Multi-player Multi-armed Bandits with Delayed Feedback

Jingqi Fan[1] , **Zilong Wang**[2] , **Shuai Li**[2*] , **Linghe Kong**[2]
[1]Northeastern University, China
[2]Shanghai Jiao Tong University
fanjingqi@stumail.neu.edu.cn, {wangzilong, shuaili8, linghe.kong}@sjtu.edu.cn

Abstract

Multi-player multi-armed bandits (MP-MAB) have been extensively studied due to their application in cognitive radio networks. In this setting, multiple players simultaneously select arms and instantly receive feedback. However, in realistic decentralized networks, feedback is often delayed due to sensing latency and signal processing. Without a central coordinator, explicit communication is impossible, and delayed feedback disrupts implicit coordination, since it depends on synchronous observations. As a result, collisions are frequent and system performance degrades significantly. In this paper, we propose an algorithm in MP-MAB with stochastic delay feedback. Each player in the algorithm independently maintains an estimate of the optimal arm set based on their own delayed rewards but only pulls arms from the set, which is, with high probability, identical to those of other players, thus avoiding collisions. The identical arm set also enables implicit communication, allowing players to utilize the exploration results of others. We establish a regret upper bound and derive a lower bound to prove the algorithm is near-optimal. Numerical experiments on both synthetic and real-world datasets validate the effectiveness of our algorithm.

1 Introduction

Multi-armed Bandits (MAB) is a classic framework widely applied in diverse fields such as online advertising, clinical trials, and recommendation systems [Lattimore and Szepesvári, 2020]. In this framework, at each time s, a single player sequentially selects an arm k from a finite set $[K] := \{1, ..., K\}$ and receives a random reward $X_k(s)$. However, in many real-world scenarios involving multiple users, the standard MAB framework may fail to capture the complexities. For instance, in cognitive radio systems designed to efficiently share spectrum resources among users, a key distinction from the traditional MAB problem is that when multiple users select the same channel, they collide and no message is transmitted. This situation motivates multi-player multi-armed bandits (MP-MAB) framework in which M players simultaneously pull arms. If two or more players pull the same arm, their rewards turn to zero which represents failed transmission.

A problem is referred to as centralized when players can freely communicate about their actions and rewards through a central coordinator at each time step. In this scenario, players coordinate their actions using shared information without cost [Komiyama *et al.*, 2015]. Yet the centralized setting incurs significant energy costs in cognitive radio networks. To address this limitation, recent research has shifted focus to the decentralized MP-MAB, where each player has access only to her own information, and explicit communication is prohibited. Many studies develop implicit communication by intentionally inducing collisions to pass binary information [Boursier and Perchet, 2019; Wang *et al.*, 2020; Shi *et al.*, 2020]. The mechanism enables players to fully utilize the information of others and achieve performance comparable to the centralized setting, but it relies on synchronous feedback from all players.

However, in practical networks, a more realistic scenario involves users experiencing delay due to various inherent factors, including spectrum analysis, heterogeneous link layer protocols, path loss, and wireless link errors [Akyildiz *et al.*, 2006; Ahmad *et al.*, 2020]. Although the MP-MAB problem has been well studied, most existing approaches struggle in the presence of delayed feedback [Shi *et al.*, 2021; Huang *et al.*, 2022]. In decentralized settings, players rely on immediate feedback to perform implicit communication, and delays break this mechanism, leading to frequent collisions and poor exploration efficiency.

While single-player MAB with delayed feedback has been extensively studied, these approaches cannot be directly applied to multi-player settings, as they may lead to competition and frequent collisions among players. Specifically, in single-player MAB, a player selects an arm but observes the reward only after a period of delay [Joulani *et al.*, 2013; Lancewicki *et al.*, 2021; Tang *et al.*, 2024]. However, when multiple players are involved, they tend to compete for the same optimal arm, leading to frequent collisions. Also note that, in single-player bandits, although delayed feedback primarily affects the timeliness of updates, each feedback is substitutable over time. Specifically, if one is delayed, subse-

*Corresponding author.

Setting	Algorithm	Regret bound
Centralized lower bound		$\Omega\left(\sum_{k>M}\frac{\log(T)}{\theta\Delta_{M,k}}+\frac{M\sum_{k>M}\Delta_{M,k}}{K}\tilde{d}_1-\frac{2}{\theta}\right)$
Centralized	DDSE	$O\left(\sum_{k>M}\frac{\log(T)}{\theta\Delta_{M,k}}+\frac{\tilde{d}_2}{\theta}+\frac{M\sum_{k>M}\Delta_{1,k}}{K-M}\mathbb{E}[d]\right)$
Decentralized	DDSE	$O\left(\sum_{k>M}\frac{\log(T)}{\theta\Delta_{M,k}}+\frac{\tilde{d}_2}{\theta}+\frac{M\sum_{k>M}\Delta_{1,k}}{K-M}\tilde{d}_3\right)$
Decentralized	DSE[1]	$O\left(\sum_{k>M}\frac{\log(T)}{\theta\Delta_{M,k}}+\frac{\tilde{d}_2\tilde{d}_3}{\theta KM}+\frac{\tilde{d}_3}{\theta KM\sum_{k>M}\Delta_{M,k}^2}+\exp(\frac{\mathbb{E}[d]}{KM}+\frac{\sigma_d^2}{K^2M^2})\right)$

Table 1: Comparison of lower bound and upper bounds of algorithms. The first row comes from Theorem 1. The second row is derived from Corollary 1, the third row is based on Theorem 2, and the last row comes from Theorem 3. Define $\tilde{d}_1 := \mathbb{E}[d]-\sqrt{\sigma_d^2\theta/(1-\theta)}$, $\tilde{d}_2 := \mathbb{E}[d]+\sqrt{\sigma_d^2\log(1/(1-\theta))}$ and $\tilde{d}_3 := \mathbb{E}[d]+\sqrt{\sigma_d^2\log(K)}$, where $\theta\in(0,1)$ is a quantile of delay distribution. σ_d^2 is the sub-Gaussian parameter of delay distribution and $\mathbb{E}[d]$ is the expectation. We also define $\Delta_{\ell,k} := \mu_{(\ell)}-\mu_{(k)}$.

quent feedback can still guide learning. In contrast, in the multi-player setting, some feedback is irreplaceable: missing it may lead to miscoordination and persistent collisions. These collisions are particularly harmful, as players pulling the same arm receive no reward, resulting in significant regret. Therefore, a key challenge for MP-MAB with delayed feedback lies in maintaining coordination even with incomplete or delayed feedback. In this paper, we address this challenge, and our contributions are presented below.

1.1 Contribution

Motivated by the pressing challenge of delay in cognitive radio networks, we propose a novel bandit framework of multi-player multi-armed bandits with stochastic delay which follows a σ^2-sub-Gaussian distribution with expectation $\mathbb{E}[d]$. For this problem, we introduce an algorithm DDSE (Delayed Decentralized Successive Elimination), where every player maintains a set of arms with the highest empirical rewards at each time. They estimate the delay and choose a set of arms updated from an earlier time slot, which is considered to be consistent with the choices of all players with high probability. By selecting arms from the identical arm set, players avoid collisions and mitigate the impact of delay. They also engage in implicit communication with other players by pulling arms from the identical arm set, thereby utilizing the exploration results of others. Therefore, this approach efficiently addresses the challenge outlined in Section 1.

Table 1 compares the regret bound of our algorithm with DSE (Decentralized Successive Elimination), which is a simplified version of DDSE. In DSE, players directly select arms in the latest updated set of arms with the highest empirical rewards, leading to a regret of $O(\exp(\mathbb{E}[d]/KM+\sigma_d^2/K^2M^2))$ which grows exponentially with increasing $\mathbb{E}[d]$ and σ_d^2. Through careful algorithm design, DDSE successfully prevents this exponential term. The term $O(\frac{M\sum_{k>M}\Delta_{1,k}}{K-M}\tilde{d}_3)$ in DDSE is the regret that players coordinate with each other to select the same set of best empirical arms. Compared with $O(\frac{M\sum_{k>M}\Delta_{1,k}}{K-M}\mathbb{E}[d])$ in the centralized upper bound, the regret of our algorithm in the decentralized setting differs by only $O(\frac{M\sum_{k>M}\Delta_{1,k}}{K-M}\sqrt{\sigma_d^2\log(K)})$, which diminishes when the delay remains stable, i.e., σ_d^2 approaches zero. Additionally, we establish a lower bound in Table 1, demonstrating that our regret bound is near-optimal. We also extensively evaluate DDSE through a large number of experiments, validating its effectiveness across various delayed feedback scenarios.

[1]Simplified version of DDSE. In this algorithm, players do not estimate delay and use the latest set of arms.

2 Related Work

The problem of multi-player multi-armed bandits has recently been studied in different settings in the existing literature, where most of the efforts have concentrated on the decentralized setting. Boursier and Perchet [2019] propose an implicit communication mechanism where players intentionally collide to signal information, achieving performance comparable to centralized approaches. Wang *et al.* [2020] improve this communication phase by electing a leader and only allowing the leader to communicate with followers. Research also has focused on heterogeneous reward settings [Besson and Kaufmann, 2018; Bistritz and Leshem, 2018; Tibrewal *et al.*, 2019; Shi *et al.*, 2021], adversarial collision scenarios [Mahesh *et al.*, 2022], incomplete feedback [Boursier and Perchet, 2019; Shi *et al.*, 2020; Lugosi and Mehrabian, 2022; Huang *et al.*, 2022], and shareable arms Wang *et al.*; Xu *et al.* [2022; 2023]. Recently, Richard *et al.* [2024] consider asynchronous multi-player bandits in the centralized setting and derive a constant or logarithmic regret.

There has been growing interest in stochastic delay in multi-armed bandits. Vernade *et al.* [2017] investigate delayed Bernoulli bandits, although their approach requires knowledge of the delay distribution. Pike-Burke *et al.* [2018] consider scenarios where a sum of observations is received after some stochastic delay. Zhou *et al.* [2019] explore contextual bandits with stochastic delay. Arm-dependent delay is discussed by Gael *et al.* [2020], and Lancewicki *et al.* [2021] later remove the restriction on delay distribution. Tang *et al.* [2024] focus on strongly reward-dependent delay and achieve near-optimal results. Yang *et al.* [2024] propose a reduction-based framework to handle delays with sub-exponential distributions.

A similar setting to ours is multi-agent bandits with delay. Existing literature has focused on decentralized coop-

erative bandits [Cesa-Bianchi *et al.*, 2016; Martínez-Rubio *et al.*, 2019], while non-cooperative game with delay is discussed in Bistritz *et al.*; Bistritz *et al.* [2019; 2022]. Zhang *et al.* [2023] consider multi-agent reinforcement learning with both finite and infinite delay. Li and Guo [2023] discuss adversarial bandit problem with delayed feedback from multiple users. Hanna *et al.* [2024] propose an algorithm in multi-agent bandits with delay and derive a sub-linear regret. However, none of these works consider collisions between players. Since collisions result in a loss of reward, current algorithms in multi-agent bandits cannot be directly applied to our problem.

3 Preliminaries

We consider a multi-player multi-armed bandits problem with delayed feedback. Denote by M the number of players and K the number of arms. Let T denote the time horizon. At each time $s \in [T]$, every player j selects an arm k and has a reward which she does not observe immediately. Denote by π_s^j the arm that is selected by player j at s. If more than one player selects the same arm at the same time, a collision occurs, and their reward turns to zero. Define $\eta_k(s) := \mathbb{1}\{|C_k(s)| > 1\}$ as the collision indicator where $C_k(s) := \{j \in [M] \mid \pi_s^j = k\}$ is the set of players who pull the same arms at time step s. After a period of delay d_s^j, at time t such that $s + d_s^j = t$, player j receives $\eta_k(s)$ and a reward $r^j(s) := X_k^j(s)\,[1 - \eta_k(s)]$, where $X_k^j(s)$ is drawn i.i.d. from an unknown fixed distribution with expectation $\mu_k \in [0, 1]$. Note that $K > M$ so that there is at least one arm available for each player without mandatory overlap or collision. Denote by $[M]$ the set of all players and $[K]$ the set of arms. The expected regret is defined as

$$R_T := T \sum_{j\in[M]} \mu_{(j)} - \mathbb{E}\left[\sum_{s=1}^{T}\sum_{j\in[M]} r^j(s)\right],$$

where $\mu_{(j)}$ is j-th order statistics of μ, i.e. $\mu_{(1)} \geq \mu_{(2)} \geq \ldots \geq \mu_{(K)}$.

Define $d(\theta) := \min\{\gamma \in \mathbb{N} | P(d \leq \gamma) \geq \theta\}$ as the quantile function of the delay distribution. We consider the following assumption.

Assumption 1. *Let $\{d_t^j\}_{t=1,j=1}^{T,M}$ denote independent non-negative random variables with sub-Gaussian distribution. Denote by σ_d^2 the sub-Gaussian parameter and $\mathbb{E}[d]$ the expectation of the distribution. Then for any $a > 0$,*

$$P(|d_t^j - \mathbb{E}[d]| \geq a) \leq 2\exp(-\frac{a^2}{2\sigma_d^2}).$$

Assumption 1 models the delay as a light-tailed random variable, where larger delays occur with lower probability. This behavior is consistent with real-world networks, where delays often arise from bounded processing times, network latency, or short-term queuing, and are typically constrained by hardware and protocol limitations [Azarfar *et al.*, 2015]. As a result, sub-Gaussian distributions provide a realistic and analytically tractable model for capturing such stochastic delays.

Algorithm 1 DDSE (Leader with $j = M$)

Input: K (number of arms), M (number of players), T
Initialize $\mathcal{M}_0^j$ randomly, $\mathcal{K} = [K]$, $p = 0$, $q_j = 0$, $\hat{\mu}_d^j = 0$, $(\hat{\sigma}_d^2)^j = 0$, $\hat{\mu}_k(t) = 0$.
while $t \leq T$ **do**
  Explore in $\mathcal{M}_{p-q_j}^j$ and $\mathcal{K}/\mathcal{M}_{p-q_j}^j$.
  if j receive a feedback from time s **then**
    $d_s^j \leftarrow t - s$.
    Update $\hat{\mu}_d^j, (\hat{\sigma}_d^2)^j, \hat{\mu}_k(t)$ accordingly.
  end if
  Remove from $\mathcal{K}$ the arm k satisfying:

$$|\{i \mid \forall i \neq k, i \in \mathcal{K} \text{ s.t. } \mathrm{LCB}_t(i) \geq \mathrm{UCB}_t(k)\}| \geq M.$$

  if $t \bmod (KM\lceil\log(T)\rceil) = 0$ **then**
    $p \leftarrow p + 1$.
    $\mathcal{M}_p^j \leftarrow \{k \mid \hat{\mu}_k(t)$ ranks in the top-M of $\hat{\mu}_i(t)$. for all $i \in \mathcal{K}\}$.
    if $\mathcal{M}_p^j \neq \mathcal{M}_{p-1}^j$ **then**
      `Communication()`.
    else `VirtualCom`$(\mathcal{M}_{p-q_j}^j)$.
    end if
    Find q_j s.t. (1).
  end if
  if $|\mathcal{K}| = M$ and $q_j = 0$ **then**
    Select $\mathcal{M}_{p_{\max}}^j(j)$ until T.
  end if
end while

Considering cognitive wireless sensor networks, sensor nodes are usually pre-deployed [Joshi *et al.*, 2013], so they are equipped with information on the total number of nodes and their ID. Consequently, we assume that each player in our algorithms is initialized with her ID among all players and is aware of the total number of players. Discussion on removing this assumption is in Appendix G.

4 Algorithm

The proposed algorithm **DDSE (Delayed Decentralized Successive Elimination)** is composed of exploration phase, communication phase and exploitation phase. Denote by $p_{\max}$ the maximum number of communication phases within a given time horizon. $\mathcal{M}_p^j$ denotes the set of the top M arms with highest empirical rewards for player j during the p-th communication phase. We have $|\mathcal{M}_p^j| = M$ for all $j \in [M]$ and $p \leq p_{\max}$. The key idea of DDSE is to estimate delay and pick $\mathcal{M}_{p-q}^j$ after the p-th communication phase, so that $\mathcal{M}_{p-q}^j = \mathcal{M}_{p-q}^\ell$ for $j \neq \ell$ with high probability. To better understand our algorithm, we also give an example in Appendix B.

4.1 Exploration Phase

We assign each player an ID j and initialize $\mathcal{M}_0^j$ for each player $j \in [M]$. The player with $j = M$ becomes the leader, and others are followers. Algorithm 1 describes DDSE from the view of the leader. The algorithm from the view of follow-

Algorithm 2 Communication (Leader with $j = M$)

Initialize i_{a^-}, a^+ by comparing $\mathcal{M}_p^j$ and $\mathcal{M}_{p-1}^j$.
Part 1: Remove Arm
for M time steps **do**
  Select $\mathcal{M}_{p-q_j}^j(i_{a^-})$.
end for
Part 2: Add Arm
for K time steps **do**
  Select a^+.
end for
Part 3: Notify End
for M time steps **do**
  if $|\mathcal{K}| = M$ **then**
    $p_{\max} \leftarrow p$.
    $i \leftarrow [t \bmod M]$.
    Select $\mathcal{M}_{p-q_j}^j(i)$.
  else
    Select $\mathcal{M}_{p-q_j}^j(j)$.
  end if
end for

Algorithm 3 Communication (Follower j)

$\texttt{Com} \leftarrow \texttt{Com} \cup \{(t, t+M+K)\}$.
$\texttt{Use} \leftarrow \texttt{Use} \cup \{p - q_j\}$.
Part 1: Remove Arm
for M time steps **do**
  $i \leftarrow [(t+j) \bmod M]$.
  Select $\mathcal{M}_{p-q_j}(i)$.
end for
Part 2: Add Arm
for K time steps **do**
  $i \leftarrow [(t+j) \bmod K]$.
  Select the i-th arm in $[K]$.
end for
Part 3: Notify End
for M time step **do**
  Select $\mathcal{M}_{p-q_j}^j(j)$.
end for
if j has received two collisions $\eta_k(s), \eta_{k'}(s')$ such that $\exists i, s < s', s \in \texttt{Com}(i)$ and $s' \in \texttt{Com}(i)$ **then**
  $t_0, t_1 \leftarrow \texttt{Com}(i)$.
  $p_0 \leftarrow t_0 / KM\lceil \log(T) \rceil$ and $u \leftarrow \texttt{Use}(i)$.
  $i_{a^-} \leftarrow$ the index of k in $\mathcal{M}_u^j$ and $a^+ \leftarrow k'$.
  $\mathcal{M}_{p_0}^j(i_{a^-}) \leftarrow a^+$.
end if

ers is in Appendix B. Define $N_t(k) := \sum_{s<t} \mathbb{1}\{\pi_s^j = k, j = M\}$ as the number of times that the leader chooses arm k before t. Also define $n_t(k) := \sum_{s<t} \mathbb{1}\{\pi_s^j = k, d_s^j + s < t, j = M\}$ as the number of received feedback of the leader from arm k before t. Denote by $\mathcal{K}$ the active arm set, and it is initialized with $\mathcal{K} = [K]$. The leader selects arms from $\mathcal{K}$ in a round-robin way. When receiving the feedback of arm k at t in the exploration phase, she updates

$$\mathrm{UCB}_t(k) := \hat{\mu}_k(t) + \sqrt{\frac{2\log(T)}{n_t(k)}},\ \mathrm{LCB}_t(k) := \hat{\mu}_k(t) - \sqrt{\frac{2\log(T)}{n_t(k)}},$$

where $\hat{\mu}_k(t) := S_k(t)/n_t(k)$ is the empirical reward of arm k and $S_k(t)$ is the sum of rewards that the leader has collected on arm k by the end of time t. During the exploration phase, the leader eliminates an arm k from $\mathcal{K}$ at t if there exist more than M arms whose lower confidence bounds are bigger than $\mathrm{UCB}_k(t)$.

When $t \bmod KM\lceil \log(T) \rceil = 0$, the leader enters a new communication phase. She sorts $\hat{\mu}_k(t)$ and places the arms corresponding to the top M highest $\hat{\mu}_k(t)$ into $\mathcal{M}_p^j$. If $\mathcal{M}_p^j$ has changed, the leader communicates the update to the followers by sending collisions in `Communication`. Otherwise, she runs a `VirtualCom`, during which no collisions are sent. Details about `VirtualCom` are in Appendix B. From the perspective of a follower ℓ, it takes time for her to receive the update because of the delay. However, since the leader only communicates with followers after $\mathcal{M}_p^j$ has changed, if follower ℓ has not received the update, $\mathcal{M}_p^j \neq \mathcal{M}_p^\ell$. Therefore, players need to pick a previous set $\mathcal{M}_{p-q_j}^j$, which is considered to be received by all players with high probability. Specifically, by the sub-Gaussian property of delay, we know that when

$$t - pKM\log(T) > \mathbb{E}[d] + \sqrt{2\sigma_d^2 \log(M-1)(K+2M)(T)},$$

$\mathcal{M}_p^j$ is received by all followers with high probability. When a player j receive a feedback at time t, she update the estimation of $\mathbb{E}[d]$ and σ_d^2 with

$$\hat{\mu}_d^j(t) := \frac{\sum_{s\le t}\left(d_s^j \mathbb{1}\{s + d_s^j \le t\}\right)}{\sum_{s\le t} \mathbb{1}\{s + d_s^j \le t\}},$$

$$(\hat{\sigma}_d^2)^j(t) := \frac{\sum_{s\le t}\left([d_s^j - \hat{\mu}_d^j(t)]\mathbb{1}\{s + d_s^j \le t\}\right)^2}{\sum_{s\le t} \mathbb{1}\{s + d_s^j \le t\}}.$$

Thus, each player j aims to find $q_j \in \mathbb{N}$ such that

$$\begin{aligned} q_j = \arg\min_q \Big\{ q \mid t > \hat{\mu}_d^j(t) + (p-q)KM\log(T) \\ + \sqrt{2(\hat{\sigma}_d^2)^j(t)\log\left((M-1)(K+2M)(T)\right)} \Big\}. \end{aligned} \tag{1}$$

Then in exploration phase, followers only select arms from $\mathcal{M}_{p-q_j}^j$ in a round-robin way. To avoid collision with followers and ensure sufficient exploration, the leader first selects arms in $\mathcal{M}_{p-q_j}^j$ with followers. Then she selects other arms in $\mathcal{K}$ in a round-robin way while skipping arms in $\mathcal{M}_{p-q_j}^j$. In other words, the leader constantly explores all arms except those that have been eliminated. An example of the process of arm selection is in Appendix B.

Sub-optimal arms in $\mathcal{K}$ are gradually eliminated by the leader. When $|\mathcal{K}| = M$, she tells the final result to followers in the next communication phase. After that, the leader remains in the exploration phase until $q_j = 0$, at which point she moves to the exploitation phase and pulls $\mathcal{M}_{p_{\max}}^j(j)$. When a follower j receives the final update and finds $q_j = 0$, she also begins the exploitation phase and continuously selects arm $\mathcal{M}_{p_{\max}}^j(j)$ until T.

4.2 Communication Phase

Players begin a communication phase at time t such that $t = KM\lceil\log(T)\rceil$. The length of each communication phase is $K + 2M$. Motivated by Wang *et al.* [2020], the communication phase in our algorithm is divided into three parts. The first and second parts are used for removing and adding an arm in $\mathcal{M}_p$. The third part is used for the leader to send the ending signal. Let a^- denote the arm to remove and a^+ the arm to add. Given a list $\mathcal{A}$, let i_k denotes the index of item k in $\mathcal{A}$ such that $\mathcal{A}(i_k) = k$.

Part 1: Remove Arm

The leader firstly initializes i_{a^-} and a^+ by comparing the difference between $\mathcal{M}_p^j$ and $\mathcal{M}_{p-1}^j$. Followers save the time duration of current communication phase to Com and save $p - q_j$ to Use. Then in this part, the leader selects arm $\mathcal{M}_{p-q_j}^M(i_{a^-})$ for M consecutive rounds. Meanwhile, followers select arms from $\mathcal{M}_{p-q_j}^j$ in a round-robin way, ensuring that each follower collides once with the leader. The round of collision represents the index of the arm to be removed. Since $\mathcal{M}_p$ is ordered for all $p \le p_{\max}$, followers can receive the update of the leader to remove arm a^- from $\mathcal{M}_p^j$ by selecting arms in $\mathcal{M}_{p-q_j}^j$. Thus, the information is passed successfully even if $\mathcal{M}_p^j$ is incomplete for follower j, allowing our algorithm to adapt to large delays.

Part 2: Add Arm

In this part, the leader continuously selects a^+ for K rounds while followers select arms in $[K]$ in a round-robin way. Each follower also collides once with the leader. The collision denotes the arm to be added. Later, if a follower j receives two collisions $\eta_s(k)$ and $\eta_{s'}(k')$ where $s < s'$ and they belong to the same $\texttt{Com}(i)$, we say that she has completely received an update of a communication phase in which the updated result should be stored in $\mathcal{M}_{p_0}$ with $p_0 = t_0/KM\lceil\log(T)\rceil$ and $t_0 \in \texttt{Com}(i)$. Since $\texttt{Use}$ saves the index $(p - q_j)$ of $\mathcal{M}_{p-q_j}^j$ that player j uses in each communication phase $p \le p_{\max}$, the index at communication phase p_0 is $u := \texttt{Use}(i)$. Thus, follower j finds the index of k in $\mathcal{M}_u^j$ is i_{a^-} and then places a^+ in the position of $\mathcal{M}_{p_0}^j(i_{a^-})$, which does not break the order of $\mathcal{M}_{p_0}^j$.

Part 3: Notify End

If $|\mathcal{K}| = M$, it indicates that all sub-optimal arms have been eliminated and the leader selects arms in $\mathcal{M}_{p-q_j}^j$ sequentially, while followers continuously select arm $\mathcal{M}_{p-q_j}^j(j)$ for M times. Otherwise, the leader does not send collisions by selecting $\mathcal{M}_{p-q_i}^j(j)$ for M times. Finally, each follower receives a collision, which is a symbol of the end of exploration.

The reason why the beginning of our communication phase is fixed rather than starts as $\mathcal{M}_p^j$ changes is that players need to ensure synchronization with others. In Wang *et al.* [2020], the leader sends a collision to followers as the beginning signal of communication. However, when the feedback of this collision is delayed, followers hardly receive it at the same time and then stagger with the leader. Once players are not aligned with others, followers may receive incorrect information during the communication phase. Furthermore, since communication and exploration are alternating, players might end up selecting the same arm during the exploration phase, resulting in collisions.

Denote by p' a communication phase such that $\mathcal{M}_{p'}^j$ is the most recent to have been completely received by player j. If the delay is sufficiently small, players can receive the feedback from the p-th communication phase before the $(p+1)$-th communication begins. Then q is equal to zero in our algorithm, and players continue selecting arms from $\mathcal{M}_p^j$. We discuss DSE, which is a simplified version of DDSE, where players directly pull arms in $\mathcal{M}_{p'}^j$. This version does not estimate $\hat{\mu}_d^j$ or $(\hat{\sigma}_d^2)^j$ and also does not coordinate players to pull in the same set of best empirical arms. Details on DSE are presented in Appendix B.

5 Theoretical Analysis

In this section, we provide a comprehensive analysis of our algorithms, examining both the decentralized and centralized settings. Additionally, we derive a lower bound to demonstrate the near-optimality of the regret upper bound.

5.1 Centralized Lower Bound

Note that a key objective in decentralized MP-MAB is to match the performance of the centralized setting. In the centralized case, a central coordinator enables communication without incurring any regret. In contrast, players in the decentralized setting rely on implicit communication, typically through intentional collisions, which inevitably induces some regret. The goal is to ensure that this communication-induced regret remains a constant, independent of the time horizon T. Therefore, comparing the regret of decentralized algorithms against the centralized lower bound is both meaningful and standard [Boursier and Perchet, 2019; Shi *et al.*, 2020; Shi *et al.*, 2021]. Theorem 1 presents a centralized lower bound for MP-MAB with delays. Full proof of this theorem is provided in Appendix F.

Theorem 1. *For any sub-optimal gap set $S_\Delta = \{\Delta_{M,k} \mid \Delta_{M,k} = \mu_{(M)} - \mu_{(k)} \in [0,1]\}$ of cardinality $K - M$ and a quantile $\theta \in (0,1)$, there exists an instance with an order on S_Δ and a delay distribution under Assumption 1 such that*

$$R_T \ge \underbrace{\sum_{k>M} \frac{(1-o(1))\log(T)}{2\theta\Delta_{M,k}}}_{\text{term I}} + \underbrace{\left(\mathbb{E}[d] - \sigma_d\sqrt{\frac{\theta}{1-\theta}}\right)\frac{M}{2K}\sum_{k>M}\Delta_{M,k} - \frac{2}{\theta}}_{\text{term II}}, \tag{2}$$

5.2 Regret of DDSE

Theorem 2 presents our main result, which establishes the regret upper bound of DDSE in the decentralized setting. The complete proof of Theorem 2 is provided in Appendix C.

Theorem 2. *In decentralized setting, for delay distribution under Assumption 1, given any K, M, μ and a quantile $\theta \in (0,1)$, the regret of DDSE satisfies*

$$R_T \leq \underbrace{\sum_{k>M} \frac{323\log(T)}{\theta\Delta_{M,k}}}_{\text{term A}} + \underbrace{\frac{15}{\theta}\left(\mathbb{E}[d] + \sigma_d\sqrt{2\log(\frac{1}{1-\theta})}\right)}_{\text{term B}} + \underbrace{\left(2\mathbb{E}[d] + \sigma_d\sqrt{3\log(K)}\right)\frac{M}{K-M}\sum_{k>M}\Delta_{1,k}}_{\text{term C}} + \underbrace{\frac{656\sqrt{2}\sigma_d^2}{\theta K^2M^2} + 3\sqrt{6}\sigma_d}_{\text{term D}} + \underbrace{C_1}_{\text{term E}}, \tag{3}$$

where $C_1 = \sum_{k>M}\frac{195}{\theta\Delta_{M,k}^2} + \frac{4Me^{-\delta^2/2}}{\delta^2}$.

Only the first terms in (2) and (3) are related to T. Term A is aligned with term I up to constant factors. Term E arises due to the decentralized environment and is not related to delay. Regarding delay, a comparison of term II with terms B, C, and D reveals that the difference on K and M is only $O(\frac{1}{1-M/K})\sqrt{\log(K)}$. This indicates that the regret caused by delay does not increase rapidly as K and M increase. Therefore, our result is near-optimal.

Moreover, since the decentralized setting is generally more challenging than the centralized one, the near-optimality of our result (Theorem 2) with respect to the centralized lower bound further highlights the strength of our approach. We also establish a centralized upper bound for DDSE in Corollary 1, with the proof provided in Appendix D.

We also analyze the regret of DDSE in the centralized setting, where players can freely exchange information. The regret upper bound is presented in Corollary 1.

Corollary 1. *In centralized setting, for delay distribution under Assumption 1, given any K, M, μ and a quantile $\theta \in (0,1)$, the regret of DDSE satisfies*

$$R_T \leq \underbrace{\sum_{k>M}\frac{323\log(T)}{\theta\Delta_{M,k}}}_{\text{term A}} + \underbrace{\frac{9}{\theta}\left(\mathbb{E}[d] + \sqrt{2\sigma_d^2\log\left(\frac{1}{1-\theta}\right)}\right)}_{\text{term F}} + \underbrace{\mathbb{E}[d]\frac{M}{K-M}\sum_{k>M}\Delta_{1,k}}_{\text{term G}} + \underbrace{\frac{656\sqrt{2}\sigma_d^2}{\theta K^2M^2} + 3\sqrt{6}\sigma_d}_{\text{term D}}.$$

When DDSE is executed in the centralized setting, followers are immediately aware of the leader's latest exploration results. Once the leader identifies the optimal set of arms $\mathcal{M}^*$, exploitation begins without incurring the regret of $O(\sigma_d\log(K))$ in term C. Comparing with term II in Theorem 1 and term F, G, D, we note that the regret due to delay differs by $O(\frac{1}{1-M/K})$ on K and M. Proof of Corollary 1 is provided in Appendix D.

5.3 Regret of DSE

This section presents the regret of DSE, where players do not adjust to the same $\mathcal{M}^j_{p-q_j}$. Theorem 3 shows that it is critical to maintain consistency between players. The detailed proof is included in Appendix E.

Theorem 3. *In decentralized setting, for delay distribution under Assumption 1, given any K, M, μ and a quantile $\theta \in (0,1)$, the regret of DSE is bounded by*

$$R_T \leq \sum_{k>M}\frac{323\log(T)}{\theta\Delta_{M,k}} + \sigma_d\left(3\sqrt{6} + \frac{12}{\theta}\sqrt{2\log(\frac{1}{1-\theta})}\right) + \left(9 + \frac{6}{\theta} + \frac{M\sum_{k>M}\Delta_{1,k}}{K-M}\right)\mathbb{E}[d] + \frac{656\sqrt{2}\sigma_d^2}{\theta K^2M^2} + \underbrace{O\left(\frac{\tilde{d}_2\tilde{d}_3}{KM} + \frac{\tilde{d}_3}{\theta KM\sum_{k>M}\Delta_{M,k}^2}\right)}_{\text{term H}} + C_2 + \underbrace{\exp\left(\frac{\mathbb{E}[d]}{KM} + \frac{\sigma_d^2}{2K^2M^2}\right)}_{\text{term J}}.$$

If players do not estimate the delay and pull arms in $\mathcal{M}^j_{p'}$, followers will collide with the leader after each communication phase ends. This happens because the leader begins communication after she updates $\mathcal{M}^j_p$, while the followers have not yet received this update, ultimately contributing to term H. Additionally, followers receive incorrect information during the communication phase if $\mathcal{M}^j_{p'} \neq \mathcal{M}^\ell_{p'}$, which leads to an exponential regret in term J.

Compare Theorem 3 with Theorem 2 and we find by using $\mathcal{M}^j_{p-q_j}$ instead of $\mathcal{M}_{p'}$, players will not collide with each other after the communication ends, thereby avoiding term H, which could be large when $\Delta^2_{M,k}$ is sufficiently small. Moreover, since $\mathcal{M}p - q_j{}^j = \mathcal{M}p - q_\ell{}^\ell$ for all $j, \ell \in [M]$ with high probability, followers receive consistent information from the leader. As a result, the regret due to asynchronous feedback and coordination errors is reduced from $O(\exp(\mathbb{E}[d]))$ in term J to $O(\mathbb{E}[d])$ in Theorem 2, highlighting the effectiveness of our delay-aware coordination mechanism.

Remark 1 (On Relaxing Assumption 1)**.** The analysis relies on Assumption 1, which is used to estimate the delay quantile. Specifically, the sub-Gaussian assumption allows us to derive high-probability bounds on delay, which are essential for player coordination. Similar guarantees can also be obtained under sub-exponential delay distributions.

More generally, since our goal is to estimate a delay threshold d such that most delays fall below it, the delay quantile can be estimated directly without assuming a specific distribution. Given a quantile level θ, the number of observed delays below d follows a Binomial(n, θ) distribution, where n is the number of samples. This enables us to bound the quantile estimation error using standard binomial tail inequalities in a distribution-free manner. We believe this quantile-based approach offers a promising direction for extending our algorithm to settings with heavy-tailed or unknown delay distributions, which we leave as future work.

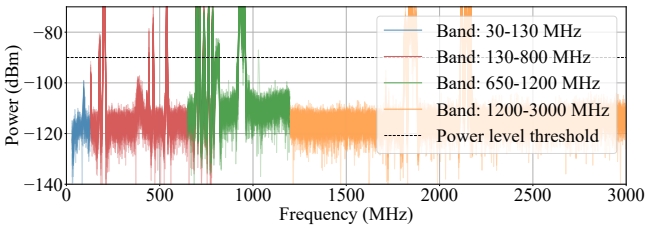


Figure 1: Captured spectrum data from paging frequency bands.

6 Experiments

We conduct experiments to validate our theoretical results. Let $\overline{\Delta} := \sum_{k=1}^{K-1} \frac{\mu_{(k)} - \mu_{(k+1)}}{K-1}$ denote the average reward gap between consecutive arms. Each experiment runs for $T = 300{,}000$ rounds and is averaged over 20 trials. Default parameters are $K = 20$, $M = 10$, $\mathbb{E}[d] = 200$, $\sigma_d = 100$, and $\overline{\Delta} = 0.05$. We assume Gaussian rewards and compare DDSE with DSE; SIC-MMAB [Boursier and Perchet, 2019], MCTopM, RandTopM, Selfish [Besson and Kaufmann, 2018]; Game of Throne [Bistritz and Leshem, 2018]; and ESER [Tibrewal *et al.*, 2019]. Real-world results are reported below, while numerical simulations are deferred to Appendix A.

We evaluate our algorithms using real-world spectrum data collected in Finland by the 5G-Xcast project[2]. Figure 1 shows a sample of power measurements across four bands. In cognitive radio networks, secondary users share spectrum with primary users without causing interference. Multi-player bandit algorithms help secondary users identify available channels. A channel is considered occupied by a primary user if its power measurement exceeds the threshold of -90 dBm, as in Alipour-Fanid *et al.* [2022]. We assess performance using cumulative throughput and collisions [Wang *et al.*, 2021; Alipour-Fanid *et al.*, 2022]. Throughput is computed using Shannon's formula:

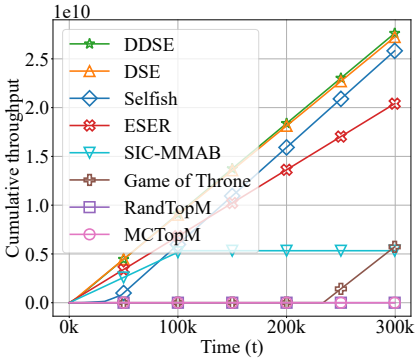
$$B = W \log_2(1 + SNR)\,,$$

where W denotes bandwidth and SNR denotes signal to noise ratio. No throughput is achieved when a user selects a busy channel or experiences a collision.

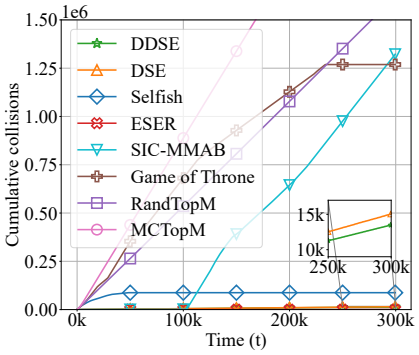
Figure 2(a) illustrates the cumulative throughput over time for different algorithms. Our algorithm, DDSE, achieves the highest throughput, demonstrating its effectiveness. DSE performs slightly worse than DDSE due to the inconsistent $\mathcal{M}_p^j$. Selfish also shows an increasing throughput as time progresses. However, as the name suggests, players in the Selfish algorithm explore independently without communication. In contrast, our DDSE algorithm leverages an adjusted form of implicit communication that is robust to incomplete feedback, making it more effective in the presence of delays.

Figure 2(b) compares the cumulative collisions across various algorithms. Notably, DDSE achieves a remarkably low level of cumulative collisions due to the carefully designed nature of our algorithm. It is worth mentioning that ESER

[2]The full dataset used in this experiment is publicly available at https://zenodo.org/records/1293283.



(a) Throughput



(b) Collisions

Figure 2: Real-world simulation results across different algorithms.

experiences nearly zero collisions, thanks to its unique mechanism, where players select fixed arms in a round-robin fashion during the exploration phase and pull their optimal arms during the exploitation phase. In contrast, DDSE allows most players (i.e., followers) to exploit arms in $\mathcal{M}_p^j$, while the leader explores and updates her results for the followers. Although this method incurs slightly higher collisions than ESER, DDSE achieves a higher throughput, demonstrating its superior performance in terms of overall efficiency.

7 Conclusion

In this paper, we proposed the algorithm DDSE for multi-player bandits with delayed feedback and derive a regret upper bound. Rather than allowing players to update blindly, coordinating to maintain consistency with other players significantly improves performance and reduces regret. The lower bound in the centralized setting further shows that our algorithm is near-optimal. Practical simulations have also validated the superiority of our algorithm.

A promising direction for future work is to relax the sub-Gaussian delay assumption. As discussed in Remark 1, our framework can potentially be extended to handle heavy-tailed or unknown delay distributions using a quantile-based estimation approach. Another direction is to study player-dependent delays in MP-MAB, as delays in cognitive networks often depend on user-specific factors such as location and device capability.

Acknowledgements

The corresponding author Shuai Li is supported by National Natural Science Foundation of China (62376154).

References

[Ahmad *et al.*, 2020] Wan Siti Halimatul Munirah Wan Ahmad, Nurul Asyikin Mohamed Radzi, Faris Syahmi Samidi, Aiman Ismail, Fairuz Abdullah, Md Zaini Jamaludin, and MohdNasim Zakaria. 5g technology: Towards dynamic spectrum sharing using cognitive radio networks. *IEEE access*, 8:14460–14488, 2020.

[Akyildiz *et al.*, 2006] Ian F Akyildiz, Won-Yeol Lee, Mehmet C Vuran, and Shantidev Mohanty. Next generation/dynamic spectrum access/cognitive radio wireless networks: A survey. *Computer networks*, 50(13):2127–2159, 2006.

[Alipour-Fanid *et al.*, 2022] Amir Alipour-Fanid, Monireh Dabaghchian, Raman Arora, and Kai Zeng. Multiuser scheduling in centralized cognitive radio networks: A multi-armed bandit approach. *IEEE Transactions on Cognitive Communications and Networking*, 8(2):1074–1091, 2022.

[Azarfar *et al.*, 2015] Arash Azarfar, Jean-François Frigon, and Brunilde Sansò. Delay analysis of multichannel opportunistic spectrum access mac protocols. *IEEE Transactions on Mobile Computing*, 15(1):92–106, 2015.

[Besson and Kaufmann, 2018] Lilian Besson and Emilie Kaufmann. Multi-player bandits revisited. In *Algorithmic Learning Theory*, pages 56–92. PMLR, 2018.

[Bistritz and Leshem, 2018] Ilai Bistritz and Amir Leshem. Distributed multi-player bandits-a game of thrones approach. *Advances in Neural Information Processing Systems*, 31, 2018.

[Bistritz *et al.*, 2019] Ilai Bistritz, Zhengyuan Zhou, Xi Chen, Nicholas Bambos, and Jose Blanchet. Online exp3 learning in adversarial bandits with delayed feedback. *Advances in neural information processing systems*, 32, 2019.

[Bistritz *et al.*, 2022] Ilai Bistritz, Zhengyuan Zhou, Xi Chen, Nicholas Bambos, and Jose Blanchet. No weighted-regret learning in adversarial bandits with delays. *Journal of Machine Learning Research*, 23(139):1–43, 2022.

[Boursier and Perchet, 2019] Etienne Boursier and Vianney Perchet. Sic-mmab: Synchronisation involves communication in multiplayer multi-armed bandits. *Advances in Neural Information Processing Systems*, 32, 2019.

[Cesa-Bianchi *et al.*, 2016] Nicol'o Cesa-Bianchi, Claudio Gentile, Yishay Mansour, and Alberto Minora. Delay and cooperation in nonstochastic bandits. In *Conference on Learning Theory*, pages 605–622. PMLR, 2016.

[Gael *et al.*, 2020] Manegueu Anne Gael, Claire Vernade, Alexandra Carpentier, and Michal Valko. Stochastic bandits with arm-dependent delays. In *International Conference on Machine Learning*, pages 3348–3356. PMLR, 2020.

[Hanna *et al.*, 2024] Osama A Hanna, Merve Karakas, Lin Yang, and Christina Fragouli. Multi-agent bandit learning through heterogeneous action erasure channels. In *International Conference on Artificial Intelligence and Statistics*, pages 3898–3906. PMLR, 2024.

[Huang *et al.*, 2022] Wei Huang, Richard Combes, and Cindy Trinh. Towards optimal algorithms for multi-player bandits without collision sensing information. In *Conference on Learning Theory*, pages 1990–2012. PMLR, 2022.

[Joshi *et al.*, 2013] Gyanendra Prasad Joshi, Seung Yeob Nam, and Sung Won Kim. Cognitive radio wireless sensor networks: applications, challenges and research trends. *Sensors*, 13(9):11196–11228, 2013.

[Joulani *et al.*, 2013] Pooria Joulani, Andras Gyorgy, and Csaba Szepesvári. Online learning under delayed feedback. In *International conference on machine learning*, pages 1453–1461. PMLR, 2013.

[Komiyama *et al.*, 2015] Junpei Komiyama, Junya Honda, and Hiroshi Nakagawa. Optimal regret analysis of thompson sampling in stochastic multi-armed bandit problem with multiple plays. In *International Conference on Machine Learning*, pages 1152–1161. PMLR, 2015.

[Lancewicki *et al.*, 2021] Tal Lancewicki, Shahar Segal, Tomer Koren, and Yishay Mansour. Stochastic multi-armed bandits with unrestricted delay distributions. In *International Conference on Machine Learning*, pages 5969–5978. PMLR, 2021.

[Lattimore and Szepesvári, 2020] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[Li and Guo, 2023] Yandi Li and Jianxiong Guo. A modified exp3 and its adaptive variant in adversarial bandits with multi-user delayed feedback. *arXiv preprint arXiv:2310.11188*, 2023.

[Lugosi and Mehrabian, 2022] Gábor Lugosi and Abbas Mehrabian. Multiplayer bandits without observing collision information. *Mathematics of Operations Research*, 47(2):1247–1265, 2022.

[Mahesh *et al.*, 2022] Shivakumar Mahesh, Anshuka Rangi, Haifeng Xu, and Long Tran-Thanh. Multi-player bandits robust to adversarial collisions. *arXiv e-prints*, pages arXiv–2211, 2022.

[Martínez-Rubio *et al.*, 2019] David Martínez-Rubio, Varun Kanade, and Patrick Rebeschini. Decentralized Cooperative Stochastic Bandits. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[Pike-Burke *et al.*, 2018] Ciara Pike-Burke, Shipra Agrawal, Csaba Szepesvari, and Steffen Grunewalder. Bandits with delayed, aggregated anonymous feedback. In *International Conference on Machine Learning*, pages 4105–4113. PMLR, 2018.

[Richard *et al.*, 2024] Hugo Richard, Etienne Boursier, and Vianney Perchet. Constant or logarithmic regret in asynchronous multiplayer bandits with limited communication. In *International Conference on Artificial Intelligence and Statistics*, pages 388–396. PMLR, 2024.

[Shi *et al.*, 2020] Chengshuai Shi, Wei Xiong, Cong Shen, and Jing Yang. Decentralized multi-player multi-armed bandits with no collision information. In *International Conference on Artificial Intelligence and Statistics*, pages 1519–1528. PMLR, 2020.

[Shi *et al.*, 2021] Chengshuai Shi, Wei Xiong, Cong Shen, and Jing Yang. Heterogeneous multi-player multi-armed bandits: Closing the gap and generalization. *Advances in neural information processing systems*, 34:22392–22404, 2021.

[Tang *et al.*, 2024] Yifu Tang, Yingfei Wang, and Zeyu Zheng. Stochastic multi-armed bandits with strongly reward-dependent delays. In *International Conference on Artificial Intelligence and Statistics*, pages 3043–3051. PMLR, 2024.

[Tibrewal *et al.*, 2019] Harshvardhan Tibrewal, Sravan Patchala, Manjesh K Hanawal, and Sumit J Darak. Distributed learning and optimal assignment in multiplayer heterogeneous networks. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, pages 1693–1701. IEEE, 2019.

[Vernade *et al.*, 2017] Claire Vernade, Olivier Cappe, and Vianney Perchet. Stochastic Bandit Models for Delayed Conversions. *Conference on Uncertainty in Artificial Intelligence, Aug 2017, Sydney, Australia*, 2017.

[Wang *et al.*, 2020] Po-An Wang, Alexandre Proutiere, Kaito Ariu, Yassir Jedra, and Alessio Russo. Optimal algorithms for multiplayer multi-armed bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 4120–4129. PMLR, 2020.

[Wang *et al.*, 2021] Wenbo Wang, Amir Leshem, Dusit Niyato, and Zhu Han. Decentralized learning for channel allocation in iot networks over unlicensed bandwidth as a contextual multi-player multi-armed bandit game. *IEEE Transactions on Wireless Communications*, 21(5):3162–3178, 2021.

[Wang *et al.*, 2022] Xuchuang Wang, Hong Xie, and John Lui. Multi-player multi-armed bandits with finite shareable resources arms: Learning algorithms & applications. *arXiv preprint arXiv:2204.13502*, 2022.

[Xu *et al.*, 2023] Renzhe Xu, Haotian Wang, Xingxuan Zhang, Bo Li, and Peng Cui. Competing for shareable arms in multi-player multi-armed bandits. In *International Conference on Machine Learning*, pages 38674–38706. PMLR, 2023.

[Yang *et al.*, 2024] Yunchang Yang, Han Zhong, Tianhao Wu, Bin Liu, Liwei Wang, and Simon S Du. A reduction-based framework for sequential decision making with delayed feedback. *Advances in Neural Information Processing Systems*, 36, 2024.

[Zhang *et al.*, 2023] Yuyang Zhang, Runyu Zhang, Yuantao Gu, and Na Li. Multi-agent reinforcement learning with reward delays. In *Learning for Dynamics and Control Conference*, pages 692–704. PMLR, 2023.

[Zhou *et al.*, 2019] Zhengyuan Zhou, Renyuan Xu, and Jose Blanchet. Learning in generalized linear contextual bandits with stochastic delays. *Advances in Neural Information Processing Systems*, 32, 2019.