# Finite-Time Frequentist Regret Bounds of Multi-Agent Thompson Sampling on Sparse Hypergraphs

**Tianyuan Jin**[1], **Hao-Lun Hsu**[2], **William Chang**[3], **Pan Xu**[2]

[1]National University of Singapore
[2]Duke University
[3]University of California, Los Angeles
tianyuan@u.nus.edu, hao-lun.hsu@duke.edu, chang314@g.ucla.edu, pan.xu@duke.edu

## Abstract

We study the multi-agent multi-armed bandit (MAMAB) problem, where agents are factored into overlapping groups. Each group represents a hyperedge, forming a hypergraph over the agents. At each round of interaction, the learner pulls a joint arm (composed of individual arms for each agent) and receives a reward according to the hypergraph structure. Specifically, we assume there is a local reward for each hyperedge, and the reward of the joint arm is the sum of these local rewards. Previous work introduced the multi-agent Thompson sampling (MATS) algorithm and derived a Bayesian regret bound. However, it remains an open problem how to derive a frequentist regret bound for Thompson sampling in this multi-agent setting. To address these issues, we propose an efficient variant of MATS, the epsilon-exploring Multi-Agent Thompson Sampling ($\epsilon$-MATS) algorithm, which performs MATS exploration with probability epsilon while adopts a greedy policy otherwise. We prove that $\epsilon$-MATS achieves a worst-case frequentist regret bound that is sublinear in both the time horizon and the local arm size. We also derive a lower bound for this setting, which implies our frequentist regret upper bound is optimal up to constant and logarithm terms, when the hypergraph is sufficiently sparse. Thorough experiments on standard MAMAB problems demonstrate the superior performance and the improved computational efficiency of $\epsilon$-MATS compared with existing algorithms in the same setting.

## 1 Introduction

Reinforcement learning (RL) is a fundamental problem in machine learning, where an agent learns to make optimal decisions in an environment by trial and error. A specific instance of RL is the multi-armed bandit (MAB) problem, in which an agent must choose between a set of arms, and each of the arms has a random reward distribution. The agent's goal is to maximize its total reward over time. In standard MAB problems, an agent is provided with a set of arms $[K] := 1, 2, ..., K$, and each arm, when pulled, generates a reward following a 1-subgaussian distribution with an unknown mean. The agent's objective is to maximize its overall rewards within a specified time frame.

We consider the **m**ulti-**a**gent MAB (MAMAB) problem, where there are $m$ agents. At each round of the interaction, each agent chooses an arm from its own arm set $[K]$. We define the concatenation of these arms as the joint arm. The bandit learner aims to coordinate with all agents and choose joint arms that maximize the cumulative rewards obtained from pulling those joint arms. It is important to note that the size of the joint arm space is exponential in the number of agents, specifically $A = K^m$. This exponential growth poses computational challenges in coordination and arm selection. To address this issue, it was proposed to factor all agents into $\rho$ possibly overlapping groups (see wind farm application), which forms a hypergraph over the agents with each agent representing a node and each group representing a hyperedge (an illustration can be found in Section 2 and Figure 1). Instead of pulling the joint arm, the learner only needs to pull the local arms, where each pulled local arm is defined as the concatenation of the arms chosen by agents within the same group. We assume each group has $d$ agents, and thus the total number of local arms $A_{\text{loc}}$ is at most $\rho K^d$, which is much smaller than the number of joint arms when the groups are small. This approach gives rise to MAMAB problems with specific coordination graph structures, which have found practical applications in various domains such as traffic light control (Wiering et al. 2000), warehouse commissioning (Claes et al. 2017), and wind farm control (Gebraad and van Wingerden 2015; Verstraeten et al. 2019).

We evaluate a learning strategy based on its cumulative rewards obtained by interacting with the environment for a total of $T$ rounds. This evaluation can be equivalently measured by calculating the regret of the strategy compared to an oracle algorithm that always selects the arm with the highest reward. Mathematically, the regret is defined as $R_T = T\mu^* - \mathbb{E}[\sum_{t=1}^{T} f(\boldsymbol{A}_t)]$, where $\mu^*$ is the mean of the optimal arm and $f(\boldsymbol{A}_t)$ represents the reward obtained when pulling the joint arm $\boldsymbol{A}_t$ at time $t$ according to the given strategy. The goal of the algorithm (or learner) is to coordinate with all agents to determine the joint arm to pull in order to minimize this regret.

Thompson sampling (Thompson 1933), introduced by Thompson in 1933, has emerged as an attractive algorithm for bandit problems. It is favored for its simplicity of implementation, good empirical performance, and strong theoretical guarantees (Chapelle and Li 2011; Agrawal and Goyal 2017; Jin et al. 2021a). The key idea behind Thompson sampling is to sample reward estimates for each possible arm from a posterior distribution and select the arm with the highest estimated value for pulling. In the single-agent setting, Thompson sampling has been shown to achieve near-optimal regret with respect to the worst possible bandit instance (Agrawal and Goyal 2017). In the context of multi-agent MAB with a coordination graph, the MATS (Multi-Agent Thompson Sampling) algorithm was proposed by Verstraeten et al. (2020). Unlike traditional Thompson sampling, where estimated rewards are sampled for each joint arm, MATS samples rewards for each local arm. This approach reduces the computational complexity, particularly in cases where the coordination hypergraph is sparse. Verstraeten et al. (2020) provided a Bayesian regret bound for MATS, which measures the average performance given the probability kernel of the environment. However, in practical scenarios, it may not always be feasible for the learner to possess knowledge or access to the probability kernel of the environment. In such cases, the frequentist regret bound, which measures the worst-case performance across all environments, is often considered. It is worth noting that a frequentist regret upper bound implies a Bayesian regret bound, but not vice versa. Deriving a frequentist regret bound for the MATS algorithm in the multi-agent MAB problem with a coordination hypergraph remains an open question.

There are several technical challenges in the analysis of the frequentist regret for MATS. The first challenge emerges when applying the regret analysis of single-agent Thompson Sampling (Agrawal and Goyal 2012) to our context. This occurs due to a dependence issue among different joint arms. Although rewards for each local arm are independently drawn from their respective reward distributions, the average rewards of the joint arms might be influenced by the other joint arms when they share some local arms (see Section 4 for detailed discussion). As a result, it is difficult to analyze the distribution of the average reward of the optimal arm or apply any concentration/anti-concentration inequalities, while all existing frequentist regret analyses of Thompson sampling (Agrawal and Goyal 2012, 2017; Jin et al. 2021a, 2022; Korda, Kaufmann, and Munos 2013; Kaufmann, Korda, and Munos 2012) heavily rely on the specific form of the distribution of the average reward of the optimal arm. A naive method of removing the dependence involves maintaining a posterior distribution for each joint arm and updating the distribution only when this joint arm is pulled. However, this method could result in significant computational complexity and regret due to the large joint arm space.

In this paper, we tackle the issue using two strategies: 1) We carefully partition the entire arm set into subsets, ensuring each arm within a subset shares the same local arms with the optimal arm, and 2) We conduct a regret analysis at the level of local arms. Specifically, let $\mathbf{1}$ denote the optimal joint arm. We consider two events: 1) The local arm $\mathbf{1}^e$ of the optimal arm $\mathbf{1}$ is not underestimated, meaning the posterior sample of $\mathbf{1}^e$ is larger than $\mathbf{1}^e - \Delta/\rho$, and 2) The local arm $\boldsymbol{a}^e$ of the suboptimal arm $\boldsymbol{a}$ is not overestimated, meaning the posterior sample of $\boldsymbol{a}^e$ is lower than $\boldsymbol{a}^e - \Delta/\rho$. Crucially, these events ensure that the sum of posterior samples for any suboptimal joint arms is lower than the sum of posterior samples of $\mathbf{1}$, which leads to a lower regret.

Another challenge in our local arm level analysis arises when we aim to establish a lower bound for the probability that the posterior sample of all local arms of $\mathbf{1}$ exceeds their means by $\Delta/\rho$. Leveraging the original Thompson Sampling analysis, we can establish this probability's lower bound as $(\Delta/\rho)^{2\rho}$, leading to $(\rho/\Delta)^{2\rho}$ suboptimal arm pulls. In terms of worst-case regret, this amounts to $O\big(T^{\frac{2\rho-1}{2\rho}}\big)$. We improve this result by applying two innovative techniques (for ease of presentation, these are elaborated in full detail in Section 4), reducing the number of pulls to $C^\rho$, where $C$ is a universal constant. Using these novel techniques, we are able to offer a $\sqrt{T}$-type worst-case regret.

**Main contributions.** We summarize our main contributions as follows.

- We propose the $\epsilon$-exploring Multi-Agent Thompson Sampling ($\epsilon$-`MATS`) algorithm, which only samples from the posterior distribution with probability $\epsilon$ and acts greedily with probability $1-\epsilon$. Note that even the local arm size is exponentially large and thus $\epsilon$-`MATS` is much more computationally efficient than MATS in practice.
- We establish a frequentist regret bound for $\epsilon$-`MATS` in the order of $\tilde{O}(\sqrt{C^\rho A_{\text{loc}} T})$, where $C$ is some universal constant and $\tilde{O}(\cdot)$ ignores constant and logarithmic factors. Here $\rho$ denotes the number of hyperedges, $A_{\text{loc}}$ represents the total number of local arms, and $T$ is the time horizon. Remarkably, when $\epsilon = 1$, our result provides the first frequentist regret bound for MATS (Verstraeten et al. 2020). Despite having $A$ joint arms, our regret bound grows as $O(\sqrt{A_{\text{loc}}})$, which is much smaller than the total number of joint arms.
- We also derive a lower bound in the order of $\Omega(\sqrt{A_{\text{loc}} T/\rho})$ in the worst-case regret bound for our setting. This lower bound implies that $\epsilon$-`MATS` is optimal up to constant and logarithmic factors when the the number of groups $\rho$ is small. Besides, we derive a lower bound for the original Multi-agent Thompson Sampling. The lower bound shows that $C^\rho$ regret is unprohibited for original Multi-Agent Thompson Sampling, which further proves the optimality of our regret bound $\tilde{O}(\sqrt{C^\rho A_{\text{loc}} T})$.
- We further conduct extensive experiments on various MAMAB problems, including the Bernoulli 0101-Chain, the Poisson 0101-Chain, and the Gem Mining problem (Roijers, Whiteson, and Oliehoek 2015; Bargiacchi et al. 2018; Verstraeten et al. 2020). Through empirical evaluation, we demonstrate that the regret of $\epsilon$-`MATS` can be significantly lower compared to MATS as $\epsilon$ decreases, outperforming existing methods in the same setting. We also find that $\epsilon$-`MATS` exhibits improves computational efficiency compared to MATS.

## 2 Preliminary and Background

In this section, we present the preliminary details of our setting. We also provide a notation table in Table 1 for the convenience of our readers. We adopt the MAMAB (Multi-Agent Multi-Armed Bandit) framework introduced by Verstraeten et al. (2020), where there are $m$ different agents, who are grouped into $\rho$ potentially overlapping groups. Each group can be represented as a hyperedge in a hypergraph, where the agents correspond to the nodes. Figure 1 provides an example for easier visualization. During each round, every agent $i \in [m]$ selects an arm from their respective arm set $\mathcal{A}_i$, which is referred to as the ”*individual*” arm played by agent $i$. For simplicity, we assume that each agent $i$ has the same number of arms, denoted as $K = |\mathcal{A}_i|$. However, it is straightforward to extend the framework to accommodate varying numbers of arms $|\mathcal{A}_i|$. The arms chosen by all agents are concatenated to form a ”*joint*” arm denoted by $\boldsymbol{a}$, which belongs to the set $\mathcal{A}_1 \times \cdots \times \mathcal{A}_m$. Consequently, the total number of joint arms is defined as $A := |\mathcal{A}_1 \times \cdots \times \mathcal{A}_m|$.

We define a ”*local*” arm as the concatenation of individual arms for a specific group $e \in [\rho]$. In other words, if agents $i_1, \ldots, i_d \in [m]$ form a hyperedge, then the local arm $\boldsymbol{a}^e \in \mathcal{A}_{i_1} \times \cdots \times \mathcal{A}_{i_d}$ represents the $d$-tuple of arms selected by these agents. We shall denote the set of local arms for group $e$ as $\mathcal{A}^e$. Let $A_{\text{loc}}$ be the total number of local arms. It is straightforward to see that $A_{\text{loc}} \leq \rho K^d$, with equality when the groups don't overlap. It is important to note that the arm space grows exponentially with the number of agents, leading to computational challenges in arm selection. To address this combinatorial complexity, we employ variable elimination techniques, which will be further explained in the subsequent sections.

In this paper, the global reward $f(\boldsymbol{a})$ associated with each joint arm $\boldsymbol{a}$ is decomposed into $\rho$ local rewards $f^e(\boldsymbol{a}^e)$, where $\boldsymbol{a}^e$ represents the local arm for group $e$. This decomposition takes advantage of the hypergraph structure. Specifically, for a given hypergraph with $\rho$ hyperedges, we have the relationship $f(\boldsymbol{a}) = \sum_{e=1}^{\rho} f^e(\boldsymbol{a}^e)$. The mean reward of a group $e$ is denoted as $\mu_{\boldsymbol{a}^e} = \mathbb{E}[f^e(\boldsymbol{a}^e)]$. Consequently, the mean reward of a joint arm $\boldsymbol{a}$ is given by $\mu_{\boldsymbol{a}} = \sum_{e=1}^{\rho} \mu_{\boldsymbol{a}^e} = \mathbb{E}[f(\boldsymbol{a})]$. We assume the local rewards $f^e(\boldsymbol{a}^e)$ to be 1-subgaussian, i.e. $\mathbb{P}(|f^e(\boldsymbol{a}^e)| \geq \epsilon) \leq 2e^{-\epsilon^2}$. As a result, the global reward $f(\boldsymbol{a})$ is $\sqrt{\rho}$-subgaussian.

Our objective is to maximize the expected cumulative global rewards obtained over a horizon of $T$ rounds of interaction with the environment. Without loss of generality, we assume that $\mathbf{1}$ is the optimal joint arm that yields the highest expected global reward. It is important to note that the goal is defined based on the performance of the best joint arm. In other words, even if a local arm $\boldsymbol{a}^e$ has a high local reward, it may not be selected frequently by an optimal policy if it is not part of joint arms with high mean rewards. To quantify the performance of a bandit strategy, we use the concept of regret denoted by $R_T$, defined as the expected difference between the cumulative rewards obtained by always selecting the optimal joint arm $\mathbf{1}$ and the actual rewards obtained by following a specific strategy. Mathematically, the regret $R_T$ is given by:

$$R_T = \mathbb{E}\big[\textstyle\sum_{t=1}^{T}(\mu_{\mathbf{1}} - f(\boldsymbol{A}_t))\big] = \textstyle\sum_{t=1}^{T}(\mu_{\mathbf{1}} - \mu_{\boldsymbol{A}_t}), \quad (2.1)$$

where $\boldsymbol{A}_t$ represents the joint arm selected at round $t$. The regret captures the deviation from the cumulative rewards that would have been obtained if the optimal joint arm was chosen at each round. Minimizing regret is a key objective in designing effective strategies for the hypergraph MAMAB problem.

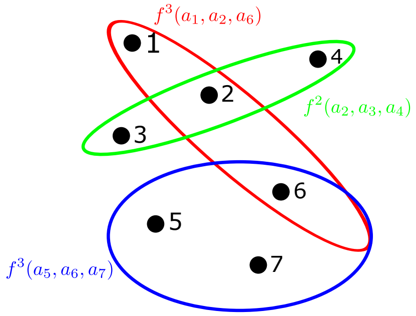


Figure 1: The hypergraph representation of a bandit environment with 8 agents and 3 groups. Each agent is represented by a vertex numbered by $\{1, 2, \ldots 7\}$ and each group is represented by a hyperedge. In this case, there are three hyperedges with each with size 3. Letting $a_i$ be the action taken by player $i$, the reward for the joint action $(a_1, \ldots, a_7)$ is decomposed as $f(a_1, a_2, \ldots, a_7) = f^1(a_1, a_2, a_6) + f^2(a_2, a_3, a_4) + f^3(a_5, a_6, a_7)$, where $a_i$ is the individual arm picked by agent $i$.

**Remark 2.1.** Although the results of our paper hold true regardless of the coordination hypergraph structure between the agents, they are most meaningful when the graph is sparse (i.e. the number of agents in each group is small). In particular, as there are $A$ joint arms, if one were to consider a different reward function for each arm, the regret and implementation complexity would be on the order of $A$. However, our results exploit the fact that there are only $A_{\text{loc}}$ local reward functions, and thus our regret bound is in terms of $A_{\text{loc}}$, which is much smaller than $A$ when the groups are small.

## 3 The $\epsilon$-Exploring Multi-Agent Thompson Sampling Algorithm

In this section, we present the $\epsilon$-exploring Multi-Agent Thompson Sampling Algorithm ($\epsilon$-`MATS`), whose pseudocode of $\epsilon$-`MATS`is displayed in Algorithm 1. $\epsilon$-`MATS` is a combination of the MATS algorithm (Verstraeten et al. 2020) and a greedy policy. The idea of adding a greedy policy to Thompson Sampling was initially proposed in Jin et al. (2022) and subsequently explored in Jin et al. (2023). In $\epsilon$-`MATS`, at each round $t \in [T]$, similar to MATS, Algorithm 1 maintains a posterior distribution $\mathcal{N}(\widehat{\mu}_{\boldsymbol{a}^e}(t), \frac{c}{n_{\boldsymbol{a}^e}(t)})$ for each local arm $\boldsymbol{a}^e$, $e \in [\rho]$, where $\widehat{\mu}_{\boldsymbol{a}^e}(t)$ is the average reward of arm $\boldsymbol{a}^e$, $n_{\boldsymbol{a}^e}(t)$ represents the number of pulls of arm $\boldsymbol{a}^e$, and $c$ is a scaling parameter. Both MATS and $\epsilon$-`MATS` maintain estimated rewards $\theta_{\boldsymbol{a}^e}(t)$ for each local arm, and select the joint arm that yields the highest sum of estimated local rewards, i.e., $\boldsymbol{A}_t = \arg\max_{\boldsymbol{a} \in \mathcal{A}} \sum_{e \in [\rho]} \theta_{\boldsymbol{a}^e}(t)$. After receiving the true rewards, the algorithms update the average reward $\widehat{\mu}_{\boldsymbol{a}^e}(t)$ and the number of pulls of $\boldsymbol{A}_t$ accordingly.

The difference between MATS and $\epsilon$-`MATS` lies in the way they construct the estimated rewards $\theta_{\boldsymbol{a}^e}(t)$ for each local arm. In particular, MATS samples $\theta_{\boldsymbol{a}^e}(t)$ from the respective posterior distribution for local arm $\boldsymbol{a}^e$. In contrast, the proposed $\epsilon$-`MATS` algorithm only samples $\theta_{\boldsymbol{a}^e}(t)$ from the posterior distribution with a probability of $\epsilon$, and it directly sets $\theta_{\boldsymbol{a}^e}(t) = \widehat{\mu}_{\boldsymbol{a}^e}(t)$, i.e., as the empirical mean reward. Here $\epsilon \in (0, 1]$ is a user-specified parameter that controls the level of exploration. For small values of $\epsilon$, $\epsilon$-MATS significantly reduces the level of exploration in MATS, which leads to improved computational efficiency.

**$\epsilon$-Exploring.** The idea of $\epsilon$-exploring is inspired from the recent work of Jin et al. (2023). We prove in the next section that $\epsilon$-MATS achieves the same order of finite-time regret bound as the MATS algorithm, even though it only needs to perform a small fraction of TS-type exploration. Furthermore, this algorithm runs faster since it doesn't have to sample each local arm from the Gaussian distribution as frequently as MATS. We also show that for specific applications the regret of $\epsilon$-MATS converges much faster than MATS and other algorithms in the same setting.

**Variable Elimination.** In Line 9 of Algorithm 1, $\epsilon$-MATS needs to find the joint arm $\boldsymbol{a}$ that maximizes the sum of the estimated local rewards. However, this step can be computationally expensive if naively implemented, as it would require considering all possible joint arms, resulting in a complexity of $O(K^m)$ since the joint space size is $K^m$. Following (Verstraeten et al. 2020), we use variable elimination (Guestrin, Koller, and Parr 2001) to reduce this computation burden. The key idea behind variable elimination is to optimize over one agent at a time instead of summing all estimated local rewards for each joint arm and then performing the maximization. By doing so, we can significantly reduce the computational burden. To explain how variable elimination works, let us rewrite the maximum sum of the local estimates as follows:

$$\max_{\boldsymbol{a}} f(\boldsymbol{a}) = \max_{\boldsymbol{a}} \sum_{e=1}^{\rho} f^e(\boldsymbol{a}^e) = \max_{\boldsymbol{a}} \Bigg[ \underbrace{\sum_{\boldsymbol{a}^e \in \boldsymbol{a} : a_m \notin \boldsymbol{a}^e} f^e(\boldsymbol{a}^e)}_{I_1}$$

$$+ \underbrace{\max_{\boldsymbol{a}^e : a_m \in \boldsymbol{a}^e} \sum_{\boldsymbol{a}^e \in \boldsymbol{a} : a_m \in \boldsymbol{a}^e} f^e(\boldsymbol{a}^e)}_{I_2} \Bigg], \tag{3.1}$$

where $a_m$ represents an individual arm of agent $m$. In Equation (3.1), we decompose the sum of the rewards into two cases based on the optimization variable $\boldsymbol{a}$. In $I_1$, we consider all the groups $\boldsymbol{a}^e$ that do not contain agent $m$. The maximization in this case is performed independently of the selection of individual arm $a_m$. Thus, the remaining agents can be optimized separately, resulting in a smaller optimization problem involving at most $m-1$ agents. In $I_2$, we focus on the groups that contain agent $m$. Here, we aim to find the individual arm $a_m$ that maximizes the sum of the local rewards for the joint arms containing $a_m$. This sum depends on the individual arms of the other agents that share a group with agent $m$. After determining the optimal $a_m$, the rest of the maximization is performed independently on the remaining agents in $I_1$. For more examples and details on variable elimination, please refer to (Guestrin, Koller, and Parr 2001).

We have the following result for variable elimination.

**Lemma 3.1.** Let $G_1, \ldots, G_\rho$ be the set of agents that belong to group $1, \ldots, \rho$ respectively. Then we have $A_{\text{loc}} = \sum_{e=1}^{\rho} \prod_{i \in G_e} |\mathcal{A}_i|$. At every round in Algorithm 1, following the above variable elimination procedure, the complexity of searching for the optimal arm is $O(A_{\text{loc}}) = O\left(\sum_{e=1}^{\rho} \prod_{i \in G_e} |\mathcal{A}_i|\right)$.

Algorithm 1: $\epsilon$-Exploring Multi-Agent Thompson Sampling

```
1:  **Input**: number of agents $m$, joint arm set $\times_{i=1}^{m} \mathcal{A}_i$, hyperparameters $c$ and $\epsilon$
2:  **for** $e \in [\rho], \boldsymbol{a}^e \in \mathcal{A}^e$ **do**
3:      Set $n_{\boldsymbol{a}^e}(1) = 0$ and $\widehat{\mu}_{\boldsymbol{a}^e}(1) = 0$
4:  **end for**
5:  **for** $t = 1, ..., T$ **do**
6:      **for** $e \in [\rho], \boldsymbol{a}^e \in \mathcal{A}^e$ **do**
7:          $\theta_{\boldsymbol{a}^e}(t) = \begin{cases} \sim \mathcal{N}(\widehat{\mu}_{\boldsymbol{a}^e}(t), \frac{c}{n_{\boldsymbol{a}^e}(t)+1}) & \text{w.p. } \epsilon \\ = \widehat{\mu}_{\boldsymbol{a}^e}(t) & \text{w.p. } 1-\epsilon \end{cases}$
8:      **end for**
9:      Pick $\boldsymbol{A}_t = \mathrm{argmax}_{\boldsymbol{a} \in \times_{i=1}^{m} \mathcal{A}_i} \sum_{e=1}^{\rho} \theta_{\boldsymbol{a}^e}(t)$
10:     Observe rewards $f^e(\boldsymbol{A}_t^e)$ for all $e \in [\rho]$
11:     **for** $e \in [\rho]$ **do**
12:         Update $\widehat{\mu}_{\boldsymbol{A}_t^e}(t) = \big(n_{\boldsymbol{A}_t^e}(t)\widehat{\mu}_{\boldsymbol{A}_t^e}(t) + f^e(\boldsymbol{A}_t^e)\big)/(n_{\boldsymbol{A}_t^e}(t)+1)$
13:         Set $n_{\boldsymbol{A}_t^e}(t) = n_{\boldsymbol{A}_t^e}(t) + 1$
14:     **end for**
15: **end for**
```

As we discussed, without variable elimination, one would naively add up all the estimated local rewards $\theta_{\boldsymbol{a}^e}$ for each joint arm $\boldsymbol{a}$ and find the joint arm with the largest posterior $\theta_{\boldsymbol{a}}$, leading to computational complexity in the order of $O(A) := O\left(\prod_{i=1}^{\rho} |\mathcal{A}_i|\right)$ at each round, which grows exponentially in the number of agents. In contrast, Lemma 3.1 indicates that by using variable elimination, $\epsilon$-MATS only needs $A_{\text{loc}}$ computation to find the joint arm with the largest estimated reward. Note that this theoretical guarantee is of independent interest to MATS as well since none was given in the original paper (Verstraeten et al. 2020).

## 4 Finite-Time Frequentist Regret Analysis

In this section, we present the proof of the frequentist regret bound for $\epsilon$-MATS.

### Finite-Time Frequentist Regret Bound of $\epsilon$-MATS

For convenience, we use $\Delta_{\boldsymbol{a}} = \mu_{\mathbf{1}} - \mu_{\boldsymbol{a}}$ to denote the suboptimality gap between joint arm $\boldsymbol{a}$ and the optimal joint arm. We let $\Delta_{\min} = \min_{\boldsymbol{a} \in \times_{i=1}^{m} \mathcal{A}_i \setminus \{\mathbf{1}\}} \Delta_{\boldsymbol{a}}$ and $\Delta_{\max} = \max_{\boldsymbol{a} \in \times_{i=1}^{m} \mathcal{A}_i} \Delta_{\boldsymbol{a}}$ be the minimum and maximum gap respectively. Moreover, let $\Delta_{\boldsymbol{a}^e} = \min\{\Delta_{\boldsymbol{a}} \mid \boldsymbol{a}^e \in \boldsymbol{a}\}$ be the minimum reward gap between joint arm $\boldsymbol{a}$ which contains $\boldsymbol{a}^e$ and $\mathbf{1}$. We present the regret of $\epsilon$-MATS as follows.

**Theorem 4.1.** Let $c = \log T$. The regret of $\epsilon$-MATS satisfies the following results.

1. There exists some universal constant $C_1$ such that

$$R_T \le C_1 (C_1/\epsilon)^{\rho} \rho^2 \log^2(T A_{\text{loc}})/\Delta_{\min}$$

$$+ C_1 \sum_{e\in[\rho]} \sum_{\boldsymbol{a}^e\in\mathcal{A}^e\setminus\{\mathbf{1}^e\}} \frac{\rho^2 \log^2(TA_{\mathrm{loc}})}{\Delta_{\boldsymbol{a}^e}} + C_1\Delta_{\max}.$$

2. There exists some universal constant $C_2$ such that

$$R_T \le C_2\Delta_{\max} + C_2\rho\sqrt{((C_2/\epsilon)^\rho + A_{\mathrm{loc}})\, T \log^2(TA_{\mathrm{loc}})}.$$

Note that when $\epsilon = 1$, $\epsilon$-MATS reduces to MATS, which gives the first frequentist regret bound of MATS. In particular, our bound is in the same order as the Bayesian regret bound (Theorem 1 (Verstraeten et al. 2020)) in terms of the order of $T$ and $A_{\mathrm{loc}}$. Compared with the Bayesian regret of (Verstraeten et al. 2020), our worst-case regret has an additional $\sqrt{\log T}$ factor because we inflate the variance of posterior distribution by $\log T$, which is a common trick in deriving the worst case regret bound of Thompson sampling (Agrawal and Goyal 2017; Jin et al. 2021a). The derivation of $(C_2/\epsilon)^\rho$ is provided in the following subsection.

## Technical Challenges in Frequentist Regret Analysis and the Proof Outline

For simplicity, this part assumes $\epsilon = 1$ (which reduces to MATS given in Verstraeten et al. (2020)). First, let's introduce some notations to simplify our discussion. We denote $S_r$ as the set of joint arms with gaps in the interval $(2^{-r}, 2^{-r+1}]$ and let $\delta_r = 2^{-(r+2)}$. The regret incurred by pulling the arms in $S_r$ is represented as $R(S_r)$. Furthermore, we define $S_r(t)$ as the set of joint arms not overestimated at time $t$, formally given by:

$$\begin{aligned} S_r(t) = \{&\boldsymbol{a} \mid \boldsymbol{a} \in S_r, \forall e \in [\rho] \\ &\text{and } \boldsymbol{a}^e \ne \mathbf{1}^e, \theta_{\boldsymbol{a}^e}(t) \le \mu_{\boldsymbol{a}^e} + \delta_r/\rho\}. \end{aligned}$$

The regret $R(S_r)$ can be expressed as

$$\begin{aligned} &R(S_r) \\ &\le 8\delta_r \cdot \big(\underbrace{\textstyle\sum_{t=1}^T \mathbb{1}\{\boldsymbol{A}_t \in S_r(t)\}}_{I_1} + \underbrace{\textstyle\sum_{t=1}^T \mathbb{1}\{\boldsymbol{A}_t \notin S_r(t)\}}_{I_2}\big). \end{aligned}$$

The term $I_2$ is relatively straightforward. It's important to note that whenever we pull the arm $\boldsymbol{A}_t \notin S_r(t)$, we inevitably pull a local arm whose posterior sample has not yet converged to its true mean. After sufficient pulls of each local arm (due to pull $\boldsymbol{A}_t$ with $\boldsymbol{A}_t \notin S_r(t)$), and by employing the maximal inequality for the reward distribution along with the concentration bound for the posterior distribution, we can demonstrate that the event $\boldsymbol{A}_t \notin S_r(t)$ occurs with an exceedingly small probability.

We now discuss how to bound term $I_1$ and the associated challenges. In the regret analysis for single agent Thompson Sampling (TS) by Agrawal and Goyal (2012); Jin et al. (2022), the term $I_1$ is bounded as follows.

$$\sum_{t=1}^T \mathbb{1}\{\boldsymbol{A}_t \in S_r(t)\} \le \mathbb{E}_{\hat\mu_{\mathbf{1},s}}\Big[1/\mathbb{P}(\theta_{\mathbf{1},s} \ge \mu_{\mathbf{1}} - \delta_r)\Big], \qquad (4.1)$$

where $\hat\mu_{\mathbf{1},s}$ is the empirical mean of arm $\mathbf{1}$ after being pulled $s$ times, $\theta_{\mathbf{1},s}$ is the posterior sample from $\mathcal{N}(\hat\mu_{\mathbf{1},s}, c\rho/s)$, and $1/\mathbb{P}(\theta_{\mathbf{1},s} \ge \mu_{\mathbf{1}} - \delta_r)$ represents the expected maximum number of posterior samples from $\mathcal{N}(\hat\mu_{\mathbf{1},s}, c\rho/s)$ such that one sample is larger than $\mu_{\mathbf{1}} - \delta_r$.

However, in a multi-agent setting, we can't decompose it in the same way due to two main reasons:

1. Since arm $\mathbf{1}$ might share some local arms with other joint arms, the number of pulls of each local arm of $\mathbf{1}$ could be different at time $t$. This contrasts with $\hat\mu_{\mathbf{1},s}$ in Equation (4.1), where we assume that arm 1 is pulled exactly $s$ times.
2. More importantly, while the samples of each local arm are independently drawn from their respective reward distributions, a dependency issue arises in the case of the joint arm. To explain, if each local arm $\mathbf{1}^e$ is pulled a fixed number of times, $N_e$, the mean reward of $\mathbf{1}$ follows the $\Big(\sqrt{\sum_{e\in[\rho]}(N_e)^{-1}}\Big)$-subgaussian with a mean of $\sum_{e\in[\rho]}\hat\mu_{\mathbf{1}^e,N_e}$. Leveraging the properties of subgaussian random variables, it can be demonstrated that the mean reward of joint arm $\mathbf{1}$ converges to its true mean as the number of pulls increases. However, this is not true when the pulls of local arms are history-dependent. In such cases, MATS is more likely to pull suboptimal arms that share overestimated local arms of $\mathbf{1}$ (the posterior samples from these local arms could surpass those from other underestimated local arms of $\mathbf{1}$). If this situation occurs, $\hat\mu_{\mathbf{1}}(t)$ is likely to be underestimated, making its distribution challenging to ascertain. Therefore, it will be difficult to derive the concentration results for $\hat\mu_{\mathbf{1}}(t)$ and consequently, the probability of $\theta_{\mathbf{1}}(t) \ge \mu_{\mathbf{1}} - \delta_r$ would also be hard to establish.

We solve the above issues by 1: carefully dividing $S_r$ into subsets, where the arms in each subset share the same local arms with $\mathbf{1}$ (total $2^\rho$ subsets); and 2: bounding term $I_1$ in local arms level. These two methods allow us to prove that

$$\begin{aligned} &\sum_{t=1}^T \mathbb{1}\{\boldsymbol{A}_t \in S_r(t)\} \\ &\le 2^\rho \sum_{s=1}^\tau \prod_{e=1}^\rho \mathbb{E}_{\hat\mu_{\mathbf{1}^e,s}}\left[1/\mathbb{P}(\theta_{\mathbf{1}^e,s} \ge \mu_{\mathbf{1}^e} - \delta_r/\rho)\right] + \Theta(1). \qquad (4.2) \end{aligned}$$

In the right hand of inequality, $2^\rho$ is due to the existence of $2^\rho$ subsets, and $\tau$ is defined as $\Theta(\rho^2(\log(TA_{\mathrm{loc}}))^2/(\delta_r)^2)$. The term $\Theta(1)$ exists because after each local arm is pulled more than $\tau$ times, the event $\mathbb{P}\left(\theta_{\mathbf{1}^e,s} \ge \mu_{\mathbf{1}^e} - \delta_r/\rho\right)$ is highly likely to occur. The cost for the non-occurrence of this event can be bounded by $\Theta(1)$. Follow Agrawal and Goyal (2012); Jin et al. (2022), one can show that

$$\prod_{e=1}^\rho \mathbb{E}_{\hat\mu_{\mathbf{1}^e,s}}\left[1/\mathbb{P}\left(\theta_{\mathbf{1}^e,s} \ge \mu_{\mathbf{1}^e} - \frac{\delta_r}{\rho}\right)\right] \le \big(\frac{\rho\sqrt{\log T}}{\delta_r}\big)^{2\rho}.$$

The above results are underwhelming, particularly in regards to the worst-case regret, which is $\tilde O(T^{2\rho/(2\rho-1)})$. In order to enhance these outcomes, we are introducing two innovative techniques:

1. First, deriving from the concentration bound, we obtain that with high probability $\hat\mu_{\mathbf{1}^e,s} \ge \mu_{\mathbf{1}^e} - \sqrt{2\log T/s}$. Instead of considering the expectation over the entire real line for $\hat\mu_{\mathbf{1}^e,s}$, we confine $\hat\mu_{\mathbf{1}^e,s}$ to the interval $(\mu_{\mathbf{1}^e} - \sqrt{2\log T/s}, \infty)$.
2. Secondly, we marginally increase the variance of the posterior distribution by $\log T$. According to the anti-concentration bound of the Gaussian posterior, the likelihood of $\theta_{\mathbf{1}^e,s}$ exceeding $\hat\mu_{\mathbf{1}^e,s} + \sqrt{2\log T/s}$ remains constant. In conjunction with the condition $\hat\mu_{\mathbf{1}^e,s} \ge \mu_{\mathbf{1}^e} -$

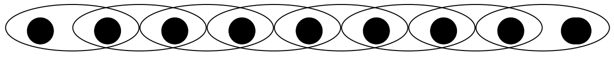


Figure 2: The hypergraph on a 0101-Chain with 10 agents. Each hyperedge (a group of two agents) is denoted by a black oval.

$\sqrt{2\log T/s}$, we can ascertain that $\mathbb{P}(\theta_{\mathbf{1}^e,s} \geq \mu_{\mathbf{1}^e})$ is also a constant.

With the above two methods, we can prove $I_1 \leq \tau \cdot C^\rho$, where $C$ is some constant. Finally, by summing over all possible values of $r$ (i.e., $\sum_r R(S_r)$), we derive our regret bound.

### Lower Bound on the Worst-Case Regret Bound

We now present some lower-bound results on the worst-case regret in our setting. The first theorem states a lower bound in terms of the horizon length $T$ and the total number of local arms across all groups $A_{\text{loc}}$ when $\rho > 0$ is treated as a fixed constant.

**Theorem 4.2.** For any policy $\pi$, there exists a mean vector $\mu \in [0,1]^{A_{\text{loc}}}$ (where each component corresponds to the mean of a local arm) such that $R_n(\pi, \nu_\mu) = \Omega(\sqrt{A_{\text{loc}}T/\rho})$.

Recall that our worst-case regret bound in Theorem 4.1 is $\tilde{O}(\sqrt{C^\rho A_{\text{loc}}T})$, with $C$ representing a universal constant. According to Theorem 4.2, when the number of groups $\rho$ is constant, indicating a sparse hypergraph, our worst-case regret for $\epsilon$-MATS is nearly optimal up to constant and logarithmic terms.

The next theorem shows the worst possible dependence of the regret bound of $\epsilon$-MATS on the number of groups, i.e., $\rho$.

**Theorem 4.3.** For $c = \epsilon = 1$, there is a bandit instance such that the regret of Algorithm 1 is $\Omega(C^\rho)$, where $C > 1$ is some constant.

Theorem 4.3 shows that $C^\rho$ regret is unavoidable for original Multi-Agent Thompson Sampling, which further proves the optimality of our regret bound $\tilde{O}(\sqrt{C^\rho A_{\text{loc}}T})$.

## 5 Experiments

In this section, we evaluate the proposed $\epsilon$-MATS algorithm on several benchmark MAMAB problems including Bernoulli 0101-Chain, Poisson 0101-Chain, and Gem Mining (Roijers, Whiteson, and Oliehoek 2015; Bargiacchi et al. 2018; Verstraeten et al. 2020). We compare $\epsilon$-MATS with the vanilla MATS (Verstraeten et al. 2020), MAUCE (Bargiacchi et al. 2018), and the random policy. We also provide a thorough ablation study of $\epsilon$-MATS to find the optimal tradeoff between greedy and Thompson sampling exploration in practice. We run all our experiments on Nvidia RTX A5000 with 24GB RAM and each experiment setting is averaged over 50 trials. Please refer to Appendix G for detailed experimental setup, ablation studies, and more experimental results.

### Bernoulli and Poisson 0101-Chain

In this subsection, we conduct experiments on the Bernoulli and Poisson 0101-Chain problems, which are commonly studied in the MAMAB literature (Bargiacchi et al. 2018; Verstraeten et al. 2020). An illustration is provided in Figure 2, where the agents are positioned along a 1-dimensional path forming a graph. Specifically, the graph consists of $m$ agents and $m-(d-1)$ edges (or groups), where $d$ is the number of agents within each hyperedge. The agents $i$ to $i+d-1$ in the group $i$ are connected to a local reward function $f^i(a_i, a_{i+1}, ..., a_{i+d-1})$, where $a_i$ denotes the individual arm of agent $i$. Each agent can has two arms: 0 and 1.

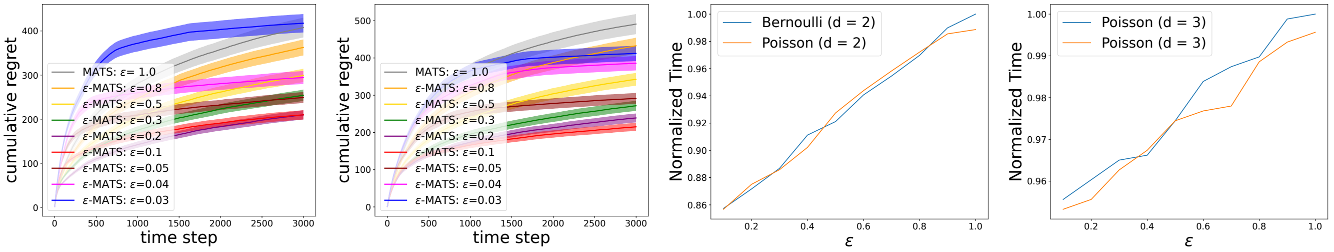
We consider two settings where $d = 2$ and $d = 3$ respectively. For each setting, we conduct experiments for two types of reward distributions (Bernoulli and Poisson), which results in the **Bernoulli 0101-Chain** problem and the **Poisson 0101-Chain** problem respectively. Due to the space limit, we defer the details of the reward generation to Appendix G.

We first perform an ablation study to show the effect of different $\epsilon$ on the performance of $\epsilon$-MATS. The results are presented in Figures 3(a) and 3(b). It can be seen that with $\epsilon$ decreasing from 1 (this corresponds to the MATS algorithm) to $0.1$, the cumulative regret of $\epsilon$-MATS also becomes lower. When $\epsilon$ gets smaller than $0.1$, the regret rapidly increases due to insufficient exploration. In the rest of the experiments in this subsection, we fix $\epsilon = 0.1$ for the best performance. We also compare the runtime complexity of $\epsilon$-MATS for different $\epsilon$, which is presented in Figures 3(c) and 3(d) for the setting $d = 2$ and $d = 3$ respectively. In particular, we calculate the ratio between the runtime of $\epsilon$-MATS and MATS ($\epsilon = 1$) for running 1000 iterations. Figure 3(c) shows that lower value of $\epsilon$ decreases the runtime complexity of $\epsilon$-MATS in both Bernoulli 0101 and Poisson 0101 problems. Comparing Figure 3(c) and Figure 3(d), we can see that the computational efficiency is adversely affected by the size of each group. In Figure 4, we compare the regret of $\epsilon$-MATS with MATS, MAUCE, and Random for both Bernoulli 0101 and Poisson 0101 tasks, which demonstrate that $\epsilon$-MATS can significantly outperform baseline methods.

## 6 Conclusion and Future Work

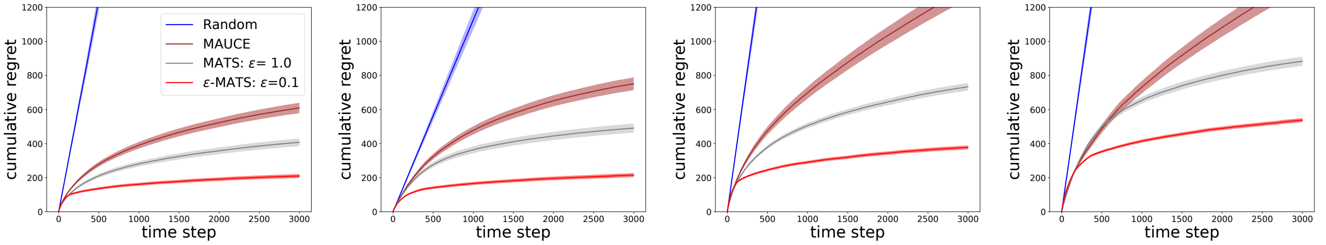
In this paper, we studied the problem of multi-agent multi-armed bandits. We proposed $\epsilon$-MATS which combines the MATS exploration with probability $\epsilon$ and greedy exploitation with probability $1-\epsilon$. We provided a frequentist finite time regret bound for $\epsilon$-MATS, which is in the order of $\tilde{O}(\sqrt{C^\rho A_{\text{loc}}T})$. When $\epsilon = 1$, our result yields the first frequentist regret bound for MATS in the coordination hypergraph setting (Verstraeten et al. 2020). We also derived a lower bound for this environment in the order of $\Omega(\sqrt{A_{\text{loc}}T/\rho})$, implying $\epsilon$-MATS is near optimal when $\rho$ is assumed to be small, i.e., the coordination hypergraph is sparse. Empirical evaluations demonstrate the superior performance of $\epsilon$-MATS compared with existing algorithms for MAMAB problems with a coordination hypergraph.

Our experimental findings present a notable observation: the performance of $\epsilon$-MATS frequently surpasses that of MATS. Nevertheless, the regret bound presented in our main theorem suggests that MATS has a slightly better worst-



(a) Bernoulli 0101 (b) Poisson0101 (c) Bernoulli 0101 and Poisson0101 with $d = 2$ (d) Bernoulli 0101 and Poisson0101 with $d = 3$

Figure 3: Ablation study on $\epsilon$-MATS. (a) and (b): Regret performance ($m = 10$, and $d = 2$) with different $\epsilon$ in Bernoulli 0101 and Poisson 0101 tasks. Note when $\epsilon = 1.0$, $\epsilon$-MATS reduces to MATS. (c) and (d): The relative computational time of $\epsilon$-MATS with different $\epsilon$ compared with MATS.



(a) Bernoulli 0101: $m = 10$, $d = 2$ (b) Poisson 0101: $m = 10$, $d = 2$ (c) Bernoulli 0101: $m = 10$, $d = 3$ (d) Poisson 0101: $m = 10$, $d = 3$

Figure 4: Regret performance compared with other algorithm baselines in Bernoulli 0101 and Poisson 0101 with different agents in a group ($d = 2$ or $d = 3$).

case regret bound compared to $\epsilon$-MATS. This discrepancy offers a compelling avenue for future research to explore the potential for $\epsilon$-MATS to achieve the same regret bound as MATS. Additionally, while this paper focuses on the worst-case regret bound, it would be intriguing to investigate if $\epsilon$-MATS could exhibit a more favorable regret bound in some easier bandit instances, leading to the analysis of instance-dependent regret bounds. Finally, it would be intriguing to investigate the potential of applying the core concepts of coordination hypergraph and epsilon-exploring in our paper to enhance Thompson sampling-based algorithms within more complex settings, such as linear bandits, neural contextual bandits, and Markov decision processes (Xu et al. 2022b,a; Zhang et al. 2020; Ishfaq et al. 2023).

## Broader Impact

This work has the potential to positively influence practical implementations in algorithms for multi-agent decision systems such as wind farm management. The algorithm demonstrates effective handling of the exponentially large joint action space, which could be leveraged to mitigate complexity in other combinatorial problems. Thus the insight of the proposed algorithm might help reduce the human resource and computational resources in such problems.

## Acknowledgements

We thank the anonymous reviewers for their helpful comments. This research is supported by the National Research Foundation, Singapore under its AI Singapore Program (AISG Award No: AISG-PhD/2021-01-004[T]), by Singapore Ministry of Education Academic Research Fund Tier 3 under MOEs official grant number MOE2017-T3-1-007, and by the Whitehead Scholars Program, and by the US National Science Foundation Award 2323112. In particular, T. Jin is supported by the National Research Foundation, Singapore under its AI Singapore Program (AISG Award No: AISG-PhD/2021-01-004[T]) and by Singapore Ministry of Education Academic Research Fund Tier 3 under MOEs official grant number MOE2017-T3-1-007. P. Xu is supported by the Whitehead Scholars Program and the US National Science Foundation Award 2323112. The views and conclusions in this paper are those of the authors and should not be interpreted as representing any funding agencies.

## References

Abramowitz, M.; and Stegun, I. A. 1964. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, volume 55. US Government printing office.

Agarwal, M.; Aggarwal, V.; and Azizzadenesheli, K. 2022. Multi-Agent Multi-Armed Bandits with Limited Communication. *Journal of Machine Learning Research*, 23(212): 1–24.

Agrawal, S.; and Goyal, N. 2012. Analysis of Thompson Sampling for the Multi-Armed Bandit Problem. In *Confer-*

*ence on Learning Theory*, 39–1. JMLR Workshop and Conference Proceedings.

Agrawal, S.; and Goyal, N. 2017. Near-Optimal Regret Bounds for Thompson Sampling. *Journal of the ACM (JACM)*, 64(5): 1–24.

Auer, P.; Cesa-Bianchi, N.; Freund, Y.; and Schapire, R. E. 2002. The Nonstochastic Multiarmed Bandit Problem. *SIAM journal on computing*, 32(1): 48–77.

Bargiacchi, E.; Verstraeten, T.; Roijers, D.; Nowé, A.; and Hasselt, H. 2018. Learning to Coordinate with Coordination Graphs in Repeated Single-Stage Multi-Agent Decision Problems. In *International Conference on Machine Learning*, 482–490. PMLR.

Besson, L.; and Kaufmann, E. 2018. Multi-Player Bandits Revisited. In *Algorithmic Learning Theory*, 56–92. PMLR.

Boursier, E.; and Perchet, V. 2020. Selfish Robustness and Equilibria in Multi-Player Bandits. In *Conference on Learning Theory*, 530–581. PMLR.

Chang, W.; Jafarnia-Jahromi, M.; and Jain, R. 2022. Online Learning for Cooperative Multi-Player Multi-Armed Bandits. In *IEEE Conference on Decision and Control (CDC)*, 7248–7253. IEEE.

Chapelle, O.; and Li, L. 2011. An Empirical Evaluation of Thompson Sampling. In *Advances in Neural Information Processing Systems*, 2249–2257.

Claes, D.; Oliehoek, F.; Baier, H.; and Tuyls, K. 2017. Decentralised Online Planning for Multi-Robot Warehouse Commissioning. In *Conference on Autonomous Agents and MultiAgent Systems*, 492–500.

De Hauwere, Y.-M.; Vrancx, P.; and Nowé, A. 2010. Learning Multi-Agent State Space Representations. In *International Conference on Autonomous Agents and Multiagent Systems: Volume 1-Volume 1*, 715–722.

Deng, Y.; Zhang, R.; Xu, P.; Ma, J.; and Gu, Q. 2023. PhyGCN: Pre-trained Hypergraph Convolutional Neural Networks with Self-supervised Learning. *bioRxiv*, 2023–10.

Elahi, S.; Atalar, B.; Öğüt, S.; and Tekin, C. 2021. Contextual Combinatorial Volatile Bandits with Satisfying via Gaussian Processes. *arXiv preprint arXiv:2111.14778*.

Gebraad, P. M.; and van Wingerden, J.-W. 2015. Maximum Power-Point Tracking Control for Wind Farms. *Wind Energy*, 18(3): 429–447.

Gentile, C.; Li, S.; and Zappella, G. 2014. Online Clustering of Bandits. In *International Conference on Machine Learning*, 757–765. PMLR.

Guestrin, C.; Koller, D.; and Parr, R. 2001. Multiagent Planning with Factored MDPs. *Advances in neural information processing systems*, 14.

Gupta, S.; Chaudhari, S.; Joshi, G.; and Yağan, O. 2021. Multi-Armed Bandits with Correlated Arms. *IEEE Transactions on Information Theory*, 67(10): 6711–6732.

Han, B.; and Arndt, C. 2021. Budget Allocation as a Multi-Agent System of Contextual & Continuous Bandits. In *ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2937–2945.

Hayes, C. F.; Verstraeten, T.; Roijers, D. M.; Howley, E.; and Mannion, P. 2022. Multi-Objective Coordination Graphs for the Expected Scalarised Returns with Generative Flow Models. *arXiv preprint arXiv:2207.00368*.

Huang, W.; Combes, R.; and Trinh, C. 2022. Towards Optimal Algorithms for Multi-Player Bandits without Collision Sensing Information. In *Conference on Learning Theory*, 1990–2012. PMLR.

Ishfaq, H.; Lan, Q.; Xu, P.; Mahmood, A. R.; Precup, D.; Anandkumar, A.; and Azizzadenesheli, K. 2023. Provable and Practical: Efficient Exploration in Reinforcement Learning via Langevin Monte Carlo. *arXiv preprint arXiv:2305.18246*.

Jin, T.; Xu, P.; Shi, J.; Xiao, X.; and Gu, Q. 2021a. MOTS: Minimax Optimal Thompson Sampling. In *International Conference on Machine Learning*, 5074–5083. PMLR.

Jin, T.; Xu, P.; Xiao, X.; and Anandkumar, A. 2022. Finite-Time Regret of Thompson Sampling Algorithms for Exponential Family Multi-Armed Bandits. *Advances in Neural Information Processing Systems*, 35: 38475–38487.

Jin, T.; Xu, P.; Xiao, X.; and Gu, Q. 2021b. Double Explore-Then-Commit: Asymptotic Optimality and Beyond. In *Conference on Learning Theory*, 2584–2633. PMLR.

Jin, T.; Yang, X.; Xiao, X.; and Xu, P. 2023. Thompson Sampling with Less Exploration is Fast and Optimal. In *International Conference on Machine Learning*, 15239–15261. PMLR.

Kaufmann, E.; Korda, N.; and Munos, R. 2012. Thompson Sampling: An Asymptotically Optimal Finite-Time Analysis. In *Algorithmic Learning Theory*, 199–213. Springer.

Kocák, T.; and Garivier, A. 2021. Best Arm Identification in Spectral Bandits. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, 2220–2226.

Kok, J. R.; and Vlassis, N. 2006. Collaborative Multiagent Reinforcement Learning by Payoff Propagation. *Journal of Machine Learning Research*, 7: 1789–1828.

Korda, N.; Kaufmann, E.; and Munos, R. 2013. Thompson Sampling for 1-Dimensional Exponential Family Bandits. *Advances in neural information processing systems*, 26.

Landgren, P.; Srivastava, V.; and Leonard, N. E. 2016. Distributed Cooperative Decision-Making in Multiarmed Bandits: Frequentist and Bayesian Algorithms. In *IEEE Conference on Decision and Control (CDC)*, 167–172. IEEE.

Li, S.; Karatzoglou, A.; and Gentile, C. 2016. Collaborative Filtering Bandits. In *ACM SIGIR Conference on Research and Development in Information Retrieval*, 539–548.

Lugosi, G.; and Mehrabian, A. 2022. Multiplayer Bandits without Observing Collision Information. *Mathematics of Operations Research*, 47(2): 1247–1265.

Ma, Y.; Huang, T.-K.; and Schneider, J. G. 2015. Active Search and Bandits on Graphs Using Sigma-Optimality. In *Uncertainty in Artificial Intelligence*, volume 542, 551.

Magesh, A.; and Veeravalli, V. V. 2019. Multi-User MABs with User Dependent Rewards for Uncoordinated Spectrum

Access. In *Asilomar Conference on Signals, Systems, and Computers*, 969–972. IEEE.

Mehrabian, A.; Boursier, E.; Kaufmann, E.; and Perchet, V. 2020. A Practical Algorithm for Multiplayer Bandits When Arm Means Vary among Players. In *International Conference on Artificial Intelligence and Statistics*, 1211–1221. PMLR.

Nayyar, N.; Kalathil, D.; and Jain, R. 2016. On Regret-Optimal Learning in Decentralized Multiplayer Multiarmed Bandits. *IEEE Transactions on Control of Network Systems*, 5(1): 597–606.

Pal, S.; Suggala, A. S.; Shanmugam, K.; and Jain, P. 2023. Optimal Algorithms for Latent Bandits with Cluster Structure. *arXiv preprint arXiv:2301.07040*.

Roijers, D. M.; Whiteson, S.; and Oliehoek, F. A. 2015. Computing Convex Coverage Sets for Faster Multi-Objective Coordination. *Journal of Artificial Intelligence Research*, 52: 399–443.

Sankararaman, A.; Ganesh, A.; and Shakkottai, S. 2019. Social Learning in Multi Agent Multi Armed Bandits. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 3(3): 1–35.

Scharpff, J.; Roijers, D.; Oliehoek, F.; Spaan, M.; and de Weerdt, M. 2016. Solving Transition-Independent Multi-Agent MDPs with Sparse Interactions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Shi, C.; Xiong, W.; Shen, C.; and Yang, J. 2020. Decentralized Multi-Player Multi-Armed Bandits with No Collision Information. In *International Conference on Artificial Intelligence and Statistics*, 1519–1528. PMLR.

Shi, C.; Xiong, W.; Shen, C.; and Yang, J. 2021. Heterogeneous Multi-Player Multi-Armed Bandits: Closing the Gap and Generalization. *Advances in Neural Information Processing Systems*, 34: 22392–22404.

Stranders, R.; Tran-Thanh, L.; Fave, F. M. D.; Rogers, A.; and Jennings, N. R. 2012. DCOPs and Bandits: Exploration and Exploitation in Decentralised Coordination. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, 289–296.

Szorenyi, B.; Busa-Fekete, R.; Hegedus, I.; Ormándi, R.; Jelasity, M.; and Kégl, B. 2013. Gossip-Based Distributed Stochastic Bandit Algorithms. In *International Conference on Machine Learning*, 19–27. PMLR.

Thaker, P.; Malu, M.; Rao, N.; and Dasarathy, G. 2022. Maximizing and Satisficing in Multi-Armed Bandits with Graph Information. *Advances in Neural Information Processing Systems*, 35: 2019–2032.

Thompson, W. R. 1933. On the Likelihood That One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika*, 25(3-4): 285–294.

Valko, M.; Munos, R.; Kveton, B.; and Kocák, T. 2014. Spectral Bandits for Smooth Graph Functions. In *International Conference on Machine Learning*, 46–54. PMLR.

Verstraeten, T.; Bargiacchi, E.; Libin, P. J.; Helsen, J.; Roijers, D. M.; and Nowé, A. 2020. Multi-Agent Thompson Sampling for Bandit Applications with Sparse Neighbourhood Structures. *Scientific reports*, 10(1): 1–13.

Verstraeten, T.; Daems, P.-J.; Bargiacchi, E.; Roijers, D. M.; Libin, P. J.; and Helsen, J. 2021. Scalable Optimization for Wind Farm Control Using Coordination Graphs. In *International Conference on Autonomous Agents and MultiAgent Systems*, 1362–1370.

Verstraeten, T.; Nowé, A.; Keller, J.; Guo, Y.; Sheng, S.; and Helsen, J. 2019. Fleetwide Data-Enabled Reliability Improvement of Wind Turbines. *Renewable and Sustainable Energy Reviews*, 109: 428–437.

Wang, P.-A.; Proutiere, A.; Ariu, K.; Jedra, Y.; and Russo, A. 2020a. Optimal Algorithms for Multiplayer Multi-Armed Bandits. In *International Conference on Artificial Intelligence and Statistics*, 4120–4129. PMLR.

Wang, Y.; Hu, J.; Chen, X.; and Wang, L. 2020b. Distributed Bandit Learning: Near-optimal Regret with Efficient Communication. In *International Conference on Learning Representations*.

Wei, L.; and Srivastava, V. 2018. On Distributed Multi-Player Multiarmed Bandit Problems in Abruptly Changing Environment. In *IEEE Conference on Decision and Control (CDC)*, 5783–5788. IEEE.

Wiering, M. A.; et al. 2000. Multi-Agent Reinforcement Learning for Traffic Light Control. In *International Conference on Machine Learning*, 1151–1158.

Xu, P.; Wen, Z.; Zhao, H.; and Gu, Q. 2022a. Neural Contextual Bandits with Deep Representation and Shallow Exploration. In *International Conference on Learning Representations*.

Xu, P.; Zheng, H.; Mazumdar, E. V.; Azizzadenesheli, K.; and Anandkumar, A. 2022b. Langevin monte carlo for contextual bandits. In *International Conference on Machine Learning*, 24830–24850. PMLR.

Yang, K.; Toni, L.; and Dong, X. 2020. Laplacian-Regularized Graph Bandits: Algorithms and Theoretical Analysis. In *International Conference on Artificial Intelligence and Statistics*, 3133–3143. PMLR.

Zhang, W.; Zhou, D.; Li, L.; and Gu, Q. 2020. Neural Thompson Sampling. In *International Conference on Learning Representations*.

Zhang, Y.; Qu, G.; Xu, P.; Lin, Y.; Chen, Z.; and Wierman, A. 2023. Global convergence of localized policy iteration in networked multi-agent reinforcement learning. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 7(1): 1–51.

Zhu, Z.; Zhu, J.; Liu, J.; and Liu, Y. 2021. Federated Bandit: A Gossiping Approach. In *Abstract Proceedings of the 2021 ACM SIGMETRICS/International Conference on Measurement and Modeling of Computer Systems*, 3–4.