{u_{n,\max}} \right) \varepsilon^{J_0-J_1} A(\varepsilon,N,M)} =$$
$$\frac{\prod_{n=1}^{N} \frac{u_{n,\max}}{u_n(\boldsymbol{a}^*)}}{(1-\varepsilon^\alpha)(1-\varepsilon^c)} \frac{\left(\frac{|Z|}{G^*} \max_{z_r,l,q} a_{l,q}^{z_r} \right) \left(\frac{M\left(M^N-1\right)}{(M-1)^2} \varepsilon^{c-J_0+J_1} + \varepsilon^{J_1-J_2} \right)}{(1-\varepsilon^c)^{(N-1)(|Z_0|-1)} \prod_{d'=1}^{N} (1-\varepsilon^c)^{(N-d')|Z_{d'}|}} \underset{(b)}{\leq}$$
$$\frac{\prod_{n=1}^{N} \frac{u_{n,\max}}{u_n(\boldsymbol{a}^*)}}{(1-\varepsilon^\alpha)(1-\varepsilon^c)^{(N-1)2^N M^N}} \left(2^N M^N \frac{\max_{z_r,l,q} a_{l,q}^{z_r}}{G^*} \right) \left(3M^{N-1} \varepsilon^{c-J_0+J_1} + \varepsilon^{J_1-J_2} \right) \tag{62}$$

where in (a) we used (60) for the numerator, where G^* denotes the number of maximal trees of z^* (from Subsection IX-A4) for the denominator. Inequality (b) follows since $\sum_{d'=1}^{N} (N-d') |Z_{d'}| \leq (N-1)(|Z|-|Z_0|)$ and $|Z| = 2^N M^N$. We also used $\frac{M\left(M^N-1\right)}{(M-1)^2} \leq 3M^{N-1}$, which holds for all $M > 1$.

We conclude that if $c > \sum_n u_{n,\max} - J_1$ then (62) vanishes to zero as $\varepsilon \to 0$. Hence, z^* is the most likely state in the stationary distribution of Z for small enough ε. Specifically, having

$$\varepsilon < \min \left\{ \left(\prod_{n=1}^{N} \frac{u_n(\boldsymbol{a}^*)}{u_{n,\max}} (2M)^{-N} \frac{3G^*}{8 \max_{z_r,l,q} a_{l,q}^{z_r}} \right)^{\frac{1}{J_1-J_2}}, \left(\prod_{n=1}^{N} \frac{u_n(\boldsymbol{a}^*)}{u_{n,\max}} (2M)^{-N} M^{-N+1} \frac{G^*}{8 \max_{z_r,l,q} a_{l,q}^{z_r}} \right)^{\frac{1}{c-J_0+J_1}} \right\} \tag{63}$$

together with

$$\varepsilon < \min \left\{ \frac{1}{10^{1/\alpha}}, \left(1 - \left(\frac{9}{10} \right)^{\frac{1}{(N-1)2^N M^N}} \right)^{1/c} \right\} \tag{64}$$

is enough to ensure that $\pi_{z^*} > \frac{1}{2}$. Note that (63) is typically much stricter than (64). Also, $\prod_{n=1}^{N} \frac{u_n(\boldsymbol{a}^*)}{u_{n,\max}}$ is typically not much below one. Evaluating the factor $\frac{G^*}{\max_{z_r,l,q} a_{l,q}^{z_r}}$ involves counting the number of trees in each category and can be done by a computer program.

References

[1] I. Bistritz and A. Leshem, "Distributed multi-player bandits-a game of thrones approach," in *Advances in Neural Information Processing Systems*, 2018, pp. 7222–7232.

[2] J.-Y. Audibert, S. Bubeck, and G. Lugosi, "Regret in online combinatorial optimization," *Mathematics of Operations Research*, vol. 39, no. 1, pp. 31–45, 2013.

[3] N. Cesa-Bianchi, G. Lugosi, and G. Stoltz, "Regret minimization under partial monitoring," *Mathematics of Operations Research*, vol. 31, no. 3, pp. 562–580, 2006.

[4] E. V. Denardo, H. Park, and U. G. Rothblum, "Risk-sensitive and risk-neutral multiarmed bandits," *Mathematics of Operations Research*, vol. 32, no. 2, pp. 374–394, 2007.

[5] V. Krishnamurthy and B. Wahlberg, "Partially observed Markov decision process multiarmed bandits-structural results," *Mathematics of Operations Research*, vol. 34, no. 2, pp. 287–302, 2009.

[6] D. Russo and B. Van Roy, "Learning to optimize via posterior sampling," *Mathematics of Operations Research*, vol. 39, no. 4, pp. 1221–1243, 2014.

[7] J. Y. Yu, S. Mannor, and N. Shimkin, "Markov decision processes with arbitrary reward processes," *Mathematics of Operations Research*, vol. 34, no. 3, pp. 737–757, 2009.

[8] S. Agrawal, V. Avadhanula, V. Goyal, and A. Zeevi, "Mnl-bandit: a dynamic learning approach to assortment selection," 2017, arXiv preprint arXiv:1706.03880.

[9] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in applied mathematics*, vol. 6, no. 1, pp. 4–22, 1985.

[10] S. Shahrampour, A. Rakhlin, and A. Jadbabaie, "Multi-armed bandits in multi-agent networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, 2017, pp. 2786–2790.

[11] E. Hillel, Z. S. Karnin, T. Koren, R. Lempel, and O. Somekh, "Distributed exploration in multi-armed bandits," in *Advances in Neural Information Processing Systems*, 2013, pp. 854–862.

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

[13] N. Cesa-Bianchi, C. Gentile, Y. Mansour, and A. Minora, "Delay and cooperation in nonstochastic bandits," in *Conference on Learning Theory*, 2016, pp. 605–622.

[14] N. Korda, B. Szörényi, and L. Shuai, "Distributed clustering of linear bandits in peer to peer networks," in *International Conference on Machine Learning*, vol. 48, 2016, pp. 1301–1309.

[15] B. Szorenyi, R. Busa-Fekete, I. Hegedus, R. Ormándi, M. Jelasity, and B. Kégl, "Gossip-based distributed stochastic bandit algorithms," in *International Conference on Machine Learning*, 2013, pp. 19–27.

[16] M. Dimakopoulou and B. Van Roy, "Coordinated exploration in concurrent reinforcement learning," in *International Conference on Machine Learning*, 2018, pp. 1270–1278.

[17] J. Rosenski, O. Shamir, and L. Szlak, "Multi-player bandits–a musical chairs approach," in *International Conference on Machine Learning*, 2016, pp. 155–163.

[18] N. Nayyar, D. Kalathil, and R. Jain, "On regret-optimal learning in decentralized multi-player multi-armed bandits," *IEEE Transactions on Control of Network Systems*, vol. 5, no. 1, pp. 597–606, 2018.

[19] D. Kalathil, N. Nayyar, and R. Jain, "Decentralized learning for multiplayer multiarmed bandits," *IEEE Transactions on Information Theory*, vol. 60, no. 4, pp. 2331–2345, 2014.

[20] K. Liu and Q. Zhao, "Distributed learning in multi-armed bandit with multiple players," *IEEE Transactions on Signal Processing*, vol. 58, no. 11, pp. 5667–5681, 2010.

[21] S. Vakili, K. Liu, and Q. Zhao, "Deterministic sequencing of exploration and exploitation for multi-armed bandit problems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 5, pp. 759–767, 2013.

[22] L. Lai, H. Jiang, and H. V. Poor, "Medium access in cognitive radio networks: A competitive multi-armed bandit framework," in *Signals, Systems and Computers, 2008 42nd Asilomar Conference on*, 2008, pp. 98–102.

[23] A. Anandkumar, N. Michael, A. K. Tang, and A. Swami, "Distributed algorithms for learning and cognitive medium access with logarithmic regret," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 4, pp. 731–745, 2011.

[24] H. Liu, K. Liu, and Q. Zhao, "Learning in a changing world: Restless multiarmed bandit with unknown dynamics," *IEEE Transactions on Information Theory*, vol. 59, no. 3, pp. 1902–1916, 2013.

[25] O. Avner and S. Mannor, "Concurrent bandits and cognitive radio networks," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2014, pp. 66–81.

[26] N. Evirgen and A. Kose, "The effect of communication on noncooperative multiplayer multi-armed bandit problems," in *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2017.

[27] C. H. Papadimitriou and K. Steiglitz, *Combinatorial optimization: algorithms and complexity.* Courier Corporation, 1998.

[28] D. P. Bertsekas, "The auction algorithm: A distributed relaxation method for the assignment problem," *Annals of operations research*, vol. 14, no. 1, pp. 105–123, 1988.

[29] M. M. Zavlanos, L. Spesivtsev, and G. J. Pappas, "A distributed auction algorithm for the assignment problem," in *Decision and Control, 2008. CDC 2008. 47th IEEE Conference on*, 2008, pp. 1212–1217.

[30] O. Naparstek and A. Leshem, "Fully distributed optimal channel assignment for open spectrum access," *IEEE Transactions on Signal Processing*, vol. 62, no. 2, pp. 283–294, 2013.

[31] O. Avner and S. Mannor, "Multi-user lax communications: a multi-armed bandit approach," in *INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications, IEEE*, 2016, pp. 1–9.

[32] S. Zafaruddin, I. Bistritz, A. Leshem, and D. Niyato, "Distributed learning for channel allocation over a shared spectrum," *to appear in IEEE Journal on Selected Areas in Communications, arXiv:1902.06353*, 2019.

[33] B. S. Pradelski and H. P. Young, "Learning efficient Nash equilibria in distributed systems," *Games and Economic behavior*, vol. 75, no. 2, pp. 882–897, 2012.

[34] J. R. Marden, H. P. Young, and L. Y. Pao, "Achieving pareto optimality through distributed learning," *SIAM Journal on Control and Optimization*, vol. 52, no. 5, pp. 2753–2770, 2014.

[35] A. Menon and J. S. Baras, "Convergence guarantees for a decentralized algorithm achieving pareto optimality," in *American Control Conference (ACC), 2013*, 2013.

[36] J. V. Uspensky, *Introduction to mathematical probability.* McGraw-Hill Book Company, New York, 1937.

[37] M. I. Freidlin and A. D. Wentzell, *Random Perturbations of Dynamical Systems.* Springer, 1998.

[38] K.-M. Chung, H. Lam, Z. Liu, and M. Mitzenmacher, "Chernoff-Hoeffding bounds for Markov chains: Generalized and simplified," in *29th International Symposium on Theoretical Aspects of Computer Science*, 2012.

[39] C. Tekin and M. Liu, "Online learning of rested and restless bandits," *IEEE Transactions on Information Theory*, vol. 58, no. 8, pp. 5588–5611, 2012.

[40] T. L. Lai and H. Robbins, "Asymptotically optimal allocation of treatments in sequential experiments," *Design of Experiments: Ranking and Selection*, pp. 127–142, 1984.