Best Arm Identification in Multi-Agent Multi-Armed Bandits

Filippo Vannella [1 2] **Alexandre Proutiere** [1] **Jaeseong Jeong** [2]

Abstract

We investigate the problem of best arm identification in Multi-Agent Multi-Armed Bandits (MAMABs) where the rewards are defined through a factor graph. The objective is to find an optimal global action with a prescribed level of confidence and minimal sample complexity. We derive a tight instance-specific lower bound of the sample complexity and characterize the corresponding optimal sampling strategy. Unfortunately, this bound is obtained by solving a combinatorial optimization problem with a number of variables and constraints exponentially growing with the number of agents. We leverage Mean Field (MF) techniques to obtain, in a computationally efficient manner, an approximation of the lower bound. The approximation scales at most as ρK^d (where ρ, K, and d denote the number of factors in the graph, the number of possible actions per agent, and the maximal degree of the factor graph). We devise MF-TaS (Mean-Field-Track-and-Stop), an algorithm whose sample complexity provably matches our approximated lower bound. We illustrate the performance of MF-TaS numerically using both synthetic and real-world experiments (e.g., to solve the antenna tilt optimization problem in radio communication networks).

1. Introduction

Best arm identification with fixed confidence in stochastic bandits (Lai & Robbins, 1985) refers to the problem of finding the arm with the highest expected reward with a prescribed level of certainty, while minimizing the number of samples or arm draws, i.e., the sample complexity. When the arm rewards are unrelated, the sample complexity needs to typically scale linearly with the number of arms, and the task quickly becomes intractable as this number grows large. To reduce the sample complexity, the learner may leverage any underlying structure tying up the expected rewards of the various arms together. Most efforts in this direction have focused on linear structures (Degenne et al., 2020; Fiez et al., 2019; Jedra & Proutiere, 2020; Karnin, 2016; Soare et al., 2014; Tao et al., 2018), see (Wang et al., 2021) and references therein. Exploiting the underlying structure is critical when the number of arms becomes extremely large, as in combinatorial bandit problems (Chen et al., 2014; Jourdan et al., 2021). To solve these problems, the learner faces a statistical efficiency issue (she has to control the sample complexity), but also needs to account for inherent computational limits (she will typically have to solve combinatorial optimization problems over the set of possible arms).

We investigate the best arm identification problem in the stochastic Multi-Agent Multi-Armed Bandits (MAMABs) – referred to as M-BAI for short. M-BAI is a particular instance of the best arm identification problem in combinatorial bandits, where (i) a global arm or action is defined by the actions individually selected by the various agents, and (ii) the reward function is defined through a factor graph. This reward structure arises naturally in networks where agents interact with their neighbors in the graph, and need to coordinate toward a common goal. The paper was actually motivated by the problem of learning to coordinate base-stations in radio communication networks (see §7.2).

Contributions. For the M-BAI problem, we present a statistically and computationally efficient algorithm. The algorithm has provable performance guarantees (in the form of sample complexity upper bounds), and can be applied in large-scale MAMABs. More precisely, our contributions are as follows.

1) Sample complexity lower bound. We derive an instance-specific lower bound on the sample complexity satisfied by any algorithm. The bound is defined through an optimization problem, whose solution provides an optimal sampling strategy. Unfortunately, because of the factored reward structure, this optimization problem contains an exponential number of variables and constraints, and is hence difficult to exploit in practice.

2) Mean-Field-based approximation of the lower bound optimization problem. We propose a tight approximation of the lower bound optimization problem obtained by com-

[1]KTH Royal Institute of Technology, Stockholm, Sweden [2]Ericsson, Stockholm, Sweden. Correspondence to: Filippo Vannella <vannella@kth.se>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

bining a Mean Field (MF) technique, to reduce the number of variables, and Factored Constraint Reduction (FCR), a procedure inspired by methods in probabilistic graphical models, to reduce the number of constraints. We show that the resulting optimization problem is equivalent to a convex program that can be solved efficiently. The resulting approximated lower bound scales at most as ρK^d, where ρ is the number of factors (or groups), K is the number of actions per agent and d is the maximal degree of the factor graph. This scaling illustrates the gains one may achieve by exploiting the factor graph structure (without leveraging the structure, the sample complexity would well scale as K^N, where N is the number of agents).

3) The MF-TaS algorithm. We devise MF-TaS (Mean-Field-Track-and-Stop), an algorithm whose sample complexity provably matches our approximated lower bound. The algorithm, based on the popular Track and Stop (TaS) algorithm, selects actions suggested by the solution of the approximated lower bound optimization problem and decides to stop gathering data according to the result of a classical Generalized Likelihood Ratio Test (GLRT).

4) Synthetic and real-world experiments. First, we test the performance of MF-TaS numerically using synthetic experiments. We then apply the algorithm to solve the problem of learning to coordinate the antenna tilts at various base stations in a radio communication network. In both sets of experiments, we show that MF-TaS can solve very large problems with millions of global actions in a sample and computationally efficient manner.

2. Related Work

Most existing works on multi-agent bandit models focus on the so-called *multi-player bandits* (Besson & Kaufmann, 2018; Rosenski et al., 2016; Shi et al., 2021; Wang et al., 2020). There, agents have access to the same set of actions and interact through collisions (if two agents select the same action, no reward is collected by both agents). Our setting is different and assumes that the global reward is a sum over groups of local rewards which depend on the actions selected by related agents.

A few papers investigate MAMABs with the same factored reward structure as ours (Bargiacchi et al., 2018; 2022; Stranders et al., 2012; Verstraeten et al., 2020). While (Bargiacchi et al., 2018; Stranders et al., 2012; Verstraeten et al., 2020) focus on regret minimization, our paper studies best arm identification. The closest related work is (Bargiacchi et al., 2022). There, the goal is to identify a global arm or action that is ε-optimal and with error probability bounded by δ. Targeting ε-optimal arms greatly simplifies the problem and the analysis, and removes the need for an adaptive stopping rule (the number of samples is fixed a-priori). In turn, the algorithm proposed in (Bargiacchi et al., 2022) does not adapt to the hardness of the problem. This hardness is captured through $\Delta_{\min}$, the gap between the best and the second-best arm. Our algorithm is learning this gap and adapting its sampling strategy accordingly.

Another closely related line of work is best arm identification in combinatorial bandits with semi-bandit feedback (Chen et al., 2014; Du et al., 2021; Jourdan et al., 2021; Wagenmaker et al., 2020), which encompasses the MAMABs setting as a particular case (see App. G). These works do not explicitly consider multi-agent problems and focus on devising computationally and statistically efficient algorithms. The closest work here is (Jourdan et al., 2021), which leverages a game interpretation of the lower bound optimization problem to devise asymptotically optimal meta-algorithms using online optimization methods. Their method requires the existence of a "best-response oracle", which is computationally inefficient for factored rewards models with combinatorial action spaces. In contrast, our algorithm leverages an MF approximation to reduce the number of variables and constraints in the lower bound optimization problem and leads to a more efficient algorithm.

A few related works investigate *regret minimization* in combinatorial semi-bandit feedback settings (Cuvelier et al., 2021a;b; Wagenmaker et al., 2020). The authors of (Cuvelier et al., 2021b) derive a lower bound on the regret that has a similar structure and presents the same challenges, as the one we derive for best arm identification: the bound is obtained by solving an optimization problem with an exponentially large number of variables and constraints (see App. F for details). By smartly rewriting the optimization problem, the authors of (Cuvelier et al., 2021b) manage to devise an asymptotically optimal algorithm. Unfortunately, the algorithm relies on Assumption 6 in (Cuvelier et al., 2021b) which, in general, does not hold in our setting (Wainwright & Jordan, 2008). It is hence impossible to follow a similar approach in the MAMAB setting.

3. Problem Setting

Model and reward structure. We consider the generic MAMAB model with factored structure introduced in (Bargiacchi et al., 2018). The model is defined by the tuple $\langle \mathcal{S}, \mathcal{A}, r \rangle$, where:

1. $\mathcal{S} = [N] \triangleq \{1, \dots, N\}$ is a set of N agents,
2. $\mathcal{A} = \times_{i \in [N]} \mathcal{A}_i$ is a set of global actions, which is the Cartesian product over i of the set $\mathcal{A}_i$ of actions available to the agent i. We assume w.l.o.g. that $|\mathcal{A}_i| = K$, for all $i \in [N]$, and define $A \triangleq |\mathcal{A}| = K^N$,
3. r is the reward function mapping the global action to the collected reward.

We now describe the reward collected when a global action $a \in \mathcal{A}$ is selected. We assume that there are ρ *groups* of possibly overlapping subsets of agents $(\mathcal{S}_e)_{e\in[\rho]}$, with $\mathcal{S}_e \subseteq \mathcal{S}$, and $|\mathcal{S}_e| = N_e$. Each group generates rewards. The *local* reward generated by group e depends on group actions $a_e \triangleq (a_i)_{i\in\mathcal{S}_e} \in \mathcal{A}_e \triangleq \times_{i\in\mathcal{S}_e}\mathcal{A}_i$ only. More precisely, each time a_e is selected, the collected local rewards are i.i.d. copies of a random variable $r_e(a_e) \sim \mathcal{N}(\theta_e(a_e), 1)$. Rewards collected in various groups are independent. The global reward for action a is then $r(a) = \sum_{e\in[\rho]} r_e(a_e)$, a random variable with expectation $\theta(a) = \sum_{e\in[\rho]} \theta_e(a_e)$. The number of possible group actions in group e is $A_e \triangleq |\mathcal{A}_e| = K^{N_e}$, and we define $\tilde{A} \triangleq \sum_{e\in[\rho]} A_e$.

Factor graph representation. The reward structure can be represented as a factor graph (Wainwright & Jordan, 2008). Factor graphs are bipartite graphs with two types of node: N *action nodes*, one for each agent (represented by circles), and ρ *factor nodes*, one for each group (represented by squares). An edge between a factor r_e and an agent i exists if the action a_i selected by the agent i is an input of r_e, i.e., $i \in \mathcal{S}_e$. Fig. 1 shows an example of a factor graph with $N = 4$ agents and $\rho = 4$ factors in which the reward is factored additively as $r(a) = r_1(a_1, a_2) + r_2(a_2, a_4) + r_3(a_1, a_3) + r_4(a_3, a_4)$. In this example, when each agent has K actions available, we have $A = K^4$ and $\tilde{A} = 4K^2$.

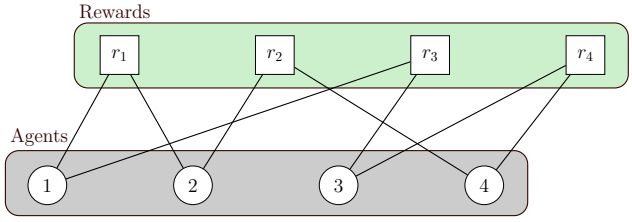


Figure 1. Example of a factor graph.

Sequential decision process. In M-BAI, the decision maker sequentially selects global actions based on the history of previous observations and receives a set of samples of the local rewards associated to the various groups. Specifically, in each round $t \geq 1$, the decision maker selects a global action $a_t = (a_{t,1}, \ldots, a_{t,N})$ and observes the local rewards $r_t = (r_{t,1}, \ldots, r_{t,\rho})$ from each group. The global action a_{t+1} is selected based on the history of observations $\mathcal{H}_t = (a_s, r_s)_{s\in[t]}$. This type of interaction is known as *semi-bandit* feedback (see App. G for details).

Best Arm Identification. In this work, we study the problem of M-BAI in the fixed confidence setting. The goal is to devise an algorithm that returns, using as few rounds as possible, the best global action with a fixed confidence level. This action is defined as $a^\star_\theta \in \arg\max_{a\in\mathcal{A}} \theta(a)$. Throughout the paper, we assume that $a^\star_\theta$ is unique. Furthermore, when there is no ambiguity, we use $a^\star$ and $a^\star_\theta$ interchangeably. In this setting, an algorithm is defined through a *sampling rule*, a *stopping rule*, and a *recommendation rule*, described as follows:

(i) *Sampling rule:* it specifies the global action selected in each round. The sampling rule is defined as a sequence of actions $(a_t)_{t\geq1}$, where $a_t \in \mathcal{A}$ may depend on past observations. Formally, a_t is $\mathcal{F}_{t-1}$-measurable, where $\mathcal{F}_t$ is the σ-algebra generated by $\mathcal{H}_t$, the history up to time t.

(ii) *Stopping rule:* it controls the end of the data acquisition phase and is defined as a stopping time τ with respect to the filtration $(\mathcal{F}_t)_{t\geq1}$.

(iii) *Recommendation rule:* in round τ, after the data acquisition phase ends, it returns an estimated best global action $\hat{a}_\tau \in \mathcal{A}$.

We denote by $\mathbb{P}_\theta$ the probability measure of the observations generated under the parameter θ, and by $\mathbb{E}_\theta$ the respective expectation. With these definitions, the objective is to devise a δ-PAC algorithm, as defined below, with minimal expected sample complexity $\mathbb{E}_\theta[\tau]$.

Definition 3.1. Let $\delta \in (0, 1)$. An algorithm is δ-PAC if $\forall\theta$, $\mathbb{P}_\theta(a^\star_\theta \neq \hat{a}_\tau) \leq \delta$ and $\mathbb{P}_\theta(\tau < \infty) = 1$.

4. Sample Complexity Lower Bound

We present a lower bound on the sample complexity satisfied by any δ-PAC algorithm. The lower bound is obtained using classical change-of-measure arguments introduced in (Kaufmann et al., 2016; Lai & Robbins, 1985). To state the lower bound, we introduce the following notations.

Notations. Define the *marginal polytope* as:

$$\tilde{\Lambda} = \left\{ \tilde{w} \in \mathbb{R}^{\tilde{A}} : \exists w \in \Lambda, \forall e \in [\rho], a_e \in \mathcal{A}_e, \tilde{w}_{e,a_e} = \sum_{b\in\mathcal{A}:b_e=a_e} w_b, \right\},$$

where $\Lambda = \{w \in \mathbb{R}^A_+ : \sum_{a\in\mathcal{A}} w_a = 1\}$ is the $(A-1)$-dimensional simplex. The set $\tilde{\Lambda}$ contains *group allocations* $\tilde{w} = (\tilde{w}_e)_{e\in[\rho]}$, where $\tilde{w}_e = (\tilde{w}_{e,a_e})_{a_e\in\mathcal{A}_e}$. In other words, $\tilde{\Lambda}$ contains group probability measures $\tilde{w}$ which satisfy a consistency constraint w.r.t. the global allocations $w \in \Lambda$. Such constraints encode the condition that there exists a global allocation w having marginal group allocations $\tilde{w}$. Let $\mathrm{kl}(x, y)$ be the Kullback–Leibler divergence between two Bernoulli distributions of mean x and y, and $\Delta(a) = \theta(a^\star) - \theta(a)$ be the sub-optimality gap for a sub-optimal action $a \neq a^\star_\theta$.

Theorem 4.1. *The sample complexity of any δ-PAC algorithm satisfies $\forall \theta$, $\mathbb{E}_\theta[\tau] \geq 2T_\theta^\star \mathrm{kl}(1-\delta, \delta)$, where*

$$T_\theta^\star = \inf_{\tilde{w} \in \tilde{\Lambda}} \max_{a \neq a^\star} \frac{\sum_{e \in [\rho]: a_e \neq a_e^\star} \left(\tilde{w}_{e,a_e^\star}^{-1} + \tilde{w}_{e,a_e}^{-1} \right)}{\Delta(a)^2} \quad (1)$$

The proof of the above theorem is given for completeness in App. A.1. $T_\theta^\star$ is referred to as *characteristic time* and represents the hardness of the M-BAI problem for a MAMAB instance θ. Furthermore, an allocation $\tilde{w} \in \tilde{\Lambda}$ attaining $T_\theta^\star$ is optimal: an algorithm relying on a sampling strategy realizing $\tilde{w}$ such that for all $e \in [\rho]$ and $a_e \in \mathcal{A}_e$, $\tilde{w}_{e,a_e} = \mathbb{E}_\theta[N_{\tau,e,a_e}]/\mathbb{E}_\theta[\tau]$ (where $N_{t,e,a_e} = \sum_{s \in [t]} \mathbb{1}_{\{a_{s,e}=a_e\}}$ is the number of times the group action a_e is selected up to time t) would yield the lowest possible sample complexity.

5. Lower Bound Approximation

Solving the lower bound optimization problem (1) is hard for the following reasons: (i) *exponential variable space*: global allocations w lie in $\Lambda \subset \mathbb{R}_+^A$; (ii) *exponentially large number of constraints*: $T_\theta^\star$ is defined as a maximum over a set of $K^N - 1$ actions ($a \neq a^\star$), which in turn corresponds to solving a problem with an equal number of constraints (see (4)). To circumvent these issues, we propose an approximation of the lower bound optimization problem, which allows to reduce the number of variables and constraints. We will then leverage this approximation in the design of an efficient M-BAI algorithm.

Mean-field variable reduction. In order to reduce the variable space, we consider an MF approximation (Wainwright & Jordan, 2008), which restricts the allocations $w \in \Lambda$ to the set factored distributions. To define this set, let $v_i = (v_{i,a_i})_{a_i \in \mathcal{A}_i}$ denote the local allocation of agent i. v_{i,a_i} is the proportion of time agent i selects $a_i \in \mathcal{A}_i$, and hence $v_i \in \Lambda_i = \{v_i \in \mathbb{R}_+^K : \sum_{a_i \in \mathcal{A}_i} v_{i,a_i} = 1\}$. The set of factored distributions is then defined as:

$$\Lambda_{\text{MF}} = \left\{ w \in \Lambda : \forall i \in [N], \exists v_i \in \Lambda_i, w_a = \prod_{i \in [N]} v_{i,a_i} \right\}.$$

We also denote by $\tilde{\Lambda}_{\text{MF}}$ the corresponding marginal MF polytope. The lower bound approximation is essentially obtained replacing $\tilde{\Lambda}$ by $\tilde{\Lambda}_{\text{MF}}$ in (1). Then, the corresponding *MF characteristic time* is:

$$T_\theta^{\text{MF}} = \inf_{\tilde{w} \in \tilde{\Lambda}_{\text{MF}}} \max_{a \neq a^\star} \frac{\sum_{e \in [\rho]: a_e \neq a_e^\star} \left(\tilde{w}_{e,a_e}^{-1} + \tilde{w}_{e,a_e^\star}^{-1} \right)}{\Delta_{\min}^2}, \quad (2)$$

where $\Delta_{\min} = \min_{a \neq a^\star} \Delta(a)$. The following lemma proves that T_θ^{MF} is an upper bound of $T_\theta^\star$ and provides a gap-dependent scaling.

Lemma 5.1. *$\forall \theta$, $T_\theta^\star \leq T_\theta^{MF} \leq \frac{2\tilde{A}}{\Delta_{\min}^2}$, and we have:*

$$T_\theta^{MF} = \inf_{(v_i)_{i \in [N]}} \max_{a \neq a^\star} \frac{\sum_{e: a_e \neq a_e^\star} \left(\prod_{i \in \mathcal{S}_e} v_{i,a_i}^{-1} + \prod_{i \in \mathcal{S}_e} v_{i,a_i^\star}^{-1} \right)}{\Delta_{\min}^2}. \quad (3)$$

The proof can be found in App. A.2. Note that the dimension of the variables involved in (2) is KN, whereas solving (1) would involve K^N-dimensional variables $w \in \Lambda$. The lemma also provides a worst-case scaling of T_θ^{MF}: it scales at most with the sum of the group action sets size $\tilde{A} = \sum_{e \in [\rho]} K^{N_e}$. Note that there are trivial factor graphs for which $T_\theta^\star$ scales as $\tilde{A}/\Delta_{\min}^2$ (this is the case when each agent is involved in a single factor).

Notice that the program in (2) is seemingly non-convex. In fact, it is generally known that MF approximations lead to non-convex programs due to the geometry of Λ_{MF} (Wainwright & Jordan, 2008). Despite the non-convexity, we show, in App. B, that (2) can be reformulated as a Geometric Program (GP), which can be reduced to a (non-linear) convex program by a change of variables.

Factored constraint reduction. After the variable reduction step, the main challenge in solving (2) is the $\max_{a \neq a^\star}$. We can rewrite (2) in epigraph form (Boyd & Vandenberghe, 2004) as:

$$\inf_{\tilde{w} \in \tilde{\Lambda}_{\text{MF}}, z \in \mathbb{R}} z \quad \text{s.t.} \quad z \geq \sum_{e \in [\rho]} f_e^{\tilde{w}_e}(a_e), \forall a \neq a^\star, \quad (4)$$

where $f_e^{\tilde{w}_e}(a_e) = (\tilde{w}_{e,a_e}^{-1} + \tilde{w}_{e,a_e^\star}^{-1})/\Delta_{\min}^2 \mathbb{1}_{\{a_e \neq a_e^\star\}}$. In the following, we will omit the superscript $\tilde{w}_e$ for clarity. Hence, the $\max_{a \neq a^\star}$ operator in (2) is equivalent to considering a set of $K^N - 1$ non-linear constraints: $\mathcal{C} = \left\{ z \geq \sum_{e \in [\rho]} f_e(a_e), \forall a \neq a^\star \right\}$.

The objective is to obtain a compact representation of $\mathcal{C}$ that avoids an explicit enumeration of the exponentially-many actions. To this aim, we adapt a popular method used in factored Markov Decision Processes (Guestrin et al., 2001; 2003). This method, which we refer to as Factored Constraint Reduction (FCR), reduces a constraint set described by an exponential number of constraints with a factored structure to a provably equivalent set with a reduced number of constraints. The method is inspired by the Variable Elimination (VE) procedure in graphical models (Dechter, 1999). Specifically, FCR considers constraints of the type:

$$z \geq \sum_{e \in [\rho]} p_e(a_e), \quad \forall a \in \mathcal{A},$$

where $p_e(\cdot)$ is a factor function mapping group actions a_e to real values, and z is a real variable, and construct an equivalent set of constraints $\mathcal{K}$. We present the pseudo-code of FCR in Alg. 1 and describe its steps below.

Algorithm 1 FCR

Input: Elimination order $\mathcal{O}$, factors $\mathcal{F}$
Initialize $\mathcal{K} = \emptyset$
for $i = 1, \dots, N$ **do**
  $l \leftarrow \mathcal{O}(i)$
  $\mathcal{F}_l \leftarrow \{p \in \mathcal{F} : l \in \text{SC}(p)\}$
  $\mathcal{K} \leftarrow \mathcal{K} \cup \left\{ u^{p_l}_{a_{\text{SC}(p_l)}} \geq \sum_{p \in \mathcal{F}_l} u^{p}_{a_{\text{SC}(p)}}, \forall a_{\text{SC}(p_l)}, a_l \right\}$
  $\mathcal{F} \leftarrow \mathcal{F} \cup \{p_l\} \setminus \mathcal{F}_l$
end for
$\mathcal{K} \leftarrow \mathcal{K} \cup \{z \geq u^{p_{\mathcal{O}(N)}}\}$
Return $\mathcal{K}$

FCR takes as input an initial set of factors $\mathcal{F} = \{p_e\}_{e \in [\rho]}$, and an ordered elimination set $\mathcal{O}$. For a factor $p \in \mathcal{F}$, we define its *scope* $\text{SC}(p) \subseteq [N]$ as the set of agents involved in p. We also associate a real variable $u^{p}_{a_{\text{SC}(p)}}$ to each factor $p \in \mathcal{F}$. After initializing the output constraint set as $\mathcal{K} = \emptyset$, the algorithm proceeds in an iterative manner. At each iteration $i = 1, \dots, N$, we set $l = \mathcal{O}(i)$ (the i^{th} element of $\mathcal{O}$), and define $\mathcal{F}_l = \{p \in \mathcal{F} : l \in \text{SC}(p)\}$. We then introduce a new factor p_l having scope $\text{SC}(p_l) = \cup_{p \in \mathcal{F}_l} \{\text{SC}(p)\} \setminus \{l\}$, and we associate the variable $u^{p_l}_{a_{\text{SC}(p_l)}}$ to p_l. We include in $\mathcal{K}$ a new set of constraints

$$u^{p_l}_{a_{\text{SC}(p_l)}} \geq \sum_{p \in \mathcal{F}_l} u^{p}_{a_{\text{SC}(p)}}, \quad \forall a_{\text{SC}(p_l)}, a_l.$$

We further include the new factor variable p_l in the set of factors $\mathcal{F}$ and remove all factors in $\mathcal{F}_l$ from it, i.e., $\mathcal{F} = \mathcal{F} \cup \{p_l\} \setminus \mathcal{F}_l$. At $l = \mathcal{O}(N)$, we introduce the constraint $z \geq u^{p_{\mathcal{O}(N)}}$ into $\mathcal{K}$, where $p_{\mathcal{O}(N)}$ is the last generated factor and has empty scope.

As shown in App. B, the set of constraints in $\mathcal{K}$, constructed through FCR, are equivalent to the ones in $\mathcal{C}$, i.e., an assignment of variables satisfies the constraints in $\mathcal{C}$ if and only if it satisfies the constraints in $\mathcal{K}$. Furthermore, the number of constraints in $\mathcal{K}$ set scales as $O(NK^{A_\mathcal{O}})$, where $A_\mathcal{O} = \max_{i \in [N]} |\text{SC}(p_{\mathcal{O}(i)})|$ is the size of the maximum scope induced by the chosen order of elimination $\mathcal{O}$ (see App. E for details).

6. The MF-TaS Algorithm

Our algorithm, MF-TaS, identifies and tracks the sampling allocation solving the approximated lower bound optimization problem (3). Such problem (3) depends on the unknown parameter θ through the optimal action $a^\star_\theta$ and $\Delta_{\min}$. The algorithm hence consists in (i) estimating the unknown parameter θ, (ii) plugging this estimator in (3) to compute the corresponding optimal sampling rule, and (iii) tracking this sampling rule and stopping when enough information has been gathered. We present MF-TaS in Alg. 2 and detail its step in the remainder of this section.

Algorithm 2 MF-TaS

Input: Confidence δ, exploration set $\mathcal{A}_0$
Initialize $\forall e \in [\rho],\ N_{0,e} = 0, \hat{\theta}_{0,e} = 0, U_{0,e} = \mathcal{A}_e$, $\forall i \in [N],\ N_{0,i} = 0, t = 1$
while $t < T^{\text{MF}}_{\hat{\theta}_t} \beta(\delta, t)$ **do**
  if $\exists e : U_{t,e} \neq \emptyset$ **then**
    $b_{t,e} \leftarrow \arg\min_{a_e \in U_{t,e}} N_{t,e,a_e}$
    $a_t \leftarrow a_t \in \mathcal{A}_0 : b_{t,e} = a_{t,e}$
  else
    $a_{t,i} \leftarrow \arg\max_{a_i \in \mathcal{A}_i} t v_{t,i,a_i} - N_{t,i,a_i}$
    $a_t \leftarrow (a_{t,i})_{i \in [N]}$
  end if
  $t \leftarrow t + 1$
  Update $(N_{t,i})_{i \in [N]}, (N_{t,e}, \hat{\theta}_{t,e}, U_{t,e})_{e \in [\rho]}$,
  $(v_{t,i})_{i \in [N]} \leftarrow$ Solve (3) with $\hat{\theta}_t$ plugged in
end while
Return $\hat{a}_\tau \leftarrow \arg\max_{a \in \mathcal{A}} \sum_{e \in [\rho]} \hat{\theta}_{t,e}(a_e)$

6.1. Parameter estimation

In principle, the estimation of θ would require evaluating an exponentially large number of components, i.e., $\theta(a), \forall a \in \mathcal{A}$. Instead, by leveraging the factored reward structure, we can focus on estimating group parameters $(\theta_e)_{e \in [\rho]} \in \mathbb{R}^{\tilde{A}}$. We define the estimate at time t, group e, and action a_e as:

$$\hat{\theta}_{t,e,a_e} = \frac{1}{N_{t,e,a_e}} \sum_{s \in [t]} r_{s,e} \mathbb{1}_{\{a_{s,e} = a_e\}}.$$

We then define $\hat{\theta}_{t,a} = \sum_{e \in [\rho]} \hat{\theta}_{t,e,a_e}$, for all $a \in \mathcal{A}$.

6.2. Sampling Rule

The sampling rule is inspired by the D-tracking rule in (Garivier & Kaufmann, 2016) and alternates between *forced exploration* and *tracking* steps as described below.

Forced Exploration. During forced exploration steps, we select arms to ensure convergence of the group parameter estimates $(\hat{\theta}_{t,e})_{e \in [\rho]}$ as $t \to \infty$. Define a set of exploratory global actions $\mathcal{A}_0 \subseteq \mathcal{A}$, which is chosen in such a way that it covers all possible group actions, i.e., $\mathcal{A}_0$ is such that $\forall e \in [\rho], \forall a_e \in \mathcal{A}_e, \exists b \in \mathcal{A}_0 : b_e = a_e$ (see App. C for an algorithm to select $\mathcal{A}_0$ efficiently). Further define the set of under-explored actions at group e and time t as: $U_{t,e} = \left\{ a_e \in \mathcal{A}_e : N_{t,e,a_e} < \sqrt{t/|\mathcal{A}_0|} \right\}$. At time t, if there is a group e such that $U_{t,e} \neq \emptyset$, the algorithm executes a forced exploration step. In such steps, we first compute the most under-explored group arm $a_{t,e} = \arg\min_{a_e \in U_{t,e}} N_{t,e,a_e}$ (with ties breaking arbitrarily), and then select an action from the exploratory action set $b_t \in \mathcal{A}_0$ such that $a_{t,e} = b_{t,e}$ (with ties breaking arbitrarily).

Tracking. In the tracking phase, at time t, we solve the optimization problem (3) with the estimated parameter $\hat{\theta}_t$, and derive the optimal estimated allocations $(v_{t,i})_{i\in[N]}$, where $v_{t,i} = (v_{t,i,a_i})_{a_i\in\mathcal{A}_i}$. Then, the action selected by agent i at a tracking time step t is:

$$a_{t,i} = \arg\max_{a_i\in\mathcal{A}_i} tv_{t,i,a_i} - N_{t,i,a_i}.$$

The global action is simply selected as $a_t = (a_{t,i})_{i\in[N]}$. Note that tracking single-agent allocations v_{t,i,a_i} rather than the global allocations $w_{t,a} = \prod_{i\in[N]} v_{t,i,a_i}$ allows us to reduce the search space for the tracking action from a combinatorial set of actions $a \in \mathcal{A}$ to local sets $a_i \in \mathcal{A}_i$, for all agents $i \in [N]$.

6.3. Stopping Rule

For the stopping rule, we use the classical GLRT as in previous works (Garivier & Kaufmann, 2016; Wang et al., 2021). Specifically, the test consists in comparing $T^{\text{MF}}_{\hat{\theta}_t}$ to an *exploration threshold* $\beta(\delta, t)$ as

$$\tau = \inf\left\{t \geq 1 : t \geq T^{\text{MF}}_{\hat{\theta}_t}\beta(\delta, t)\right\}. \tag{5}$$

The conditions that the exploration threshold must satisfy are the same as Sec. 3.2 of (Wang et al., 2021) (see App. D for details). An exploration threshold satisfying these conditions is presented in (Kaufmann & Koolen, 2021). Unless otherwise mentioned, we will use such a threshold.

6.4. Decision Rule

The decision rule selects the best empirical action:

$$\hat{a}_\tau = \arg\max_{a\in\mathcal{A}} \sum_{e\in[\rho]} \hat{\theta}_{\tau,e,a_e}.$$

Note that, in principle, computing $\hat{a}_\tau$ would require a $\max$ operation over an exponential number of actions $a \in \mathcal{A}$. However, due to the factored structure, we can implement the decision rule efficiently through VE (Dechter, 1999), an important sub-routine presented in Alg. 3 and detailed in the remainder of this section.

Algorithm 3 VE

Input: Elimination order $\mathcal{O}$, factors $\mathcal{R}$
for $i = 1, \dots, N$ **do**
  $l = \mathcal{O}(i)$
  $\mathcal{R}_l = \{r_e \in \mathcal{R} : l \in \text{Sc}(r_e)\}$
  $p_l(a_{e\setminus l}) = \max_{a_l\in\mathcal{A}_l} \sum_{r_e\in\mathcal{R}_l} r_e(a_l, a_{e\setminus l})$
  $\mathcal{R} \leftarrow \mathcal{R} \cup \{p_l(a_{e\setminus l})\} \setminus \mathcal{R}_l$
end for
Return $\sum_{p\in\mathcal{R}} p_{\mathcal{O}(N)}$

Variable Elimination. Similarly to FCR, VE follows an elimination order $\mathcal{O}$, where $\mathcal{O}(i)$ is the i^{th} variable to be eliminated. The algorithm takes as input a set of factorized reward functions $\mathcal{R} = \{r_e\}_{e\in[\rho]}$. The algorithm proceeds iteratively for $i = 1, \dots, N$, by eliminating variable $l = \mathcal{O}(i)$ in each round. At round l, all the factors in $\mathcal{R}$ containing variable l in their scopes are collected in the set $\mathcal{R}_l$. Subsequently, the (marginal) best response is computed as $p_l(a_{e\setminus l}) = \max_{a_l\in\mathcal{A}_l} \sum_{r_e\in\mathcal{R}_l} r_e(a_l, a_{e\setminus l})$, where $a_{e\setminus l}$ corresponds to the action a_e corresponds to the action a_e with the l-th component removed. The set of factors is then updated as $\mathcal{R} \leftarrow \mathcal{R}\cup\{p(a_{e\setminus l})\}\setminus\mathcal{R}_l$. At this point, every factor containing l in its scope is eliminated. At the next iteration, the algorithm selects the next variable to be eliminated until $i = N$. Finally, it returns the optimal value $\sum_{p\in\mathcal{R}} p_{\mathcal{O}(N)}$.

Note that VE, applied to $\mathcal{R} = \{\hat{\theta}_{\tau,e}\}_{e\in[\rho]}$ returns the highest global estimated reward $\hat{\theta}_{\tau,\hat{a}_\tau}$. A backward pass of the VE algorithm allows to recover the optimal arm $\hat{a}_\tau$. The time and memory complexity of VE is $O(NK^{A_{\mathcal{O}}})$ (see App. E).

6.5. Sample complexity guarantees

We establish that the MF-TaS algorithm achieves a sample complexity, matching the approximated lower bound $T^{\text{MF}}_\theta \text{kl}(1-\delta, \delta)$ asymptotically (as $\delta \to 0$). The proof is given in App. D.

Theorem 6.1. *MF-TaS is δ-PAC, and its sample complexity satisfies, $\forall\theta$, $\mathbb{P}_\theta\left(\limsup_{\delta\to 0} \frac{\tau}{\log(\frac{1}{\delta})} \leq T^{MF}_\theta\right) = 1$, and $\limsup_{\delta\to 0} \frac{\mathbb{E}_\theta[\tau]}{\log(\frac{1}{\delta})} \leq T^{MF}_\theta$.*

7. Experiments

In this section, we assess the performance of MF-TaS. We propose two sets of experiments through numerical experiments. We apply MF-TaS to synthetic MAMABs with different levels of complexity in §7.1, and to a timely industrial use-case from the radio communication domain: *antenna tilt optimization*, in §7.2. Additional experiments are reported in App. J, and the code is available at this link.

7.1. Synthetic Experiments

Problem instance. We consider a ring factor graph, depicted in Fig. 2, in which the reward is described as $r(a) = \sum_{i\in[N-1]} r_i(a_i, a_{i+1}) + r_N(a_1, a_N)$, for $a_i \in [K]$. The local expected rewards are selected at random as $\theta_i(a_i, a_{i+i}) \sim \mathcal{U}(0, M)$, for all $i \in [N]$ and for some $M > 0$. We propose two sets of synthetic experiments to test different levels of complexity: (i) we vary the number of agents $N \in \{3, 5, 10\}$ while keeping a fixed $K = 10$ and $\delta = 0.01$, and (ii) we vary the confidence $\delta \in \{10^{-i}\}_{i\in[5]}$, while keeping $K = 10$ and $N = 3$ fixed.

Table 1. Experimental results for the synthetic experiments with varying N (with fixed $K = 10$, $\delta = 0.01$).

Problem instance					Sample complexity			Computational complexity [s]	
N	A	$\tilde{A}$	$T_\theta^{\text{MF}} \log\left(\frac{1}{\delta}\right)$	$\frac{4\tilde{A}}{\Delta_{\min}^2} \log\left(\frac{1}{\delta}\right)$	Oracle	MF-TaS	Random	T_θ^{MF}	$T_\theta^\star$
3	10^3	300	158.3	209.0	223	299.6 ± 107.5	644.7 ± 140.2	0.09 ± 0.03	6.87 ± 1.12
5	10^5	500	270.5	358.0	385	356.8 ± 150.0	708.3 ± 132.9	0.54 ± 0.13	2375.82 ± 5.32
10	10^{10}	1000	305.9	405.0	533	411.1 ± 193.5	800.4 ± 177.9	0.81 ± 0.31	> 10800 (3 h)

Implementation details. We execute our experiments for $N_{\text{sim}} = 100$ runs. Following previous works (Kaufmann & Koolen, 2021; Wang et al., 2021), the exploration threshold is selected as $\beta(\delta, t) = \log(\log(t) + 1)/\delta)$. The elimination order for both VE and FCR is chosen as $\mathcal{O} = \{N, N-1, \dots, 1\}$. We implement the solver for the lower bound optimization problems using CVXPY (Diamond & Boyd, 2016), with a MOSEK solver. The experiments run on a MacBook Pro 2.6 GHz 6-Core Intel Core i7 processor. We use this setup in all of our experiments.

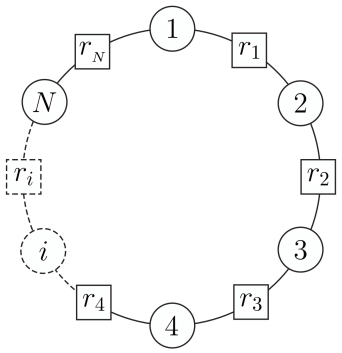


Figure 2. Ring factor graph used in the synthetic experiments.

Results. The results are presented in Tab. 1 for the experiments with varying N, and in Tab. 2 for the experiments with varying δ, with the sample complexity box-plots reported in Fig. 3. The sample complexity corresponds to the stopping time averaged over the various runs. The performance of MF-TaS is compared to that of an oracle algorithm aware of the problem parameters θ (obtained by replacing $\hat{\theta}_t$ by θ in MF-TaS), and to a random strategy selecting actions uniformly at random. The computational complexity is the average running time (in seconds) to solve one instance of the lower bound optimization problem for $T_\theta^\star$ and T_θ^{MF}.

Table 2. Experimental results for the synthetic experiments with varying δ (with fixed $K = 10$, $N = 3$).

Problem instance		Sample complexity		
δ	$T_\theta^{\text{MF}} \log\left(\frac{1}{\delta}\right)$	Oracle	MF-TaS	Random
10^{-1}	79.1	154	223.5 ± 82.4	635.5 ± 135.5
10^{-2}	158.3	223	317.5 ± 133.2	649.4 ± 142.6
10^{-3}	237.4	303	397.6 ± 129.0	668.2 ± 131.5
10^{-4}	316.5	384	493.9 ± 153.8	695.5 ± 141.4
10^{-5}	395.7	464	535.4 ± 154.2	793.4 ± 190.8

We note that MF-TaS exhibits a sample complexity close to the proposed MF approximation of the lower bound $T_\theta^{\text{MF}} \log(1/\delta)$ (they differ from a small multiplicative constant) for all values of N. We also observe that, as expected, MF-TaS outperforms the random sampling strategy and is competitive with the oracle strategy. In terms of computational complexity, the average running time to solve T_θ^{MF} is significantly lower than that to solve $T_\theta^\star$, which becomes quickly untractable even for a small number of agents.

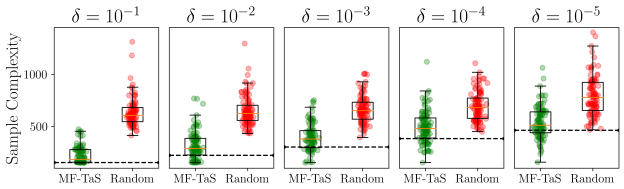


Figure 3. Sample complexity boxplots for the experiments with varying δ (dashed line represents the oracle performance).

7.2. Antenna Tilt Optimization

Next, we test MF-TaS on the antenna tilt optimization problem. The task consists in controlling the vertical antenna tilt at different network base stations to optimize the network throughput. In the following, we detail the network model, our simulation setup, and present our experimental results. Additional details are presented in App. I.

Network Model. We consider a sectorized mobile network consisting of a set of *sectors* $\mathcal{S} = [N]$. The set of sectors corresponds to the set of agents in our M-BAI model. Since each sector is associated to a unique antenna, we will use the terms *sector* and *antenna* interchangeably. We assume that each sector $i \in \mathcal{S}$ serves (on the downlink) a fixed set of Users Equipments (UEs) $\mathcal{U}_i$ (each UE is associated with a unique antenna, that from which it receives the strongest signal). The set of UEs in the network is $\mathcal{U} = \cup_{i \in \mathcal{S}} \mathcal{U}_i$.

Factor graph. We model the observed reward in the network as a factor graph with $N = |\mathcal{S}|$ agent nodes and $\rho = |\mathcal{S}|$ factor nodes. Each sector is associated with a unique factor, which models the rewards observed in that sector. We build the factor graph based on the interference pattern of the antennas, i.e., antennas that can interfere with each other are connected to common factors. Fig. 4 shows an example of such a graph on a network with $|\mathcal{S}| = 15$.

Table 3. Experimental results for the antenna tilt optimization experiments.

Problem instance				Sample complexity			Computational complexity [s]	
N	A	$\tilde{A}$	$T_\theta^{\text{MF}} \log\left(\frac{1}{\delta}\right)$	Oracle	MF-TaS	Random	T_θ^{MF}	$T_\theta^\star$
6	$2.4 \cdot 10^2$	324	129.2	203	286.4	341.1	0.93 ± 0.27	4.34 ± 0.92
12	$5.9 \cdot 10^4$	1124	472.1	524	623.5	813.9	1.32 ± 0.62	2778.53 ± 5.32
15	$1.4 \cdot 10^7$	1799	568.9	729	913.7	1262.1	3.24 ± 0.91	> 10800 (3 h)

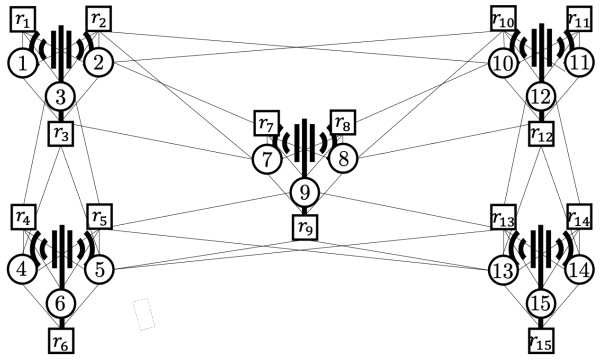


Figure 4. Network factor graph.

Table 4. Simulator parameters.

PARAMETER	SYMBOL	VALUE
Number of sectors	$\lvert\mathcal{S}\rvert$	$\{6, 12, 15\}$
Number of UEs	$\lvert\mathcal{U}\rvert$	1000
Antenna tilt values	$\mathcal{A}_i$	$\{3°, 6°, 12°\}$
Carrier frequency	f	1800 MHz
Antenna height	h	32 m
Network size	M	2 km^2

Actions. The action $a_{t,i}$ represents the antenna tilt for sector $i \in \mathcal{S}$ and at time t. For simplicity, it is chosen from a discrete set of K tilts, i.e., $a_{t,i} \in \{\alpha_1, \dots, \alpha_K\}$. The antenna tilt for a group of sectors e is denoted by a_e.

Rewards. Rewards are based on the throughput of UEs in sector i, which depends on the actions of a group of agents a_e: $r_e(a_e) = \sum_{u \in \mathcal{U}_i} T_{i,u}(a_e)$, where $T_{i,u}$ is the throughput of an UE u associated to sector i. Hence, the global reward for a tilt configuration $a \in \mathcal{A}$ is $r(a) = \sum_{i \in [N]} \sum_{u \in \mathcal{U}_i} T_{i,u}(a_e)$. The throughput $T_{i,u}$ depends on channel conditions (or *fading*) between the base station antenna and the user. These conditions rapidly evolve over time around their mean. The fadings between pairs of (antenna, user) are typically stochastically independent across users and antennas (Tse & Viswanath, 2009). More precisely, since the sets of $(\mathcal{U}_i)_{i \in [N]}$ form a partition, they do not overlap, and the random variables $r_e(a_e)$ are independent across groups and we assume that can be modeled as independent Gaussian r.v.

Simulator. We run our experiments in a proprietary mobile network simulator in an urban environment. The simulation parameters used in our experiments are reported in Tab. 4. Based on the user positions and network parameters, the simulator computes the path loss in the network environment using a BEZT propagation model (Rappaport, 2001) and returns the throughput for each sector by conducting user association and resource allocation in a full-buffer traffic demand scenario. Given a user configuration, the goal is to identify the best global tilt configuration in the network, i.e., the one which maximizes the overall network throughput.

Results. We test our algorithm with the same experimental conditions in §7.1. The sample complexity and the computational complexity of MF-TaS are presented in Tab. 3. The results are in line with the experimental findings of the previous section. However, due to the higher degree of the factor graph, the MF-TaS running time is higher.

8. Conclusions

In this paper, we investigated the M-BAI problem: we derived a sample complexity lower bound, proposed a Mean Field approximation of it, and devised MF-TaS, an algorithm achieving this limit in a computationally efficient manner. MF-TaS is statistically and computationally efficient on both synthetic examples and the antenna tilt optimization problem. The algorithm runs fast and identifies the best global action using a limited number of samples, even for scenarios with a very large number of actions.

Interesting future research directions include (i) the analysis of the sample complexity lower bound and its Mean Field approximation depending on the factor graph topology, (ii) extending the analysis towards tighter lower bound approximations, (iii) proposing efficient distributed implementations of MF-TaS with a specific focus on its communication complexity, and (iv) investigating representation learning problems in M-BAI where the underlying factor graph is initially unknown and needs to be learned.

Acknowledgement

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

References

Bargiacchi, E., Verstraeten, T., Roijers, D., Nowé, A., and van Hasselt, H. Learning to coordinate with coordination graphs in repeated single-stage multi-agent decision problems. In *Proc. of ICML*, 2018.

Bargiacchi, E., Verstraeten, T., Roijers, D., Nowé, A., and van Hasselt, H. Multi-agent RMax for multi-agent multi-armed bandits. In *Proc. of Adaptive and Learning Agents Worksh.*, 2022.

Berge, C. Topological spaces. *Proceedings of the Edinburgh Mathematical Society*, 1963.

Besson, L. and Kaufmann, E. Multi-Player Bandits Revisited. In *Proc. of ALT*, 2018.

Boyd, S. and Vandenberghe, L. *Convex Optimization*. Cambridge University Press, 2004.

Boyd, S. P., Kim, S.-J., Vandenberghe, L., and Hassibi, A. A tutorial on geometric programming. *Optimization and Engineering*, 2007.

Chen, S., Lin, T., King, I., Lyu, M. R., and Chen, W. Combinatorial pure exploration of multi-armed bandits. In *Proc. of NeurIPS*, 2014.

Cuvelier, T., Combes, R., and Gourdin, E. Statistically efficient, polynomial-time algorithms for combinatorial semi-bandits. *Proc. ACM Meas. Anal. Comput. Syst.*, 2021a.

Cuvelier, T., Combes, R., and Gourdin, E. Asymptotically optimal strategies for combinatorial semi-bandits in polynomial time. In *Proc. of ALT*, 2021b.

Dechter, R. Bucket elimination: A unifying framework for reasoning. *Artif. Intell.*, 1999.

Degenne, R., Ménard, P., Shang, X., and Valko, M. Gamification of pure exploration for linear bandits. In *Proc. of ICML*, 2020.

Diamond, S. and Boyd, S. CVXPY: A Python-embedded modeling language for convex optimization. *JMLR*, 2016.

Du, Y., Kuroki, Y., and Chen, W. Combinatorial pure exploration with full-bandit or partial linear feedback. In *Proc. of AAAI*, 2021.

Fiez, T., Jain, L., Jamieson, K. G., and Ratliff, L. Sequential experimental design for transductive linear bandits. In *Proc. of NeurIPS*, 2019.

Garivier, A. and Kaufmann, E. Optimal best arm identification with fixed confidence. In *Proc. of COLT*, 2016.

Guestrin, C., Koller, D., and Parr, R. Max-norm projections for factored MDPs. In *Proc. of IJCAI*, 2001.

Guestrin, C., Koller, D., Parr, R., and Venkataraman, S. Efficient solution algorithms for factored mdps. *J. Artif. Int. Res.*, 2003.

Jedra, Y. and Proutiere, A. Optimal best-arm identification in linear bandits. In *Proc. of NeurIPS*, 2020.

Jourdan, M., Mutnỳ, M., Kirschner, J., and Krause, A. Efficient pure exploration for combinatorial bandits with semi-bandit feedback. In *Proc. of ALT*, 2021.

Karnin, Z. S. Verification based solution for structured mab problems. In *Advances in Neural Information Processing Systems*, 2016.

Kaufmann, E. and Koolen, W. M. Mixture martingales revisited with applications to sequential tests and confidence intervals. *JMLR*, 2021.

Kaufmann, E., Cappé, O., and Garivier, A. On the complexity of best-arm identification in multi-armed bandit models. *JMLR*, 2016.

Kiefer, J. and Wolfowitz, J. The equivalence of two extremum problems. *Canad. Journ. of Math.*, 1960.

Lai, T. L. and Robbins, H. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 1985.

Lawler, E. A procedure for computing the k best solutions to discrete optimization problems and its application to the shortest path problem. *Management Science*, 1972.

Nilsson, D. An efficient algorithm for finding the m most probable configurationsin probabilistic expert systems. *Statistics and Computing*, 1998.

Rappaport, T. *Wireless Communications: Principles and Practice*. Prentice Hall PTR, 2001.

Rosenski, J., Shamir, O., and Szlak, L. Multi-player bandits – a musical chairs approach. In *Proc. of ICML*, 2016.

Shi, C., Xiong, W., Shen, C., and Yang, J. Heterogeneous multi-player multi-armed bandits: Closing the gap and generalization. In *Proc. of NeurIPS*, 2021.

Soare, M., Lazaric, A., and Munos, R. Best-arm identification in linear bandits. In *Proc. of NeurIPS*, 2014.

Stranders, R., Fave, F. M. D., Rogers, A., and Jennings, N. R. DCOPs and bandits: exploration and exploitation in decentralised coordination. In *Proc. of AAMAS*, 2012.

Tao, C., Blanco, S., and Zhou, Y. Best arm identification in linear bandits with linear dimension dependency. In *Proc. of ICML*, 2018.

Tse, D. N. C. and Viswanath, P. Fundamentals of wireless communication. *IEEE Trans. Inf. Theory*, 2009.

Verstraeten, T., Bargiacchi, E., Libin, P. J. K., Helsen, J., Roijers, D. M., and Nowé, A. Multi-agent thompson sampling for bandit applications with sparse neighbourhood structures. *Scientific Reports*, 2020.

Wagenmaker, A., Katz-Samuels, J., and Jamieson, K. Experimental design for regret minimization in linear bandits, 2020.

Wainwright, M. and Jordan, M. *Graphical Models, Exponential Families, and Variational Inference*. Now Publishers Inc., 2008.

Wang, P.-A., Proutiere, A., Ariu, K., Jedra, Y., and Russo, A. Optimal algorithms for multiplayer multi-armed bandits. In *Proc. of AISTATS*, 2020.

Wang, P.-A., Tzeng, R.-C., and Proutiere, A. Fast pure exploration via frank-wolfe. In *Proc. of NeurIPS*, 2021.

A. Lower Bound Proofs

In this appendix, we prove Theorem 4.1 (in App. A.1) and Lemma 5.1 (in App. A.2). Finally, we state a result bounding the approximation ratio $T_\theta^{\text{MF}}/T_\theta^\star$ in App. A.3.

A.1. Proof of Theorem 4.1

Proof. The proof leverages the classical change-of-measure argument (Lai & Robbins, 1985) and the transportation lemma from Lemma 19 in (Kaufmann et al., 2016) to accommodate the MAMAB setting. We divide the proof into 5 steps. The first four steps are standard, and consist in relating the log-likelihood ratio of the observations under two models to the expected sample complexity. They are given for completeness. The last step is specific to MAMABs and deals with optimizing the resulting lower bounds.

1) The log-likelihood ratio. Let $\mathcal{M} = \{\theta \in \mathbb{R}^A : \exists (\theta_e(a_e))_{e\in[\rho], a_e\in\mathcal{A}} : \forall a \in \mathcal{A}, \theta(a) = \sum_{e\in[\rho]} \theta_e(a_e), a_\theta^\star \text{ is unique}\}$ denote the set of possible parameters describing a MAMAB. Let $\theta, \mu \in \mathcal{M}$. For any (global) action $a \in \mathcal{A}$ denote by f_a^θ the density (w.r.t. the Lebesgue measure) of the reward distribution of action a. For any (group) action $a_e \in \mathcal{A}_e$, denote by $\nu_{a_e}^{\theta_e} = \mathcal{N}(\theta_e(a_e), 1)$ the distribution of the corresponding reward, and by $f_{a_e}^{\theta_e}$ its density. For the parameters θ, μ, the log-likelihood ratio of the observations under a M-BAI algorithm up to round T can be written as

$$\begin{aligned}
L_T^{\theta,\mu}(a_{1,[\rho]}, r_{1,[\rho]}, \dots, a_{T,[\rho]}, r_{T,[\rho]}) &= \log\left(\frac{f^\theta(a_{1,[\rho]}, r_{1,[\rho]}, \dots, a_{T,[\rho]}, r_{T,[\rho]})}{f^\mu(a_{1,[\rho]}, r_{1,[\rho]}, \dots, a_{T,[\rho]}, r_{T,[\rho]})}\right) \\
&= \sum_{t\in[T]} \log\left(\prod_{e\in[\rho]} \frac{f^{\theta_e}(r_{t,e}|a_{t,e})}{f^{\mu_e}(r_{t,e}|a_{t,e})}\right) \\
&= \sum_{t\in[T]} \sum_{e\in[\rho]} \log\left(\frac{f^{\theta_e}(r_{t,e}|a_{t,e})}{f^{\mu_e}(r_{t,e}|a_{t,e})}\right) \\
&= \sum_{t\in[T]} \sum_{e\in[\rho]} \sum_{a_e\in\mathcal{A}_e} \mathbb{1}_{\{a_{t,e}=a_e\}} \log\left(\frac{f_{a_e}^{\theta_e}(r_{t,e})}{f_{a_e}^{\mu_e}(r_{t,e})}\right).
\end{aligned} \tag{6}$$

2) The change-of-measure argument. Define the set of confusing parameters as $B(\theta) = \left\{\mu \in \mathcal{M} : a_\mu^\star \neq a_\theta^\star\right\}$, and the event $\mathcal{E} = \{\hat{a}_\tau = a_\theta^\star\}$. Under any δ-PAC algorithm, it holds that

$$\forall \theta \in \mathcal{M},\ \mathbb{P}_\theta(\mathcal{E}) \geq 1-\delta, \text{ and } \forall \mu \in B(\theta), \mathbb{P}_\mu(\mathcal{E}) \leq \delta.$$

Therefore, by applying Lemma 19 in (Kaufmann et al., 2016), we obtain

$$\forall \mu \in B(\theta), \qquad \mathbb{E}\left[L_\tau^{\theta,\mu}\right] \geq \mathrm{kl}\left(1-\delta, \delta\right). \tag{7}$$

3) The expected log-likelihood ratio. We apply (6) to $T = \tau$, the stopping time. Taking the expectation of (6), we get

$$\begin{aligned}
\mathbb{E}_\theta\left[L_\tau^{\theta,\mu}\right] &= \mathbb{E}_\theta\left[\sum_{t=1}^\infty \mathbb{1}_{\{\tau>t-1\}} \sum_{e\in[\rho]} \sum_{a_e\in\mathcal{A}_e} \mathbb{1}_{\{a_{t,e}=a_e\}} \log\left(\frac{f_{a_e}^{\theta_e}(r_{t,e})}{f_{a_e}^{\mu_e}(r_{t,e})}\right)\right] \\
&= \mathbb{E}_\theta\left[\sum_{t=1}^\infty \mathbb{E}_\theta\left[\sum_{e\in[\rho]} \sum_{a_e\in\mathcal{A}_e} \mathbb{1}_{\{\tau>t-1\}} \mathbb{1}_{\{a_{t,e}=a_e\}} \log\left(\frac{f_{a_e}^{\theta_e}(r_{t,e})}{f_{a_e}^{\mu_e}(r_{t,e})}\right) \middle| \mathcal{F}_{t-1}\right]\right] \\
&= \mathbb{E}_\theta\left[\sum_{t=1}^\infty \sum_{e\in[\rho]} \sum_{a_e\in\mathcal{A}_e} \mathbb{1}_{\{\tau>t-1\}} \mathbb{1}_{\{a_{t,e}=a_e\}} \mathbb{E}_\theta\left[\log\left(\frac{f_{a_e}^{\theta_e}(r_{t,e})}{f_{a_e}^{\mu_e}(r_{t,e})}\right) \middle| \mathcal{F}_{t-1}\right]\right] \\
&= \sum_{e\in[\rho]} \sum_{a_e\in\mathcal{A}_e} \mathbb{E}_\theta\left[N_{\tau,e,a_e}\right] \mathrm{KL}(\nu_{a_e}^{\theta_e}, \nu_{a_e}^{\mu_e}),
\end{aligned}$$

where $\mathrm{KL}(\nu, \nu')$ is the KL divergence for distributions ν and ν'. Now, since the distributions are Gaussians, we get:

$$\mathbb{E}_\theta\left[L_\tau^{\theta,\mu}\right] = \sum_{e\in[\rho]}\sum_{a_e\in\mathcal{A}_e} \mathbb{E}_\theta\left[N_{\tau,e,a_e}\right]\frac{\left(\theta_e(a_e)-\mu_e(a_e)\right)^2}{2}. \tag{8}$$

4) Optimizing the lower bound. By combining (7) and (8), we obtain

$$\inf_{\mu\in B(\theta)} \sum_{e\in[\rho]}\sum_{a_e\in\mathcal{A}_e} \mathbb{E}_\theta\left[N_{\tau,e,a_e}\right]\frac{\left(\theta_e(a_e)-\mu_e(a_e)\right)^2}{2} \geq \mathrm{kl}(1-\delta,\delta). \tag{9}$$

Dividing both sides of (9) by $\mathbb{E}_\theta[\tau]$, defining the group allocations $\tilde{w}_{e,a_e} = \mathbb{E}_\theta\left[N_{\tau,e,a_e}\right]/\mathbb{E}_\theta[\tau]$ for all $e\in[\rho]$, $a_e\in\mathcal{A}_e$, and optimizing over the set of possible allocations $\tilde{w}\in\tilde{\Lambda}$, we get:

$$\mathbb{E}_\theta[\tau] \geq \frac{2\mathrm{kl}(1-\delta,\delta)}{\sup_{\tilde{w}\in\tilde{\Lambda}}\inf_{\mu\in B(\theta)}\sum_{e\in[\rho]}\sum_{a_e\in\mathcal{A}_e}\tilde{w}_{e,a_e}\left(\theta_e(a_e)-\mu_e(a_e)\right)^2}. \tag{10}$$

5) Solving the optimization problem. Consider the optimization problem at the denominator of (10):

$$\inf_{\mu\in B(\theta)} \sum_{e\in[\rho]}\sum_{a_e\in\mathcal{A}_e} \tilde{w}_{e,a_e}\frac{\left(\theta_e(a_e)-\mu_e(a_e)\right)^2}{2}. \tag{11}$$

Note that the set of confusing parameters defined before can be expressed as:

$$\begin{aligned} B(\theta) &= \left\{\mu\in\mathcal{M} : a_\mu^\star \neq a_\theta^\star\right\} \\ &= \bigcup_{a\neq a_\theta^\star}\left\{\mu\in\mathcal{M} : \mu(a_\theta^\star)\leq\mu(a)\right\} \\ &= \bigcup_{a\neq a_\theta^\star}\left\{(\mu_e)_{e\in[\rho]} : \sum_{e\in[\rho]}\mu_e(a_e^\star)\leq\sum_{e\in[\rho]}\mu_e(a_e)\right\}. \end{aligned}$$

Then, using this decomposition, (11) can be rewritten as:

$$\begin{aligned} \min_{a\neq a^\star}\inf_{(\mu_e)_{e\in[\rho]}} &\sum_{e\in[\rho]}\sum_{a_e\in\mathcal{A}_e}\tilde{w}_{e,a_e}\frac{\left(\theta_e(a_e)-\mu_e(a_e)\right)^2}{2} \\ \text{subject to } &\sum_{e\in[\rho]}\left(\mu_e(a_e^\star)-\mu_e(a_e)\right)\leq 0. \end{aligned} \tag{12}$$

Now, (12) is a convex program and it can be easily verified that Slater's conditions hold (Boyd & Vandenberghe, 2004). The Lagrangian associated to (12) is:

$$\mathcal{L}((\mu_e)_{e\in[\rho]},\lambda) = \sum_{e\in[\rho]}\sum_{a_e\in\mathcal{A}_e}\tilde{w}_{e,a_e}\frac{\left(\theta_e(a_e)-\mu_e(a_e)\right)^2}{2} - \lambda\left(\sum_{e\in[\rho]}\left(\mu_e(a_e^\star)-\mu_e(a_e)\right)\mathbb{1}_{\{a_e\neq a_e^\star\}}\right),$$

where $\lambda\in\mathbb{R}$ is the Lagrange multiplier associated with the inequality constraint. The optimality conditions impose:

$$\begin{cases} \tilde{w}_{e,a_e}(\mu_e(a_e)-\theta_e(a_e)) + \lambda\mathbb{1}_{\{a_e\neq a_e^\star\}} = 0, \text{ for } e\in[\rho] \\ \tilde{w}_{e,a_e^\star}(\mu_e(a_e^\star)-\theta_e(a_e^\star)) - \lambda\mathbb{1}_{\{a_e\neq a_e^\star\}} = 0, \text{ for } e\in[\rho] \\ \lambda\left(\sum_{e\in[\rho]}\left(\mu_e(a_e^\star)-\mu_e(a_e)\right)\mathbb{1}_{\{a_e\neq a_e^\star\}}\right) = 0 \\ \lambda\geq 0 \end{cases}. \tag{13}$$

It can be directly verified that the solution of the set of equations in (13) is attained at:

$$\mu_e(a_e) = \theta_e(a_e) - \frac{\lambda}{\tilde{w}_{e,a_e}},\quad \mu_e(a_e^\star) = \theta_e(a_e^\star) + \frac{\lambda}{\tilde{w}_{e,a_e^\star}},\quad \lambda = \frac{\sum_{e\in[\rho]}(\theta_e(a_e^\star)-\theta_e(a_e))}{\sum_{e\in[\rho]}\left(\frac{1}{\tilde{w}_{e,a_e}}+\frac{1}{\tilde{w}_{e,a_e^\star}}\right)\mathbb{1}_{\{a_e\neq a_e^\star\}}} > 0.$$

Hence, by substitution of $(\mu_e(a_e))_{e\in[\rho],a_e\in\mathcal{A}_e}$ into the objective of (12), and by (10), we get the result. □

A.2. Proof of Lemma 5.1

Proof. We first show that $\forall \theta \in \mathcal{M}$, $T^\star_\theta \leq T^{\text{MF}}_\theta$. This simply follows by noting that $1/\Delta(a) \leq 1/\Delta_{\min}$, and $\Lambda_{\text{MF}} \subseteq \Lambda$, and hence $\tilde{\Lambda}_{\text{MF}} \subseteq \tilde{\Lambda}$. Combining these facts, we can write:

$$\begin{aligned} T^\star_\theta &= \inf_{\tilde{w}\in\tilde{\Lambda}} \max_{a\neq a^\star} \frac{\sum_{e\in[\rho]:a_e\neq a^\star_e} \left(\tilde{w}^{-1}_{e,a^\star_e} + \tilde{w}^{-1}_{e,a_e}\right)}{\Delta(a)^2} \\ &\leq \inf_{\tilde{w}\in\tilde{\Lambda}} \max_{a\neq a^\star} \frac{\sum_{e\in[\rho]:a_e\neq a^\star_e} \left(\tilde{w}^{-1}_{e,a^\star_e} + \tilde{w}^{-1}_{e,a_e}\right)}{\Delta^2_{\min}} \\ &\leq \inf_{\tilde{w}\in\tilde{\Lambda}_{\text{MF}}} \max_{a\neq a^\star} \frac{\sum_{e\in[\rho]:a_e\neq a^\star_e} \left(\tilde{w}^{-1}_{e,a^\star_e} + \tilde{w}^{-1}_{e,a_e}\right)}{\Delta^2_{\min}} = T^{\text{MF}}_\theta \end{aligned}$$

Next, we show that (2) and (3) are equivalent. First, recall the expression of the MF marginal polytope:

$$\tilde{\Lambda}_{\text{MF}} = \left\{ \tilde{w} \in \mathbb{R}^{\tilde{A}} : \exists w \in \Lambda_{\text{MF}}, \forall e \in [\rho], a_e \in \mathcal{A}_e, \tilde{w}_{e,a_e} = \sum_{b\in\mathcal{A}:b_e=a_e} w_b, \right\}.$$

Note that, for any $\tilde{w} \in \tilde{\Lambda}_{\text{MF}}$, we must have that:

$$\tilde{w}_{e,a_e} = \sum_{b\in\mathcal{A}:b_e=a_e} w_b = \sum_{b\in\mathcal{A}:b_e=a_e} \prod_{i\in[N]} v_{i,b_i} = \prod_{i\in\mathcal{S}_e} v_{i,a_i} \sum_{b\in\mathcal{A}:b_e=a_e} \prod_{j\in[N]\setminus\mathcal{S}_e} w_{j,b_j} = \prod_{i\in\mathcal{S}_e} v_{i,a_i}. \tag{14}$$

Hence, by substituting (14) into (2), and noticing that T^{MF}_θ involves only local allocation variables $(v_i)_{i\in[N]}$, we have that:

$$T^{\text{MF}}_\theta = \inf_{\tilde{w}\in\tilde{\Lambda}_{\text{MF}}} \max_{a\neq a^\star} \frac{\sum_{e:a_e\neq a^\star_e} \left(\tilde{w}^{-1}_{e,a^\star_e} + \tilde{w}^{-1}_{e,a_e}\right)}{\Delta^2_{\min}} = \inf_{(v_i\in\Lambda_i)_{i\in[N]}} \max_{a\neq a^\star} \frac{\sum_{e:a_e\neq a^\star_e} \left(\prod_{i\in\mathcal{S}_e} v^{-1}_{i,a_i} + \prod_{i\in\mathcal{S}_e} v^{-1}_{i,a^\star_i}\right)}{\Delta^2_{\min}}.$$

Note that, if $(v^\star_i)_{i\in[N]}$ is an optimal solution to T^{MF}_θ, we can express such solution it in terms of the optimal group allocations as in (14), i.e., $\tilde{w}^\star = \prod_{i\in\mathcal{S}_e} v^\star_{i,a_i}$.

Finally, we show that, $\forall \theta \in \mathcal{M}$, $T^{\text{MF}}_\theta \leq \frac{4\tilde{A}}{\Delta^2_{\min}}$. The bound is proved using techniques from combinatorial bandits with semi-bandit feedback. This class of bandit problems encompasses MAMAB as a particular case in (we provide clarification for this connection in App. G). Specifically, the proof relies on two known results from (Wagenmaker et al., 2020) that we report below. Lemma A.1 is a reformulation of the MAMAB sample complexity lower bound in the combinatorial semi-bandit feedback setting, while Lemma A.2 is an adaptation of the celebrated Kiefer-Wolfowitz Equivalence Theorem (Kiefer & Wolfowitz, 1960) for the case of semi-bandit feedback problems.

Lemma A.1 ((Wagenmaker et al., 2020), Theorem 6)**.** *For any $\theta \in \mathcal{M}$, we have that*

$$T^\star_\theta = \inf_{w\in\Lambda} \max_{a\neq a^\star} \frac{\|\phi(a^\star) - \phi(a)\|^2_{A_{semi}(w)^{-1}}}{(\tilde{\theta}^\top(\phi(a^\star) - \phi(a))^2}, \tag{15}$$

where

- $\tilde{\theta} = (\theta_e(a_e))_{e\in[\rho],a_e\in\mathcal{A}_e}$
- $\phi(a) \in \{0,1\}^{\tilde{A}}$*, is the binary vector containing* $\phi(a) = [\phi_e(b_e)]_{e\in[\rho],b_e\in\mathcal{A}_e}$ *such that* $\phi_e(b_e) = 1_{\{a_e=b_e\}}$*,*
- $A_{semi}(w) = \text{diag}(\sum_{a\in\mathcal{A}} w_a \phi(a)\phi(a)^\top)$*, where* $\text{diag}(X)$ *is the operator which sets all elements in a matrix X not on the diagonal to* 0

Lemma A.2 ((Wagenmaker et al., 2020), Proposition 9)**.** $\inf_{w\in\Lambda}\max_{a\in\mathcal{A}}\|\phi(a)\|^2_{A_{semi}(w)^{-1}} = \tilde{A}$.

By Lemma A.1, we can equivalently characterize the MF characteristic time as:

$$T^{\text{MF}}_\theta = \inf_{w\in\Lambda_{\text{MF}}}\max_{a\neq a^\star}\frac{\|\phi(a^\star)-\phi(a)\|^2_{A_{\text{semi}}(w)^{-1}}}{(\tilde{\theta}^\top(\phi(a^\star)-\phi(a))^2}.$$

To verify this, notice that $\tilde{\theta}^\top(\phi(a^\star)-\phi(a)) = \sum_{e\in[\rho]}(\theta_e(a^\star_e)-\theta_e(a_e))$. Hence, to show the equivalence, it is sufficient to show that

$$\|\phi(a^\star)-\phi(a)\|^2_{A_{\text{semi}}(w)^{-1}} = \sum_{e\in[\rho]:a_e\neq a^\star_e} 1/\tilde{w}_{e,a_e} + 1/\tilde{w}_{e,a^\star_e}.$$

By definition of $(\phi(a))_{a\in\mathcal{A}}$, and $\forall\tilde{w}\in\tilde{\Lambda}$, we have that $\|\phi(a^\star)-\phi(a)\|^2_{A_{\text{semi}}(w)^{-1}} = \|\phi(a^\star)-\phi(a)\|^2_{M(\tilde{w})^{-1}}$, where $M(\tilde{w})\in\mathbb{R}^{\tilde{A}\times\tilde{A}}$ is the diagonal matrix containing the vector $\tilde{w}$ on its diagonal. Hence, we get:

$$\begin{aligned}
\|\phi(a^\star)-\phi(a)\|^2_{A_{\text{semi}}(w)^{-1}} &= (\phi(a^\star)-\phi(a))^\top\left(\text{diag}\left(\sum_{a\in\mathcal{A}} w_a\phi(a)\phi(a)^\top\right)\right)^{-1}(\phi(a^\star)-\phi(a)) \\
&= (\phi(a^\star)-\phi(a))^\top M(\tilde{w})^{-1}(\phi(a^\star)-\phi(a)) \\
&= \sum_{e\in[\rho]:a_e\neq a^\star_e} 1/\tilde{w}_{e,a_e} + 1/\tilde{w}_{e,a^\star_e}.
\end{aligned}$$

Then, we can upper bound the characteristic time as:

$$\begin{aligned}
T^{\text{MF}}_\theta = T^{\text{MF}}_{\tilde{\theta},\text{semi}} &= \frac{1}{\Delta^2_{\min}}\inf_{w\in\Lambda_{\text{MF}}}\max_{a\neq a^\star}\|\phi(a^\star)-\phi(a)\|^2_{A_{\text{semi}}(w)^{-1}} \\
&\overset{(i)}{\leq} \frac{4}{\Delta^2_{\min}}\inf_{w\in\Lambda_{\text{MF}}}\max_{a\in\mathcal{A}}\|\phi(a)\|^2_{A_{\text{semi}}(w)^{-1}} \\
&\overset{(ii)}{=} \frac{4\tilde{A}}{\Delta^2_{\min}}
\end{aligned}$$

In the above steps, (i) follows by an application of the triangular inequality. Note that, in general, showing (ii) is a non-trivial step due to the non-convexity of the MF approximation. By Lemma A.2, we have that:

$$\inf_{w\in\Lambda}\max_{a\in\mathcal{A}}\|\phi(a)\|^2_{A_{\text{semi}}(w)^{-1}} = \tilde{A},$$

and it can be directly verified that the optimal value is attained at the point $w^* = (1/A,\dots,1/A)\in\Lambda$. Now, since $\Lambda_{\text{MF}}\subseteq\Lambda$, in order to show (ii), it suffices to verify that $w^*\in\Lambda_{\text{MF}}$. Specifically, we need to show that there exists a set of local allocations $(v^*_i)_{i\in[N]}$, where $v^*_i\in\Lambda_i$, for all $i\in[N]$, such that $w^*_a = \prod_{i\in[N]} v^*_{i,a_i}$, for all $a\in\mathcal{A}$. This is easily verified by the vector of marginal allocations $v^*_i = (1/K,\dots,1/K)$. □

A.3. Quantifying the approximation ratio

Lemma A.3. *For any θ, we have that $1\leq\frac{T^{MF}_\theta}{T^\star_\theta}\leq\tilde{A}/\sqrt{\rho}$.*

Proof. The lower bound $\frac{T^{\text{MF}}_\theta}{T^\star_\theta}\geq 1$ is obvious from Lemma 5.1. For the upper bound, we have

$$\frac{T^{\text{MF}}_\theta}{T^\star_\theta}\overset{(i)}{\leq}\frac{2\tilde{A}}{\Delta^2_{\min}}\frac{1}{T^\star_\theta}\overset{(ii)}{\leq}\frac{2\tilde{A}}{\Delta^2_{\min}}\frac{\sqrt{\rho}\Delta^2_{\min}}{2\rho} = \frac{\tilde{A}}{\sqrt{\rho}},$$

where (i) follows from Lemma 5.1, and (ii) follows directly from an application of Lemma 2 of (Soare et al., 2014). □

B. Properties of T_θ^{MF}

In this appendix, we provide important properties of T_θ^{MF}. We first establish, in App. B.1, that the optimization leading to T_θ^{MF} is equivalent to a non-linear convex program, which ensures that it can be computed efficiently. We then show, in App. B.2, the continuity of functions involved in the definition of T_θ^{MF}. These continuity arguments will be needed in the sample complexity analysis for our algorithm.

B.1. An equivalent convex program

We establish in Proposition B.1 below that the optimization problem defining the MF characteristic time T_θ^{MF} (2) is equivalent to a convex program. Note that this is a non-trivial result and does not hold in general. Indeed, it is known that in general, MF approximations lead to non-convex optimization problems (see e.g., (Wainwright & Jordan, 2008)) in App. D.

Proposition B.1. *The optimization problem* (2) *is equivalent to a (non-linear) convex program.*

The proof relies on the application of Lemma B.2 given below. Specifically, in Lemma B.2, we show that (2) can be reformulated as a GP in standard form. The proof of the Lemma relies on the fact that relaxing the constraints that are not in standard GP form does not affect optimality, while leading to a GP in standard form. Then, in Proposition B.1, we show that the GP (3) can be reformulated as a non-linear convex optimization program by a change of variables.

Lemma B.2. *The optimization problem* (2) *is equivalent to a geometric program.*

Proof. Recall that, for a positive variable $x \in \mathbb{R}^I_{>0}$, a GP in standard form is described as (Boyd et al., 2007):

$$\begin{array}{ll} \inf_x & f_0(x) \\ \text{subject to} & f_i(x) \leq 1, \quad i = 1, \dots, m \\ & h_i(x) = 1, \quad i = 1, \dots, p \end{array} \tag{16}$$

where $f_0, \dots, f_m$ are posynomials and $h_1, \dots, h_p$ are monomials. Recall that for $x \in \mathbb{R}^I_{>0}$, $\alpha \in \mathbb{R}^I$, and $\gamma > 0$, a *monomial* is a function of the type $h(x) = \gamma \prod_{i\in[I]} x_i^{\alpha_i}$ and a *posynomial* is a positive sum of monomials: $f(x) = \sum_{j\in[J]} \xi_j \prod_{i\in[I]} x_i^{\beta_{i,j}}$, for some $\beta \in \mathbb{R}^{I\times J}$ and $\xi \in \mathbb{R}^J_{>0}$.

By Lemma 5.1, and using an epigraph representation (Boyd & Vandenberghe, 2004), we can rewrite the optimization problem (2) describing T_θ^{MF}, for a positive variable $z > 0$ and $v_i \in \mathbb{R}^K_{>0}$, for all $i \in [N]$, as:

$$\inf_{z,(v_i)_{i\in[N]}} \quad z \tag{17a}$$

$$\text{subject to} \quad \frac{1}{\Delta_{\min}^2} \sum_{e\in[\rho]: a_e \neq a_e^\star} \left(\prod_{i\in\mathcal{S}_e} v_{i,a_i}^{-1} z^{-1} + \prod_{i\in\mathcal{S}_e} v_{i,a_i^\star}^{-1} z^{-1} \right) \leq 1, \forall a \neq a^\star \tag{17b}$$

$$\sum_{a_i\in\mathcal{A}_i} v_{i,a_i} = 1, \forall i \in [N]. \tag{17c}$$

Note that, at this stage, (17) cannot be described as a GP in standard form. Indeed, although the objective (17a) is a monomial and the set of inequality constraints (17b) are posynomials, the set of constraints (17c) are posynomial equality constraints, which do not comply with standard GP requirements. The problem (17) is in fact generally known as a *signomial program*, and is hard to solve in general (Boyd et al., 2007). The rest of the proof will be aimed at showing that (17) can be transformed into an equivalent GP.

In order to show that T_θ^{MF} can be actually casted as a GP, we can apply the method described in (Boyd et al., 2007) to relax the set of constraints (17c). Define the relaxed GP as:

$$\begin{aligned}\inf_{z,(v_i)_{i\in[N]}} \quad & z \\ \text{subject to} \quad & \frac{1}{\Delta_{\min}^2}\sum_{e\in[\rho]:a_e\neq a_e^\star}\left(\prod_{i\in\mathcal{S}_e} v_{i,a_i}^{-1}z^{-1}+\prod_{i\in\mathcal{S}_e} v_{i,a_i^\star}^{-1}z^{-1}\right)\leq 1, \forall a\neq a^\star \\ & \sum_{a_i\in\mathcal{A}_i} v_{i,a_i}\leq 1, \forall i\in[N].\end{aligned} \tag{18}$$

As described in (Boyd et al., 2007) (Sec. 7.4), we have that (18) and (17) are equivalent if and only if we have: for all $i\in[N]$, there exists a $k\in[N]$, such that:

(a) the variable v_{i,a_k} does not appear in any of the monomial equality constraints;

(b) the objective and the inequality constraint functions are all monotone decreasing in v_{i,a_k}, i.e., if we increase v_{i,a_k} (holding all other variables constant), the inequality constraint functions decrease, or remain constant;

(c) the posynomial function in the equality constraints are monotone strictly increasing in v_{i,a_k}, i.e., if we increase v_{i,a_k} (holding all other variables constant), the posynomial function increases.

In our setting, these assumptions hold naturally: (a) is satisfied since the original optimization problem does not involve any monomial equality constraints; (b) holds since the functions on the LHS of (17b) are monotone decreasing in $v_{i,a_k}, \forall a\in\mathcal{A}$; (c) holds since the function on the LHS of (17c) are monotone strictly increasing in $v_{i,a_k}, \forall i\in[N]$. □

Proof of Proposition B.1. By Lemma B.2, the optimization problem leading to T_θ^{MF} is equivalent to (18). Let $y_z=\log(z)$, $y_{v_{i,a_i}}=\log(v_{i,a_i})$, for all $i\in[N]$, $a_i\in\mathcal{A}_i$. By applying this change of variables, we can rewrite (18) as:

$$\begin{aligned}\inf_{y_z\in\mathbb{R}_{>0},(y_{v_i}\in\mathbb{R}^K_{>0})_{i\in[N]}} \quad & e^{y_z} \\ \text{subject to} \quad & \frac{1}{\Delta_{\min}^2}\sum_{e:a_e\neq a_e^\star}\exp\left(-\sum_{i\in\mathcal{S}_e} y_{v_{i,a_i}}y_z\right)+\exp\left(-\sum_{i\in\mathcal{S}_e} y_{v_{i,a_i^\star}}y_z\right)\leq 1, \forall a\neq a^\star, \\ & \sum_{a_i\in\mathcal{A}_i} e^{y_{v_{i,a_i}}}\leq 1, \forall i\in[N].\end{aligned} \tag{19}$$

By the monotonicity of the logarithm function, we have that (19) is also equivalent to:

$$\begin{aligned}\inf_{y_z\in\mathbb{R}_{>0},(y_{v_i}\in\mathbb{R}^K_{>0})_{i\in[N]}} \quad & y_z \\ \text{subject to} \quad & \log\left(\frac{1}{\Delta_{\min}^2}\sum_{e:a_e\neq a_e^\star}\exp\left(-\sum_{i\in\mathcal{S}_e} y_{v_{i,a_i}}y_z\right)+\exp\left(-\sum_{i\in\mathcal{S}_e} y_{v_{i,a_i^\star}}y_z\right)\right)\leq 0, \forall a\neq a^\star, \\ & \log\left(\sum_{a_i\in\mathcal{A}_i}\exp\left(y_{i,w_{a_i}}\right)\right)\leq 0, \forall i\in[N].\end{aligned} \tag{20}$$

It follows directly from the convexity of the Log-Sum-Exp function that (20) is a convex program. □

B.2. Continuity arguments

This section presents continuity arguments on functions related to the optimization problem (2). Define, for $\theta \in \mathcal{M}$ and $\tilde{w} \in \tilde{\Lambda}_{\text{MF}}$, the function

$$\psi(\theta, \tilde{w}) = \min_{a \neq a^\star} \frac{\Delta_{\min}^2}{\sum_{e \in [\rho]: a_e \neq a_e^\star} \left(\tilde{w}_{e,a_e}^{-1} + \tilde{w}_{e,a_e^\star}^{-1} \right)}.$$

Further define $\psi^\star(\theta) = \sup_{\tilde{w} \in \tilde{\Lambda}_{\text{MF}}} \psi(\theta, \tilde{w})$, and $\tilde{w}^\star = \arg\max_{\tilde{w}} \psi(\theta, \tilde{w})$. In the following, we shall assume w.l.o.g. that $\tilde{w}^\star$ is unique, i.e., the set of optimal allocations is a singleton: $C^\star(\theta) = \{\tilde{w} \in \tilde{\Lambda}_{\text{MF}} : \psi(\theta, \tilde{w}) = (T_\theta^{\text{MF}})^{-1}\} = \{\tilde{w}^\star\}$. This may be proved using similar techniques as those of Theorem 5 in (Garivier & Kaufmann, 2016). Note that if this was not the case, one may reason in terms of the objective function as, e.g., in (Jedra & Proutiere, 2020) for linear bandits.

Lemma B.3. *The function $\psi(\theta, \tilde{w})$ is continuous in both $\tilde{w}$ and θ, and $\psi^\star(\theta)$ is continuous in θ. Furthermore there exists a $\tilde{w}^\star \in \tilde{\Lambda}_{MF}$ such that $\tilde{w}^\star \in \arg\max_{\tilde{w} \in \tilde{\Lambda}} \psi(\theta, \tilde{w})$.*

Proof. The proof of Lemma B.3 follows from standard continuity arguments and is similar to that of the related results in (Jedra & Proutiere, 2020).

First, we prove that $\psi(\theta, \tilde{w})$ is continuous in both θ and $\tilde{w}$. Let $(\theta_t, \tilde{w}_t)_{t \geq 1}$ be a sequence taking values in $\mathcal{M} \times \tilde{\Lambda}_{\text{MF}}$, and converging to $(\theta, \tilde{w})$. Recall the definition of the set of confusing parameters with respect to $\theta \in \mathcal{M}$,

$$B(\theta) = \{\mu \in \mathcal{M} : \exists a \neq a_\theta^\star : \mu(a) > \mu(a_\theta^\star)\}.$$

and define

$$f(\theta, \mu, \tilde{w}) = \sum_{e \in [\rho]} \sum_{a_e \in \mathcal{A}_e} \tilde{w}_{e,a_e} \frac{\left(\theta_e(a_e) - \mu_e(a_e)\right)^2}{2}.$$

Since $(\theta_t, \tilde{w}_t)$ converges to $(\theta, \tilde{w})$, and $a_\theta^\star$ is unique, there exists $\varepsilon > 0$ and a $t_1 \geq 1$ such that $\forall t \geq t_1$, $\|(\theta, \tilde{w}^\star) - (\theta_t, \tilde{w}_t)\| < \varepsilon$ and such that $B(\theta) = B(\theta_t)$. Further note that, since $f(\theta, \mu, \tilde{w})$ is a polynomial in $\theta, \mu, \tilde{w}$, is continuous. As a consequence, there exists $t_2 \geq 1 : \forall t \geq t_2$, such that for all $\mu \in \mathcal{M}$, we have $|f(\theta_t, \mu, \tilde{w}_t) - f(\theta, \mu, \tilde{w})| \leq \varepsilon f(\theta, \mu, \tilde{w})$. Thus, for all $t \geq \max\{t_1, t_2\}$, we get

$$\begin{aligned} |\psi(\theta, \tilde{w}) - \psi(\theta_t, \tilde{w}_t)| &= \left| \min_{\mu \in B(\theta)} f(\theta, \mu, \tilde{w}) - \min_{\mu \in B(\theta)} f(\theta_t, \mu, \tilde{w}_t) \right| \\ &\leq \varepsilon \, |f(\theta, \mu, \tilde{w})| \leq \varepsilon |\psi(\theta, \tilde{w})|. \end{aligned}$$

The continuity of $\psi^\star(\theta)$ and existence of a solution $\tilde{w}^\star$ follows from Berge's maximum theorem (Berge, 1963). □

C. Sampling Rule Analysis

In this appendix, we present various results on the sampling rule of MF-TaS. We divide the analysis for the forced exploration (in App. C.1) and tracking (in App. C.2). Recall that, the forced exploration relies on the set $\mathcal{A}_0$, as described in §6.2. This set may be difficult to construct due to the combinatorial nature of the action set $\mathcal{A}$. To address this issue, we provide, in App. C.3, a simple and efficient algorithm to build $\mathcal{A}_0$.

C.1. Forced Exploration

In this section, we state and prove Lemma C.1. It shows that each group arm is sampled sufficiently often. This, in turn, ensures that the estimators of the group means $\hat{\theta}_{t,e}$ converge to θ_e, for all groups $e \in [\rho]$, and hence that $\hat{\theta}_t \to \theta$ a.s..

Lemma C.1. *The sampling rule of MF-TaS ensures that exists a finite $t' > 0$, such that $\forall t \geq t'$ and $\forall e \in [\rho], a_e \in \mathcal{A}_e$, we have that*

$$N_{t,e,a_e} \geq \sqrt{\frac{t - |\mathcal{A}_0| - 1}{|\mathcal{A}_0|}}.$$

Proof. Recall the expression for the set of under-explored actions at group e and time t:

$$U_{t,e} = \left\{ a_e \in \mathcal{A}_e : N_{t,e,a_e} < \sqrt{\frac{t}{|\mathcal{A}_0|}} \right\}.$$

The proof is inspired by that of Lemma 5 in (Jedra & Proutiere, 2020). The main idea is to show that, if at some time $t_0 + 1$ the condition $\exists e \in [\rho], U_{t_0+1,e} \neq \emptyset$ is violated, then the number of rounds needed to satisfy the condition again cannot exceed $|\mathcal{A}_0|$ rounds. This follows by the definition of $\mathcal{A}_0$ and of the forced exploration rule.

By construction, we have that

$$\inf\{t \geq 1 : \forall e \in [\rho], U_{t,e} = \emptyset\} \triangleq T_0 \leq |\mathcal{A}_0|.$$

Now, if there exists $t_0 \geq T_0$ such that $\forall e \in [\rho], U_{t_0,e} = \emptyset$ and $\exists e \in [\rho]$ such that $U_{t_0+1,e} \neq \emptyset$, we may define

$$t_1 = \inf\{t > t_0 : \forall e \in [\rho], U_{t,e} = \emptyset\}.$$

Observe that for all $t_0 \leq t \leq t_1$, and for all $e \in [\rho]$, we have,

$$N_{t,e,a_e} \geq N_{t_0,e,a_e} \geq \sqrt{t_0/|\mathcal{A}_0|}.$$

Furthermore, if $t_1 \geq t_0 + |\mathcal{A}_0| + 1$ then we have, $\forall e \in [\rho]$,

$$N_{t_1,e,a_e} \geq N_{t_0+|\mathcal{A}_0|+1,e,a_e} \geq N_{t_0,e,a_e} + 1 \geq \sqrt{t_0/|\mathcal{A}_0|} + 1.$$

However, we have:

$$t_0 \geq \frac{1}{4|\mathcal{A}_0|} \Rightarrow \sqrt{\frac{t_0}{|\mathcal{A}_0|}} + 1 \geq \sqrt{\frac{t_0 + |\mathcal{A}_0| + 1}{|\mathcal{A}_0|}}.$$

Therefore if $t_0 \geq \frac{1}{4|\mathcal{A}_0|}$ then $t_1 \leq t_0 + |\mathcal{A}_0| + 1$. In other words, we have shown that for all $t \geq \frac{1}{4|\mathcal{A}_0|} + |\mathcal{A}_0| + 1$, we have for all $e \in [\rho]$, for all $a_e \in \mathcal{A}_e$

$$N_{t,e,a_e} \geq \sqrt{\frac{t - |\mathcal{A}_0| - 1}{|\mathcal{A}_0|}}.$$

□

C.2. Tracking

In this section we state and prove Lemma C.2 and Lemma C.3. Lemma C.2 shows that MF-TaS can correctly track a changing sequence that concentrates. Lemma C.3 (and Corollary C.4) shows that tracking local allocations ensures that global (and group) allocations are correctly tracked. Finally, in Proposition C.5, we show that MF-TaS ensures that the empirical allocations converge to the optimal allocation.

Lemma C.2. *Define, for all $i \in [N]$, the sequence $(v_{t,i})_{t\geq 1}$, where $v_{t,i} = (v_{t,i,a_i})_{a_i \in \mathcal{A}_i}$, taking values in Λ_i used in the tracking rule of MF-TaS. Let, for all $i \in [N]$, $v_i^\star \in \Lambda_i$. Then, for all $\varepsilon > 0$, and for all t_0, there exists $t_\varepsilon > t_0$ such that*

$$\sup_{t\geq t_0} \max_{i\in[N],a_i\in\mathcal{A}_i} \left|v_{t,i,a_i} - v^\star_{i,a_i}\right| \leq \varepsilon \Rightarrow \sup_{t\geq t_\varepsilon} \max_{i\in[N],a_i\in\mathcal{A}_i} \left|\frac{N_{t,i,a_i}}{t} - v^\star_{i,a_i}\right| \leq 3(K-1)\varepsilon.$$

Proof. The proof is quite similar to the one of Lemma 7 in (Garivier & Kaufmann, 2016) for D-tracking. We adapt the proof to allow our sampling rule to track local allocations for each agent, as opposed to tracking arm allocations as in (Garivier & Kaufmann, 2016). Also, there is an asymmetry in the forced exploration that slightly complicates the analysis.

For all $i \in [N], a_i \in \mathcal{A}_i$, define $X_{t,i,a_i} = N_{t,i,a_i} - tv^\star_{i,a_i}$, and note that $\forall i \in [N]$, we have:

$$\sum_{a_i\in\mathcal{A}_i} X_{t,i,a_i} = \sum_{a_i\in\mathcal{A}_i} (N_{t,i,a_i} - tv^\star_{i,a_i}) = t - t\left(\sum_{a_i\in\mathcal{A}_i} v^\star_{i,a_i}\right) = 0. \tag{21}$$

Furthermore, we have that $\max_{i\in[N],a_i\in\mathcal{A}_i} |X_{t,i,a_i}| \leq N(K-1)\max_{i\in[N],a_i\in\mathcal{A}_i} X_{t,i,a_i}$. To see this, note that for every $i \in [N], a_i \in \mathcal{A}_i$, we have that $X_{t,i,a_i} \leq \max_{k\in[N]} \max_{a_k\in\mathcal{A}_k} X_{t,k,a_k}$, and that, by (21), we have

$$X_{t,i,a_i} = -\sum_{j\neq i} X_{t,j,a_i} \geq -\sum_{j\neq i} \max_{k\in[N]} X_{t,k,a_i} = -(K-1)\max_{k\in[N]} X_{t,k,a_i}.$$

The remaining part of the proof will aim at determining an upper bound on $\max_{i\in[N],a_i\in\mathcal{A}_i} X_{t,i,a_i}$, for t large enough. Now, let $t_0' \geq t_0$ be such that

$$\forall t \geq t_0', \quad \sqrt{t/|\mathcal{A}_0|} \leq 2t\varepsilon \quad \text{and} \quad 1/t \leq \varepsilon.$$

We will show that for $t \geq t_0'$, and for all $i \in [N]$,

$$\{a_{t+1,i} = a_i\} \subseteq \{X_{t,i,a_i} \leq 2t\varepsilon\}. \tag{22}$$

We analyze two (mutually exclusive) cases: (i) forced exploration and (ii) tracking.

(i) if at $t \geq t_0'$, $\exists e \in [\rho] : N_{t,e,a_e} \leq \sqrt{t}/|\mathcal{A}_0|$, MF-TaS is in a forced exploration step. This, in turn, implies that $\exists i \in \mathcal{S}_e$ such that $N_{t,i,a_i} \leq \sqrt{t}/|\mathcal{A}_0|$. Hence we have that:

$$X_{t,i,a_i} \leq \sqrt{t}/|\mathcal{A}_0| - tv^\star_{i,a_i} \leq \sqrt{t}/|\mathcal{A}_0| \leq 2t\varepsilon,$$

where the last inequality follows by definition of t_0'.

(ii) If MF-TaS is in a tracking step at $t \geq t_0'$, for all $i \in [N], a_i \in \mathcal{A}_i$, we have,

$$\begin{aligned} N_{t,i,a_i} - tw_{t,i,a_i} &= \min_{b_i\in\mathcal{A}_i} (N_{t,i,b_i} - tw_{t,i,b_i}) \\ X_{t,i,a_i} + (w^\star_{i,a_i} - w_{t,i,a_i}) &= \min_{b_i\in\mathcal{A}_i} \left(X_{t,i,b_i} + t(w^\star_{i,b_i} - w_{t,i,b_i})\right). \end{aligned}$$

Now, for all $t \geq t_0$, we have that $X_{t,i,a_i} + (w^\star_{i,a_i} - v_{t,i,a_i}) \leq \min_{b_i\in\mathcal{A}_i}(X_{t,i,b_i} + t\varepsilon) \leq t\varepsilon$, where we used that, $\forall i \in [N]$, $\min_{b_i\in\mathcal{A}_i} X_{t,i,b_i} \leq 0$ by (21). Now, by assumption, we have that for all $t \geq t_0' \geq t_0$, $|v^\star_{i,a_i} - v_{t,i,a_i}| \leq \varepsilon$, and we can conclude that (22) holds.

Note that, for all $i \in [N]$, we have $X_{t+1,i,a_i} = X_{t,i,a_i} + \mathbb{1}_{\{a_{t+1,i}=a_i\}} - v^\star_{i,a_i}$. Hence, for $t \geq t_0'$, we have

$$X_{t+1,i,a_i} \leq X_{t,i,a_i} + \mathbb{1}_{\{X_{t,i,a_i}\leq 2t\varepsilon\}} - v^\star_{i,a_i}.$$

We now prove by induction that for all $t \geq t_0'$, we have

$$X_{t,i,a_i} \leq \max\{X_{t_0',i,a_i}, 2t\varepsilon + 1\}. \tag{23}$$

The base case $t = t_0'$ automatically holds. Now, assume that, at $t \geq t_0'$, (23) holds. If $X_{t,i,a_i} \leq 2t\varepsilon$, we have

$$X_{t+1,i,a_i} \leq 2t\varepsilon + 1 - v^\star_{i,a_i} \leq 2t\varepsilon + 1 \leq \max\{X_{t_0',i,a_i}, 2t\varepsilon + 1\} \leq \max\{X_{t_0',i,a_i}, 2(t+1)\varepsilon + 1\}.$$

On the other hand, if $X_{t,i,a_i} > 2t\varepsilon$, we have $X_{t+1,i,a_i} \leq \max\{X_{t_0',i,a_i}, 2t\varepsilon + 1\} - v^\star_{i,a_i} \leq \max\{X_{t_0',i,a_i}, 2(t+1)\varepsilon + 1\}$. This concludes the induction step. Now, for $t \geq t_0'$, using that $X_{t_0',i,a_i} \leq t_0'$ and $1/t \leq \varepsilon$, we have

$$\max_{i\in[N], a_i \in \mathcal{A}_i} \left| \frac{X_{t,i,a_i}}{t} \right| \leq (K-1)\max\{2\varepsilon + 1/t, t_0'/t\} \leq (K-1)\max\{3\varepsilon, t_0'/t\}.$$

Hence, we can conclude that exists $t_1(\varepsilon) \geq t_0'$ such that for all $t \geq t_1(\varepsilon)$,

$$\max_{i\in[N], a_i \in \mathcal{A}_i} \left| \frac{X_{t,i,a_i}}{t} \right| \leq 3(K-1)\varepsilon.$$

□

Lemma C.3. *Let $w^{(1)}, w^{(2)} \in \Lambda_{MF}$, and let $v^{(1)}, v^{(2)}$ be the corresponding local allocations. For all $\varepsilon \geq 0$, we have that*

$$\max_{i\in[N], a_i \in \mathcal{A}_i} |v^{(1)}_{i,a_i} - v^{(2)}_{i,a_i}| \leq \varepsilon \iff \max_{a\in\mathcal{A}} |w^{(1)}_a - w^{(2)}_a| \leq N\varepsilon.$$

Proof. We have that

$$\varepsilon \geq \max_{i\in[N], a_i \in \mathcal{A}_i} |v^{(1)}_{i,a_i} - v^{(2)}_{i,a_i}| \geq \frac{1}{N}\sum_{i\in[N]} \max_{a_i \in \mathcal{A}_i} |v^{(1)}_{i,a_i} - v^{(2)}_{i,a_i}| \geq \frac{1}{N}\max_{a\in\mathcal{A}} \sum_{i\in[N]} |v^{(1)}_{i,a_i} - v^{(2)}_{i,a_i}|, \tag{24}$$

where the first inequality follows by hypothesis and the second and third by inequalities on the max. Next, we show by induction that, $\forall a \in \mathcal{A}$,

$$\sum_{i\in[N]} |v^{(1)}_{i,a_i} - v^{(2)}_{i,a_i}| \geq \left| \prod_{i\in[N]} v^{(1)}_{i,a_i} - \prod_{i\in[N]} v^{(2)}_{i,a_i} \right|.$$

Let $n \in [N]$, and define $W^{(1)}_n = \prod_{i\in[n]} v^{(1)}_{i,a_i}$ and $W^{(2)}_n = \prod_{i\in[n]} v^{(2)}_{i,a_i}$. The base case $n = 1$ obviously holds true since. Suppose that $|W^{(1)}_n - W^{(2)}_n| \leq \sum_{i\in[n]} |v^{(1)}_{i,a_i} - v^{(2)}_{i,a_i}|$, for any $n \in [N-1]$. Then we have:

$$|W^{(1)}_{n+1} - W^{(2)}_{n+1}| \leq |v^{(1)}_{n+1,a_{n+1}} - v^{(2)}_{n+1,a_{n+1}}||W^{(1)}_n| + |W^{(1)}_n - W^{(2)}_n||v^{(2)}_{n+1,a_{n+1}}| \leq \sum_{i\in[n+1]} |v^{(1)}_{i,a_i} - v^{(2)}_{i,a_i}|,$$

where the inequalities follow by applying a triangular inequality and by the fact that $|v_{i,a_i}| \leq 1, \forall i \in [N], a_i \in \mathcal{A}_i$. This concludes the induction step. Hence, by (24), we get:

$$\varepsilon \geq \frac{1}{N}\max_{a\in\mathcal{A}} \sum_{i\in[N]} |v^{(1)}_{i,a_i} - v^{(2)}_{i,a_i}| \geq \frac{1}{N}\max_{a\in\mathcal{A}} \left| \prod_{i\in[N]} v^{(1)}_{i,a_i} - \prod_{i\in[N]} v^{(2)}_{i,a_i} \right| = \frac{1}{N}\max_{a\in\mathcal{A}} |w^{(1)}_a - w^{(2)}_a|.$$

□

By following a similar approach, we can prove the following corollary.

Corollary C.4. *Let $\tilde{w}^{(1)}, \tilde{w}^{(2)} \in \tilde{\Lambda}_{MF}$ and let $v^{(1)}, v^{(2)}$ be the corresponding local allocations. For all $\varepsilon > 0$, we have that*

$$\max_{i\in[N], a_i \in \mathcal{A}_i} |v^{(1)}_{i,a_i} - v^{(2)}_{i,a_i}| \leq \varepsilon \iff \max_{e\in[\rho], a_e \in \mathcal{A}_e} |\tilde{w}^{(1)}_{e,a_e} - \tilde{w}^{(2)}_{e,a_e}| \leq N\varepsilon.$$

Proposition C.5. *Let $\tilde{w}^\star = \arg\max_{\tilde{w}\in\tilde{\Lambda}_{MF}} \psi(\theta, \tilde{w})$, and let $(v_i^\star)_{i\in[N]}$ be the corresponding optimal local allocations. The MF-TaS sampling rule satisfies $\forall i \in [N], a_i \in \mathcal{A}_i$*

$$\mathbb{P}\left(\lim_{t\to\infty} \frac{N_{t,i,a_i}}{t} = v^\star_{i,a_i}\right) = 1.$$

Proof. Let $\mathcal{E} = \left\{\hat{\theta}_t \overset{t\to\infty}{\longrightarrow} \theta\right\}$, and note that $\mathbb{P}(\mathcal{E}) = 1$. In fact, by Lemma C.1, there exists a finite $t_0 \geq 1$ such that for all $t \geq t_0$, we have $\min_{e\in[\rho], a_e\in\mathcal{A}_e} N_{t,e,a_e} \geq \sqrt{\frac{t-|\mathcal{A}_0|-1|}{|\mathcal{A}_0|}}$. Hence, by the law of large numbers (each group action will be played infinite times), we have that $\hat{\theta}_{t,e} \overset{t\to\infty}{\longrightarrow} \theta_e$ a.s., and hence $\hat{\theta}_t \overset{t\to\infty}{\longrightarrow} \theta$ a.s..

Note that under the event $\mathcal{E}$, by continuity of $\tilde{w}^\star$ (Lemma B.3), we have that there exists $t_0(\varepsilon) \geq 1$ such that:

$$\sup_{t\geq t_0(\varepsilon)} \max_{e\in[\rho], a_e\in\mathcal{A}_e} |\tilde{w}^\star_{e,a_e} - \tilde{w}_{t,e,a_e}| \leq \frac{N\varepsilon}{3(K-1)}.$$

Furthermore, by Corollary C.4, we have that $\sup_{t\geq t_0(\varepsilon)} \max_{i\in[N], a_i\in\mathcal{A}_i} |v^\star_{i,a_i} - v_{t,i,a_i}| \leq \frac{\varepsilon}{3(K-1)}$. Hence, using Lemma C.2, there exists $t_\varepsilon \geq t_0(\varepsilon)$ such that for all $t \geq t_\varepsilon$,

$$\sup_{t\geq t_\varepsilon} \max_{i\in[N], a_i\in\mathcal{A}_i} \left|\frac{N_{t,i,a_i}}{t} - v^\star_{i,a_i}\right| \leq \varepsilon.$$

□

C.3. An algorithm for selecting $\mathcal{A}_0$

We present in Alg. 4, the pseudocode of a simple procedure for selecting $\mathcal{A}_0$. It takes as input the set of global actions $\mathcal{A}$ and the set of group actions $\mathcal{A}_e$ for all $e \in [\rho]$. Let $I_{e,a_e} = \sum_{b\in\mathcal{A}_0} \mathbb{1}_{\{a_e=b_e\}}$ be the counter of group actions $a_e \in \mathcal{A}_e$ in $\mathcal{A}_0$, and define $I_e = [I_{e,a_e}]_{a_e\in\mathcal{A}_e} \in \mathbb{N}^{A_e}$. To describe the algorithm, we assume that $\mathcal{A}$ is an ordered set, and denote by $\mathcal{A}(i)$ is the i-th global action. First, the algorithm initializes $\mathcal{A}_0 \leftarrow \emptyset$, and $I_e = 0 \in \mathbb{N}^{A_e}, \forall e \in [\rho]$. Then, the algorithm iterates over groups $e \in [\rho]$ and groups' actions $b_e \in \mathcal{A}_e$, and iteratively includes arms in $\mathcal{A}$ into the set $\mathcal{A}_0$ which are never observed in previous iterates. By construction, Alg. 4, ensures that every group arm is observed at least once in $\mathcal{A}_0$.

Algorithm 4 BUILD $\mathcal{A}_0$

Input: Global actions $\mathcal{A}$, group actions $(\mathcal{A}_e)_{e\in[\rho]}$
Initialize: $\mathcal{A}_0 \leftarrow \emptyset$, $I_e = 0 \in \mathbb{N}^{A_e}, \forall e \in [\rho]$,
for $e \in [\rho]$ **do**
  $i \leftarrow 1$
  while $\min_{a_e\in\mathcal{A}_e} I_{e,a_e} = 0$ **do**
    $a \leftarrow \mathcal{A}(i)$
    for $b_e \in \mathcal{A}_e$ **do**
      if $b_e = a_e$ and $I_{e,b_e} = 0$ **then**
        $\mathcal{A}_0 \leftarrow \mathcal{A}_0 \cup \{a\}$
        $I_{e,b_e} \leftarrow I_{e,b_e} + 1$
      end if
    end for
    $i \leftarrow i + 1$
  end while
end for
Return $\mathcal{A}_0$

Note that Alg. 4 may be improved with more precise search strategies, at the cost of increased computational complexity. For example, the algorithm could include in $\mathcal{A}_0$, at each step, global arms $a \in \mathcal{A}$ which maximizes the number of non-observed group actions corresponding to the global actions in $\mathcal{A}_0$.

Example 1 ($\mathcal{A}_0$ action choice). Consider the example in Fig. 1 with $K = 2$ local actions and $N = 4$ agents (i.e., $A = 16$ actions). In such setting, we can select $\mathcal{A}_0 = \{a_{0000}, a_{0110}, a_{1001}, a_{1111}\}$. Running Alg. 4 on this instance instead produces the set $\mathcal{A}_0 = \{a_{0000}, a_{0001}, a_{0110}, a_{0111}, a_{1000}, a_{1010}, a_{1100}\}$.

D. Asymptotic Sample Complexity Upper Bound

In this section, we present the proof of Theorem 6.1. The proofs of the results are quite similar to those of the related results in (Garivier & Kaufmann, 2016). We divide the proof for the guarantee in probability (a.s.) and in expectation. In the remainder of this section, we will make use of the following technical lemma.

Lemma D.1 (Lemma 18 in (Garivier & Kaufmann, 2016)). *For every $\alpha \in [1, e/2]$, and for any two constants $c_1, c_2 > 0$,*

$$x = \frac{1}{c_1}\left[\log\left(\frac{c_2 e}{c_1^\alpha}\right) + \log\log\left(\frac{c_2}{c_1^\alpha}\right)\right]$$

is such that $c_1 x \geq \log(c_2 x^\alpha)$.

D.1. Almost Sure Upper Bound

Proof. Define the event

$$\mathcal{E} = \left\{\forall i \in [N], a_i \in \mathcal{A}_i,\ \frac{N_{t,i,a_i}}{t} \overset{t\to\infty}{\longrightarrow} \tilde{w}^\star_{i,a_i},\ \hat{\theta}_t \overset{t\to\infty}{\longrightarrow} \theta\right\}.$$

Observe that $\mathbb{P}(\mathcal{E}) = 1$. This follows directly from Lemma C.1 and Proposition C.5. Note that, under $\mathcal{E}$, we also have that $\forall e \in [\rho]$, $a_e \in \mathcal{A}_e$, $\frac{N_{t,e,a_e}}{t} \overset{t\to\infty}{\longrightarrow} \tilde{w}^\star_{e,a_e}$. Let $\tilde{w}_t = (N_{t,e,a_e}/t)_{e\in[\rho],a_e\in\mathcal{A}_e}$. As described in §6.3, the conditions that the stopping threshold $\beta(\delta, t)$ must satisfy are the same as in Sec. 3.2 of (Wang et al., 2021):

$$\forall t \geq 1, t\psi(\tilde{w}_t, \hat{\theta}_t) \geq \beta(\delta, t) \Rightarrow \mathbb{P}[a^\star_\theta \neq \hat{a}_t] \leq \delta, \tag{25a}$$

$$\exists c_1, c_2 > 0 : \forall t \geq c_1, \beta(\delta, t) \leq \log(c_2 t/\delta). \tag{25b}$$

Note that a threshold satisfying these properties exists (see e.g., (Kaufmann & Koolen, 2021)). In the following, we assume these conditions hold. Let $\varepsilon > 0$. Under the event $\mathcal{E}$, by continuity of ψ (Lemma B.3), we have that there exists t_1 such that for all $t \geq t_1$, $\psi\big(\hat{\theta}_t, \tilde{w}_t\big) \geq (1-\varepsilon)\psi\left(\theta, \tilde{w}^\star\right)$. This implies that for all $t \geq t_0$ we have:

$$\begin{aligned}
\tau &= \inf\left\{t \in \mathbb{N}_{>0} : t\psi(\hat{\theta}_t, \tilde{w}_t) > \beta(\delta, t)\right\} \\
&\leq t_1 \vee \inf\left\{t \in \mathbb{N}_{>0} : t(1-\varepsilon)\psi\left(\theta, \tilde{w}^\star\right) > \beta(\delta, t)\right\} \\
&\leq t_1 \vee \inf\left\{t \in \mathbb{N}_{>0} : t \geq \frac{T^{\text{MF}}_\theta \beta(\delta, t)}{1-\varepsilon}\right\} \\
&\leq c_1 \vee t_1 \vee \inf\left\{t \in \mathbb{N}_{>0} : t \geq \frac{\log(c_2 t) T^{\text{MF}}_\theta}{(1-\varepsilon)\delta}\right\},
\end{aligned}$$

where we used the fact that $\psi(\theta, \tilde{w}^\star) = (T^{\text{MF}}_\theta)^{-1}$ and the assumption on the stopping threshold (25b).

Now, by applying Lemma D.1, for $\alpha = 1$, $c_1 = \frac{(1-\varepsilon)\delta}{T^{\text{MF}}_\theta}$ we get:

$$\tau \leq c_1 + t_1 + \frac{T^{\text{MF}}_\theta}{(1-\varepsilon)\delta}\left[\log\left(\frac{T^{\text{MF}}_\theta c_2 e}{(1-\varepsilon)\delta}\right) + \log\log\left(\frac{T^{\text{MF}}_\theta c_2}{(1-\varepsilon)\delta}\right)\right].$$

Hence, for all $\delta \in (0, 1)$, we have that:

$$\mathbb{P}_\theta\left(\limsup_{\delta\to 0} \frac{\tau}{\log(1/\delta)} \leq T^{\text{MF}}_\theta\right) = 1.$$

□

D.2. Expected Upper Bound

Proof. Let $\varepsilon > 0$. By Lemma B.3, we have that $\tilde{w}^\star$ is continuous in θ, and hence there exists $\xi(\varepsilon, \theta)$ such that

$$\mathcal{I}_\varepsilon = \times_{a\in\mathcal{A}}[\theta(a) - \xi, \theta(a) + \xi],$$

is such that for all $\theta' \in \mathcal{I}_\varepsilon$, we have

$$\max_{e\in[\rho], a_e\in\mathcal{A}_e} |\tilde{w}^\star_{e,a_e}(\theta') - \tilde{w}^\star_{e,a_e}(\theta)| \leq \varepsilon.$$

Note that, when $\hat{\theta}_t \in \mathcal{I}_\varepsilon$, we have $\hat{a}_t = a^\star_\theta$. Define, for $T \in \mathbb{N}$, $h(T) = T^{1/4}$, and let the "good event" be defined as:

$$\mathcal{E}_T(\varepsilon) = \bigcap_{t=h(T)}^{T} \left\{\hat{\theta}_t \in \mathcal{I}_\varepsilon\right\}.$$

We will now show that:

(i) $\mathbb{P}[\mathcal{E}_T^c] \leq BT\exp(-CT^{1/8})$, for some constants B, C (which depend on θ and ε).

(ii) There exists a T_ε such that for all $T \geq T_\varepsilon$, it holds, under the event $\mathcal{E}_T$, that

$$\forall t \geq \sqrt{T}, \max_{i\in[N], a_i\in\mathcal{A}_i} \left|\frac{N_{t,i,a_i}}{t} - w^\star_{i,a_i}\right| \leq 3(K-1)\varepsilon.$$

We shall prove (i) first. By a union bound on $\mathcal{E}_T^c$, we have

$$\mathbb{P}\left(\mathcal{E}_T^c\right) \leq \sum_{t=h(T)}^{T} \mathbb{P}\left(\hat{\theta}_t \notin \mathcal{I}_\varepsilon\right) = \sum_{t=h(T)}^{T}\sum_{a\in\mathcal{A}} \left[\mathbb{P}\left(\hat{\theta}_{t,a} \leq \theta(a) - \xi\right) + \mathbb{P}\left(\hat{\theta}_{t,a} \geq \theta_a + \xi\right)\right].$$

Now, let T be such that $h(T) \geq \frac{|\mathcal{A}_0|(|\mathcal{A}_0|+1)}{(|\mathcal{A}_0|-1)}$. Then, for $t \geq h(T)$, we have $\forall e \in [\rho], a_e \in \mathcal{A}_e$, $N_{t,e,a_e} \geq \sqrt{t}/|\mathcal{A}_0|$ by Lemma C.1. Applying a union bound and Chernoff's inequality, we get:

$$\begin{aligned}
\mathbb{P}\left(\hat{\theta}_{t,a} \leq \theta(a) - \xi\right) &= \mathbb{P}\left(\sum_{e\in[\rho]} \hat{\theta}_{t,e,a_e} \leq \sum_{e\in[\rho]} \theta_e(a_e) - \xi\right) \\
&\leq \sum_{e\in[\rho]} \mathbb{P}\left(\hat{\theta}_{t,e,a_e} - \theta_{e,a_e} \leq -\xi/\rho\right) \\
&= \sum_{e\in[\rho]} \mathbb{P}\left(\hat{\theta}_{t,e,a_e} \leq \theta_e(a_e) - \xi/\rho, N_{t,e,a_e} \geq \sqrt{t}/|\mathcal{A}_0|\right) \\
&\leq \sum_{e\in[\rho]} \sum_{s=\sqrt{t}/|\mathcal{A}_0|}^{t} \mathbb{P}\left(\hat{\theta}_{s,e,a_e} - \theta_e(a_e) \leq \xi/\rho\right) \\
&\leq \sum_{e\in[\rho]} \sum_{s=\sqrt{t}/|\mathcal{A}_0|}^{t} \exp\left(-s\mathrm{kl}(\hat{\theta}_{s,e,a_e} - \theta_e(a_e) \leq \xi/\rho\right) \\
&\leq \sum_{e\in[\rho]} \frac{\exp\left(-\sqrt{t}/|\mathcal{A}_0|\mathrm{kl}(\theta_e(a_e) - \xi/\rho, \theta_e(a_e))\right)}{1 - \exp\left(-\mathrm{kl}(\theta_e(a_e) - \xi/\rho, \theta_e(a_e))\right)}.
\end{aligned}$$

Similarly, we can prove that

$$\mathbb{P}\left(\hat{\theta}_{t,a} \geq \theta(a) + \xi\right) \leq \sum_{e\in[\rho]} \frac{\exp\left(-\sqrt{t}/|\mathcal{A}_0|\mathrm{kl}(\theta_e(a_e) + \xi/\rho, \theta_e(a_e))\right)}{1 - \exp\left(-\mathrm{kl}(\theta_e(a_e) + \xi/\rho, \theta_e(a_e))\right)}.$$

Hence, by letting

$$B = \sum_a \sum_e \frac{\exp\left(-\mathrm{kl}(\theta_e(a_e) + \xi/\rho, \theta_e(a_e))/|\mathcal{A}_0|\right)}{1 - \exp\left(-\mathrm{kl}(\theta_e(a_e) + \xi/\rho, \theta_e(a_e))\right)} + \frac{\exp\left(-\mathrm{kl}(\theta_e(a_e) - \xi/\rho, \theta_e(a_e))/|\mathcal{A}_0|\right)}{1 - \exp\left(-\mathrm{kl}(\theta_e(a_e) - \xi/\rho, \theta_e(a_e))\right)},$$

$$C = \min_a \left\{ \sum_e \mathrm{kl}(\theta_e(a_e) + \xi/, \theta_e(a_e)) \wedge \sum_e \mathrm{kl}(\theta_e(a_e) + \xi/, \theta_e(a_e)) \right\},$$

we have $\mathbb{P}(\mathcal{E}_T^c) \leq \sum_{h(T)}^T B \exp(-C\sqrt{t}) \leq BT \exp(-C\sqrt{h(T)}) \leq BT \exp(-CT^{1/8})$.

Note that (ii) follows directly from the definition of $\mathcal{I}_\varepsilon$ and Lemma C.1. For $T \geq T_\varepsilon$, define the constant

$$C_\varepsilon^\star(\theta) = \inf_{\substack{\theta' : \|\theta' - \theta\| \leq \xi(\varepsilon) \\ \tilde{w}' : \|\tilde{w}' - \tilde{w}^\star\| \leq 3(K-1)\varepsilon}} \psi\left(\theta', \tilde{w}'\right)$$

Now, on the event $\mathcal{E}_T(\varepsilon)$, it holds that for every and for all $t \geq h(T)$, we have that $\hat{a}_t = a_\theta^\star$ and $\psi(\hat{\theta}_t, \tilde{w}_t) \geq C_\varepsilon^\star(\theta)$. Let $T \geq T_\varepsilon$, on $\mathcal{E}_T(\varepsilon)$, we have

$$\begin{aligned}
\min\{\tau, T\} &\leq \sqrt{T} + \sum_{t=\sqrt{T}}^T \mathbb{1}_{\{\tau > t\}} \leq \sqrt{T} + \sum_{t=\sqrt{T}}^T \mathbb{1}_{\{t\psi(\hat{\theta}_t, \tilde{w}_t) \leq \beta(\delta, t)\}} \\
&\leq \sqrt{T} + \sum_{t=\sqrt{T}}^T \mathbb{1}_{\{tC_\varepsilon^\star(\theta) \leq \beta(\delta, T)\}} \leq \sqrt{T} + \frac{\beta(\delta, T)}{C_\varepsilon^\star(\theta)}.
\end{aligned}$$

Let us introduce $T_0(\delta) = \inf\left\{T \in \mathbb{N} : \sqrt{T} + \frac{\beta(\delta,T)}{C_\varepsilon^\star(\theta)} \leq T\right\}$. For every $T \geq \max\{T_0(\delta), T_\varepsilon\}$, we have that $\mathcal{E}_T(\varepsilon) \subseteq \{\tau \leq T\}$. Hence, we get

$$\mathbb{P}(\tau > T) \leq \mathbb{P}(\mathcal{E}_T(\varepsilon)) \leq BT \exp(-CT^{1/8}), \quad \text{and} \quad \mathbb{E}[\tau] \leq T_0(\delta) + T_\varepsilon + \sum_{T=1}^\infty BT \exp(-CT^{1/8}).$$

We now provide an upper bound on $T_0(\delta)$. For $\xi > 0$, let

$$C(\xi) = \inf\{T \in \mathbb{N} : T - \sqrt{T} \geq T/(1+\xi)\}$$

Using the upper bound (25b) on the threshold $\beta(\delta, T)$, we have

$$T_0^\varepsilon(\delta) \leq c_1 + C(\xi) + \inf\left\{T \in \mathbb{N} : \frac{\ln\left(\frac{c_2 T}{\delta}\right)}{C_\varepsilon^\star(\theta)} \leq \frac{T}{1+\xi}\right\}.$$

Using Prop. 8 in (Kaufmann & Koolen, 2021), it follows that

$$T_0(\delta) \leq c_1 + C(\xi) + \frac{(1+\xi)}{C_\varepsilon^\star(\theta)} \left[\ln\left(\frac{(1+\xi)c_2}{C_\varepsilon^\star(\theta)\delta}\right) + \ln\left(\ln\left(\frac{(1+\xi)c_2}{C_\varepsilon^\star(\theta)\delta}\right) + \sqrt{2\ln\left(\frac{(1+\xi)c_2}{C_\varepsilon^\star(\theta)\delta}\right) - 2}\right)\right].$$

Hence for any $\xi > 0$ and $\varepsilon > 0$, it holds that $\limsup_{\delta \to 0} \frac{\mathbb{E}_\theta[\tau]}{\ln(1/\delta)} \leq \frac{(1+\xi)}{C_\varepsilon^\star(\theta)}$. Letting $\varepsilon \to 0$ and $\xi \to 0$ and by continuity of ψ, we get

$$\lim_{\varepsilon \to 0} C_\varepsilon^\star(\theta) = (T_\theta^{\text{MF}})^{-1},$$

which implies

$$\limsup_{\delta \to 0} \frac{\mathbb{E}[\tau]}{\log(1/\delta)} \leq T_\theta^{\text{MF}}.$$

□

E. Examples and results on VE and FCR

In this section, we present examples of the application of VE (see Alg. 3) and FCR (see Alg. 1) on a factor graph, in order to clarify their use. We also report results that ensure the correctness and bound the complexity of these methods.

E.1. Variable Elimination

We illustrate the use of VE to compute a joint optimal arm $a^\star_\theta$.

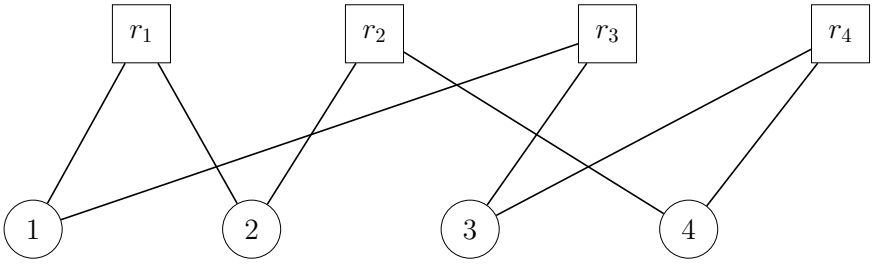


Figure 5. Factor graph from example 2.

Example 2. Consider the factor graph in Fig. 5 with $N = \rho = 4$. The average reward is described as:

$$\theta(a) = \theta_1(a_1, a_2) + \theta_2(a_2, a_4) + \theta_3(a_1, a_3) + \theta_4(a_3, a_4).$$

The key idea in VE is that, rather than summing all reward functions and then doing the maximization, we fix an ordering for the variables, and we maximize over variables one at a time, according to the predefined ordering. For example, fix the ordering as $\mathcal{O} = \{a_4, a_3, a_2, a_1\}$. Starting from a_4, we get

$$\max_{a\in\mathcal{A}} \theta(a) = \max_{a_1,a_2,a_3} \theta_1(a_1, a_2) + \theta_3(a_1, a_3) + \max_{a_4} \theta_2(a_2, a_4) + \theta_4(a_3, a_4).$$

Agent 4 can summarize the value that it brings to the system when varying (a_2, a_3) using a new function $p_4(a_2, a_3) = \max_{a_4} \theta_2(a_2, a_4) + \theta_4(a_3, a_4)$. Note that p_4 represents the best response of agent 4 conditioned on the actions played by agents 2, 3. We may also denote $a^\star_4(a_2, a_3) = \arg\max_{a_4} \theta_2(a_2, a_4) + \theta_4(a_3, a_4)$ as the best action for agent 4 conditioned on the actions of agent 2, 3. Hence, we get

$$\max_{a\in\mathcal{A}} \theta(a) = \max_{a_1,a_2,a_3} \theta_1(a_1, a_2) + \theta_3(a_1, a_3) + p_4(a_2, a_3).$$

Next, we do the same for agent 3, where we denote by $p_3(a_1, a_2) = \max_{a_3} \theta_3(a_1, a_3) + p_4(a_2, a_3)$, $a^\star_3(a_1, a_2) = \arg\max_{a_3} \theta_3(a_1, a_3) + p_4(a_2, a_3)$, and we reduce the problem to

$$\max_{a\in\mathcal{A}} \theta(a) = \max_{a_1,a_2} \theta_1(a_1, a_2) + p_3(a_1, a_2).$$

Next, agent 2 computes her response $p_2(a_1) = \max_{a_2} \theta_1(a_1, a_2) + p_3(a_1, a_2)$, and $a^\star_2(a_1) = \arg\max_{a_2} \theta_1(a_1, a_2) + p_3(a_1, a_2)$. Hence agent a_1 can simply select her action a_1 that maximizes $p_1 = \max_{a_1} p_2(a_1)$.

We can recover the best joint action $a^\star = (a^\star_1, a^\star_2, a^\star_3, a^\star_4)$ by performing the entire process in reverse order: $a^\star_1 = \arg\max_{a_1} p_2(a_1)$, $a^\star_2 = \arg\max_{a_2} \theta_1(a^\star_1, a_2) + p_3(a^\star_1, a_2)$, $a^\star_3 = \arg\max_{a_3} \theta_3(a^\star_1, a_3) + p_4(a^\star_2, a_3)$, and $a^\star_4 = \arg\max_{a_4} \theta_2(a^\star_2, a_4) + \theta_4(a^\star_3, a_4)$.

Complexity of VE.VE is guaranteed to return the optimal global arm in $O(NK^{A_\mathcal{O}+1})$ operations (Dechter, 1999), where $A_\mathcal{O} = \max_{i\in[N]} |\text{SC}(p_{\mathcal{O}(i)})|$ is the size of the largest factor generated when using elimination order $\mathcal{O}$. The complexity of VE depends on the elimination order $\mathcal{O}$ and is linear in the maximum size of the scope of "best-response functions" introduced in the elimination process.

Remark E.1. Note that the complexity is linear in N for any order of elimination $\mathcal{O}$. However, finding the optimal ordering, i.e., the one minimizing $A_\mathcal{O}$, is an NP-hard problem (Dechter, 1999). This issue has been addressed successfully for a large variety of graph structures in the graphical model community, where there exists a variety of good heuristics for the VE ordering problem (Wainwright & Jordan, 2008). In addition, there are approximate (and more efficient) alternatives to VE (e.g., the max-plus algorithm (Wainwright & Jordan, 2008)), but using those methods invalidates the correctness of VE.

E.2. Factored Constraint Reduction

We provide an example of the application of FCR to reduce a combinatorial number of constraints. We consider set of constraints

$$z \geq \frac{1}{\Delta_{\min}^2} \sum_{e \in [\rho]: a_e \neq a_e^\star} \underbrace{\tilde{w}_{e,a_e}^{-1} + \tilde{w}_{e,a_e^\star}^{-1}}_{f_e(a_e)}, \ \forall a \neq a^\star.$$

These constraints can be equivalently rewritten as $z \geq \max_{a \in \mathcal{A}} \sum_{e \in [\rho]} f_e(a_e)$.

Example 3. For the graph in Ex. 2 we have:

$$z \geq \max_{a_1, a_2, a_3, a_4} f_1(a_1, a_2) + f_2(a_2, a_4) + f_3(a_1, a_3) + f_4(a_3, a_4).$$

We introduce a set of variables $(u_{a_e}^{f_e})_{e \in [\rho], a_e \in \mathcal{A}_e}$, and the equality constraints:

$$u_{a_e}^{f_e} = \tilde{w}_{e,a_e}^{-1}, \forall e \in [\rho], a_e \in \mathcal{A}_e.$$

Note that we can rewrite $f_e(a_e) = u_{a_e}^{f_e} + u_{a_e^\star}^{f_e}$. Fix the elimination ordering $\mathcal{O} = \{4, 3, 2, 1\}$ and let $\mathcal{F} = \emptyset$. Now we introduce a new "function" p_l into $\mathcal{F}$ by eliminating a variable $l = \mathcal{O}(i)$. For $i = 1$, we have $\mathcal{O}(1) = 4$ and $\mathcal{F}_{\mathcal{O}(1)} = \{f_2(a_2, a_4), f_4(a_3, a_4)\}$, and a variable associated to this function $u_{a_2, a_3}^{p_4}$, for all a_2, a. We introduce a set of constraints:

$$u_{a_2,a_3}^{p_4} \geq u_{a_4,a_2}^{f_2} + u_{a_4^\star,a_2^\star}^{f_2} + u_{a_4,a_3}^{f_4} + u_{a_4^\star,a_3^\star}^{f_4}, \forall a_2, a_3, a_4$$

and we include these in the constraints set $\mathcal{K}$. We further exclude the function f_2 and f_4 from the set $\mathcal{F}$, while including $p_4(a_2, a_3)$. Subsequently, we consider $\mathcal{O}(2) = 3$. Then $\mathcal{F}_3 = \{p_4(a_2, a_3), f_3(a_3, a_1)\}$. We introduce the new constraints:

$$u_{a_1,a_2}^{p_3} \geq u_{a_2,a_3}^{p_4} + u_{a_2^\star,a_3^\star}^{p_4} + u_{a_3,a_1}^{f_3} + u_{a_3^\star,a_1^\star}^{f_3}, \forall a_1, a_2, a_3,$$

and we add them to the constraint set $\mathcal{K}$. We proceed to eliminate p_4 and f_3 from $\mathcal{F}$ and include p_3. We then move to $\mathcal{O}(3) = 2$ and define $\mathcal{F}_2 = \{f_1(a_1, a_2), p_3(a_1, a_2)\}$. The set of constraints introduced at this step are:

$$u_{a_1}^{p_2} \geq u_{a_1,a_2}^{p_3} + u_{a_1^\star,a_2^\star}^{p_3} + u_{a_1,a_2}^{f_1} + u_{a_1^\star,a_2^\star}^{f_1}, \forall a_1, a_2,$$

and similarly to the previous steps we add these constraints to $\mathcal{K}$ and eliminate the variables p_3 and f_1 from $\mathcal{F}$, while including p_2. The last step at $\mathcal{O}(4) = 1$ consists of simply including in $\mathcal{K}$ the constraints

$$u^{p_1} \geq u_{a_1}^{p_2}, \forall a_1 \in \mathcal{A}_1.$$

Finally we add to $\mathcal{K}$ the constraint $z \geq u^{p_1}$, and output $\mathcal{K}$.

Correctness and Complexity of FCR. The number of constraints in $\mathcal{K}$ set scales as $O(NK^{A_\mathcal{O}})$, where $A_\mathcal{O} = \max_{i \in [N]} |\text{Sc}(p_{\mathcal{O}(i)})|$ is the size of the maximum scope induced by the chosen order of elimination $\mathcal{O}$. Note that FCR also includes $O(NK^{A_\mathcal{O}})$ new variables in the optimization problem. Hence, similarly to VE, the number of constraints and variables to represent an exponentially large set depends linearly in N and exponentially only on the width of the induced graph, i.e., $O(N \exp(A_\mathcal{O}))$. Furthermore, the following lemma, adapted from Theorem 4.4 in (Guestrin et al., 2003), establishes the correctness of the FCR algorithm.

Lemma E.2 ((Guestrin et al., 2003), Theorem 4.4)**.** *Let $\mathcal{K} = FCR(\mathcal{C})$. Then $\mathcal{C}$ and $\mathcal{K}$ are equivalent, that is, an assignment of variables $(p, z, \tilde{w})$ is feasible for $\mathcal{K}$ if and only if $(z, \tilde{w})$ is feasible for $\mathcal{C}$.*

E.3. m-BEST algorithm

In this section, we discuss an algorithm to find the m-best global arms. As explained in App. H, a set of tighter approximations can be built by considering an ordering of the first m smallest gaps and hence requires to compute the $m+1$ global arms with highest expected rewards. The Lawler and Nilsson's m-BEST algorithm (Lawler, 1972; Nilsson, 1998), briefly described in the remainder of this section, will serve this purpose.

The procedure was originally devised to compute the m most probable configurations in graphical models. The main idea is the following: At each step, the m-BEST find the best solution to a re-formulation of the original problem that excludes the solutions already discovered. Specifically, at each time iteration $j < m$, the algorithm runs VE excluding the first j most probable configurations. The Lawler's algorithm (Lawler, 1972) starts by computing the best global action $a^{(1)}$ by applying VE (with elimination order $\mathcal{O}$) over the combinatorial action space $\mathcal{A}$ by applying VE N times. To determine the second best action $a^{(2)}$, the algorithm searches over the set $\mathcal{A}_{(2)} = \mathcal{A} \setminus \{a^{(1)}\}$. More generally, at iteration j, the algorithm finds the j^{th} best global action $a^{(j)}$ by running VE over the sets $\mathcal{A}_{(j)} = \mathcal{A} \setminus \cup_{k\in[j]}\{a^{(k)}\}$.

This procedure provably identifies the m-best global actions with complexity $O(mN^2K^{A_{\mathcal{O}}+1})$. By leveraging similar ideas and using a junction tree representation of the graph, Nilsson (Nilsson, 1998) improves over this procedure leading to an m-best algorithm with complexity $O(mNK^{A_{\mathcal{O}}+1})$.

F. Regret

In this section, we consider the MAMAB model as in §3 in the regret setting. We provide a lower bound on the regret and an approximation of such lower bound, using similar techniques as the ones used in the BAI setting.

In regret minimization, the goal is to devise an algorithm π to minimize the *regret* up to time $T \geq 1$, defined as

$$R^\pi(T) = \mathbb{E}\left[\sum_{t=1}^{T} \theta(a^\star) - \theta(a_t)\right] = \sum_{a\neq a^\star} (\theta(a^\star) - \theta(a))\mathbb{E}[N_{T,a}],$$

where $a_t \in \mathcal{A}$ is the action selected by algorithm π at time t.

F.1. Regret lower bound

Now, we give an instance-specific lower bound on the regret in the MAMAB setting. The regret lower bound is stated on the class of uniformly good algorithms, according to the following definition.

Definition F.1. An algorithm π is uniformly good algorithm if for all θ, we have that $R^\pi(T) = o(T^\alpha), \forall \alpha > 0$.

The following theorem gives a lower bound on the regret of any consistent algorithm. It is a direct consequence of the analogous lower bound in the combinatorial semi-bandit feedback setting, given e.g., in Theorem 1 in (Cuvelier et al., 2021b) or Theorem 12 in (Wagenmaker et al., 2020).

Theorem F.2. *Let π be a uniformly good algorithm. Then $\forall\theta$,*

$$\liminf_{T\to\infty} \frac{R^\pi(T)}{\log(T)} \geq c^\star_\theta, \tag{26}$$

where $c^\star_\theta$ is the value of the optimization problem:

$$\begin{array}{ll} \min_{\tilde{w}\in\mathbb{R}^{\tilde{A}}_{\geq 0}, w\in\mathbb{R}^{A}_{\geq 0}} \sum_{e\in[\rho],a_e\in\mathcal{A}_e} \tilde{w}_{e,a_e}(\theta_e(a^\star_e) - \theta_e(a_e)) & \\ \textit{subject to} \quad \sum_{e\in[\rho]:a_e\neq a^\star_e} \tilde{w}^{-1}_{e,a_e} \leq \Delta(a)^2/2, \forall a \in \mathcal{A} & \\ \tilde{w}_{e,a_e} = \sum_{b\in\mathcal{A}:b_e=a_e} w_b, \forall e \in [\rho], a_e \in \mathcal{A}_e & \end{array} \tag{27}$$

Note that the structure of the regret lower bound is similar to the one for M-BAI. The challenges are also similar: the optimization problem (27) has a combinatorial number of variables and constraints.

G. Connection to Combinatorial Semi-bandit Feedback Bandits

The MAMAB setting can be regarded as a specific instance of a combinatorial semi-bandit feedback setting (Cuvelier et al., 2021a;b; Wagenmaker et al., 2020). In the following, we present an equivalent characterization of the MAMAB problem to clarify its connection to the combinatorial semi-bandit feedback setting.

We first describe the interaction model in the generic (linear) combinatorial semi-bandit feedback setting. In such a setting, at each time step $t \geq 1$, the learner selects an action from a combinatorial set $a_t \in \{0,1\}^d$, and, given an unknown parameter $\tilde{\theta} \in \mathbb{R}^d$, she observes:

$$r_{t,i} = \tilde{\theta}_i + \eta_{t,i}, \forall i \in [d] : a_{t,i} = 1,$$

where $\eta_{t,i} \sim \mathcal{N}(0,1)$, for all $i \in [d]$, are independent Gaussian noise samples.

Recall that, since in MAMAB the set of global actions is defined as $\mathcal{A} = \times_{i\in[N]}\mathcal{A}_i$, the problem is not directly interpretable in the semi-bandit feedback setting. We show that a simple map from actions in $\mathcal{A}$ to binary vectors in the $\tilde{A}$-dimensional space can reformulate the MAMAB problem to the semi-bandit feedback setting.

Let $\phi(\cdot) : \mathcal{A} \to \{0,1\}^{\tilde{A}}$ be a function mapping global actions to binary vectors in the $\tilde{A}$-dimensional space. In MAMAB, the vector ϕ has a block structure: it can be decomposed as $\phi(a) = [\phi_e(b_e)]_{e\in[\rho],b_e\in\mathcal{A}_e}$, where $\phi_e(b_e) \in \{0,1\}^{A_e}$ is a group vector $\phi_e(b_e) = 1_{\{a_e=b_e\}}$, i.e., containing 1 in correspondence of the activated group action a_e. Further define $\tilde{\theta} = [\theta_e(a_e)]_{e\in[\rho],a_e\in\mathcal{A}_e} \in \mathbb{R}^{\tilde{A}}$, i.e., $\tilde{\theta}$ is the vector containing the local mean parameters. At round $t \geq 1$, a global action $a_t \in \mathcal{A}$ is selected by the learner, and she observes:

$$r_{t,e,a_e} = \tilde{\theta}_e(a_e), \forall e \in [\rho] : \phi_e(a_{t,e}) = 1.$$

In other words, in the semi-bandit feedback setting, a (joint) action $a \in \mathcal{A}$ is selected and the learner observes a vector of rewards $[r_e(a_e)]_{e\in[\rho],a_e\subseteq a}$, where $r_e(a_e) = \theta_e(a_e)^\top \phi_e(a_e) + \eta_e$, where $\eta_e \sim \mathcal{N}(0,1)$ is i.i.d. Gaussian Noise. Note that the feature vectors satisfy $\|\phi(a)\|_0 = \rho, \forall a \in \mathcal{A}$ and $\|\phi_e(a_e)\|_0 = 1, \forall e \in [\rho], a_e \in \mathcal{A}_e$. In order to further clarify the connection to the semi-bandit feedback, we provide a concrete example below.

Example 4. Consider the factor graph in Fig. 6 with $N = 3$ agents, $\rho = 2$ groups, and $K = 2$ actions. The reward can be written as $r(a_1, a_2, a_3) = r_1(a_1, a_2) + r_2(a_2, a_3)$. Let $a_i \in \{0,1\}$, for all $i \in [N]$. The average reward can be expressed by the vector $\tilde{\theta} = \left[[\theta_1(a_1,a_2)]_{(a_1,a_2)\in\{0,1\}^2}, [\theta_2(a_2,a_3)]_{(a_2,a_3)\in\{0,1\}^2}\right] \in \mathbb{R}^8$, where

$$[\theta_1(a_1,a_2)]_{(a_1,a_2)\in\{0,1\}^2} = [\theta_1(0,0), \theta_1(0,1), \theta_1(1,0), \theta_1(1,1)]$$
$$[\theta_2(a_2,a_3)]_{(a_2,a_3)\in\{0,1\}^2} = [\theta_2(0,0), \theta_2(0,1), \theta_2(1,0), \theta_2(1,1)].$$

For example, selecting action $a = (0,0,0)$, corresponds to the feature vector $\phi(a) = (1,0,0,0,1,0,0,0)$, while selecting action $b = (0,1,0)$ corresponds to the feature vector $\phi(b) = (0,1,0,0,0,0,1,0)$.

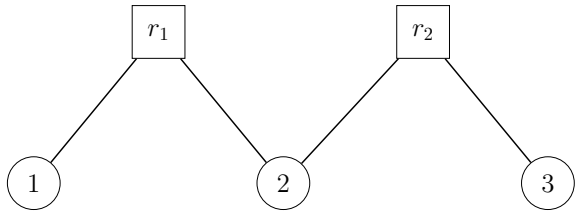


Figure 6. Factor graph from Example 4.

H. Tighter Approximations

H.1. Tighter constraints reduction

In this section, we propose tighter constraint approximations by leveraging an ordering of the arms and of the sub-optimality gaps. For $m \in [K^N]$, let $a^{(m)}$ be the m^{th} best arm and, for $m \in [K^N - 1]$, let $\Delta_m = \theta(a^\star_\theta) - \theta(a^{(m+1)})$ the m^{th} minimal non-zero gap (with ties breaking arbitrarily).

Lemma H.1. *Let $m \in [K^N - 1]$, and $T^{MF}_\theta(m)$ be the solution to the following optimization problem:*

$$\begin{aligned} \inf_{\tilde{w} \in \tilde{\Lambda}_{MF}, z \in \mathbb{R}} \quad & z \\ \textit{subject to} \quad & z \geq \frac{1}{\Delta_j^2} \sum_{e \in [\rho]: a_e^{(j+1)} \neq a_e^\star} \tilde{w}^{-1}_{e,a_e^\star} + \tilde{w}^{-1}_{e,a_e^{(j+1)}}, \forall j \in [m] \\ & z \geq \frac{1}{\Delta_m^2} \sum_{e \in [\rho]: a_e \neq a_e^\star} \tilde{w}^{-1}_{e,a_e^\star} + \tilde{w}^{-1}_{e,a_e}, \forall a \in \mathcal{A} \setminus \cup_{j \in [m]} \{a^{(j+1)}\}. \end{aligned} \tag{28}$$

Then, $T^{MF}_\theta(m+1) \leq T^{MF}_\theta(m)$, $\forall m \in [K^N - 2]$, and $T^\star_\theta \leq T^{MF}_\theta(m)$, $\forall m \in [K^N - 1]$.

Note that $T^{\text{MF}}_\theta(1) = T^{\text{MF}}_\theta$, and for $m > 1$, $T^{\text{MF}}_\theta(m)$ provides provably tighter approximations. This approximation gain comes at the cost of increased computational complexity. Indeed, to solve $T^{\text{MF}}_\theta(m)$, one needs to compute the first $m+1$ best arms and the m minimal gaps. To solve this task, there exist algorithm having complexity $O((m+1)NK^{A_\mathcal{O}+1})$ (see Sec. E.3).

This result shows that for increasingly larger values of m, the approximation gets tighter, but the computational complexity increases. Hence, the approximations $T^{\text{MF}}_\theta(m)$ allow for an interplay between sample complexity and computational complexity when varying m. Sec. J.1 provides numerical results on this trade-off.

There exists an interesting trade-off between sample complexity and computational complexity. The smallest sample complexity achievable is the true lower bound constant, i.e. $T^\star_\theta$, which is the solution to an optimization problem with $O(K^N)$ variables and constraints. The approximation T^{MF}_θ has generally higher complexity, but its computational complexity is characterized by $O(mNK^{A_\mathcal{O}})$, where $A_\mathcal{O}$ is typically much smaller than N. As explained above, for $m = 1$ $T^{\text{MF}}_\theta(m)$ reduces to the MF approximation T^{MF}_θ. For illustration purposes we also report the sample complexity $2\tilde{A}/\Delta^2_{\min}$ which is achieved by the random allocation $w = (1/A)_{a \in \mathcal{A}} \in \Lambda$.

Proof of Lemma H.1. First, we shall prove $T^{\text{MF}}_\theta(m+1) \leq T^{\text{MF}}_\theta(m), \forall m \in [K^N - 1]$, by induction. The base case, for $m = 1$, is $T^{\text{MF}}_\theta(2) \leq T^{\text{MF}}_\theta(1)$. Note that $T^{\text{MF}}_\theta(1)$ can be written as

$$\inf_{\tilde{w} \in \tilde{\Lambda}_{\text{MF}}, z \in \mathbb{R}} \quad z \tag{29}$$

$$\text{subject to} \quad z \geq \frac{1}{\Delta_1^2} \sum_{e \in [\rho]: a_e^{(2)} \neq a_e^\star} \tilde{w}^{-1}_{e,a_e^\star} + \tilde{w}^{-1}_{e,a_e^{(2)}} \tag{30}$$

$$z \geq \frac{1}{\Delta_1^2} \sum_{e \in [\rho]: a_e \neq a_e^\star} \tilde{w}^{-1}_{e,a_e^\star} + \tilde{w}^{-1}_{e,a_e}, \forall a \in \mathcal{A} \setminus \{a^{(1)}, a^{(2)}\}, \tag{31}$$

while $T^{\text{MF}}_\theta(2)$ is defined as

$$\inf_{\tilde{w} \in \tilde{\Lambda}_{\text{MF}}, z \in \mathbb{R}} \quad z \tag{32}$$

$$\text{subject to} \quad z \geq \frac{1}{\Delta_1^2} \sum_{e \in [\rho]: a_e^{(2)} \neq a_e^\star} \tilde{w}^{-1}_{e,a_e^\star} + \tilde{w}^{-1}_{e,a_e^{(2)}} \tag{33}$$

$$z \geq \frac{1}{\Delta_2^2} \sum_{e \in [\rho]: a_e^{(3)} \neq a_e^\star} \tilde{w}^{-1}_{e,a_e^\star} + \tilde{w}^{-1}_{e,a_e^{(3)}} \tag{34}$$

$$z \geq \frac{1}{\Delta_2^2} \sum_{e \in [\rho]: a_e \neq a_e^\star} \tilde{w}^{-1}_{e,a_e^\star} + \tilde{w}^{-1}_{e,a_e}, \forall a \in \mathcal{A} \setminus \{a^{(1)}, a^{(2)}, a^{(3)}\}. \tag{35}$$

The constraints (30) and (33) are identical. The constraints (34)-(35) can be simply written as

$$z \geq \frac{1}{\Delta_2^2} \sum_{e \in [\rho]: a_e \neq a_e^\star} \tilde{w}_{e,a_e^\star}^{-1} + \tilde{w}_{e,a_e}^{-1}, \forall a \in \mathcal{A} \setminus \{a^{(1)}, a^{(2)}\},$$

which corresponds to (31) with the exception of the term $\frac{1}{\Delta_1^2}$ in place of $\frac{1}{\Delta_2^2}$. As $\Delta_1 \leq \Delta_2$, we naturally conclude that $T_\theta^{\text{MF}}(2) \leq T_\theta^{\text{MF}}(1)$.

Now, assume that $T_\theta^{\text{MF}}(m+1) \leq T_\theta^{\text{MF}}(m)$ holds for $m-1$, to complete the induction we need to show that $T_\theta^{\text{MF}}(m+1) \leq T_\theta^{\text{MF}}(m)$. By following a similar approach to the base case, we can show that the only difference in the optimization problems defining $T_\theta^{\text{MF}}(m)$ and $T_\theta^{\text{MF}}(m+1)$ is in the last set of constraints: for $T_\theta^{\text{MF}}(m)$ these constraints are

$$z \geq \frac{1}{\Delta_m^2} \sum_{e \in [\rho]: a_e \neq a_e^\star} \tilde{w}_{e,a_e^\star}^{-1} + \tilde{w}_{e,a_e}^{-1}, \forall a \in \mathcal{A} \setminus \cup_{j \in [m]} \{a^{(j+1)}\},$$

while for $T_\theta^{\text{MF}}(m+1)$ they can be written as

$$z \geq \frac{1}{\Delta_{m+1}^2} \sum_{e \in [\rho]: a_e \neq a_e^\star} \tilde{w}_{e,a_e^\star}^{-1} + \tilde{w}_{e,a_e}^{-1}, \forall a \in \mathcal{A} \setminus \cup_{j \in [m]} \{a^{(j+1)}\}.$$

As we have $\Delta_m \leq \Delta_{m+1}$, we can conclude that $T_\theta^{\text{MF}}(m+1) \leq T_\theta^{\text{MF}}(m)$.

Now, to complete the proof, we show that $T_\theta^\star \leq T_\theta^{\text{MF}}(m), \forall m \in [K^N - 1]$. In light of the fact that $T_\theta^{\text{MF}}(m+1) \leq T_\theta^{\text{MF}}(m)$, it is sufficient to prove that $T_\theta^\star \leq T_\theta^{\text{MF}}(K^N - 1)$. It is easy to check that $T_\theta^{\text{MF}}(K^N - 1)$ can be written as

$$\inf_{\tilde{w} \in \tilde{\Lambda}_{\text{MF}}} \max_{a \neq a^\star} \frac{\sum_{e \in [\rho]: a_e \neq a_e^\star} \tilde{w}_{e,a_e^\star}^{-1} + \tilde{w}_{e,a_e}^{-1}}{\Delta(a)^2}. \tag{36}$$

Eq. (36) is essentially the same as $T_\theta^\star$ in (1), with the difference that the allocations variables are in $\tilde{\Lambda}_{\text{MF}}$. As $\tilde{\Lambda}_{\text{MF}} \subseteq \tilde{\Lambda}$ we conclude that $T_\theta^\star \leq T_\theta^{\text{MF}}(K^N - 1)$. □

H.2. Tighter variables reduction: Structured Mean Field

In this section we present tighter variable reduction schemes that consider a different factorization factorization w.r.t. allocation distributions named Structured Mean Field (SMF). Let $\mathcal{G}$ be a set of subsets of $[N]$ and consider the following factorization

$$w_a = \prod_{g \in \mathcal{G}} w_{g,a_g}^{\gamma_g}, \forall a \in \mathcal{A},$$

where a_g denotes the sub-vector of actions corresponding to indices $g \in \mathcal{G}$, and $\gamma_g > 0$. Let $\Lambda_{\mathcal{G}}$ denote the set of distributions satisfying this factorization, and $\tilde{\Lambda}_{\mathcal{G}}$ the corresponding marginal polytope. The approximation is essentially obtained replacing $\tilde{\Lambda}$ by $\tilde{\Lambda}_{\mathcal{G}}$ in (4.1). Note that if $\mathcal{G} = \{\{1\}, ..., \{N\}\}$, and $\alpha_g = 1$, $\forall g \in \mathcal{G}$, we recover the MF approximation T_θ^{MF}. In general tighter (or even exact approximations) are possible.

Lemma H.2. *For any valid factorization $\mathcal{G}$, we have that $T_\theta^\star \leq T_\theta^{\mathcal{G}}$, where*

$$T_\theta^{\mathcal{G}} = \inf_{\tilde{w} \in \tilde{\Lambda}_{\mathcal{G}}} \max_{a \neq a^\star} \frac{\sum_{e \in [\rho]: a_e \neq a_e^\star} \tilde{w}_{e,a_e^\star}^{-1} + \tilde{w}_{e,a_e}^{-1}}{\left(\sum_{e \in [\rho]} \theta_e(a_e^\star) - \theta_e(a_e) \right)^2}. \tag{37}$$

I. Details on antenna tilt optimization experiments

In this section, we present details on the antenna tilt optimization experiments.

Throughput. The throughput $T_{i,u}$, is formally defined in terms of the Signal-to-Interference-plus-Noise Ratio (SINR), a metric that measures the quality of a signal in the presence of interference and noise. Let $a_{\mathcal{N}_i}$ be the group vector containing

Specifically, the SINR of a UE $u \in \mathcal{U}$ connected to cell $c \in \mathcal{C}$ is defined as:

$$\text{SINR}_{i,u}(a_i) = \frac{P_i G_{i,u}(a_i) L_{i,u}(a_i)}{\sum_{k \in \mathcal{N}_i} P_k G_{k,u}(a_k) L_{k,u}(a_k) + \sigma},$$

where P_i, $G_{i,u}$, and $L_{i,u}$ are the transmitter antenna power, the gain of the transmitter antenna, and path loss for UE u connected to cell i, respectively. The gain is influenced by antenna parameters such as tilt and azimuth, and the path loss accounts for the transmission medium and obstacles (e.g., buildings, atmospherical conditions, vegetation, etc.). The throughput $T_{i,u}$ experienced by UE u connected to the cell i is then expressed as a function of the SINR and available bandwidth:

$$T_{i,u} = \omega_B n_{i,u}^R \log_2 \left(1 + \text{SINR}_{i,u}\right),$$

where $n_{i,u}^R$ is the number of Physical Resource Blocks (PRBs) allocated to UE u in cell i and ω_B is the bandwidth per PRB (180 kHz). We use the average throughput of a cell in our group reward definition, i.e.,

$$r_i(a_e) = \frac{1}{|\mathcal{U}_i|} \sum_{u \in \mathcal{U}_i} T_{i,u}.$$

Hence the global reward is expressed as

$$r(a) = \sum_{i \in [N]} \frac{1}{|\mathcal{U}_i|} \sum_{u \in \mathcal{U}_i} T_{i,u}(a_e).$$

On the noise independence assumption. In our experiments, each group $e \in [\rho]$ corresponds to a sector: more precisely, it consists of an antenna $i \in [N]$ serving the users $u \in \mathcal{U}_i$ connected to this sector, and the set of antennas that can interfere with the transmissions of the antenna i. Recall that the group reward is defined as $r_e(a_e) = \sum_{u \in \mathcal{U}_i} T_{i,u}(a_e)$, where a_e represents the tilts of antennas in group i. The throughput $T_{i,u}(a_e)$ is the rate at which an user u can decode transmissions from the antenna u. This rate depends on the random channel conditions (also known as *fading*) between each antenna in the group and the user i. Now, the fadings between pairs of (antenna, user) are typically stochastically independent across users and antennas (Tse & Viswanath, 2009). Since the sets of $(\mathcal{U}_i)_{i \in [N]}$ form a partition, they do not overlap, and the random variables $r_e(a_e)$ are indeed independent across groups. They can be modeled as independent Gaussian realizations in the sum-throughput over groups. For details, refer e.g., to (Tse & Viswanath, 2009).

Additional details. The set of UEs in the network is $\mathcal{U} = \cup_{i \in \mathcal{S}} \mathcal{U}_i$ as presented in Sec. 7.2. The number of UEs connected to cell i is affected by tilt variation since we assume UEs connect to the cell from which they get maximum Reference Signal Received Power (RSRP). In particular, given a tilt configuration a, the UEs in cell i are defined as

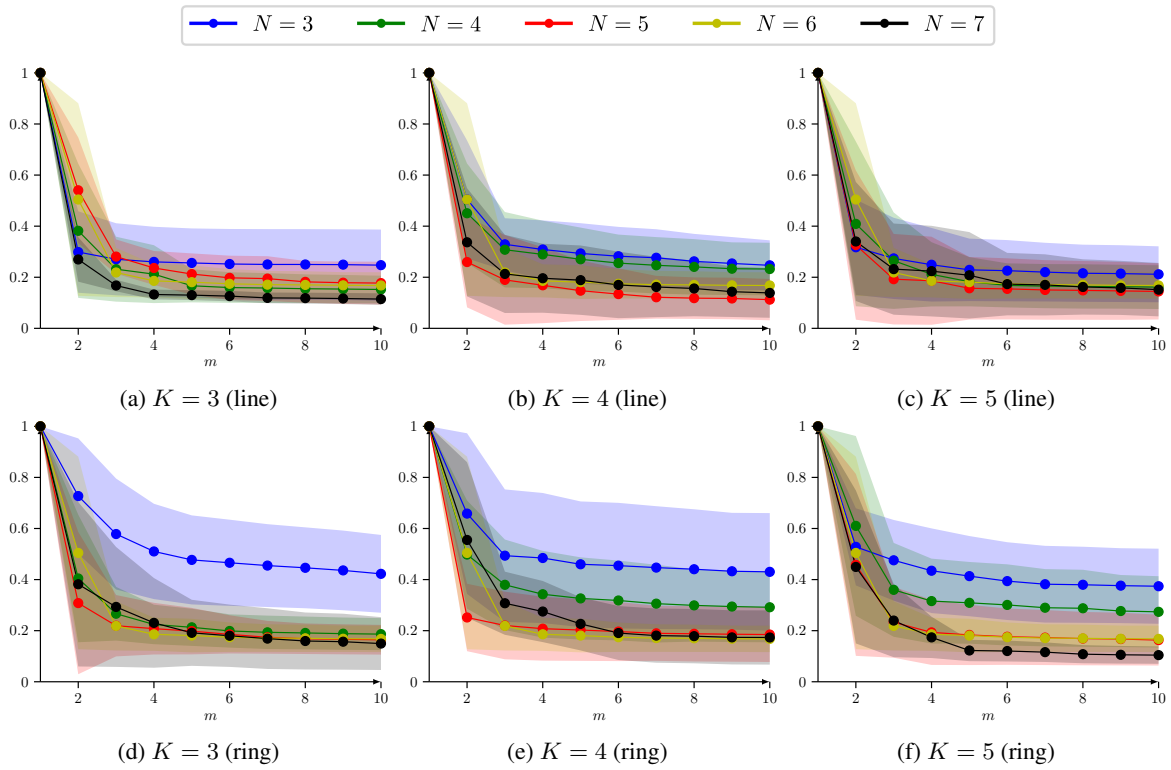
$$\mathcal{U}_i = \left\{ u \in \mathcal{U} : \arg\max_{k \in [N]} P_k G_{k,u} L_{k,u} = i \right\}.$$

There exist other methods to determine relations between antennas which rely on automated procedures, domain knowledge, and heuristics. For example, they may be based on the geographic distance between cells, on Neighbor Relations (ANR) as defined in 3GPP standards, on network planning tools for coverage prediction, or on cell handover logs (Rappaport, 2001). In addition, domain knowledge can be used to refine the graph topology by pruning or adding edges based on key feature of a city or knowledge about the terrain (if there is a natural obstacle for example). Analyzing the influence of the graph structure is not in the scope of this paper and is left as future work.

J. Additional Experiments

J.1. Tighter bounds experiments

In this section, we propose a set of experiments to test the effect of the tighter lower bound approximations $T_\theta^{\text{MF}}(m)$, presented in App. H. We consider different instances of the line and ring graphs with varying $N \in \{3, 4, 5, 6, 7\}$, $K \in \{2, 3, 4, 5\}$, and $m \in \{1, 10\}$. The group means $(\theta_e)_{e\in[\rho]}$ are generated uniformly at random. We report in Fig. 7, the results averaged over $N_{\text{sim}} = 5$ independent runs and report the results in terms of the mean and standard deviation of the normalized $T_\theta^{\text{MF}}(m)$.



(a) $K = 3$ (line) (b) $K = 4$ (line) (c) $K = 5$ (line)

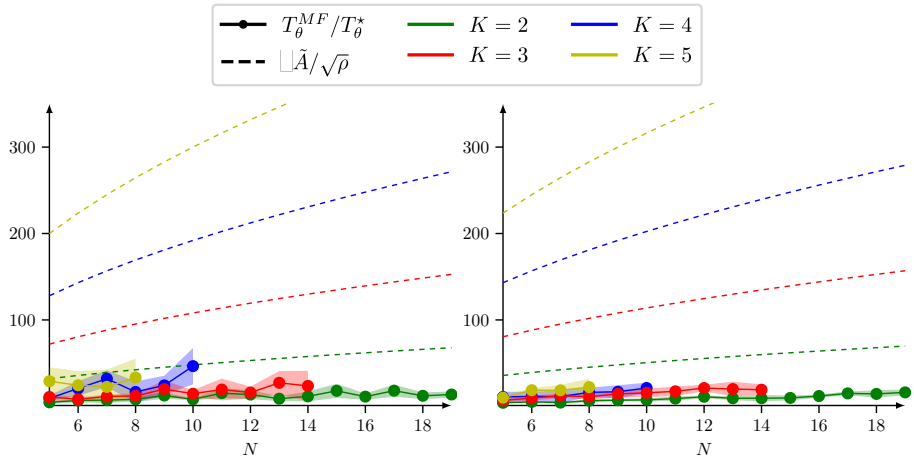
(d) $K = 3$ (ring) (e) $K = 4$ (ring) (f) $K = 5$ (ring)

Figure 7. Normalized $T_\theta^{\text{MF}}(m)$ for varying m, K, and N.

As expected, $T_\theta^{\text{MF}}(m)$ is a monotonically decreasing function of m for all considered graphs. The approximation for $m = 1$ corresponds to the MF approximation T_θ^{MF}. Hence the curves in Fig. 7 represent improvement over the MF approximations (recall that $T_\theta^{\text{MF}} \leq T_\theta^{\text{MF}}(m)$, for any m). We can observe that for increasing values of N, the improvement over T_θ^{MF} gets larger, while the effect of larger K is less significant.

J.2. Quantifying the approximation

In this section, we provide numerical results on the approximation ratio $T_\theta^{\mathrm{MF}}/T_\theta^\star$. We consider different instances of line and ring graphs with varying N and $K \in \{2,3,4,5\}$. The group means $(\theta_e)_{e\in[\rho]}$ are generated uniformly at random, as described in Sec. 7.1. We report the results averaged over $N_{\mathrm{sim}} = 5$ independent runs and report the results in terms of mean and standard deviation of the ratio $T_\theta^{\mathrm{MF}}/T_\theta^\star$, together with its upper bound $\tilde{A}/\rho$ (see Lem. A.3). The results are shown in Fig. 8. It can be observed that, as conjectured in App. A.3, the upper bound on the approximation ratio is loose for different instances, and generally the actual value of $T_\theta^{\mathrm{MF}}/T_\theta^\star$ is significantly smaller than the bound.



(a) Line graph. (b) Ring graph.

Figure 8. Approximation ratio $T_\theta^{\mathrm{MF}}/T_\theta^\star$ vs upper bound $4\tilde{A}/\rho$.