A Practical Algorithm for Multiplayer Bandits when Arm Means Vary Among Players

Etienne Boursier, Emilie Kaufmann, Abbas Mehrabian, Vianney Perchet

A Practical Algorithm for Multiplayer Bandits when Arm Means Vary Among Players

Etienne Boursier
Université Paris-Saclay, ENS Paris-Saclay
CNRS, Centre Borelli, Cachan, France
eboursie@ens-paris-saclay.fr

Emilie Kaufmann
Univ. Lille, CNRS, Inria SequeL,
UMR 9189 - CRIStAL, Lille, France
emilie.kaufmann@univ-lille.fr

Abbas Mehrabian
McGill University, Montréal, Canada
abbas.mehrabian@gmail.com

Vianney Perchet
CREST, ENSAE Paris, Palaiseau, France
Criteo AI Lab, Paris, France
vianney.perchet@normalesup.org

Abstract

We study a multiplayer stochastic multi-armed bandit problem in which players cannot communicate, and if two or more players pull the same arm, a collision occurs and the involved players receive zero reward. We consider the challenging *heterogeneous* setting, in which different arms may have different means for different players, and propose a new and efficient algorithm that combines the idea of leveraging forced collisions for implicit communication and that of performing matching eliminations. We present a finite-time analysis of our algorithm, giving the first sublinear minimax regret bound for this problem, and prove that if the optimal assignment of players to arms is unique, our algorithm attains the optimal $O(\ln(T))$ regret, solving an open question raised at NeurIPS 2018 by Bistritz and Leshem (2018).

1 Introduction

Stochastic multi-armed bandit models have been studied extensively as they capture many sequential decision-making problems of practical interest. In the simplest setup, an agent repeatedly chooses among several actions (referred to as "arms") in each round of a game. To each action i is associated a real-valued parameter μ_i. Whenever the player performs the ith action ("pulls arm i"), she receives a random reward with mean μ_i. The player's objective is to maximize the sum of rewards obtained during the game. If she knew the means associated with the actions before starting the game, she would play an action with the largest mean reward during all rounds. The problem is to design a strategy for the player to maximize her reward in the setting where the means are unknown. The *regret* of the strategy is the difference between the accumulated rewards in the two scenarios.

To minimize the regret, the player is faced with an exploration/exploitation trade-off as she should try (explore) all actions to estimate their means accurately enough but she may want to exploit the action that looks *probably* best given her current information. We refer the reader to (Bubeck and Cesa-Bianchi, 2012; Lattimore and Szepesvári, 2020) for surveys on this problem. Multi-armed bandit (MAB) has been first studied as a simple model for sequential clinical trials (Thompson, 1933; Robbins, 1952) but has also found many modern applications to online content optimization, such as the design of recommender systems (Li et al., 2010). Recently, MAB algorithms have also been investigated for cognitive radios (Jouini et al., 2009; Anandkumar et al., 2011). In this context, arms model the available radio channels on which radio devices can communicate, and the reward associated with each arm is either a binary indicator of the success of a communication on that channel or some measure of its quality.

The applications to cognitive radios have motivated the *multiplayer* bandit problem, in which several agents (devices) play on the same bandit (communicate using the same channels). If two or more agents pull the same arm, a *collision* occurs and all agents pulling that arm receive zero reward. Without communicating, each agent must adopt a strategy aimed at maximizing the global reward obtained by all agents—so, we are considering a cooperative scenario rather than a competitive one. While most previous work on this problem focuses on the case in which the means of the arms are identical across players (the homogeneous variant), in this paper we study the more challenging heterogeneous variant, in which each user may have a different utility for each arm: if player m selects arm k, she receives a reward with mean μ_k^m. This variant is more realistic for applications to cognitive radios, as the quality of each channel may vary from one user (device) to another, depending for instance on its configuration and location.

More precisely, we study the model introduced by Bistritz and Leshem (2018), which has two main characteristics:

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

first, each arm has a possibly different mean for each player; second, we are in a fully distributed setting with no communication between players. Let T denote the time horizon. Bistritz and Leshem (2018) proposed an algorithm with regret bounded by $O((\ln T)^{2+\kappa})$ (for any constant κ), proved a lower bound of $\Omega(\ln T)$ *for any algorithm*, and asked if there is an algorithm matching this lower bound. In this paper, we propose a new algorithm for this model, M-ETC-Elim, which depends on a hyperparameter c, and we upper bound its regret by $O(\ln(T)^{1+1/c})$ for any $c > 1$. We also bound its worst-case regret by $O(\sqrt{T \ln T})$, which is the first sublinear minimax bound for this problem. Moreover, if the optimal assignment of the players to the arms is unique, we prove that instantiating M-ETC-Elim with $c = 1$ yields regret at most $O(\ln(T))$, which is optimal and answers affirmatively the open question mentioned above in this particular case.[1] We present a non-asymptotic regret analysis of M-ETC-Elim leading to nearly optimal regret upper bounds, and also demonstrate the empirical efficiency of this new algorithm via simulations.

Outline In Section 2, we formally introduce the heterogeneous multiplayer multi-armed bandit model and present our contributions. These results are put in perspective by comparison with the literature given in Section 3. We describe the M-ETC-Elim algorithm in Section 4 and upper bound its regret in Section 5. Finally, we report in Section 6 results from an experimental study demonstrating the competitive practical performance of M-ETC-Elim.

2 Model and Contributions

We study a multi-armed bandit model where M players compete over K arms, with $M \leq K$. We denote by μ_k^m the mean reward (or expected utility) of arm k for player m. In each round $t = 1, 2, \ldots, T$, player m selects arm $A^m(t)$ and receives a reward

$$R^m(t) = Y^m_{A^m(t),t} \left(1 - \mathbb{1}\left(\mathcal{C}_{A^m(t),t}\right)\right),$$

where $(Y^m_{k,t})_{t=1}^{\infty}$ is an i.i.d. sequence with mean μ_k^m taking values in $[0, 1]$, $\mathcal{C}_{k,t}$ is the event that at least two players have chosen arm k in round t (i.e., a collision occurs), and $\mathbb{1}(\mathcal{C}_{k,t})$ is the corresponding indicator function. In the cognitive radio context, $Y^m_{k,t}$ is the quality of channel k for player m if she were to use this channel in isolation in round t, but her actual reward is zero if a collision occurs.

We assume that each player m in each round t observes her reward $R^m(t)$ and the collision indicator $\mathbb{1}\left(\mathcal{C}_{A^m(t),t}\right)$. Note that in the special case in which the reward distributions satisfy $\mathbb{P}(Y^m_{k,t} = 0) = 0$ (e.g., if the corresponding distribution is continuous), the collision indicator $\mathbb{1}\left(\mathcal{C}_{A^m(t),t}\right)$ can be reconstructed from the observation of $R^m(t)$. The decision of player m at round t can depend only on her past observations; that is, $A^m(t)$ is $\mathcal{F}^m_{t-1}$ measurable, where $\mathcal{F}^m_t = \sigma(A^m(1), R^m(1), \mathbb{1}\left(\mathcal{C}_{A^m(1),1}\right), \ldots, A^m(t), R^m(t), \mathbb{1}\left(\mathcal{C}_{A^m(t),t}\right))$.

Hence, our setting is fully distributed: a player cannot use extra information such as observations made by others to make her decisions. Under this constraint, we aim at maximizing the global reward collected by all players. If the mean rewards μ_k^m were known and a central controller would assign arms to players, this would boil down to finding a maximum matching between players and arms.

A *matching* is a one-to-one assignment of players to arms; formally, any one-to-one function $\pi : [M] \to [K]$ is a matching, where we use the shorthand $[n] \coloneqq \{1, \ldots, n\}$ for any integer n. The *utility* (or *weight*) of a matching π is defined as $U(\pi) \coloneqq \sum_{m=1}^{M} \mu^m_{\pi(m)}$. We denote by $\mathcal{M}$ the set of all matchings and let $U^\star \coloneqq \max_{\pi \in \mathcal{M}} U(\pi)$ denote the maximum attainable utility. A *maximum matching* (or *optimal matching)* is a matching with utility $U^\star$. The strategy maximizing the social utility of the players (i.e. the sum of all their rewards) would be to play in each round according to a maximum matching, and the *(expected) regret* with respect to that oracle is defined as

$$R_T \coloneqq TU^\star - \mathbb{E}\left[\sum_{t=1}^{T}\sum_{m=1}^{M} R^m(t)\right].$$

Our goal is to design a strategy (a sequence of arm pulls) for each player that minimizes the regret. Our regret bounds will depend on the gap between the utility of the best matching and the utility of the matching with the second best utility, defined as $\Delta \coloneqq \inf_{\pi : \Delta(\pi) > 0} \Delta(\pi)$, where $\Delta(\pi) \coloneqq U^\star - U(\pi)$. Note that $\Delta > 0$ even in the presence of several optimal matchings. In the degenerate case that $\Delta(\pi) = 0$ for all matchings π, we define $\Delta \coloneqq \infty$.

Contributions We propose an efficient algorithm for the heterogeneous multiplayer bandit problem achieving (quasi) logarithmic regret. The algorithm, called Multiplayer Explore-Then-Commit with matching Elimination (M-ETC-Elim), is described in detail in Section 4. It combines the idea of exploiting collisions for implicit communication, initially proposed by Boursier and Perchet (2019) for the homogeneous setting (which we have improved and adapted to our setting), with an efficient way to perform "matching eliminations."

M-ETC-Elim consists of several epochs combining exploration and communication, and may end with an exploitation phase if a unique optimal matching has been found. The algorithm depends on a parameter c controlling the epoch sizes and enjoys the following regret guarantees.

[1] In practice, the optimal assignment may not be unique, but the players may circumvent this by adding a tiny random bias to their observations, independent of other players, and this will make the optimal assignment unique with high probability.

Theorem 1. *(a) The M-ETC-Elim algorithm with parameter $c \in \{1, 2, \dots\}$ satisfies*

$$R_T = O\left(MK\left(\frac{M^2\ln(T)}{\Delta}\right)^{1+1/c}\right) \text{ if } \Delta \neq \infty, \text{ and}$$

$$R_T = O\left(M^3K\log(K)\sqrt{\log T} + M^2K\log(T)^{1+1/c}\right)$$

if $\Delta = \infty$.

(b) If the maximum matching is unique, M-ETC-Elim with $c = 1$ satisfies

$$R_T = O\left(\frac{M^3K\ln(T)}{\Delta}\right).$$

(c) Regardless of whether the optimal matching is unique or not, M-ETC-Elim with $c = 1$ satisfies the minimax regret bound

$$R_T = O\left(M^{\frac{3}{2}}\sqrt{KT\ln(T)}\right).$$

We emphasize that we carry out a non-asymptotic analysis of M-ETC-Elim. The regret bounds of Theorem 1 are stated with the $O(\cdot)$ notation for the ease of presentation and the hidden constants depend on the chosen parameter c only. In Theorems 3, 8 and 9 we provide the counterparts of these results with explicit constants.

A consequence of part (a) is that for a fixed problem instance, for any (arbitrarily small) κ, there exists an algorithm (M-ETC-Elim with parameter $c = \lceil 1/\kappa \rceil$) with regret $R_T = O((\ln(T))^{1+\kappa})$. This quasi-logarithmic regret rate improves upon the $O(\ln^2(T))$ regret rate of Bistritz and Leshem (2018). Moreover, we provide additional theoretical guarantees for M-ETC-Elim using the parameter $c = 1$: an improved analysis in the presence of a unique optimal matching, which yields logarithmic regret (part (b)); and a problem-independent $O(\sqrt{T\ln T})$ regret bound (part (c)), which supports the use of this particular parameter tuning regardless of whether the optimal matching is unique. This is the first sublinear minimax regret bound for this problem.

To summarize, we present a unified algorithm that can be used in the presence of either a unique or multiple optimal matchings and get a nearly logarithmic regret in both cases, almost matching the known logarithmic lower bound. Moreover, our algorithm is easy to implement, performs well in practice and does not need problem-dependent hyperparameter tuning.

3 Related Work

Centralized Variant Relaxing the decentralization assumption, i.e., when a central controller is jointly selecting $A^1(t), \dots, A^M(t)$, our problem coincides with a combinatorial bandit problem with semi-bandit feedback, which is studied by Gai et al. (2012). More precisely, introducing $M \times K$ *elementary arms* with means μ_m^k for $m \in [M]$ and $k \in [K]$, the central controller selects at each timestep M elementary arms whose indices form a matching. Then, the reward of each chosen elementary arm is observed and the obtained reward is their sum. A well-known algorithm for this setting is CUCB (Chen et al., 2013), whose regret satisfies $R_T = O\left((M^2K/\Delta)\ln(T)\right)$ (Kveton et al., 2015). Wang and Chen (2018) also proposed a Thompson sampling-based algorithm with a similar regret bound. Improved dependency in M was obtained for the ESCB algorithm (Combes et al., 2015; Degenne and Perchet, 2016), which is less numerically appealing as it requires to compute an upper confidence bound for each matching in every round. In this work, we propose an efficient algorithm with regret upper bounded by (roughly) $O\left((M^3K/\Delta)\ln(T)\right)$ for the more challenging decentralized setting.

Homogeneous Variant Back to the decentralized setting, the particular case in which all players share a common utility for all arms, i.e. $\mu_k^m = \mu_k$ for all $m \in [M]$, has been studied extensively: the first line of work on this variant combines standard bandit algorithms with an orthogonalization mechanism (Liu and Zhao, 2010; Anandkumar et al., 2011; Besson and Kaufmann, 2018a), and obtains logarithmic regret, with a large multiplicative constant due to the number of collisions. Rosenski et al. (2016) proposes an algorithm based on a uniform exploration phase in which each player identifies the top M arms, followed by a "musical chairs" protocol that allows each player to end up at a different arm quickly. Drawing inspiration from this musical chairs protocol, Boursier and Perchet (2019) recently proposed an algorithm with an $O\left(((K-M)/\Delta + KM)\ln(T)\right)$ regret bound, which relies on two other crucial ideas: exploiting collisions for communication and performing arm eliminations. Our algorithm also leverages these two ideas, with the following enhancements. The main advantage of our communication protocol over that of Boursier and Perchet (2019) is that the followers only send each piece of information once, to the leader, instead of sending it to the $M-1$ other players. Then, while Boursier and Perchet (2019) uses *arm eliminations* (coordinated between players) to reduce the regret, we cannot employ the same idea for our heterogeneous problem, as an arm that is bad for one player might be good for another player, and therefore cannot be eliminated. Our algorithm instead relies on *matching eliminations*.

Towards the Fully Distributed Heterogeneous Setting Various semi-distributed variants of our problem in which some kind of communication is allowed between players have been studied by Avner and Mannor (2016); Kalathil et al. (2014); Nayyar et al. (2018). In particular, the algorithms proposed by Kalathil et al. (2014); Nayyar et al. (2018) require a pre-determined channel dedicated to communications: in some phases of the algorithm, players in turn send information (sequences of bits) on this channel,

and it is assumed that all other players can observe the sent information.

The fully distributed setting was first studied by Bistritz and Leshem (2018), who proposed the Game-of-Thrones (GoT) algorithm and proved its regret is bounded by $O((\ln T)^{2+\kappa})$ for any given constant $\kappa > 0$, if its parameters are "appropriately tuned." In a recent preprint (Bistritz and Leshem, 2019), the same authors provide an improved analysis, showing the same algorithm (with slightly modified phase lengths) enjoys quasi-logarithmic regret $O((\ln T)^{1+\kappa})$. GoT is quite different from M-ETC-Elim: it proceeds in epochs, each consisting of an exploration phase, a so-called GoT phase and an exploitation phase. During the GoT phase, the players jointly run a Markov chain whose unique stochastically stable state corresponds to a maximum matching of the estimated means. A parameter $\varepsilon \in (0,1)$ controls the accuracy of the estimated maximum matching obtained after a GoT phase. Letting c_1, c_2, c_3 be the constants parameterizing the lengths of the phases, the improved analysis of GoT (Bistritz and Leshem, 2019) upper bounds its regret by $Mc_3 2^{k_0+1} + 2(c_1+c_2)M\log_2^{1+\kappa}(T/c_3+2)$. This upper bound is asymptotic as it holds for T *large enough*, where "how large" is not explicitly specified and *depends on* Δ.[2] Moreover, the upper bound is valid only when the parameter ε is chosen *small enough*: ε should satisfy some constraints (Equations (66)-(67)) also featuring Δ. Hence, a valid tuning of the parameter ε would require prior knowledge of arm utilities. In contrast, we provide in Theorem 3 a non-asymptotic regret upper bound for M-ETC-Elim, which holds for any choice of the parameter c controlling the epoch lengths. Also, we show that if the optimal assignment is unique, M-ETC-Elim has logarithmic regret. Besides, we also illustrate in Section 6 that M-ETC-Elim outperforms GoT in practice. Finally, GoT has several parameters to set $(\delta, \varepsilon, c_1, c_2, c_3)$, while M-ETC-Elim has only one integral parameter c, and setting $c = 1$ works very well in all our experiments.

If Δ is known, an algorithm with similar ideas to M-ETC-Elim with $O(\log T)$ regret was presented independently in the recent preprint of Magesh and Veeravalli (2019).

Finally, the recent independent preprint of Tibrewal et al. (2019) studies a slightly stronger feedback model than ours: they assume each player in each round has the option of "observing whether a given arm has been pulled by someone," without actually pulling that arm (thus avoiding collision due to this "observation"), an operation that is called "sensing." Due to the stronger feedback, communications do not need to be implicitly done through collisions and bits can be broadcast to other players via sensing. Note that it is actually possible to send a single bit of information from one player to all other players in a single round in their model, an action that requires $M-1$ rounds in our model. Still, the algorithms proposed by Tibrewal et al. (2019) can be modified to obtain algorithms for our setting, and M-ETC-Elim can also be adapted to their setting. The two algorithms proposed by Tibrewal et al. (2019) share similarities with M-ETC-Elim: they also have exploration, communication and exploitation phases, but they do not use eliminations. Regarding their theoretical guarantees, a first remark is that those proved in Tibrewal et al. (2019) only hold in the presence of a unique optimal matching, whereas our analysis of M-ETC-Elim applies in the general case. The second remark is that their regret bounds for the case in which Δ is unknown (Theorems 3(ii) and 4) feature exponential dependence on the gap $1/\Delta$, whereas our regret bounds have polynomial dependence. Finally, the first-order term of their Theorem 4 has a quadratic dependence in $1/\Delta$, whereas our Theorem 1(b) scales linearly, which is optimal (see the lower bounds section below) and allows us to get the $\widetilde{O}(\sqrt{T})$ minimax regret bound for M-ETC-Elim.

Lower Bounds The $\Omega(\ln(T))$ lower bound proven by Bistritz and Leshem (2018) hides the problem parameters; we next review the lower bounds that flesh out the dependence on K, M and Δ. In the (easier) centralized setting discussed above, an asymptotic lower bound of $\Omega((K-M)\ln(T)/\Delta)$ is proved in the homogeneous case (Anantharam et al., 1987). In the centralized heterogeneous case, Combes et al. (2015) obtain a general problem dependent lower bound for combinatorial semi-bandits of the form $c(\boldsymbol{\mu}, M)\ln(T)$ and show that $c(\boldsymbol{\mu}, M) = \Theta(K/\Delta)$ for many common combinatorial structures, including matchings. A minimax lower bound of $\Omega(\sqrt{MKT})$ was given by Audibert et al. (2014) in the same setting. These lower bounds show that the dependency in T, Δ and K obtained in Theorem 1(b),(c) are essentially not improvable, but that the dependency in M might be. However, this observation can be mitigated by noting that finding an algorithm whose regrets attain the available lower bounds for combinatorial semi-bandits is already hard even without the extra challenge of decentralization.

4 The M-ETC-Elim Algorithm

Our algorithm relies on an initialization phase in which the players elect a leader in a distributed manner. Then a communication protocol is set up, in which the leader and the followers have different roles: followers explore some arms and communicate to the leader estimates of the arm means, while the leader maintains a list of "candidate optimal matchings" and communicates to the followers the list of arms that need exploration in order to refine the list, i.e. to eliminate some candidate matchings. The algorithm is called *Multiplayer Explore-Then-Commit with matching Eliminations* (M-ETC-Elim for short). Formally, each player executes Algorithm 1 below.

[2] (Bistritz and Leshem, 2019, Theorem 4) requires T to be larger than $c_3(2^{k_0}-2)$, where k_0 satisfies Equation (16), which features κ and Δ.

Algorithm 1: M-ETC-Elim with parameter c

Input: Time horizon T, number of arms K

$R, M \longleftarrow$ INIT$(K, 1/KT)$

if $R = 1$ **then** LEADERALGORITHM(M) **else** FOLLOWERALGORITHM(R,M)

M-ETC-Elim requires as input the number of arms K (as well as a shared numbering of the arms across the players) and the time horizon T (the total number of arm selections). However, if the players know only an upper bound on T, our results hold with T replaced by that upper bound as well. If no upper bound on T is known, the players can employ a simple doubling trick (Besson and Kaufmann, 2018b): we execute the algorithm assuming $T = 1$, then we execute it assuming $T = 2 \times 1$, and so on, until the actual time horizon is reached. If the expected regret of the algorithm for a known time horizon T is R_T, then the expected regret of the modified algorithm for unknown time horizon T would be $R'_T \leq \sum_{i=0}^{\log_2(T)} R_{2^i} \leq \log_2(T) \times R_T$.

Initialization The initialization procedure, borrowed from Boursier and Perchet (2019), outputs for each player a rank $R \in [M]$ as well as the value of M, which is initially unknown to the players. This initialization phase relies on a "musical chairs" phase after which the players end up on distinct arms, followed by a "sequential hopping" protocol that permits them to know their ordering. For the sake of completeness, the initialization procedure is described in detail in Appendix A, where we also prove the following.

Lemma 2. *Fix $\delta_0 > 0$. With probability at least $1 - \delta_0$, if the M players run the* INIT(K, δ_0) *procedure, which takes $K \ln(K/\delta_0) + 2K - 2 < K \ln(e^2 K/\delta_0)$ many rounds, all players learn M and obtain a distinct ranking from 1 to M.*

Communication Phases Once all players have learned their ranks, player 1 becomes the *leader* and other players become the *followers*. The leader executes additional computations, and communicates with the followers individually, while each follower communicates only with the leader.

The leader and follower algorithms, described below, rely on several *communication phases*, which start at the same time for every player. During communication phases, the default behavior of each player is to pull her *communication arm*. It is crucial that these communication arms are distinct: an optimal way to do so is for each player to use her arm in the best matching found so far. In the first communication phase, such an assignment is unknown and players simply use their ranking as communication arm. Suppose at a certain time the leader wants to send a sequence of b bits $t_1, \dots, t_b$ to the player with ranking i and communication arm k_i. During the next b rounds, for each $j = 1, 2, \dots, b$, if $t_j = 1$, the leader pulls arm k_i; otherwise, she pulls her own communication arm k_1, while all followers stick to their communication arms. Player i can thus reconstruct these b bits after these b rounds, by observing the collisions on arm k_i. The converse communication between follower i and the leader is similar. The rankings are also useful to know *in which order communications should be performed*, as the leader successively communicates messages to the $M-1$ followers, and then the $M-1$ followers successively communicate messages to the leader.

Note that in case of unreliable channels where some of the communicated bits may be lost, there are several options to make this communication protocol more robust, such as sending each bit multiple times or using the Bernoulli signaling protocol of Tibrewal et al. (2019). Robustness has not been the focus of our work.

Leader and Follower Algorithms The leader and the followers perform distinct algorithms, explained next. Consider a bipartite graph with parts of size M and K, where the edge (m, k) has weight μ_k^m and associates player m to arm k. The weights μ_k^m are unknown to the players, but the leader maintains a set of *estimated* weights that are sent to her by the followers, and approximate the real weights. The goal of these algorithms is for the players to jointly explore the matchings in this graph, while gradually focusing on better and better matchings. For this purpose, the leader maintains a set of *candidate edges* $\mathcal{E}$, which is initially $[M]\times[K]$, that can be seen as edges that are potentially contained in optimal matchings, and gradually refines this set by performing eliminations, based on the information obtained from the exploration phases and shared during communication phases.

M-ETC-Elim proceeds in epochs whose length is parameterized by c. In epoch $p = 1, 2, \dots$, the leader weights the edges using the estimated weights. Then for every edge $(m, k) \in \mathcal{E}$, the leader computes the associated matching $\widetilde{\pi}_p^{m,k}$ defined as the estimated maximum matching containing the edge (m, k). This computation can be done in polynomial time using, e.g., the Hungarian algorithm (Munkres, 1957). The leader then computes the utility of the maximum matching and eliminates from $\mathcal{E}$ any edge for which the weight of its associated matching is smaller by at least $4M\varepsilon_p$, where

$$\varepsilon_p := \sqrt{\frac{\ln(2/\delta)}{2^{1+p^c}}}, \text{ with } \delta := \frac{1}{M^2 K T^2}. \tag{1}$$

The leader then forms the set of associated candidate matchings $\mathcal{C} := \{\widetilde{\pi}_p^{m,k}, (m, k) \in \mathcal{E}\}$ and communicates to each follower the list of arms to explore in these matchings. Then exploration begins, in which for each candidate matching every player pulls its assigned arm 2^{p^c} times and records the received reward. Then another communication phase begins, during which each follower sends her observed estimated mean for the arms to the leader. More precisely, for each explored arm, the follower truncates the estimated mean (a number in $[0, 1]$) and sends only the $\frac{p^c+1}{2}$ most significant

bits of this number to the leader. The leader updates the estimated weights and everyone proceeds to the next epoch. If at some point the list of candidate matchings $\mathcal{C}$ becomes a singleton, it means that (with high probability) the actual maximum matching is unique and has been found; so all players jointly pull that matching for the rest of the game (the exploitation phase).

Possible Exploitation Phase Note that in the presence of several optimal matchings, the players will not enter the exploitation phase but will keep exploring several optimal matchings, which still ensures small regret. On the contrary, in the presence of a unique optimal matching, they are guaranteed to eventually enter the exploitation phase.[3] Also, observe that the set $\mathcal{C}$ of candidate optimal matchings does not necessarily contain *all* potentially optimal matchings, but all the edges in those matchings remain in $\mathcal{E}$ and are guaranteed to be explored.

The pseudocode for the leader's algorithm is given below, while the corresponding follower algorithm appears in Appendix A. In the pseudocodes, `(comm.)` refers to a call to the communication protocol.

Procedure LeaderAlgorithm(M) for the M-ETC-Elim algorithm with parameter c

Input: Number of players M

$\mathcal{E} \longleftarrow [M] \times [K]$ `// list of candidate edges`
$\widetilde{\mu}_k^m \longleftarrow 0$ for all $(m,k) \in [M] \times [K]$ `// empirical estimates for utilities`
for $p = 1, 2, \ldots$ **do**
  $\mathcal{C} \longleftarrow \emptyset$ `// list of associated matchings`
  $\pi_1 \longleftarrow \text{argmax} \left\{ \sum_{n=1}^M \widetilde{\mu}^n_{\pi(n)} : \pi \in \mathcal{M} \right\}$ `// using Hungarian algorithm`
  for $(m,k) \in \mathcal{E}$ **do**
    $\pi \longleftarrow \text{argmax} \left\{ \sum_{n=1}^M \widetilde{\mu}^n_{\pi(n)} : \pi(m) = k \right\}$ `// using Hungarian algorithm`
    if $\sum_{n=1}^M \{\widetilde{\mu}^n_{\pi_1(n)} - \widetilde{\mu}^n_{\pi(n)}\} \leq 4M\varepsilon_p$ **then** add π to $\mathcal{C}$
    else remove (m,k) from $\mathcal{E}$
  end
  for *each player* $m = 2, \ldots, M$ **do**
    Send to player m the value of size$(\mathcal{C})$ `// (comm.)`
    for $i = 1, 2, \ldots, \text{size}(\mathcal{C})$ **do**
      Send to player m the arm associated to player m in $\mathcal{C}[i]$ `// (comm.)`
    end
    Send to player m the communication arm of the leader and player m, namely $\pi_1(1)$ and $\pi_1(m)$
  end
  if $size(\mathcal{C}) = 1$ **then** pull for the rest of the game the arm assigned to player 1 in the unique matching in $\mathcal{C}$ `// enter the exploitation phase`
  for $i = 1, 2, \ldots, \text{size}(\mathcal{C})$ **do**
    pull 2^{p^c} times the arm assigned to player 1 in the matching $\mathcal{C}[i]$ `// exploration`
  end
  for $k = 1, 2, \ldots, K$ **do**
    $\widetilde{\mu}_k^1 \longleftarrow$ empirically estimated utility of arm k if it was pulled in this epoch, 0 otherwise
  end
  Receive the values $\widetilde{\mu}_1^m, \widetilde{\mu}_2^m, \ldots, \widetilde{\mu}_K^m$ from each player m `// (comm.)`
end

5 Analysis of M-ETC-Elim

We may assume that $K \leq T$, otherwise all parts of Theorem 1 would be trivial, since $R_T \leq MT$ always. Theorem 3 provides a non-asymptotic upper bound on the regret of M-ETC-Elim when $\Delta \neq \infty$.

Theorem 3. *Let $\pi^{m,k}$ be the best suboptimal matching assigning arm k to player m, namely,*

$$\pi^{m,k} \in \textit{argmax } \{U(\pi) : \pi(m) = k \textit{ and } U(\pi) < U^\star\}\,.$$

For all $c \geq 1$, let $T_0(c) := \exp\left(2^{\frac{c^c}{\ln^c(1+\frac{1}{2c})}}\right)$. For all $T \geq T_0(c)$, if $\Delta \neq \infty$, the regret of M-ETC-Elim with parameter c is upper bounded as[4]

$$\begin{aligned} R_T \leq\ & 2 + MK\ln(e^2K^2T) + 6M^2K\lg(K)(\lg T)^{1/c} \\ & + e^2MK(\lg T)^{1+1/c} + \frac{2M^3K\lg(K)}{\sqrt{2}-1}\sqrt{\ln(2M^2KT^2)} \\ & + \frac{2\sqrt{2}}{3-2\sqrt{2}}M^2K\sqrt{\ln(2M^2KT^2)}\lg(\ln(T)) \\ & + \frac{2\sqrt{2}-1}{\sqrt{2}-1}\sum_{(m,k)\in[M]\times[K]}\left(\frac{32M^2\ln(2M^2KT^2)}{\Delta(\pi^{m,k})}\right)^{1+1/c}. \end{aligned}$$

The first statement of Theorem 1(a) easily follows by lower bounding $\Delta(\pi^{m,k}) \geq \Delta$ for all m, k. The second statement is proved by noting that if $\Delta = \infty$, then the exploration phases incur zero regret, so in that case a variant of Theorem 3 holds without the last term on the right-hand-side. Parts (b) and (c) of Theorem 1 similarly follow respectively from Theorems 8 and 9 in Appendices C and D, with proofs similar to that of Theorem 3 presented below.

The constant $T_0(c)$ in Theorem 3 equals 252 for $c = 1$ but becomes large when c increases. Still, the condition on T is explicit and independent of the problem parameters. In the case of multiple optimal matchings, our contribution is mostly theoretical, as we would need a large enough value of c and a long time $T_0(c)$ for reaching a prescribed $\ln^{1+o(1)}(T)$ regret. However, in the case of a unique optimal matching (common in practice, and sometimes assumed in other papers), for the choice $c = 1$, the logarithmic regret upper bound stated in Theorem 8 is valid for all $T \geq 1$. Even if there are several optimal matchings, the minimax

[3] This different behavior is the main reason for the improved regret upper bound obtained when the optimal matching is unique.

[4] In this paper, $\ln(\cdot)$ and $\lg(\cdot)$ denote the natural logarithm and the logarithm in base 2, respectively.

bound of Theorem 9 gives an $O(\sqrt{T \ln T})$ regret bound that is a best-possible worst-case bound (also known as the minimax rate), up to the $\sqrt{\ln T}$ factor. Hence M-ETC-Elim with $c = 1$ is particularly good, both in theory and in practice. Our experiments also confirm that for $c = 1, 2$ the algorithm performs well (i.e., beats our competitors) even in the presence of multiple optimal matchings.

5.1 Sketch of Proof of Theorem 3

The analysis relies on several lemmas with proofs delayed to Appendix E. Let $\mathcal{C}_p$ denote the set of candidate matchings used in epoch p, and for each matching π let $\widetilde{U}_p(\pi)$ be the utility of π that the leader can estimate based on the information received by the end of epoch p. Let $\widehat{p}_T$ be the total number of epochs before the (possible) start of the exploitation phase. As $2^{\widehat{p}_T^c} \leq T$, we have $\widehat{p}_T \leq \lg(T)$. Recall that a successful initialization means all players identify M and their ranks are distinct. Define the *good event*

$$\mathcal{G}_T := \Big\{ \textsc{Init}(K, 1/KT) \text{ is successful and } \forall p \leq \widehat{p}_T, \forall \pi \in \mathcal{C}_{p+1}, |\widetilde{U}_p(\pi) - U(\pi)| \leq 2M\varepsilon_p \Big\}. \quad (2)$$

During epoch p, for each candidate edge (m, k), player m has pulled arm k at least 2^{p^c} times and the quantization error is smaller than ε_p. Hoeffding's inequality and a union bound over at most $\lg(T)$ epochs (see Appendix E.1) together with Lemma 2 yield that $\mathcal{G}_T$ holds with large probability.

Lemma 4. $\mathbb{P}(\mathcal{G}_T) \geq 1 - \frac{2}{MT}$.

If $\mathcal{G}_T$ does not hold, we may upper bound the regret by MT. Hence it suffices to bound the expected regret conditional on $\mathcal{G}_T$, and the unconditional expected regret is bounded by this value plus 2.

Suppose that $\mathcal{G}_T$ happens. First, the regret incurred during the initialization phase is upper bounded by $MK \ln(e^2 K^2 T)$ by Lemma 2. Moreover, the gap between the best estimated matching of the previous phase and the best matching is at most $2M\varepsilon_{p-1}$ during epoch p. Any single communication round then incurs regret at most $2 + 2M\varepsilon_{p-1}$, the first term being due to the collision between the leader and a follower, the second to the gap between the optimal matching and the matching used for communication. Summing over all communication rounds and epochs leads to Lemma 5 below.

Lemma 5. *The regret due to communication is bounded by*

$$3M^2 K \lg(K) \widehat{p}_T + \frac{2^c \sqrt{2}}{3 - 2\sqrt{2}} M^2 K \sqrt{\ln(2/\delta)} + MK(\widehat{p}_T)^{c+1} + \frac{2M^3 K \lg(K)}{\sqrt{2} - 1} \sqrt{\ln(2/\delta)}.$$

For large horizons, Lemma 6 bounds some terms such as $\widehat{p}_T$ and $(\widehat{p}_T)^c$. When $c = 1$, tighter bounds that are valid for any T are used to prove Theorems 1(b) and 1(c).

Lemma 6. *For any suboptimal matching π, let $P(\pi) := \inf\{p \in \mathbb{N} : 8M\varepsilon_p < \Delta(\pi)\}$. The assumption $T \geq T_0(c)$ implies that for any matching π, $\Delta(\pi) 2^{P(\pi)^c} \leq \left(\frac{32M^2 \ln(2M^2KT^2)}{\Delta(\pi)}\right)^{1+\frac{1}{c}}$. Also, $2^c \leq 2\lg(\ln(T))$, $\widehat{p}_T \leq 2(\lg T)^{1/c}$ and $(\widehat{p}_T)^c \leq e \lg T$.*

Hence for $T \geq T_0(c)$, we can further upper bound the first three terms of the sum in Lemma 5 by

$$6M^2 K \lg(K)(\lg T)^{1/c} + e^2 MK(\lg T)^{1+1/c} + \frac{2\sqrt{2}}{3 - 2\sqrt{2}} M^2 K \sqrt{\ln(2/\delta)} \lg(\ln(T)). \quad (3)$$

It then remains to upper bound the regret incurred during exploration and exploitation phases. On $\mathcal{G}_T$, during the exploitation phase the players are jointly pulling an optimal matching and no regret is incurred. For an edge (m, k), let $\widetilde{\Delta}_p^{m,k} := U^\star - U(\widetilde{\pi}_p^{m,k})$ be the gap of its associated matching at epoch p. During any epoch p, the incurred regret is then $\sum_{\pi \in \mathcal{C}_p} \Delta(\pi) 2^{p^c} = \sum_{(m,k) \in \mathcal{E}} \widetilde{\Delta}_p^{m,k} 2^{p^c}$.

Recall that $\pi^{m,k}$ is the best suboptimal matching assigning arm k to player m. Observe that for any epoch $p > P(\pi^{m,k})$, since $\mathcal{G}_T$ happens, $\pi^{m,k}$ (and any worse matching) is not added to $\mathcal{C}_p$; thus during any epoch $p > P(\pi^{m,k})$, the edge (m, k) is either eliminated from the set of candidate edges, or it is contained in some optimal matching and satisfies $\widetilde{\Delta}_p^{m,k} = 0$. Hence, the total regret incurred during exploration phases is bounded by

$$\sum_{(m,k) \in [M] \times [K]} \sum_{p=1}^{P(\pi^{m,k})} \widetilde{\Delta}_p^{m,k} 2^{p^c}. \quad (4)$$

The difficulty for bounding this sum is that $\widetilde{\Delta}_p^{m,k}$ is random since $\widetilde{\pi}_p^{m,k}$ is random. However, $\widetilde{\Delta}_p^{m,k}$ can be related to $\Delta(\pi^{m,k})$ by $\widetilde{\Delta}_p^{m,k} \leq \frac{\varepsilon_{p-1}}{\varepsilon_{P(\pi^{m,k})}} \Delta(\pi^{m,k})$. A convexity argument then allows us to bound the ratio $\frac{\varepsilon_{p-1}}{\varepsilon_{P(\pi^{m,k})}}$, which yields Lemma 7, proved in Appendix E.4.

Lemma 7. *For any edge (m, k), if $p < P(\pi^{m,k})$ then $\widetilde{\Delta}_p^{m,k} 2^{p^c} \leq \Delta(\pi^{m,k}) \frac{2^{P(\pi^{m,k})^c}}{\sqrt{2}^{P(\pi^{m,k}) - (p+1)}}$.*

By Lemma 7, $\sum_{p=1}^{P(\pi^{m,k})} \widetilde{\Delta}_p^{m,k} 2^{p^c}$ is upper bounded by $\left(\sum_{p=0}^{\infty} 1/\sqrt{2}^p\right) \Delta(\pi^{m,k}) 2^{P(\pi^{m,k})^c} + \widetilde{\Delta}_{P(\pi^{m,k})}^{m,k} 2^{P(\pi^{m,k})^c}$. As $\widetilde{\pi}_{P(\pi^{m,k})}^{m,k}$ is either optimal or its gap is larger than $\Delta(\pi^{m,k})$, Lemma 6 yields

$$\widetilde{\Delta}_{P(\pi^{m,k})}^{m,k} 2^{P(\pi^{m,k})^c} \leq \left(\frac{32M^2 \ln(2M^2KT^2)}{\Delta(\pi^{m,k})}\right)^{1+1/c}$$

in both cases. Therefore, we find that

$$\sum_{p=1}^{P(\pi^{m,k})} \widetilde{\Delta}_p^{m,k} 2^{p^c} \leq \frac{2\sqrt{2} - 1}{\sqrt{2} - 1} \left(\frac{32M^2 \ln(2M^2KT^2)}{\Delta(\pi^{m,k})}\right)^{1+1/c}.$$

Plugging this bound in (4), the bound (3) in Lemma 5 and summing up all terms yields Theorem 3.

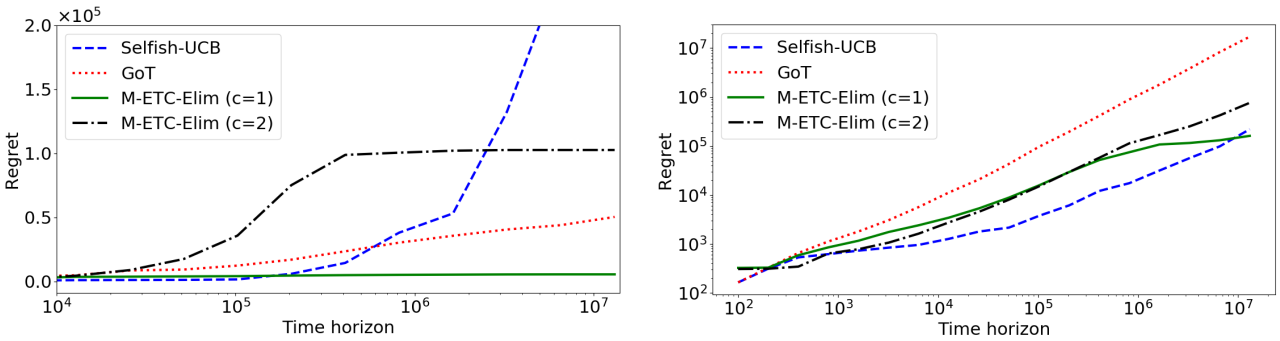

Figure 1: R_T as a function of T with reward matrices U_1 (left) and U_2 (right) and Bernoulli rewards.

5.2 Proof of Theorem 1(b), Unique Optimal Matching

The reader may wonder why can we obtain a better (logarithmic) bound if the maximum matching is unique. The intuition is as follows: in the presence of a unique optimal matching, M-ETC-Elim eventually enters the exploitation phase (which does not happen with multiple optimal matchings), and we can therefore provide a tighter bound on the number of epochs before exploitation phase compared with the one provided by Lemma 6. More precisely, in that case we have $\widehat{p}_T \leq \lg\left(64M^2\Delta^{-2}\ln(2M^2KT^2)\right)$. Moreover, another bound given by Lemma 6 can be tightened when $c = 1$ regardless of whether the optimal matching is unique or not: $\Delta(\pi)2^{P(\pi)} \leq 64M^2\ln(2M^2KT^2)/\Delta(\pi)$.These two inequalities lead to Theorem 1(b), proved in Appendix C.

5.3 Proof of Theorem 1(c), Minimax Regret Bound

Using the definition of the elimination rule, on $\mathcal{G}_T$ we have $\widetilde{\Delta}_p^{m,k} \leq 8M\varepsilon_{p-1}$. Directly summing over these terms for all epochs yields an exploration regret scaling with $\sum_{m,k}\sqrt{t_{m,k}}$, where $t_{m,k}$ roughly corresponds to the number of exploration rounds associated with edge (m,k). This regret is maximized when all $t_{m,k}$ are equal, which leads to the sublinear regret bound of Theorem 1(c). See Appendix D for the rigorous statement and proof.

6 Numerical Experiments

We executed the following algorithms:[5] M-ETC-Elim with $c = 1$ and $c = 2$, GoT (the latest version in Bistritz and Leshem (2019)) with parameters[6] $\delta = 0, \varepsilon = 0.01, c_1 = 500, c_2 = c_3 = 6000$ and Selfish-UCB, a heuristic studied by Besson and Kaufmann (2018a); Boursier and Perchet (2019) in the homogeneous setting which often performs surprisingly well despite the lack of theoretical evidence. In Selfish-UCB, each player runs the UCB1 algorithm of Auer et al. (2002) on the reward sequence $(R^m(t))_{t=1}^{\infty}$.[7] We experiment with Bernoulli rewards and the following reward matrices, whose entry (m,k) gives the value of μ_k^m:

$$U_1 = \begin{pmatrix} 0.1 & 0.05 & 0.9 \\ 0.1 & 0.25 & 0.3 \\ 0.4 & 0.2 & 0.8 \end{pmatrix}, U_2 = \begin{pmatrix} 0.5 & 0.49 & 0.39 & 0.29 & 0.5 \\ 0.5 & 0.49 & 0.39 & 0.29 & 0.19 \\ 0.29 & 0.19 & 0.5 & 0.499 & 0.39 \\ 0.29 & 0.49 & 0.5 & 0.5 & 0.39 \\ 0.49 & 0.49 & 0.49 & 0.49 & 0.5 \end{pmatrix}.$$

Figure 1 reports the algorithms' regrets for various time horizons T, averaged over 100 independent replications. The first instance (matrix U_1, left plot) has a unique optimal matching and we observe that M-ETC-Elim has logarithmic regret (as promised by Theorem 1) and largely outperforms all competitors. The second instance (matrix U_2, right plot) is more challenging, with more arms and players, two optimal matchings and several near-optimal matchings. M-ETC-Elim with $c = 1$ performs the best for large T as well, though Selfish-UCB is also competitive. Yet there is very little theoretical understanding of Selfish-UCB, and it fails badly on the other instance. Appendix B contains additional experiments corroborating our findings, where we also discuss practical aspects of implementing M-ETC-Elim.

7 Conclusion

We have presented a practical algorithm for the heterogeneous multiplayer multi-armed bandit problem, which can be used in the presence of either unique or multiple optimal matchings and get a nearly logarithmic regret in both cases, as well as a sublinear regret in the worst case. M-ETC-Elim crucially relies on the assumption that the collision indicators are observed in each round. In future work, we aim to find algorithms with logarithmic regret in the setting when the players observe their rewards $R^m(t)$ only. So far, such algorithms have been proposed only in the homogeneous setting, see (Lugosi and Mehrabian, 2018; Boursier and Perchet, 2019).

[5] The source codes are included in the supplementary material.

[6] These parameters and the reward matrix U_1 are taken from the simulations section of Bistritz and Leshem (2019).

[7] Note that this sequence is *not* i.i.d. due to some observed zeros that are due to collisions.

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., Varaya, 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.

Audibert, J., Bubeck, S., and Lugosi, G. (2014). Regret in online combinatorial optimization. *Math. Oper. Res.*, 39(1):31–45.

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

Avner, O. and Mannor, S. (2016). Multi-user lax communications: A multi-armed bandit approach. In *IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications*, pages 1–9.

Besson, L. and Kaufmann, E. (2018a). Multi-player bandits revisited. In *Proceedings of Algorithmic Learning Theory (ALT)*.

Besson, L. and Kaufmann, E. (2018b). What doubling tricks can and can't do for multi-armed bandits. *arXiv.org:1803.06971*.

Bistritz, I. and Leshem, A. (2018). Distributed multi-player bandits - a game of thrones approach. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Bistritz, I. and Leshem, A. (2019). Game of thrones: Fully distributed learning for multi-player bandits. *arXiv.org:1810.11162v3*.

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

Bubeck, S. and Cesa-Bianchi, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122.

Chen, W., Wang, Y., and Yuan, Y. (2013). Combinatorial multi-armed bandit: General framework and applications. In *International Conference on Machine Learning*.

Combes, R., Talebi, S., Proutière, A., and Lelarge, M. (2015). Combinatorial bandits revisited. In *Advances in Neural Information Processing Systems (NIPS)*.

Degenne, R. and Perchet, V. (2016). Combinatorial semi-bandit with known covariance. In *Advances in Neural Information Processing Systems (NIPS)*.

Gai, Y., Krishnamachari, B., and Jain, R. (2012). Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations. *IEEE/ACM Trans. Netw.*, 20(5):1466–1478.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58:13–30.

Jouini, W., Ernst, D., Moy, C., and Palicot, J. (2009). Multi-Armed Bandit Based Policies for Cognitive Radio's Decision Making Issues. In *International Conference Signals, Circuits and Systems*. IEEE.

Kalathil, D., Nayyar, N., and Jain, R. (2014). Decentralized learning for multiplayer multiarmed bandits. *IEEE Transactions on Information Theory*, 60(4):2331–2345.

Kveton, B., Wen, Z., Ashkan, A., and Szepesvári, C. (2015). Tight regret bounds for stochastic combinatorial semi-bandits. In *AISTATS*.

Lattimore, T. and Szepesvári, C. (2020). *Bandit Algorithms*. Cambridge University Press. Draft available at https://tor-lattimore.com/downloads/book/book.pdf.

Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010). A Contextual-Bandit Approach to Personalized News Article Recommendation. In *International Conference on World Wide Web*, pages 661–670. ACM.

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

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

Magesh, A. and Veeravalli, V. V. (2019). Multi-user MABs with user dependent rewards for uncoordinated spectrum access. *arXiv.org:1910.09091*.

Munkres, J. (1957). Algorithms for the assignment and transportation problems. *J. Soc. Indust. Appl. Math.*, 5:32–38.

Nayyar, N., Kalathil, D., and Jain, R. (2018). On regret-optimal learning in decentralized multiplayer multiarmed bandits. *IEEE Transactions on Control of Network Systems*, 5(1):597–606.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535.

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

Thompson, W. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25:285–294.

Tibrewal, H., Patchala, S., Hanawal, M. K., and Darak, S. J. (2019). Multiplayer multi-armed bandits for optimal assignment in heterogeneous networks. *arXiv.org:1901.03868v4*.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.

Wang, S. and Chen, W. (2018). Thompson sampling for combinatorial semi-bandits. In *International Conference on Machine Learning*.

A Description of the Initialization Procedure and Followers' Pseudocode

The pseudocode of the INIT(K, δ_0) procedure, first introduced by Boursier and Perchet (2019), is presented in Algorithm 2 for the sake of completeness. We now provide a proof of Lemma 2.

Algorithm 2: INIT, the initialization algorithm

Input: number of arms K, failure probability δ_0
Output: Ranking R, number of players M

```
// first, occupy a distinct arm using the musical chairs algorithm
```
$k \longleftarrow 0$
for $T_0 \coloneqq K\ln(K/\delta_0)$ *rounds* **do** // rounds $1, \dots, T_0$
  if $k = 0$ **then**
    pull a uniformly random arm $i \in [K]$
    if *no collision occurred* **then** $k \longleftarrow i$ // arm k is occupied
  else
    pull arm k
  end
end
```
// next, learn M and identify your ranking
```
$R \longleftarrow 1$
$M \longleftarrow 1$
for $2k-2$ *rounds* **do** // rounds $T_0+1, \dots, T_0+2k-2$
  pull arm k
  if *collision occurred* **then**
    $R \longleftarrow R+1$
    $M \longleftarrow M+1$
  end
end
for $i = 1, 2, \dots, K-k$ **do** // rounds $T_0+2k-1, \dots, T_0+K+k-2$
  pull arm $k+i$
  if *collision occurred* **then**
    $M \longleftarrow M+1$
  end
end
for $K-k$ *rounds* **do** // rounds $T_0+K+k-1, \dots, T_0+2K-2$
  pull arm 1
end

Let $T_0 \coloneqq K\ln(K/\delta_0)$. During the first T_0 rounds, each player tries to occupy a distinct arm using the so-called musical chairs algorithm, first introduced in Rosenski et al. (2016): she repeatedly pulls a random arm until she gets no collision, and then sticks to that arm. We claim that after T_0 rounds, with probability $1-\delta_0$ all players have succeeded in occupying some arm. Indeed, the probability that a given player $\mathcal{A}$, who has not occupied an arm so far, does not succeed in the next round is at most $1-1/K$, since there exists at least one arm that is not pulled in that round, and this arm is chosen by $\mathcal{A}$ with probability $1/K$. Hence, the probability that $\mathcal{A}$ does not succeed in occupying an arm during these T_0 rounds is not more than

$$(1-1/K)^{T_0} < \exp(-T_0/K) = \delta_0/K \leq \delta_0/M,$$

and a union bound over the M players proves the claim.

Once each player has occupied some arm, the next goal is to determine the number of players and their rankings. This part of the procedure is deterministic. The players' rankings will be determined by the indices of the arms they have occupied: a player with a smaller index will have a smaller ranking. To implement this, a player that has occupied arm $k \in [K]$ will pull this arm for $2k-2$ more rounds (the waiting period), and will then sweep through the arms $k+1, k+2, \dots, K$, and can learn the number of players who have occupied arms in this range by counting the number of collisions she gets. Moreover, she can learn the number of players occupying arms $1, \dots, k-1$ by counting the collisions during the waiting period; see

Algorithm 2 for details. The crucial observation to verify the correctness of the algorithm is that two players occupying arms k_1 and k_2 will collide exactly once, and that happens at round $T_0 + k_1 + k_2 - 2$.

Next, we present the pseudocode that the followers execute in M-ETC-Elim. Recall that `(comm.)` refers to a call to the communication protocol.

Procedure FollowerAlgorithm(R,M) for the M-ETC-Elim algorithm with parameter c

Input: Ranking R, number of players M

for $p = 1, 2, \ldots$ **do**
  Receive the value of $\text{size}(\mathcal{C})$ `// (comm.)`
  for $i = 1, 2, \ldots, \text{size}(\mathcal{C})$ **do**
    Receive the arm assigned to this player in $\mathcal{C}[i]$ `// (comm.)`
  end
  Receive the communication arm of the leader and of this player
  if $size(\mathcal{C}) = 1$ `// (enter exploitation phase)`
  then
    pull for the rest of the game the arm assigned to this player in the unique matching in $\mathcal{C}$
  end
  for $i = 1, 2, \ldots, \text{size}(\mathcal{C})$ **do**
    pull 2^{p^c} times the arm assigned to this player in the matching $\mathcal{C}[i]$
  end
  for $k = 1, 2, \ldots, K$ **do**
    $\widehat{\mu}_k^R \longleftarrow$ empirically estimated utility of arm k if arm k has been pulled in this epoch, 0 otherwise
    Truncate $\widehat{\mu}_k^R$ to $\widetilde{\mu}_k^R$ using the $\frac{p^c+1}{2}$ most significant bits
  end
  Send the values $\widetilde{\mu}_1^R, \widetilde{\mu}_2^R, \ldots, \widetilde{\mu}_K^R$ to the leader `// (comm.)`
end

B Practical Considerations and Additional Experiments

B.1 Implementation Enhancements for M-ETC-Elim

In the implementation of M-ETC-Elim, the following enhancements significantly improve the regret in practice (and have been used for the reported numerical experiments), but only by constant factors in theory, hence we have not included them in the analysis for the sake of brevity.

First, to estimate the means, the players are better off taking into account all pulls of the arms, rather than just the last epoch. Note that after the exploration phase of epoch p, each candidate edge has been pulled $N_p := \sum_{i=1}^{p} 2^{i^c}$ times. Thus, with probability at least $1 - 2\lg(T)/(MT)$, each edge has been estimated within additive error $\leq \varepsilon'_p = \sqrt{\ln(M^2TK)/2N_p}$ by Hoeffding's inequality. The players then truncate these estimates using $b := \lceil -\lg(0.1\varepsilon'_p) \rceil$ bits, adding up to $0.1\varepsilon'_p$ additive error due to quantization. They then send these b bits to the leader. Now, the threshold for eliminating a matching would be $2.2M\varepsilon'_p$ rather than $4M \times \sqrt{\ln(2M^2KT^2)/2^{1+p^c}}$ (compare with line 8 of the LeaderAlgorithm presented on page 6).

The second enhancement is to choose the set $\mathcal{C}$ of matchings to explore more carefully. Say that a matching is *good* if its estimated gap is at most $2.2M\varepsilon'_p$, and say an edge is *candidate* (lies in $\mathcal{E}$) if it is part of some good matching. There are at most MK candidate edges, and we need only estimate those in the next epoch. Now, for each candidate edge, we can choose any good matching containing it, and add that to $\mathcal{C}$. This guarantees that $|\mathcal{C}| \leq MK$, which gives the bound in Theorem 1. But to reduce the size of $\mathcal{C}$ in practice, we do the following: initially, all edges are candidate. After each exploration phase, we do the following: we mark all edges as *uncovered*. For each candidate uncovered edge e, we compute the maximum matching π' containing e (using estimated means). If this matching π' has gap larger than $2.2M\varepsilon'_p$, then it is not good hence we remove e from the set of candidate edges. Otherwise, we add π' to $\mathcal{C}$, and moreover, we mark all of its edges as *covered*. We then look at the next uncovered candidate edge, and continue similarly, until all candidate edges are covered. This guarantees that all the candidate edges are explored, while the number of explored matchings could be much smaller than the number of candidate edges, which results in faster exploration and a smaller regret in practice.

To reduce the size of $\mathcal{C}$ even further, we do the following after each exploration phase: first, find the maximum matching

(using estimated means), add it to $\mathcal{C}$, mark all its edges as covered, and only then start looking for uncovered candidate edges as explained above.

B.2 Other Reward Distributions

In our model and analysis, we have assumed $Y_{k,t}^m \in [0,1]$ for simplicity (this is a standard assumption in online learning), but it is immediate to generalize the algorithm and its analysis to reward distributions bounded in any known interval via a linear transformation. Also, we can adapt our algorithm and analysis to subgaussian distributions with mean lying in a known interval. A random variable X is σ-subgaussian if for all $\lambda \in \mathbb{R}$ we have $\mathbb{E}[e^{\lambda(X-\mathbb{E}X)}] \leq e^{\sigma^2\lambda^2/2}$. This includes Gaussian distributions and distributions with bounded support. Suppose for simplicity that the means lie in $[0,1]$. Then the algorithm need only change in two places: first, when the followers are sending the estimated means to the leader, they must send 0 and 1 if the empirically estimated mean is < 0 and > 1, respectively. Second, the definition of ε_p must be changed to $\varepsilon_p := \sqrt{\sigma^2 \ln(2/\delta)/2^{p^c-1}}$. The only change in the analysis is that instead of using Hoeffding's inequality which requires a bounded distribution, one has to use a concentration inequality for sums of subgaussian distributions, see, e.g., (Wainwright, 2019, Proposition 2.5).

We executed the same algorithms as in Section 6 with the same reward matrices but with Gaussian rewards with variance 0.05. The results are somewhat similar to the Bernoulli case and can be found in Figure 2.

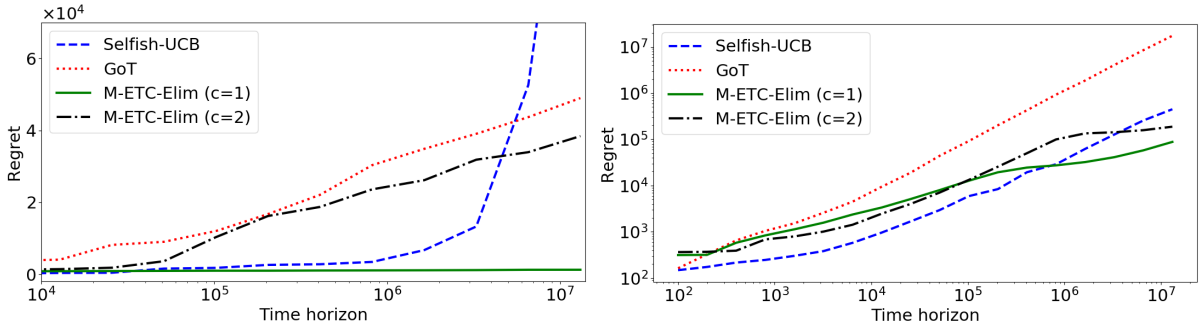

Figure 2: Numerical comparison of M-ETC-Elim, GoT and Selfish-UCB on reward matrices U_1 (left) and U_2 (right) with Gaussian rewards and variance 0.05. The x-axis has logarithmic scale in both plots. The y-axis has logarithmic scale in the right plot.

The reason we performed these Gaussian experiments is to have a more fair comparison against GoT. Indeed, the numerical experiments of Bistritz and Leshem (2019) rely on the same reward matrix U_1 and Gaussian rewards.

C Regret Analysis in the Presence of a Unique Maximum Matching

In Theorem 8 below we provide a refined analysis of M-ETC-Elim with parameter $c = 1$ if the maximum matching is unique, justifying the $O(\frac{KM^3}{\Delta}\ln(T))$ regret upper bound stated in Theorem 1(b). Its proof, given below, follows essentially the same line as the finite-time analysis given in Section 5, except for the last part. Recall that $\ln(\cdot)$ denotes the natural logarithm and $\lg(\cdot)$ denotes logarithm in base 2.

Theorem 8. *If the maximum matching is unique, for any $T > 0$ the regret of the M-ETC-Elim algorithm with parameter $c = 1$ is upper bounded by*

$$2 + MK\ln(e^2K^2T) + 3M^2K\lg(K)\lg\left(\frac{64M^2\ln(2M^2KT^2)}{\Delta^2}\right) + MK\lg^2\left(\frac{64M^2\ln(2M^2KT^2)}{\Delta^2}\right)$$
$$+ \frac{4\sqrt{2}-2}{3-2\sqrt{2}}M^3K\lg(K)\sqrt{\ln(2M^2KT^2)} + \frac{2\sqrt{2}-1}{\sqrt{2}-1}\sum_{(m,k)\in[M]\times[K]}\frac{64M^2\ln(2M^2KT^2)}{\Delta(\pi^{m,k})}.$$

Proof. The good event and the regret incurred during the initialization phase are the same as in the finite-time analysis given in Section 5. Recall the definition of P, which is $P(\pi) = \inf\{p \in \mathbb{N} : 8M\varepsilon_p < \Delta(\pi)\}$. When there is a unique optimal matching, if the good event happens, the M-ETC-Elim algorithm will eventually enter the exploitation phase, so $\widehat{p}_T$ can be

much smaller than the crude upper bound given by Lemma 6. Specifically, introducing π' as the second maximum matching so that $\Delta(\pi') = \Delta$, we have, on the event $\mathcal{G}_T$,

$$\widehat{p}_T \leq P(\pi') \leq \lg\left(\frac{64M^2\ln(2M^2KT^2)}{\Delta^2}\right).$$

Plugging this bound in Lemma 5 yields that the regret incurred during communications is bounded by

$$3M^2K\lg(K)\lg\left(\frac{64M^2\ln(2M^2KT^2)}{\Delta^2}\right) + MK\lg^2\left(\frac{64M^2\ln(2M^2KT^2)}{\Delta^2}\right) + \frac{2M^3K\lg K}{\sqrt{2}-1}\sqrt{\ln(2/\delta)} + \frac{2\sqrt{2}}{3-2\sqrt{2}}M^2K\sqrt{\ln(2/\delta)}.$$

Also, for $c = 1$ and any matching π, the definition of ε_p in (1) gives

$$P(\pi) \leq 1 + \lg\left(\frac{32M^2\ln(2M^2KT^2)}{\Delta(\pi)^2}\right).$$

In particular, $\Delta(\pi)2^{P(\pi)} \leq \frac{64M^2\ln(2M^2KT^2)}{\Delta(\pi)}$. Using the same argument as in Section 5, the regret incurred during the exploration phases is bounded by

$$\frac{2\sqrt{2}-1}{\sqrt{2}-1}\sum_{(m,k)\in[M]\times[K]}\frac{64M^2\ln(2M^2KT^2)}{\Delta(\pi^{m,k})}.$$

Summing up the regret bounds for all phases proves Theorem 8.

□

D Minimax Regret Analysis

In Theorem 9 below we provide a minimax regret bound for M-ETC-Elim with parameter $c = 1$, justifying the $O\left(M^{\frac{3}{2}}\sqrt{KT\ln(T)}\right)$ regret upper bound stated in Theorem 1(c).

Theorem 9. *For all T, the regret of the M-ETC-Elim algorithm with parameter $c = 1$ is upper bounded by*

$$2 + MK\ln(e^2K^2T) + 3M^2K\lg(K)\lg(T) + MK\lg^2(T) + \frac{4\sqrt{2}-2}{3-2\sqrt{2}}M^3K\lg(K)\sqrt{\ln(2M^2KT^2)} + \frac{8}{\sqrt{2}-1}K^{\frac{1}{2}}M^{\frac{3}{2}}\sqrt{T\ln(2M^2KT^2)}.$$

Note that the above regret bound is independent of the suboptimality gaps.

Proof. The good event and the regret incurred during the initialization phase are the same as in the finite-time analysis given in Section 5. Furthermore, using Lemma 5 stated therein and since $\widehat{p}_T \leq \lg(T)$, the regret incurred during the communication phases is bounded by

$$3M^2K\lg(K)\lg(T) + MK\lg^2(T) + \frac{4\sqrt{2}-2}{3-2\sqrt{2}}M^3K\lg(K)\sqrt{\ln(2M^2KT^2)}.$$

We next bound the exploration regret. Fix any edge (m,k), and let $\widetilde{P}^{m,k}$ be the last epoch in which this edge is explored. If this edge belongs to an optimal matching, i.e., if $\pi^{m,k}$ is optimal, we instead define $\widetilde{P}^{m,k}$ as the last epoch in which the pulled matching $\widetilde{\pi}_p^{m,k}$ associated with (m,k) is suboptimal. In either case, the contribution of the edge (m,k) to the exploration regret can be bounded by $\sum_{p=1}^{\widetilde{P}^{m,k}}\widetilde{\Delta}_p^{m,k}2^p$.

Fix an epoch $p \le \widetilde{P}^{m,k}$. Recall that $\mathcal{C}_p$ contains at least one actual maximum matching, which we denote by $\pi^\star$. Also, let $\widetilde{\pi}_p^\star$ denote the maximum empirical matching right before the start of epoch p. Since (m,k) is candidate in epoch p, we have

$$\begin{aligned}\widetilde{\Delta}_p^{m,k} &= U^\star - U_{p-1}(\pi^\star) + U_{p-1}(\pi^\star) - U_{p-1}(\widetilde{\pi}_p^{m,k}) + U_{p-1}(\widetilde{\pi}_p^{m,k}) - U(\widetilde{\pi}^{m,k}) \\ &\le (U_\star - U_{p-1}(\pi^\star)) + (U_{p-1}(\widetilde{\pi}_p^\star) - U_{p-1}(\widetilde{\pi}_p^{m,k}) + (U_{p-1}(\widetilde{\pi}_p^{m,k}) - U(\widetilde{\pi}_p^{m,k})) \\ &\le 2M\varepsilon_{p-1} + 4M\varepsilon_p + 2M\varepsilon_{p-1} \\ &\le 8M\varepsilon_{p-1} = 8M\sqrt{\frac{\ln(2/\delta)}{2^p}},\end{aligned}$$

so, the contribution of the edge (m,k) to the exploration regret can further be bounded by

$$\sum_{p=1}^{\widetilde{P}^{m,k}} \widetilde{\Delta}_p^{m,k} 2^p \le 8M\sqrt{\ln(2/\delta)}\left(\sum_{p=1}^{\widetilde{P}^{m,k}} \sqrt{2}^p\right) < \frac{8\sqrt{2}M\sqrt{\ln(2/\delta)}}{\sqrt{2}-1}\sqrt{2}^{\widetilde{P}^{m,k}}.$$

To bound the total exploration regret, we need to sum this over all edges (m,k).

Note that during each epoch $p = 1, 2, \dots, \widetilde{P}_{m,k}$, there are exactly 2^p exploration rounds associated with the edge (m,k). Since the total number of rounds is T, we find that

$$\sum_{(m,k)\in[M]\times[K]} \sum_{p=1}^{\widetilde{P}_{m,k}} 2^p \le T,$$

and in particular,

$$\sum_{(m,k)\in[M]\times[K]} 2^{\widetilde{P}_{m,k}} \le T,$$

hence by the Cauchy-Schwarz inequality,

$$\sum_{(m,k)\in[M]\times[K]} \sqrt{2}^{\widetilde{P}_{m,k}} = \sum_{(m,k)\in[M]\times[K]} \sqrt{2^{\widetilde{P}_{m,k}}} \le \sqrt{MKT},$$

so the total exploration regret can be bounded by

$$\frac{8\sqrt{2}M\sqrt{\ln(2/\delta)}}{\sqrt{2}-1} \sum_{(m,k)\in[M]\times[K]} \sqrt{2}^{\widetilde{P}^{m,k}} \le \frac{8\sqrt{2}M\sqrt{\ln(2/\delta)}}{\sqrt{2}-1}\sqrt{MKT},$$

completing the proof of Theorem 9.

□

E Proofs of Auxiliary Lemmas for Theorems 3 and 8

E.1 Proof of Lemma 4

We recall Hoeffding's inequality.

Proposition 10 (Hoeffding's inequality (Hoeffding, 1963, Theorem 2)). *Let $X_1, \dots, X_n$ be independent random variables taking values in $[0,1]$. Then for any $t \ge 0$ we have*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum X_i - \mathbb{E}\left[\frac{1}{n}\sum X_i\right]\right| > t\right) < 2\exp(-2nt^2).$$

Recall the definition of the good event

$$\mathcal{G}_T = \left\{\textsc{Init}(K, 1/KT) \text{ is successful and } \forall p \le \widehat{p}_T, \forall \pi \in \mathcal{C}_{p+1}, |\widetilde{U}_p(\pi) - U(\pi)| \le 2M\varepsilon_p\right\}.$$

and recall that $\varepsilon_p \coloneqq \sqrt{\ln(2/\delta)/2^{p^c+1}}$ and $\delta = 1/M2KT^2$. Let $\mathcal{H}$ be the event that $\text{INIT}(K, 1/KT)$ is successful for all players. Then,

$$\begin{aligned}\mathbb{P}\left(\mathcal{G}_T^c\right) &\leq \mathbb{P}\left(\mathcal{H}^c\right) + \mathbb{P}\left(\mathcal{H} \text{ happens and } \exists p \leq \widehat{p}_T, \exists \pi \in \mathcal{M} \text{ with candidate edges such that } |\widetilde{U}_p(\pi) - U(\pi)| > 2M\varepsilon_p\right) \\ &\leq \frac{1}{KT} + \mathbb{P}\left(\mathcal{H} \text{ happens and } \exists p \leq \lg(T), \exists \pi \in \mathcal{M} \text{ with candidate edges such that} |\widetilde{U}_p(\pi) - U(\pi)| > 2M\varepsilon_p\right),\end{aligned}$$

where we have used that $\widehat{p}_T \leq \lg(T)$ deterministically.

Fix an epoch p and a candidate edge (m, k). We denote by $\widehat{\mu}_k^m(p)$ the estimated mean of arm k for player m at the end of epoch p and by $\widetilde{\mu}_k^m(p)$ the truncated estimated mean sent to the leader by this player at the end of epoch p.

By Hoeffding's inequality and since this estimated mean is based on at least 2^{p^c} pulls, we have

$$\mathbb{P}\left(|\widehat{\mu}_k^m(p) - \mu_k^m| > \varepsilon_p\right) < \delta.$$

The value $\widetilde{\mu}_k^m(p) \in [0, 1]$ which is sent to the leader uses the $(p^c + 1)/2$ most significant bits. The truncation error is thus at most $2^{-(p^c+1)/2} < \varepsilon_p$, hence we have

$$\mathbb{P}\left(|\widetilde{\mu}_k^m(p) - \mu_k^m| > 2\varepsilon_p\right) < \delta.$$

Given the event $\mathcal{H}$ that the initialization is successful, the quantity $\widetilde{U}_p(\pi)$ is a sum of M values $\widetilde{\mu}_k^m(p)$ for M different edges $(m, k) \in [M] \times [K]$. Hence, we have

$$\begin{aligned}&\mathbb{P}\left(\mathcal{H} \text{ happens and } \exists \pi \in \mathcal{M} \text{ with candidate edges such that } |\widetilde{U}_p(\pi) - U(\pi)| > 2M\varepsilon_p|\right) \\ &\leq \mathbb{P}\left(\exists \text{ candidate edge } (m, k) \text{ such that } |\widetilde{\mu}_k^m(p) - \mu_k^m| > 2\varepsilon_p\right) \leq KM\delta.\end{aligned}$$

Finally, a union bound on p yields

$$\mathbb{P}\left(\mathcal{G}_T^c\right) \leq \frac{1}{KT} + \lg(T)KM\delta \leq \frac{1}{MT} + \frac{1}{MT},$$

completing the proof of Lemma 4

E.2 Proof of Lemma 5

For each epoch p, the leader first communicates to each player the list of candidate matchings. There can be up to MK candidate matchings, and for each of them the leader communicates to the player the arm she has to pull (there is no need to communicate to her the whole matching) which requires $\lg K$ bits, and there are a total of M players, so this takes at most $M^2K\lg(K)$ many rounds.[8]

At the end of the epoch, each player sends the leader the empirical estimates for the arms she has pulled, which requires at most $MK(1 + p^c)/2$ many rounds. As players use the best estimated matching as communication arms for the communication phases, a single communication round incurs regret at most $2 + 2M\varepsilon_{p-1}$, since the gap between the best estimated matching of the previous phase and the best matching is at most $2M\varepsilon_{p-1}$ conditionally to $\mathcal{G}_T$ (we define $\varepsilon_0 \coloneqq \sqrt{\frac{\ln(2/\delta)}{2}} \geq \frac{1}{2}$). The first term is for the two players colliding, while the term $2M\varepsilon_{p-1}$ is due to the other players who are pulling the best estimated matching instead of the real best one. With $\widehat{p}_T$ denoting the number of epochs before the (possible) start of the exploitation, the total regret due to communication phases can be bounded by

$$\begin{aligned}R_c &\leq \sum_{p=1}^{\widehat{p}_T} \left(2M^2K\lg(K) + MK(1 + p^c)\right)(1 + M\varepsilon_{p-1}) \\ &\leq 3M^2K\lg(K)\widehat{p}_T + MK(\widehat{p}_T)^{c+1} + M^2K\sum_{p=1}^{\widehat{p}_T}\left(2M\lg(K) + (1 + p^c)\right)\varepsilon_{p-1}.\end{aligned}$$

[8] Strictly speaking, the leader also sends her communication arm and the size of the list she is sending, but there are at most $MK - M + 1$ candidate matchings, as the best one is repeated M times. So, this communication still takes at most $M^2K\lg K$ many rounds.

We now bound the sum as:

$$\begin{aligned}\sum_{p=1}^{\widehat{p}_T}\left(2M\lg(K)+(1+p^c)\right)\varepsilon_{p-1} &= 2M\lg(K)\sqrt{\ln(2/\delta)}\sum_{p=0}^{\widehat{p}_T-1}\frac{1}{\sqrt{2}^{1+p^c}}+\sqrt{\ln(2/\delta)}\sum_{p=0}^{\widehat{p}_T-1}\frac{1+(p+1)^c}{\sqrt{2}^{1+p^c}}\\ &\le 2M\lg(K)\sqrt{\ln(2/\delta)}\sum_{n=1}^{\infty}\frac{1}{\sqrt{2^n}}+\sqrt{\ln(2/\delta)}\sum_{n=1}^{\infty}\frac{n2^c}{\sqrt{2^n}}\\ &\le 2M\lg(K)\sqrt{\ln(2/\delta)}\frac{1}{\sqrt{2}-1}+\sqrt{\ln(2/\delta)}\frac{2^c\sqrt{2}}{(\sqrt{2}-1)^2},\end{aligned}$$

completing the proof of Lemma 5.

E.3 Proof of Lemma 6

The assumption $T \ge \exp(2^{\frac{c^c}{\ln c(1+\frac{1}{2c})}})$ gives $\lg(\ln T)^{1/c} \ge \frac{c}{\ln(1+1/2c)}$. In particular, $(\lg T)^{1/c} \ge c$. We will also use the inequality

$$(x+1)^c \le e^{c/x}x^c, \tag{5}$$

which holds for all positive x, since $(x+1)^c/x^c = (1+1/x)^c \le \exp(1/x)^c = \exp(c/x)$.

Using a crude upper bound on the number of epochs that can fit within T rounds, we get $\widehat{p}_T \le 1+(\lg T)^{1/c}$. As $(\lg T)^{1/c} \ge c \ge 1$ we have $\widehat{p}_T \le 2(\lg T)^{1/c}$. Also (5) gives $(\widehat{p}_T)^c \le e\lg T$.

Also, $2\lg(\ln(T)) \ge 2c^c \ge 2^c$. It remains to show the first inequality of Lemma 6.

Straightforward calculations using the definition of ε_p in (1) give

$$P(\pi) \le 1+L(\pi)^{1/c}, \text{ where } L(\pi) \coloneqq \lg\left(\frac{32M^2\ln(2M^2KT^2)}{\Delta(\pi)^2}\right).$$

We claim that we have

$$P(\pi)^c \le \left(1+\frac{1}{2c}\right)L(\pi). \tag{6}$$

Indeed, since $\Delta(\pi) \le M$, we have $L(\pi)^{1/c} > (\lg\ln T)^{1/c} \ge \frac{c}{\ln(1+1/2c)}$ and so (5) with $x = L(\pi)^{1/c}$ gives (6). Hence,

$$\Delta(\pi)2^{P(\pi)^c} \le \Delta(\pi)\left(\frac{32M^2\ln(2M^2KT^2)}{\Delta(\pi)^2}\right)^{1+1/2c} \le \left(\frac{32M^2\ln(2M^2KT^2)}{\Delta(\pi)}\right)^{1+1/c}, \tag{7}$$

completing the proof of Lemma 6.

E.4 Proof of Lemma 7

For brevity we define, for this proof, $\Delta \coloneqq \Delta(\pi^{m,k})$, $P \coloneqq P(\pi^{m,k})$ and $\Delta_p \coloneqq \widetilde{\Delta}_p^{m,k}$. First, $\Delta > 8M\varepsilon_P$ by definition of P. Also, $\Delta_p \le 8M\varepsilon_{p-1}$ for any $p \le P-1$, otherwise the edge (m,k) would have been eliminated before epoch p. It then holds

$$\Delta_p \le \frac{\varepsilon_{p-1}}{\varepsilon_P}\Delta = \sqrt{2}^{P^c-(p-1)^c}\Delta. \tag{8}$$

It comes from the convexity of $x \mapsto x^c$ that $(p+1)^c+(p-1)^c-2p^c \ge 0$, and thus

$$P^c+(p-1)^c-2p^c \ge P^c-(p+1)^c \ge P-(p+1).$$

It then follows

$$p^c+\frac{P^c-(p-1)^c}{2} \le P^c+\frac{p+1-P}{2}.$$

Plugging this in (8) gives

$$2^{p^c}\Delta_p \le \frac{2^{P^c}}{\sqrt{2}^{P-(p+1)}}\Delta,$$

completing the proof of Lemma 7.