On No-Sensing Adversarial Multi-Player Multi-Armed Bandits With Collision Communications

Chengshuai Shi, *Graduate Student Member, IEEE*, and Cong Shen, *Senior Member, IEEE*

***Abstract*—We study the notoriously difficult no-sensing adversarial multi-player multi-armed bandits (MP-MAB) problem from a new perspective. Instead of focusing on the hardness of multiple players, we introduce a new dimension of hardness, called *attackability*. All adversaries can be categorized based on the attackability and we introduce Adversary-Adaptive Collision-Communication (A2C2), a family of algorithms with forced-collision communication among players. Both attackability-aware and unaware settings are studied, and information-theoretic tools of the Z-channel model and error-correction coding are utilized to address the challenge of implicit communication without collision information in an adversarial environment. For the more challenging attackability-unaware problem, we propose a simple method to estimate the attackability enabled by a novel error-detection repetition code and randomized communication for synchronization. Theoretical analysis proves that asymptotic attackability-dependent sublinear regret can be achieved, with or without knowing the attackability. In particular, the asymptotic regret does not have an exponential dependence on the number of players, revealing a fundamental tradeoff between the two dimensions of hardness in this problem.**

***Index Terms*—Multi-player multi-armed bandits, adversarial bandits, statistical learning.**

I. Introduction

THE DECENTRALIZED multi-player multi-armed bandits (MP-MAB) problem has received increasing interest in recent years [1]–[5]. In MP-MAB, multiple players simultaneously play the bandit game without explicit communications and interact with each other only through arm collisions. When two or more players play the same arm simultaneously, they all get a reward 0 (or equivalently loss 1) instead of the true underlying reward of that action. This model is largely motivated by practical applications such as cognitive radio [6]–[9] and wireless caching [10], where standard (single-player) MAB does not fully capture the system complexity and user interactions must be taken into account in conjunction with the bandit game.

Manuscript received October 4, 2020; revised February 28, 2021; accepted April 20, 2021. Date of publication April 27, 2021; date of current version June 21, 2021. This work was supported in part by the U.S. National Science Foundation (NSF) under Grant CNS-2002902 and Grant ECCS-2029978, and in part by the Commonwealth Cyber Initiative (CCI) Cybersecurity Research Collaboration Grant. *(Corresponding author: Cong Shen.)*

The authors are with the Charles L. Brown Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, VA 22904 USA (e-mail: cs7ync@virginia.edu; cong@virginia.edu).

Digital Object Identifier 10.1109/JSAIT.2021.3076027

Depending on how rewards are generated, the MP-MAB game can be either stochastic or adversarial, as in the single-player bandit problem. Most of the existing MP-MAB works focus on the stochastic setting, in which a well-behaved stochastic model exists for each arm (albeit unknown to the players). However, it is often difficult to determine the correct stochastic assumptions in real-world applications, and there are use cases where such assumptions do not hold. For example, in cognitive radio systems, it is common to have channel availability or signal quality fluctuations due to changing environmental conditions or bursty radio frequency interference [11], [12]. Adversarial MP-MAB is a more suitable model for such use cases, as it makes no stochastic assumptions on the rewards and assigns an arbitrary reward sequence to each arm exogenously. However, compared with stochastic MP-MAB, adversarial MP-MAB is a considerably harder problem because of the need to fight the adversary while interacting with other players.

Since the MAB problem for a single player is well understood, a predominant approach for both stochastic and adversarial MP-MAB is to let each player play the single-player MAB game while avoiding collisions for as much as possible [1], [2], [7], [13], [14]. Recently, a pioneering work [3] proposes to purposely instigate collisions as a way to share information between players. Such *implicit communication* is instrumental in breaking the performance barrier and achieving a regret that approaches the centralized multi-play MAB [15], [16]. This idea has been extended to several variants in the stochastic setting [17]–[20] as well as adversarial MP-MAB [5], [12], with improved regret performance for all models.

All the aforementioned works make an important assumption of *collision sensing* – any collision with another player is perfectly known. Such "collision indicator" plays a fundamental role in both collision avoidance and forced-collision communication. It is widely recognized that a more difficult problem in MP-MAB is the *no-sensing* scenario, in which players can only observe the final rewards but not collisions. The difficulty lies in that the zero rewards can indistinguishably come from collisions or null arm rewards. Recently, there is some progress on the stochastic no-sensing problem [4], [21]. In particular, the fundamental idea of *implicit communication* is again proved crucial in achieving regret that approaches the centralized counterpart [9].

Nevertheless, the most difficult setting of no-sensing adversarial MP-MAB in a fully decentralized setting remains wide

TABLE I
REGRET BOUNDS OF ADVERSARIAL MP-MAB ALGORITHMS

Model	Reference	Asymptotic Bound
Centralized Multiplayer, Optimal Regret	[22]	$\Theta\left(\sqrt{MKT}\right)$
Decentralized, Collision Sensing	[14]	$\tilde{O}\left(K^M M^2 T^{\frac{3}{4}}\right)$
Decentralized, Collision Sensing	[12]	$\tilde{O}\left(M^{\frac{4}{3}} K^{\frac{1}{3}} T^{\frac{2}{3}}\right)$
Decentralized, Collision Sensing, $M = 2$	[5]	$\tilde{O}\left(K^2\sqrt{T}\right)$
Decentralized, No Sensing	[5]	$\tilde{O}\left(MK^{\frac{3}{2}} T^{1-\frac{1}{2M}}\right)$
Decentralized, No Sensing, Local Attackability	α-aware A2C2 (this work)	$\tilde{O}\left(M^{\frac{4}{3}} K^{\frac{1}{3}} T^{\frac{2+\alpha+\epsilon}{3}}\right)$
Decentralized, No Sensing, Global Attackability	β-aware A2C2 (this work)	$\tilde{O}\left(M^2 K^{\frac{2}{3}} T^{\max\{\frac{1+\beta}{2}, \frac{2}{3}\}}\right)$
Decentralized, No Sensing, Local Attackability	α-unaware A2C2 (this work)	$\tilde{O}\left(M^{\frac{4}{3}} K^{\frac{1}{3}} T^{\frac{5+\alpha+\epsilon}{6}}\right)$
Decentralized, No Sensing, Global Attackability	β-unaware A2C2 (this work)	$\tilde{O}\left(M^2 K^{\frac{1}{3}} T^{\max\{\frac{2+\beta+\epsilon}{3}, \frac{3}{4}\}}\right)$

K: number of arms; M: number of players with $1 < M \leq K$; ϵ: an arbitrarily small constant;
α: local attackability (see Corollary 1); β: global attackability (see Corollary 2).
With the notation of $\tilde{O}(\cdot)$, the logarithmic factors of T and K are ignored.

open. To the best of the authors' knowledge, [5] is the only work that achieves a sublinear regret by a collision-avoidance design (i.e., no implicit communication) where "safe" arms are reserved for players. However, its asymptotic regret of $O(T^{1-\frac{1}{2M}})$ is almost linear in T when M is large, where M is the number of players and T is the time horizon of the game. We note that this exponential dependence on M reveals a particular *dimension of hardness* (multiple players) in the no-sensing adversarial MP-MAB problem.

Recent development has repeatedly demonstrated that implicit communication is crucial in achieving lower regret. However, as pointed out in [5], it is unclear how to implicitly communicate without collision information in an adversarial environment. This paper makes progress in the no-sensing adversarial MP-MAB problem by addressing the challenges in incorporating implicit communication. In particular, this work reveals a novel dimension of hardness associated with the no-sensing adversarial MP-MAB problem: *attackability* of the adversary, that is orthogonal to the multi-player dimension of hardness. Technically, we depart from the approach of [5] which always assumes the worst possible adversary while focusing on the multi-player hardness, and analyze the relationship between the attackability hardness and implicit communications. Notably, *all* possible adversaries can be classified based on this new concept of attackability, which is defined either by a local view (for a "one-time" attack) or a global view (for the cumulative attacks). The hardness of attackability may or may not be aware by the players, and we develop a suite of *Adversary-Adaptive Collision-Communication (A2C2)* algorithms under both attackability-aware and attackability-unaware settings, which adaptively adjust the implicit communication by learning the attackability of the adversary in an online manner. All of the A2C2 algorithms utilize some (common) new elements that have not been considered before in no-sensing adversarial MP-MAB, such as an information-theoretical Z-channel model and error-correction coding, to design a forced-collision communication protocol that can effectively fight against the adversary and achieve a non-dominant communication regret in the no-sensing setting. On the other hand, for the more challenging attackability-unaware setting, we show that a simple "escalation" estimation of the attackability, a novel error-detection repetition code, and randomized synchronizations are crucial to handle the unknown attackability. A key idea behind algorithms in the attackability-unaware setting is that *communication error is not bad if it happens to all players*, as such error does not affect player synchronization.

The regret analysis of the A2C2 algorithms shows that they can achieve attackability-dependent sublinear regrets asymptotically, without an exponential dependence on the number of players as in [5]. This benefit, however, does not lead to a universally lower regret. In fact, we may view A2C2 of this paper and the method of [5] as operating at two different regimes in the two-dimensional hardness space (multi-player and attackability). On one hand, the T terms in the regret of A2C2 algorithms are oblivious to the number of players (i.e., M) and the overall dependency on M is only a multiplicative factor, while the regret of [5] has an exponential dependency on M. On the other hand, A2C2 algorithms have exponential dependencies on the attackability, which does not affect the regret of [5]. Philosophically speaking, the regret comparison between A2C2 and [5] shows that *one can trade off the multi-player dimension of hardness with the attackability dimension of hardness*, which may provide insight into other relevant adversarial bandit problems. A comparison of the regret bounds are given in Table I for both collision-sensing and no-sensing adversarial MP-MAB algorithms.

The rest of the paper is organized as follows. Related works are surveyed in Section II. The no-sensing adversarial MP-MAB problem is formulated in Section III. The general algorithm structure is presented in Section IV, followed by algorithms for known (Section V) and unknown (Section VI) attackability. The regret analyses of all algorithms are given in Section VII. Discussions on some algorithmic details and future research directions are given in Section VIII. Numerical illustrations and experimental results are provided in Section IX to support the theoretical analyses. Finally, Section X concludes the paper.

II. Related Work

A. Overall Review

Collision-Sensing Stochastic and Adversarial MP-MAB: As stated in Section I, initial approaches for collision-sensing MP-MAB adopt single-player MAB algorithms with various collision-avoidance protocols. Examples include Explore-then-Commit [13], UCB [1], [2], ϵ-greedy [7] for stochastic MP-MAB, and EXP3 [14] for adversarial MP-MAB with a regret of $O(T^{\frac{3}{4}})$. Although these strategies achieve sublinear regret, their performance cannot approach the centralized counterparts. In particular, for the stochastic environment, there is a multiplicative factor M increase in the regret coefficient of $\log(T)$ compared with the natural lower bound of centralized MP-MAB [15], [16], which has long been considered fundamental due to the lack of explicit communication among players.

The idea of implicit communication with forced collisions is introduced by the SIC-MMAB algorithm [3], where bits 1 and 0 are transmitted by collision and no collision, respectively. The theoretical analysis of SIC-MMAB shows, for the first time, that the regret of decentralized MP-MAB can approach the centralized lower bound in the stochastic environment. The DPE1 algorithm [17] further improves the regret by combining the KL-UCB algorithm [23] with implicit communication. Similar ideas have also been extended to other stochastic variants, such as the heterogeneous setting [18], [20], where rewards are player-dependent. For the adversarial environment, implicit communication also proves to be effective. In particular, the C&P algorithm [12] achieves a regret of $O(T^{\frac{2}{3}})$ by invoking forced collisions to let players coordinately perform a centralized EXP3 algorithm. The performance for two players ($M = 2$) has been improved to $O(\sqrt{T\log(T)})$ in [5] by applying a filtering strategy with bandit-type information supported by implicit communication, which approaches the lower bound of $\Theta(\sqrt{T})$ [22].

No-Sensing Stochastic and Adversarial MP-MAB: No-sensing MP-MAB represents a more challenging scenario and the progress has been limited. A collision-avoidance scheme is investigated in [4] for the stochastic environment, which cannot approach the centralized lower bound. Some initial attempts to incorporate implicit communication in the no-sensing stochastic setting, e.g., sharing arm indices instead of statistics, are discussed in [3]. The EC-SIC algorithm proposed in [9] proves that it is possible to approach the centralized lower bound even without information of collision. For the most difficult case of no-sensing adversarial MP-MAB, progress is extremely limited. To the best of our knowledge, [5] is the only work studying this problem, and detailed comparisons between [5] and A2C2 are provided in the next subsection.

Cooperative MP-MAB: This is another line of MP-MAB research where explicit communications are allowed (under certain constraints) and players do not collide with each other. Such scenarios have been studied in both stochastic and adversarial environments [24]–[28], which are under a completely different framework than this work.

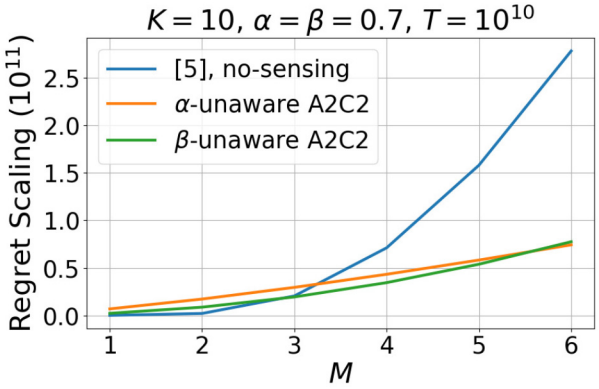


Fig. 1. Multi-player dimension of hardness.

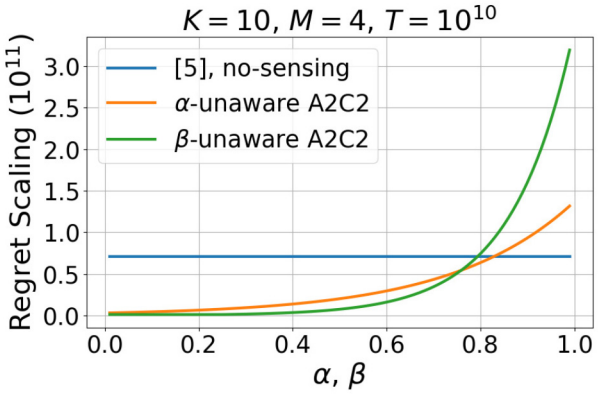


Fig. 2. Attackability dimension of hardness.

B. Comparison With [5]

As the only existing work studying the no-sensing adversarial MP-MAB problem, [5] designs a novel collision-avoidance algorithm by reserving "safe" arms for players. Nevertheless, this approach gives up sharing information through implicit communication. As a result, intuitively, when there are more players involved in the game, it becomes increasingly difficult to avoid collisions without sufficient coordination enabled by (implicit) communications. This issue is reflected in its achievable regret of $O(T^{1-\frac{1}{2M}})$, which has an exponential dependency on M. To address this critical issue of [5], this paper focuses on designing the implicit communication strategies through forced collisions and proposes the concept of attackability. The resulting suite of A2C2 algorithms have only multiplicative dependencies on M in their achievable regrets. However, this is accomplished by incurring additional communication regrets caused by collisions, which has exponential dependencies on the attackability. Thus, we conclude that A2C2 establishes an alternative dimension in the hardness space from attackability, in additional to the original dimension from multiple players in [5].

The following preview of the analytical results provides a more clear view of the two different dimensions of hardness. Fig. 1 and Fig. 2 numerically illustrate the theoretical dependencies of the asymptotical regret (i.e., scaling) of A2C2 algorithms and the no-sensing algorithm in [5] on the two

dimensions of hardness. When fixing the attackability, i.e., fixing the local attackability parameter α to be 0.7 or the global attackability parameter β to be 0.7, the regrets of α-unaware A2C2 algorithm and the β-unaware A2C2 algorithm only increase slowly with M in Fig. 1, since their T terms are oblivious to M and the overall dependency on M is only a multiplicative factor. However, the regret of [5] increases sharply with more players due to the exponential dependency. When M is large (larger than 4 in Fig. 1), the advantage of A2C2 algorithms is obvious. On the other hand, while fixing the number of players, the regret performance of [5] is immune to the change of attackability. However, A2C2 algorithms have exponential dependencies on the attackability. As a result, their performances are very good when the adversary's attackability is weak or medium, but degrade quickly when the attackability is extremely strong.

III. Problem Formulation

A. The No-Sensing Adversarial MP-MAB Problem

We focus on the following decentralized no-sensing adversarial MP-MAB model. There are K arms and $1 < M \leq K$ players in the game.[1] The arms are labeled 1 to K and the players 1 to M, respectively. There are no *explicit* communications among players, which is a key constraint of the decentralized MP-MAB problem. The time horizon T is slotted and synchronized among players, and at each time step $t \in [T]$, each player $m \in [M]$ individually chooses and pulls arm $\pi_m(t)$. Simultaneously, an adversary selects loss $l_k(t)$ for each arm $k \in [K]$. The true loss $l_k(t)$ is player-independent and has a bounded support on $[0, 1]$. A *collision* happens if more than one player pull the same arm simultaneously. If no collision happens for player m at time t, she receives the true loss $l_{\pi_m(t)}(t)$; otherwise, she always receives loss 1 (i.e., the maximum loss) regardless of $l_{\pi_m(t)}(t)$. The actual loss $s_{\pi_m(t)}(t)$ received by player m at time t can be written as

$$s_{\pi_m(t)}(t) \coloneqq \underbrace{l_{\pi_m(t)}(t)\big(1 - \eta_{\pi_m(t)}(t)\big)}_{\text{no collision}} + \underbrace{\eta_{\pi_m(t)}(t)}_{\text{collision}},$$

where $\eta_{\pi_m(t)}$ is the collision indicator defined as $\eta_k(t) \coloneqq \mathbb{1}\{|C_k(t)| > 1\}$, with $C_k(t) \coloneqq \{n \in [M] | \pi_n(t) = k\}$.

If the players have access to both $s_{\pi_m(t)}(t)$ and $\eta_{\pi_m(t)}(t)$, it is a collision-sensing problem and player m makes decision $\pi_m(t)$ with past information $\{\pi_m(v), s_{\pi_m(v)}(v), \eta_{\pi_m(v)}(v)\}_{v<t}$. When information of $\eta_{\pi_m(t)}(t)$ is unavailable and players only know $s_{\pi_m(t)}(t)$, the problem is a no-sensing one as considered in this paper. In this no-sensing setting, a loss 1 can indistinguishably come from collisions or be exogenously generated by the adversary, and player m makes decisions $\pi_m(t)$ only with $\{\pi_m(v), s_{\pi_m(v)}(v)\}_{v<t}$. The lack of information on the collision indicators complicates the MP-MAB problem in general [3], [9], and this challenge is more significant for the adversarial setting [5]. Note that if $l_k(t) \neq 1$, $\forall k, t$, the no-sensing setting is equivalent to collision-sensing.

In the adversarial MP-MAB model, the notion of regret can be generalized with respect to the best allocation of players to arms as follows [12]:

$$R(T) \coloneqq \sum_{t=1}^{T} \sum_{m\in[M]} s_{\pi_m(t)}(t) - \min_{\substack{k_1,\ldots,k_M\in[K],\\ k_p\neq k_q, \forall p\neq q}} \sum_{t=1}^{T} \sum_{m\in[M]} l_{k_m}(t). \tag{1}$$

We are interested in the *expected regret* $\mathbb{E}[R(T)]$ where the expectation is with respect to the algorithm randomization.

As shown in [5], one cannot obtain any non-trivial regret guarantees facing an adaptive adversary. This work thus focuses on the *oblivious* adversarial MP-MAB where the reward generation of the adversary is independent of the actions of players. Equivalently, the loss sequence is chosen by the adversary at the beginning of the game.

A Motivating Example: The adversarial MP-MAB problem formulated in this section captures the key characteristics of a cognitive radio system, where arms correspond to channels, and players represent the distributed devices trying to communicate over the channels. The adversarial loss captures the unpredictable time-varying channel quality, e.g., due to bursty interference [29]. When two (or more) users use the same channel simultaneously, a packet collision happens and both of their attempted communications fail (i.e., a loss of 1). This is a commonly used model for shared wireless medium, e.g., the CSMA protocol in Wi-Fi [30]. Regarding whether the collision is perceivable or not (i.e., collision-sensing or no-sensing), it is determined by the communication protocol and the sensing capabilities of devices. As stated in [2], [4], [31], the no-sensing setting is more suited to large scale Internet-of-Things (IoT) applications. Furthermore, minimizing the regret defined in Eqn. (1) is equivalent to minimizing the cumulative communication loss over the chosen channels, which is a meaningful metric for practical systems. As a final remark, since the changing of channel quality is typically caused by external factors such as the dynamic radio frequency environment instead of user actions, the assumption of an oblivious adversary is reasonable in the application of cognitive radio.

B. Attackabilities of the Adversary

To explore the idea of forced-collision communication in the no-sensing adversarial setting, the overall horizon T is divided into the exploration and communication phases, similar to the approaches in collision-sensing settings [3], [12]. Information is shared by purposely created collisions in the communication phases to maintain synchronization and coordination between players in the subsequent exploration phases. However, in the no-sensing setting, loss 1 assigned by the adversary can be viewed as a certain "attack", since players have no knowledge whether it comes from the adversary or collision. Such loss-1 attack has very different impacts on the regret in different phases:

- *Exploration Phase:* The loss-1 attack does not negatively affect the exploration phase if the preceding communication phase is successful, as it would not ruin the synchronization among players. Especially, with successful synchronization, strategies can be designed to better allocate arms and minimize the regret during exploration phases.

[1]The special case of $M = 1$ is addressed in Appendix A.

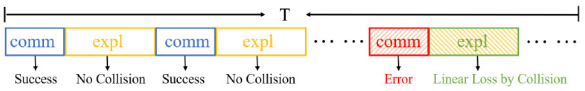


Fig. 3. Illustration of communication and exploration phases with communication errors.

- *Communication Phase:* The loss-1 attack in a communication phase requires special attention, as it may lead to communication errors for players, which jeopardize the essential synchronization among them and lead to a potential *linear* regret due to collisions in the subsequent exploration phase, as illustrated in Fig. 3.

Because of these different impacts, the worst-case adversarial attack scenario can have all-one loss sequences for all communication phases, which prevents any information sharing among players and may cause a linear loss over the entire T time slots.

From the previous studies in the stochastic setting [3], [9], it is clear that any bandit policy attempting to enable forced-collision communication in the no-sensing setting will have a dependency on the environment's ability to "attack" such communications. This naturally requires bounding the worst-case loss in communications. In the stochastic settings, this ability is characterized by a positive lower bound $\mu_{\min}$ such that $0 < \mu_{\min} \leq \min_{k\in[K]} \mu_k$, where μ_k is the mean of arm k's rewards. For example, such a lower bound is assumed to exist and be known to all players in [3], [9].

Analogous to the role of $\mu_{\min}$ in the stochastic MP-MAB models, we propose a new metric to characterize the adversarial environment, called the adversary's *attackability*, which represents the upper bound on the adversary's mechanism in generating loss 1's. This is a notable distinction to [5] where no communication is utilized in the no-sensing setting, and thus modeling the adversary's attackability is not necessary. More specifically, in this work, we define two types of attackabilities: the *local attackability* and the *global attackability*, which provide two different ways to *categorize* all the adversaries as detailed in this section.

First, the local attackability aims at modeling the one-time worst-case attack. In a communication phase shown in Fig. 3, the worst case is for this phase to see all loss 1's from the adversary, because no information can be reliably shared in such situation. In other words, the local attackability is captured by the maximum length of contiguous loss 1's assigned on the loss sequence, since it represents the longest duration that no reliable communication can happen. Without loss of generality, the local attackability is defined as follows.

Definition 1: For time horizon T, the *local attackability* $W(T)$ of the adversary is defined as $W(T) = \max_{k\in[K]} n_1^k(T)$, where $n_1^k(T)$ denotes the maximum length of the all-one loss sequences that are assigned by the adversary on arm k throughout the T time slots.

It is clear that every possible adversary is featured with a $W(T) \in [0, T]$. In the case that the adversary has $W(T) = \Omega(T)$, it may (asymptotically) attack at all time slots. If $W(T)$ of an adversary is a constant independent of T, i.e., $O(1)$, the attack is finite each time. A more general case lies in between these two extremes and any possible adversary can be characterized by the local attackability parameter α as follows.

Corollary 1: For any given adversary, there exists $\alpha \in [0, 1]$ such that the local attackability of this adversary satisfies $W(T) \leq O(T^\alpha)$.

The local attackability captures the one-time "budget" for the adversary attack. The adversaries sharing the same parameter α can be viewed as in the same category.

Another perspective is to consider the overall attacks over T as the *global attackability*. It captures the total amount of loss 1's assigned on one arm, and is defined as follows.

Definition 2: For given time horizon T, the *global attackability* $V(T)$ of the adversary is defined as $V(T) = \max_{k\in[K]} N_1^k(T)$, where $N_1^k(T)$ represents the total number of losses 1 that are assigned by the adversary on arm k throughout the T time slots.

Similar to $W(T)$, each possible adversary is featured with a $V(T) \in [0, T]$. If an adversary has $V(T) = \Omega(T)$, similar to $W(T) = \Omega(T)$, it may (asymptotically) attack at all time slots and no successful communication can happen. If $V(T)$ of an adversary is a constant independent of T, i.e., $O(1)$, the overall attacks are finite, which is negligible asymptotically and equivalent to the collision-sensing setting. Similarly, the adversaries can also be categorized with their global attackability parameter β as follows.

Corollary 2: For any given adversary, there exists $\beta \in [0, 1]$ such that the global attackability of this adversary satisfies $V(T) \leq O(T^\beta)$.

Similar to local attackability, the adversaries sharing the same parameter β can be viewed as in the same category, where the overall "budget" for the adversary attacks is of the same order. However, note that since the local attackability parameter α in Corollary 1 does not provide any bound on the overall attack budget, it is more stringent than the global attack parameter β in Corollary 2. For an adversary satisfying a sublinear local attackability, it is possible to have a global attackability of $\Omega(T)$. For example, the adversary with loss sequences of $0, 1, 0, 1, \ldots$ for all arms satisfies the local attackability of $W(T) = O(1)$, i.e., $\alpha = 0$, but also satisfies the global attackability of $V(T) = \Omega(T)$, i.e., $\beta = 1$. Finally, we note that the following sections consider either local or global attackability, but not simultaneously.

It is important to keep in mind that Corollaries 1 and 2 represent two ways of categorizing the adversaries rather than imposing constraints or requirements on them. Each category still has many adversaries as long as their scalings of attackability are the same, and every possible adversary is in a certain category. In addition, as shown in the subsequent sections, such categorization does not even need to be aware by the players. In this scenario, we do not impose more assumptions than [5]. Rather, the attackability view represents a different angle of the *same* no-sensing adversarial MP-MAB problem, and the proposed algorithms can adapt to the varying attackability in an automatic way, based on the perceived category of the adversary that it faces.

A final remark is that the proposed concepts of attackability are closely related to the different modes of external interference sources in the application of cognitive radio.

Specifically, the concept of local attackability is suitable for the low-duty-cycle interference source as it captures the essence of burst interference. On the other hand, global attackability focuses on measuring the overall interference and thus is more appropriate to model an interference source with a median or high duty cycle, or the cumulative interference from many low-duty-cycle sources.

IV. Algorithm Outline

All the algorithms proposed in this paper have two different phases: exploration phases and communication phases, and share a common leader–follower structure [12], [20]. Player 1 (leader) determines arm assignments for the remaining players (followers). The arm assignment is transmitted to each follower in the communication phases. Then, in the following exploration phases, all the players keep sampling the assigned arms. This section introduces the arm assignment procedure for the exploration phases and gives a brief introduction of the communication phases, which will be separately discussed in the following sections for different attackability scenarios.

A. Exploration Phase

Assuming explicit communications are allowed, i.e., in the centralized model, the challenge of exploration phases is how to choose M arms to explore for all the M players. We note that this is similar to the adversarial multi-play problem, where the leader (the centralized agent) chooses M arms $A(t) = \{A_1(t), \ldots, A_M(t)\}$ at each time step t for the followers to explore, i.e., player m is assigned with arm $A_m(t)$. As commonly adopted in the multi-play setting [32], each subset of M distinct arms $\{A_1, \ldots, A_M\}$ is viewed as a single meta-arm A to be chosen by the leader. The set of all meta-arms $\mathcal{K}$ is defined as:

$$\mathcal{K} \coloneqq \{\{A_1, \ldots, A_M\} \subseteq [K] | A_m \neq A_n \text{ for any } m \neq n\}.$$

An arm assignment policy that builds on [12] is proposed in this paper. At time $t-1$, players first explore the assigned arms in $A(t-1)$. Then, the leader updates an unbiased loss estimator for each arm $k \in [K]$ as $\tilde{l}_k(t-1) = M\frac{\hat{l}_{A_1(t-1)}(t-1)}{\sum_{A;k\in A\in\mathcal{K}} P_A(t-1)}\mathbb{1}\{k = A_1(t-1)\}$, where $P_A(t-1)$ is the probability that meta-arm A is chosen at time $t-1$ and $\hat{l}_{A_1(t-1)}(t-1)$ is the loss that the leader observes on her arm $A_1(t-1)$ at time $t-1$. Note that the update only requires past observations from the leader, which is designed to reduce the communication burden and to facilitate the generalization to the decentralized setting.[2]

For time t, the cumulative loss estimator $\tilde{L}_A(t)$ for each meta-arm $A \in \mathcal{K}$ is first updated as the sum of the loss estimations of its elementary arms up to time $t-1$, i.e., $\tilde{L}_A(t) = \sum_{\upsilon=1}^{t-1}\sum_{k\in A}\tilde{l}_k(\upsilon)$. Then, the EXP3 algorithm [33] is applied to the meta-arm MAB problem, so that each meta-arm $A \in \mathcal{K}$ is sampled with a probability $P_A(t)$ which is proportional to $\exp(-\eta\tilde{L}_A(t))$, as the exploration meta-arm $A(t)$ for time slot t. The loss estimator $\{\tilde{l}_k(t)\}_{k\in[K]}$ is then again updated after pulling the chosen meta-arm. At time $t+1$, the same procedures are performed to get $\{\tilde{L}_A(t+1)\}_{A\in\mathcal{K}}$, $\{P_A(t+1)\}_{A\in\mathcal{K}}$ and $A(t+1)$. As shown in [12] and Appendix B, this algorithm guarantees a regret bound of $2M\sqrt{K\log(K)T}$ when $\eta = \sqrt{\log\binom{K}{M}/MKT}$.

Furthermore, since there are $|\mathcal{K}| = \binom{K}{M}$ meta-arms, computing the probability $P_A(t)$ for each meta-arm and the marginal probability $\sum_{A:k\in A\in\mathcal{K}} P_A(t)$ for an arm k that is to be updated would lead to an exponential complexity if it is done naively. However, with a concept called K-DPPs [34], sampling and marginalization can be made more efficient. As shown in [12], the complexity of sampling a meta-arm and computing the marginal probability for a fixed arm can both be reduced to $O(KM)$ with K-DPPs, which makes it less complex for implementation.

Lastly, with the centralized algorithm described above, the key adjustment to the decentralized setting is to notify followers of their assigned arms by forced-collision communications. However, to avoid a linear communication regret due to frequently updating, the exploration phase is extended from one time slot to τ slots. This means each player is fixated on one arm for at least τ slots, and the update happens only after each exploration phase. The leader then uses her samples of losses observed during this entire exploration phase as the feedback to assign arms for the next phase. Note that although this infrequent switching reduces the communication burden, it also degrades the regret guarantees [35]; we will elaborate on this aspect in the analysis. When there is no ambiguity, the time variable in $A(t)$, $P_A(t)$, $\hat{l}_k(t)$, $\tilde{l}_k(t)$ and $\tilde{L}_A(t)$ are replaced by the corresponding phase index as $A(p)$, $P_A(p)$, $\hat{l}_k(p)$, $\tilde{l}_k(p)$ and $\tilde{L}_A(p)$ under the decentralized setting for the p-th phase.

B. Communication Phase

In the communication phases, arm assignments are transmitted from the leader to the followers with forced collisions. Functions `Send()` and `Receive()` are used in the algorithm description for the sending and receiving procedure with forced collisions. Every player is first assigned with a unique communication arm corresponding to her index, i.e., arm m for player m. In the collision-sensing setting, players only need to take predetermined turns to communicate by having the "receive" player sample her own communication arm and the "send" user either pull (create collision; bit 1) or not pull (create no collision; bit 0) the receive player's communication arm to transmit one-bit information. For a player that is not engaged in the current peer-to-peer communication, she keeps pulling her communication arm to avoid interrupting other ongoing communications. Since the collision indicator is perfectly known in the collision-sensing setting, player m can receive error-free information after implicit communication.

In the more challenging no-sensing setting, there is no information about the collision indicator, which means attacks, i.e., loss 1 assigned by the adversary, may cause communication errors and incur a linear regret in the subsequent exploration phase as shown in Fig. 3. The no-sensing settings are discussed under four different scenarios (two with

[2]We note that the recent advance [17] in stochastic MP-MAB proves that using information collected by only one leader is sufficient to have an optimal regret behaviour. However, it is unclear whether the similar argument holds for adversarial MP-MAB, which may be an interesting direction for future research.

Algorithm 1 α-Aware A2C2: Leader

Input: M, K, T

Initialize: $\tau \leftarrow \lceil M^{\frac{2}{3}} K^{-\frac{1}{3}} \log(K)^{\frac{1}{3}} T^{\frac{1+2\alpha+2\epsilon}{3}} \rceil$; $\eta \leftarrow \sqrt{\log\binom{K}{M}\tau/MKT}$

for $p = 1, 2, \ldots$ **do**

  $\forall A \in \mathcal{K}$, $\tilde{L}_A(p) \leftarrow \sum_{v=1}^{p-1} \sum_{k \in A} \tilde{l}_k(v)$

  $\forall A \in \mathcal{K}$, $P_A(p) \leftarrow \frac{e^{-\eta \tilde{L}_A(p)}}{\sum_{J \in \mathcal{K}} e^{-\eta \tilde{L}_J(p)}}$ ▷ Loss estimator

  Choose $A(p) = \{A_1(p), \ldots, A_M(p)\}$ with $P_A(p)$

  Randomly permute $A(p)$ into $\tilde{A}(p)$

▷ **Communication Phase:**

  $\forall m \in [M]$, $\text{msg}_m \leftarrow \text{rEncoder}(\tilde{A}_m(p), h(T, \alpha+\epsilon))$

  $\forall m \in [M]$, $\text{Send}\big(m, \text{msg}_m\big)$ ▷ Send Assignment

▷ **Exploration Phase:**

  Stay on arm $\tilde{A}_1(p)$ for τ time steps

  Record cumulative loss $\hat{l}_{\tilde{A}_1(p)}(p)$ ▷ Exploration

  $\forall k \in [K]$, $\tilde{l}_k(p) \leftarrow \frac{M}{\tau} \frac{\hat{l}_{\tilde{A}_1(p)}(p)}{\sum_{A: k \in A \in \mathcal{K}} P_A(p)} \mathbb{1}\{\tilde{A}_1(p) = k\}$

end for

Algorithm 2 α-Aware A2C2: Follower

Input: M, K, T, index m

Initialize: $\tau \leftarrow \lceil M^{\frac{2}{3}} K^{-\frac{1}{3}} \log(K)^{\frac{1}{3}} T^{\frac{1+2\alpha+2\epsilon}{3}} \rceil$; $\eta \leftarrow \sqrt{\log\binom{K}{M}\tau/MKT}$

for $p = 1, 2, \ldots$ **do**

▷ **Communication Phase:**

  $\text{msg}_m \leftarrow \text{Receive}(h(T, \alpha+\epsilon))$ ▷ Receive Assignment

  $\tilde{A}_m(p) \leftarrow \text{rDecoder}\big(\text{msg}_m, h(T, \alpha+\epsilon)\big)$

▷ **Exploration Phase:**

  Stay on arm $\tilde{A}_m(p)$ for τ time steps ▷ Exploration

end for

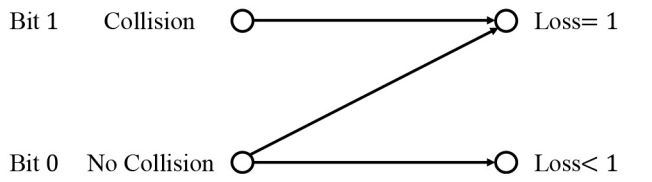


Fig. 4. The Z-channel model for implicit communications in no-sensing adversarial MP-MAB.

knowledge of attackability while the other two without such knowledge) in the following sections, respectively.[3]

- *α-Aware and β-Aware:* Section V-A (resp. Section V-B) proposes the α-aware (resp. β-aware) A2C2 algorithm with Corollary 1 (resp. 2) and players having knowledge of α (resp. β).
- *α-Unaware and β-Unaware:* Similar to the above but players have no knowledge of α (resp. β). These are reported in Sections VI-A and VI-B, respectively. This is the more challenging case, and the main focus of this work.

V. Attackability Known to Players

In this section, the attackability is assumed to be perfectly known by all players, but such information does not tell the players how and when the attacks would happen. The two definitions of attackability lead to two algorithms, which also serve as the building blocks of subsequent algorithm designs on the attackability-unaware setting.

A. α-Aware

Although the local attackability is theoretically more stringent than the global attackability, it is relatively easier to handle. The α-aware A2C2 is presented in Algorithm 1 (leader) and Algorithm 2 (follower). Two information-theoretic concepts that are intimately related to reliable communications but less utilized in the bandit literature are introduced; similar ideas are also adopted in algorithm designs under other scenarios.

Z-channel Model: There exists an asymmetry in collision communication of MP-MAB: bit 1 (collision) is always received correctly, while bit 0 (no collision) can be potentially corrupted by a loss 1 from the adversary. In other words, the adversary attack is *asymmetric* – she can attack bit 0 but not bit 1. From an information-theoretic point of view, this corresponds to a Z-channel model [36] as shown in Fig. 4. We note that this connection to the Z-channel model was first utilized in [9] to study stochastic no-sensing MP-MAB. A key challenge in the adversary setting as compared to [9], however, is that a *fixed* crossover probability does not exist.

Error-Correction Code With Long Blocklength: The idea of utilizing error-correction code naturally arises under the formulation of a Z-channel model. With the knowledge of α, the key idea to overcome the full attackability is to "overpower" the adversary via codes that have sufficient error-correction capabilities [37]. Different error-correction codes, ranging from simple repetition code to more complex (and powerful) algebraic and nonlinear codes, can be adopted in the proposed algorithms. To facilitate the regret analysis, we choose to use the repetition code [38] and functions `rEncoder()` and `rDecoder()` are used in the algorithms as the encoder and decoder.[4] At the encoder, each information bit is expanded to a string of length $h(T, \alpha+\epsilon) = \Theta(T^{\alpha+\epsilon}) = \omega(T^{\alpha})$, where $\epsilon > 0$ is a fixed constant which can be arbitrarily small.[5] Then, the coded bits are sent via forced collisions. If the received bit sequence of length $h(T, \alpha+\epsilon)$ is all-one, then the decoder outputs bit 1. Otherwise, it outputs bit 0 (i.e., as long as there exists at least one received bit 0 in the sequence). With $h(T, \alpha+\epsilon) = \omega(T^{\alpha})$, which implies $h(T, \alpha+\epsilon) = \omega(W(T))$, this error-correction code is guaranteed to overpower the local attackability *asymptotically*. Thus, all the communications are guaranteed to be successful *asymptotically*.

B. β-Aware

In the α-aware case, although the local attackability is bounded, the adversary can attack *arbitrarily many* times. This

[3]The discussions focus on the no-sensing setting, but the proposed algorithms can be easily modified to cover the collision-sensing setting, with details provided in Appendix A.

[4]Note that more advanced codes can be used, but our regret analysis shows that the regret scaling is not affected.

[5]$f(T) = \Theta(g(T))$ iff $f(T) = O(g(T))$ and $f(T) = \Omega(g(T))$; $f(T) = \omega(g(T))$ iff $\liminf_{T\to\infty} \frac{f(T)}{g(T)} = \infty$.

is the fundamental reason why each of the communication phases needs protection. However, with an upper bound of global attackability, the adversary has an overall budget for attacking rather than a one-time budget. The α-aware A2C2 algorithm can still be applied by replacing α with β, but it is an overkill to prevent the global attackability in *every* communication phase.

A more efficient algorithm, β-aware A2C2, is proposed with a holistic consideration for the total budget of the adversary. Details of β-aware A2C2 can be found in Appendix D, and we only highlight the key design philosophies here. Because of the iteration between exploration and communication phases, if the adversary succeeds in attacking one communication phase, the immediately following exploration phase of length τ could be out of synchronization. However, after time τ, players enter the communication phase again and a new iteration starts. In other words, attacks in one communication phase can only cause a one-time linear loss in the next (finite) τ time slots, while the adversary has a reduced total budget for future attacks. By the time that the adversary runs out of budget, no more communication errors can happen. Thus, as long as the loss caused by the global attackability does not dominate the regret, a certain amount of errors are tolerable. This is a key observation because it means that the power of error-correction coding does not need to be too strong. Technically, instead of coding with a long blocklength of $h(T, \alpha+\epsilon) = \Theta(T^{\alpha+\epsilon})$, a much shorter coded length of $k(T, \nu) = \Theta(T^{\nu})$, where $\nu = \max\{\frac{3\beta-1}{2}, 0\}$, is sufficient of achieving a sublinear regret that is better than α-aware A2C2. In fact, a closer look reveals a very surprising result: for $\beta \leq \frac{1}{3}$, we have $\nu = 0$, which means there is no need for coding at all in communication phases (see Section VII for details).

VI. Attackability Unknown to Players

In this section, the assumption of the knowledge of attackability is removed. No information on the parameter α or β is revealed to the players. The α-unaware and β-unaware settings are tackled separately in the subsequent subsections.

A. α-Unaware

With no information of α, the main difficulty lies in how to prepare for the worst case without incurring a linear loss. All the key features in α-aware A2C2 are still applied in the adaptive algorithm called α-unaware A2C2, but several new ideas are needed: an error-detection code to estimate α, and a synchronization procedure with randomized length to synchronize the estimation update among players. The algorithms for the leader and followers are presented in Algorithms 3 and 4, respectively.

Estimation of α: Without the knowledge of α, no effective prevention is possible for communications in the worst-case scenario. We propose to adaptively estimate α in an escalation fashion. The interval [0, 1] (support of α) is uniformly divided into sub-intervals with length ϵ, where $\epsilon > 0$ is an arbitrarily small constant. The estimated value α' starts with $\alpha' = 0$, and increases with a step size of ϵ while a communication failure is observed until the upper limit is reached. As we see in the regret analysis, this seemingly naive estimation works very well.

Error-Detection Repetition Code: The aforementioned escalation mechanism to estimate α relies on knowing when communication failure happens, which is non-trivial. This leads to the second idea of utilizing a special kind of error-detection code for the Z-channel, called the constant weight code [39]. Codewords in one constant weight code share

Algorithm 3 α-Unaware A2C2: Leader

Input: M, K, T
Initialize: α estimation: $\alpha' \leftarrow 0$; error flag: $F \leftarrow 0$
for $p = 1, 2, \ldots$ **do**
  $\tau \leftarrow \lceil M^{\frac{2}{3}} K^{-\frac{1}{3}} \log(K)^{-\frac{1}{3}} T^{\frac{2+\alpha'}{3}} \rceil$
  $\eta \leftarrow \sqrt{\log\binom{K}{M}\tau/MKT}$; $\xi \leftarrow \frac{1-\alpha'}{2}$; $F \leftarrow 0$
  $\forall A \in \mathcal{K}, \tilde{L}_A(p) \leftarrow \sum_{v=1}^{p-1} \sum_{k \in A} \tilde{l}_k(v)$
  $\forall A \in \mathcal{K}, P_A(p) \leftarrow \frac{e^{-\eta \tilde{L}_A(p)}}{\sum_{J \in \mathcal{K}} e^{-\eta \tilde{L}_J(p)}}$ ▷ Loss estimator
  Choose $A(p) = \{A_1(p), \ldots, A_M(p)\}$ with $P_A(p)$
  Randomly permute $A(p)$ into $\tilde{A}(p)$
 ▷ **Communication Phase:**
  $\forall m \in [M]$, $\text{msg}_m \leftarrow \text{eEncoder}(\tilde{A}_m(p), h(T, \alpha'))$
  $\forall m \in [M]$, $\text{Send}\big(m, \text{msg}_m\big)$ ▷ Send Assignment
  for $q = 1, 2, \ldots, N(\xi)$ **do** ▷ Synchronization
    $\text{msg}_F \leftarrow \text{Receive}\big(h\big(T, \alpha'\big)\big)$
    $F \leftarrow \text{rDecoder}\big(\text{msg}_F, h\big(T, \alpha'\big)\big)$ ▷ Uplink: $F = 0/1$
    $\text{msg}_F \leftarrow \text{rEncoder}\big(F, h(T, \alpha')\big)$
    $\forall m \in [M]$, $\text{Send}\big(m, \text{msg}_F\big)$ ▷ Downlink: $F = 0/1$
  end for
  $\alpha' \leftarrow \alpha' + F\epsilon$ ▷ Update Estimation
 ▷ **Exploration Phase:**
  Stay on arm $\tilde{A}_1(p)$ for τ time steps
  if $F = 0$ **then**
    Record cumulative loss $\hat{l}_{\tilde{A}_1(p)}(p)$
    $\forall k \in [K], \tilde{l}_k(p) \leftarrow \frac{M}{\tau} \frac{\hat{l}_{\tilde{A}_1(p)}(p)}{\sum_{A: k \in A \in \mathcal{K}} P_A(p)} \mathbb{1}\{\tilde{A}_1(p) = k\}$
  else $\forall k \in [K], \tilde{l}_k(p) \leftarrow 0$
  end if
end for

Algorithm 4 α-Unaware A2C2: Follower

Input: M, K, T, index m
Initialize: α estimation: $\alpha' \leftarrow 0$; error flag: $F \leftarrow 0$; exploration set: $\mathcal{S} \leftarrow \emptyset$
for $p = 1, 2, \ldots$ **do**
  $\tau \leftarrow \lceil M^{\frac{2}{3}} K^{-\frac{1}{3}} \log(K)^{-\frac{1}{3}} T^{\frac{2+\alpha'}{3}} \rceil$
  $\eta \leftarrow \sqrt{\log\binom{K}{M}\tau/MKT}$; $\xi \leftarrow \frac{1-\alpha'}{2}$; $F \leftarrow 0$
 ▷ **Communication Phase:**
  $\text{msg}_m \leftarrow \text{Receive}\big(h(T, \alpha')\big)$
  $\mathcal{S} \leftarrow \text{eDecoder}\big(\text{msg}_m, h(T, \alpha')\big)$ ▷ Receive Assignment
  Randomly choose in $\mathcal{S}$ for $\tilde{A}_m(p)$
  $F \leftarrow \mathbb{1}\{|\mathcal{S}| > 1\}$ ▷ Comm Error or not
  for $q = 1, 2, \ldots, N(\xi)$ **do**
    $\text{msg}_F \leftarrow \text{rEncoder}\big(\text{F}, h(T, \alpha')\big)$
    $\text{Send}(1, \text{msg}_F)$ ▷ Uplink: $F = 0/1$
    $\text{msg}_F \leftarrow \text{Receive}\big(h(T, \alpha')\big)$
    $F \leftarrow \text{rDecoder}\big(\text{msg}_F, h(T, \alpha')\big)$ ▷ Downlink: $F = 0/1$
  end for
  $\alpha' \leftarrow \alpha' + F\epsilon$ ▷ Update Estimation
 ▷ **Exploration Phase:**
  Stay on arm $\tilde{A}_m(p)$ for τ time steps
end for

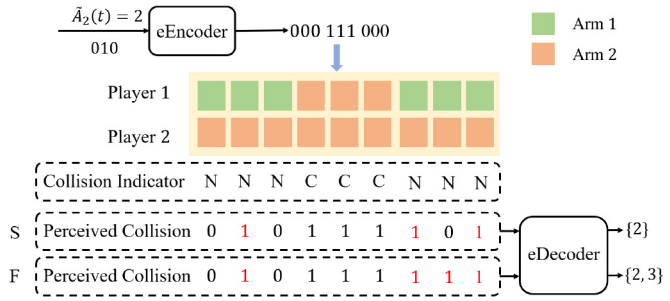


Fig. 5. Example of one communication phase and error-detection coding with $M = 2$, $K = 3$ and $h(T, \alpha') = 3$.

the same Hamming weight, which enables error detection. As noted in [39], a constant weight code can detect any number of asymmetric errors, and the maximal number of constant-weight codewords of length n can be attained by taking all codewords of weight $\lceil \frac{n}{2} \rceil$ or $\lfloor \frac{n}{2} \rfloor$. Thus, codeword length of $O(\log(K))$ is theoretically sufficient to enable a constant-weight code with $O(K)$ codewords.[6]

To facilitate the discussion, a particular kind of constant weight code is adopted. As shown in Fig. 5, while transmitting arm k's index from the leader, it is represented by a bit sequence of length K in which the k-th bit is 1 while all other bits are 0. Then, each bit of this sequence is repeated $h(T, \alpha') = \Theta(T^{\alpha'})$ times. Thus, all the codewords share the weight of $h(T, \alpha')$. The resulting coded bit sequence is then transmitted with forced collisions. Upon receiving, the entire bit string is divided into K blocks and each block is processed separately. The decoder outcome $\mathcal{S}$ is the (possibly multiple) indices of the blocks that have all-ones.

This is a very effective error *detection* method because, based on the property of the Z-channel, the source index k is always decoded correctly. Thus, if the decoder outputs more than one indexes, there must be a communication error (example 'F' in Fig. 5), meaning the current communication protocol with α' is not strong enough to overcome the attacks. Otherwise, there is only one index as the decoder output, which means the communication is successful (example 'S' in Fig. 5). This suggests that it is sufficient to maintain the current estimation α'. Once $\alpha' = \nu\epsilon > \alpha$, we have $h(T, \alpha') = \Theta(T^{\alpha'}) = \omega(W(T))$, which means there will be no more communication errors (asymptotically) and α' can be used for the remainder of time slots. Functions `eEncoder()` an `eDecoder()` are used in the algorithms for the corresponding error-detection encoder and decoder, respectively.

Synchronization With Randomized Length: Error-detection repetition code allows each follower to decide whether the α-estimation needs to be updated. However, the leader does not have access to this information and such estimation *across* players may not be synchronized, which poses a significant challenge that calls for communication for synchronization. Note that unlike communications for arm assignment, not all errors in this procedure are equally bad. The worst case during synchronization is *uneven attacks*: attacks that happen only on a subgroup of players, i.e., some players receive incorrect signals and update the α estimation, while the others do not. On the other hand, if *all* players receive wrong signals simultaneously, they are still synchronous because they all will increase α' to the next value.

To solve this problem, we return to the fundamental technique of (single-player) adversarial bandits and introduce *randomness* to the synchronous procedure. After communication of arm assignment in each round, the followers report to the leader whether communication errors occurred in this round at the same time. The followers send bit 1 representing error and bit 0 representing no error with the same error-correction repetition code of length $h(T, \alpha')$. Borrowing the terminology from wireless communications, this process is referred to as 'uplink', and as long as one follower has communication errors, the leader can correctly receive bit 1 (signal for updating). The uplink is robust to the attacks since there is only one receiver (the leader); even if an attack is successful, the need for updating is still conveyed correctly. After the uplink, if bit 1 (no matter from followers or the adversary) is received by the leader, she sends back bit 1 to every follower; otherwise, bit 0 is sent. This is called 'downlink'.[7] If bit 1 is sent, all the followers receive it correctly over the Z-channel. However, in the case of bit 0, uneven attacks on partial followers may happen, which leads to a failed synchronization. To keep synchronization successful with a high probability, the 'uplink-downlink' procedure keeps iterating for a *random* number of rounds $N(\xi)$, which is generated by a shared randomness source among players[8] and uniformly distributed in $[0, \lceil T^{\xi} \rceil]$. If the follower detects communication errors during the assignment or receives bit 1 in the preceding downlink, she keeps sending bit 1 to the leader in the following uplink cycles until the procedure ends. For this protocol, the adversary has to *exactly* attack the last round of downlink to destroy synchronization, since attacks at other rounds are broadcasting. As analyzed in Section VII, this procedure with a carefully chosen $\xi = \frac{1-\alpha'}{2}$ (which changes with α') is crucial in maintaining a sub-linear regret.

In addition to updating the estimation, the synchronization also plays an important role in maintaining unbiased loss estimations. With communication errors, potential collisions may happen on the assigned arm of the leader, thus an unbiased estimation is impossible if the leader uses the observed rewards of this round as feedback. The algorithm allows the leader to only use collected cumulative rewards when the signals from followers indicate correct communications.

With full details given in Algorithms 3 and 4, we offer a summary overview of the algorithmic structure of α-unaware A2C2 as follows.

[6]This observation can be verified by invoking the Stirling's formula: $n! \approx \sqrt{2\pi n}(\frac{n}{e})^n$.

[7]Note that 'uplink' denotes that followers communicate to the leader, while 'downlink' denotes the opposite direction, i.e., the leader communicates to the followers.

[8]Specifically, we can assume that players have access to a shared random seed in advance, e.g., the starting time of the game, and then use this random seed to generated $N(\xi)$ for later synchronizations.

- *Step I (Meta-Arm Selection):* With the arm assignment policy specified in Section IV-A, a meta-arm is chosen by the leader;
- *Step II (Communication Phase for Arm Assignment):* The leader encodes the indices of the chosen arms with the error-detection code. Then, she informs each follower of her assigned arm by implicitly communicating the coded arm index. Each follower decodes the received message, and detects whether a communication error has occurred.
- *Step III (Communication Phase for Synchronization):* A random length of "uplink-downlink" iteration is performed, which composes of the synchronization procedure. After synchronization, each player updates the estimation of the attackability parameter α'.
- *Step IV (Exploration Phase):* The players pull the assigned arms for a certain duration. The leader counts the cumulative losses which will be used by the loss estimators.

Similar structures also apply to the previously discussed α/β-aware A2C2 algorithms by removing the communication phase for synchronization. For the β-unaware A2C2 algorithm, detailed changes are discussed in the next subsection.

B. β-Unaware

Similar ideas of the previous design can be applied to the β-unaware setting, and the resulting β-unaware A2C2 algorithm is presented in Appendix F. Some important differences to α-unaware A2C2 are explained in this subsection.

Unlike the α-estimation starting from $\alpha' = 0$, the estimation β' starts with $\frac{1}{4}$ as analyzed in Section VII. In each communication phase, the arm assignment is similarly encoded with the error-detection code but the codeword length for each bit is adjusted to $k(T, \nu') = \Theta(T^{\nu'})$, where $\nu' = \frac{4\beta'-1}{3}$. The increased coding rate is due to the same reason described in Section V-B, i.e., a certain amount of communication errors are tolerable within the global attackability bound.

As for the feedback from the followers, two different uplink operations are performed: one to maintain the unbiased loss estimations and the other to keep players synchronized with the estimated β'. First, after each arm assignment, followers immediately notify the leader of communication errors for her to maintain an unbiased loss estimation. This uplink does not indicate the need of updating β' since β' is an estimation of the *overall* budget and there is no following downlink. Since communication errors in this uplink can only influence the subsequent exploration phase, it is performed with a repetition code of length $k(T, \nu') = \Theta(\nu')$. Then, for the update of β', each player keeps counting the overall number of attacks on her communication arm and reports to the leader when the estimated budget is exceeded. To reduce the communication burden, the update of β' and the synchronization procedure is performed only at potential updating time slots rather than after each communication phase. Specifically, similar iterations of uplink and downlink for synchronization happen every $\lceil T^{\beta'}/k(T, \nu') \rceil$ phases, with length $k(T, \beta') = \Theta(T^{\beta'})$ in each round and a random number of rounds $N(\xi) \in [0, \lceil T^{\xi} \rceil]$. The choice of length $k(T, \beta')$ is because that synchronization error may influence the entire remaining time slots. Similar to analysis in Section VI-A, the adversary must attack *exactly* the last round of synchronization to succeed in breaking the coordination among players.

VII. Performance Analysis

This section is devoted to the theoretical analyses for all proposed A2C2 algorithms. Detailed proofs can be found in Appendices C to F.

α/β-Aware: The regret of α-aware A2C2 and β-aware A2C2 algorithms are first presented in Theorems 1 and 2, respectively.

Theorem 1 (α-Aware): With $\tau = \lceil M^{\frac{2}{3}} K^{-\frac{1}{3}} \log(K)^{\frac{1}{3}} T^{\frac{1+2\alpha+2\epsilon}{3}} \rceil$, $\eta = \sqrt{\log \binom{K}{M} \tau / MKT}$, the expected regret of α-aware A2C2 algorithm is bounded by

$$\mathbb{E}[R_1(T)] \leq O\Big(M^{\frac{4}{3}} K^{\frac{1}{3}} \log(K)^{\frac{2}{3}} T^{\frac{2+\alpha+\epsilon}{3}}\Big),$$

where $\epsilon > 0$ is an arbitrarily small constant.

Theorem 2 (β-Aware): With $\tau = \lceil K^{\frac{1}{3}} \log(K)^{-\frac{1}{3}} T^{\max\{\beta, \frac{1}{3}\}} \rceil$, $\eta = \sqrt{\log \binom{K}{M} \tau / MKT}$, $\nu = \max\{\frac{3\beta-1}{2}, 0\}$, the expected regret of β-aware A2C2 algorithm is bounded by

$$\mathbb{E}[R_2(T)] \leq O\left(M^2 K^{\frac{2}{3}} \log(K)^{\frac{1}{3}} T^{\max\left\{\frac{1+\beta}{2}, \frac{2}{3}\right\}}\right).$$

It is worth noting that for $\beta \leq \frac{1}{3}$, we have $\nu = 0$ which means there is no need for coding at all. Another note is that for $\alpha = 0$ or $\beta \leq \frac{1}{3}$, the known regret in collision-sensing setting $O(T^{\frac{2}{3}})$ [12] is recovered. When $\alpha = 1$ or $\beta = 1$, which means $W(T) = \Omega(T)$ or $V(T) = \Omega(T)$, the adversary can asymptotically attack all time slots to prevent any communication, and thus our regret becomes $O(T)$.

α/β-Unaware: Without knowledge of the attackability, the performance of α-unaware A2C2 and β-unaware A2C2 algorithms are guaranteed in Theorems 3 and 4, and analyzed subsequently.

Theorem 3 (α-Unaware): With $\tau = \lceil M^{\frac{2}{3}} K^{-\frac{1}{3}} \log(K)^{-\frac{1}{3}} T^{\frac{2+\alpha'}{3}} \rceil$, $\eta = \sqrt{\log \binom{K}{M} \tau / MKT}$ and $\xi = \frac{1-\alpha'}{2}$ under the estimation α', the expected regret of the α-unaware A2C2 algorithm is bounded by

$$\mathbb{E}[R_3(T)] \leq O\Big(M^{\frac{4}{3}} K^{\frac{1}{3}} \log(K)^{\frac{1}{3}} T^{\frac{5+\alpha+\epsilon}{6}}\Big),$$

where $\epsilon > 0$ is an arbitrarily small constant.

Theorem 4 (β-Unaware): With $\tau = \lceil K^{-\frac{1}{3}} \log(K)^{-\frac{1}{3}} T^{\frac{1+2\beta'}{3}} \rceil$, $\eta = \sqrt{\log \binom{K}{M} \tau / MKT}$, $\xi = \frac{1+2\beta'}{3}$ and $\nu' = \frac{4\beta'-1}{3}$ under estimation β' starting from $\frac{1}{4}$, the expected regret of the β-unaware A2C2 algorithm is bounded by

$$\mathbb{E}[R_4(T)] \leq O\left(M^2 K^{\frac{1}{3}} \log(K)^{\frac{1}{3}} T^{\max\left\{\frac{2+\beta+\epsilon}{3}, \frac{3}{4}\right\}}\right),$$

where $\epsilon > 0$ is an arbitrarily small constant.

Similar to the β-aware case, we have $\nu' = 0$ for $\beta \leq \frac{1}{4}$ in the β-unaware A2C2 algorithm, which indicates there is no need of coding for assigning arms and reporting communication errors. Compared with Theorem 1, the lack of knowledge of α worsens the regret by a factor of $O(T^{\frac{1-\alpha}{6}})$ in Theorem 3.

Similarly, the lack of knowledge of β degrades the regret by $O(T^{\frac{1-\beta}{6}})$. It is also worth noting that Theorems 2 and 4 provide better dependencies on T than Theorems 1 and 3, respectively, which reinforces the intuition that the local attackability parameter α in Corollary 1 is more stringent than the global attackability parameter β in Corollary 2.

Compared with the regret of $O(MK^{\frac{3}{2}}T^{1-\frac{1}{2M}})$ in [5], it can be observed that the regret results of A2C2 have an exponential dependence on the attackability rather than the number of players M, which could be an advantage while dealing with a large number of players. From another perspective, these two different dependencies reveal two orthogonal "dimensions of hardness" in the no-sensing adversarial MP-MAB problem: multiple players and attackability. As no information sharing among players is utilized in [5], the coordination is limited and the difficulty of the problem grows exponentially with the number of players. In our work, forced collisions are used for communications and coordination among players is established. As a result, the regret shifts the exponential dependence from number of players (M) to attackability (α or β), and the dependence on M is only a multiplicative factor.

A. Proof Sketch: α-Unaware

Regret of the α-unaware A2C2 algorithm can be decomposed as $R_3(T) = R_3^{expl}(T) + R_3^{comm}(T) + R_3^{sync}(T)$, corresponding to the exploration regret, communication regret, and synchronization error regret, respectively. The remaining of this subsection is devoted to a brief discussion of each regret term in the decomposition and its corresponding upper bound.

First, the exploration regret $R_3^{expl}(T)$ can be further divided into two parts depending on whether the preceding communication phase is successful or not. The first part has a regret caused by exploration without collision. The second part is caused by potential collisions due to the failed communications, which only occur with an underestimated attackability. Lemma 1 presents an upper bound for the overall exploration regret.

Lemma 1: Denoting $\zeta = \mathbb{1}\{\text{successful preceding communication}\}$, the expected exploration regret of the α-unaware A2C2 algorithm is bounded by:

$$\begin{aligned}\mathbb{E}\Big[R_3^{expl}(T)\Big] &= \mathbb{E}\Big[R_3^{expl}(T|\zeta=1)\Big] + \mathbb{E}\Big[R_3^{expl}(T|\zeta=0)\Big] \\ &\le O\Big(M^{\frac{4}{3}}K^{\frac{1}{3}}\log(K)^{\frac{1}{3}}T^{\frac{5+\alpha+\epsilon}{6}}\Big).\end{aligned}$$

The communication regret of α-unaware A2C2 consists of two parts: the first from arm assignments and the second from synchronizations, which can be bounded as follows.

Lemma 2: The expected communication regret of α-unaware A2C2 is bounded by

$$\mathbb{E}\big[R_3^{comm}(T)\big] \le O\Big(M^{\frac{4}{3}}K^{\frac{1}{3}}\log(K)^{\frac{1}{3}}T^{\frac{5+\alpha+\epsilon}{6}}\Big).$$

Finally, the regret $R_3^{sync}(T)$ caused by the risk of losing synchronization during updates is bounded in the following lemma, where the worst case is assumed such that linear regret is caused by collisions. Note that this term is unique in the attackability-unaware setting as there is no synchronization performed in the attackability-aware setting.

Lemma 3: The expected regret caused by potential synchronization errors of α-unaware A2C2 is bounded by:

$$\mathbb{E}\big[R_3^{sync}(T)\big] \le O\Big(M^{\frac{1}{3}}K^{\frac{1}{3}}\log(K)^{\frac{1}{3}}T^{\frac{5+\alpha+\epsilon}{6}}\Big).$$

Combining Lemmas 1 to 3, the overall regret in Theorem 3 is proved. Note that all three component regret bounds share the same order of T and similar factors of M and K, which is not a coincidence. As the analysis in Appendix E shows, these three terms can be further decomposed based on the estimation α', and they are all dominated by the elements associated with the final (largest) estimation. Optimizations over τ and ξ are carried out based on these elements, which carefully control the regret associated with the communication error and the synchronization error to avoid one component regret dominating others.

B. Proof Sketch: β-Unaware

The regret of β-unaware A2C2 can be decomposed as $R_4(T) = R_4^{expl}(T) + R_4^{err}(T) + R_4^{comm}(T) + R_4^{sync}(T)$, which refers to the exploration regret with successful preceding communication phases, exploration regret with failed preceding communication phases, communication regret, and synchronization error regret, respectively. The decomposed regret components $R_4^{expl}(T)$, $R_4^{comm}(T)$ and $R_4^{sync}(T)$ are similar to the corresponding components in the proof of α-unaware A2C2, which are bounded in Lemmas 4 to 6.

Lemma 4: The expected regret caused by explorations after successful communications in β-unaware A2C2 is bounded by:

$$\mathbb{E}\Big[R_4^{expl}(T)\Big] \le O\Bigg(MK^{\frac{1}{3}}\log(K)^{\frac{1}{3}}T^{\max\left\{\frac{2+\beta+\epsilon}{3},\frac{3}{4}\right\}}\Bigg).$$

Lemma 5: The expected communication regret of β-unaware A2C2 is bounded by:

$$\mathbb{E}\big[R_4^{comm}(T)\big] \le O\Bigg(M^2K^{\frac{1}{3}}\log(K)^{\frac{1}{3}}T^{\max\left\{\frac{2+\beta+\epsilon}{3},\frac{3}{4}\right\}}\Bigg).$$

Lemma 6: The expected regret caused by potential synchronization errors of β-unaware A2C2 is bounded by:

$$\mathbb{E}\big[R_4^{sync}(T)\big] \le O\Bigg(MK^{\frac{1}{3}}\log(K)^{\frac{1}{3}}T^{\max\left\{\frac{2+\beta+\epsilon}{3},\frac{3}{4}\right\}}\Bigg).$$

Since β-unaware A2C2 is designed to be capable of handling a certain amount of communication errors, the exploration regret associated with failed communications (i.e., $R_4^{err}(T)$) is separately analyzed, as opposed to being lump summed into one exploration loss as in the local attackability analysis.

Lemma 7: The expected regret caused by explorations after failed communications in β-unaware A2C2 is bounded by:

$$\mathbb{E}\big[R_4^{err}(T)\big] \le O\Bigg(M^2K^{\frac{1}{3}}\log(K)^{-\frac{1}{3}}T^{\max\left\{\frac{2+\beta+\epsilon}{3},\frac{3}{4}\right\}}\Bigg).$$

In addition to similar optimizations over τ and ξ as in the analysis of α-unaware-A2C2, the choice of ν' is also optimized so that the exploration regret caused by the global attackability, i.e., $\mathbb{E}[R_4^{err}(T)]$, does not dominate the total regret.

VIII. Discussion

We discuss some algorithmic details and a few directions for potential future research.

Sublinear Regrets: As shown in Theorems 1 to 4, the regrets of the A2C2 algorithms can be sublinear as long as the adversary cannot (asymptotically) attack all time slots, i.e., $\alpha < 1$ or $\beta < 1$, which is similar to the no-sensing stochastic setting that sublinear regrets can be achieved only when $\mu_{\min} > 0$. Further, as stated in Section III-B, it is possible that $\beta = 1$ while $\alpha < 1$, and in such cases, α-(un)aware A2C2 can still achieve a sublinear regret.

The Choice of ϵ: With an unknown attackability, i.e., α or β-unaware A2C2, ϵ can be an arbitrarily small constant in the asymptotic case. In the finite case, this choice influences both the convergence of estimation and the regret behavior. A larger ϵ can let the algorithm escalate faster but with a potential overestimation and a larger regret. With a smaller ϵ, although the escalation is slower, better performance can be achieved with a more precise estimation.

Implicit Communication of Loss Information: As stated in Section IV-A, the choices of meta-arms in A2C2 algorithms are only based on observations of the leader. While this simplifies communications, an open problem is how to transmit the collected loss information between players through collisions. The methods of arm index sharing proposed in this work is a starting point, but more careful designs may further reduce the regret, e.g., quantizing the losses.

Combining A2C2 and [5]: Without a mature method of statistics sharing, another idea to increase the information usage is to have multiple leaders instead of one, which alludes to a hybrid algorithm of this work and [5]. Specifically, each leader with a certain number of followers can form as a meta-player. Among the meta-players, the method from [5] can be adopted since it does not require information sharing. Within each meta-player, A2C2 algorithms can be performed to let the leader coordinate the followers. It would be interesting to see a rigorous analysis of this hybrid algorithm, especially its dependence on M.

Dynamic Setting: While this work assumes the players are fixed in the entire game, it is an interesting (but also very challenging) open problem to consider the dynamic setting, where players can enter and leave the game freely. Some attempts have been made in the stochastic MP-MAB [3], [13] and the collision-sensing adversarial MP-MAB [14], while the no-sensing adversarial dynamic environment remains largely open. For the attackability-aware setting (i.e., α/β-aware A2C2), algorithmic extensions are relatively easy since the new players can be synchronized with the already existing players using the knowledge of attackability, but the theoretical analysis for this case may be difficult. If the new players do not have information of the current estimation of attackability (i.e., α/β-unaware A2C2), this extension becomes difficult. With some constraint on the frequency of players entering or leaving the game, it is conceivable to run α/β-unaware A2C2 independently on each of these intervals, which is similar to [13]. Nevertheless, its regret analysis would be highly nontrivial.

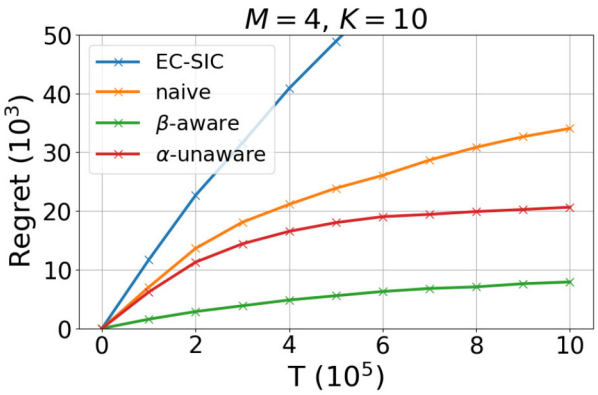


Fig. 6. Regret comparison where the loss generations are non-stationary in the entire horizon.

IX. Experiments

More numerical illustrations comparing A2C2 and the algorithm in [5] are provided in Appendix G. The A2C2 algorithms are also empirically evaluated against a naive adversarial algorithm as well as the EC-SIC design in [9], which is a stochastic no-sensing algorithm. In the naive adversarial algorithm, each player individually runs EXP3 [33] using her own observed loss information, which results in M parallel EXP3 algorithms. The EC-SIC algorithm is the state-of-the-art stochastic no-sensing algorithm, which also utilizes error-correction coding but is confined to stochastic loss distributions. The testing environments have $M = 4$ and $K = 10$ (same as the numerical illustrations in Fig. 2) and the step-size parameter ϵ is set as 0.01. The following figures are plotted by running the algorithms under different time horizons and the regret result of each horizon is the average over 50 runs.

The first loss sequence for evaluation is generated similar to [14]. Specifically, for arm k, a random variable c_k is first sampled from the uniform distribution on $[0.2, 0.9]$, then $l_k(t)$ is drawn from the uniform distribution on $[c_k, 0.9]$. This results in a non-stationary environment. After the generation, to explicitly characterize the adversarial attackability, contiguous loss-1 sequences of length 50 are randomly spread in the whole loss sequence to replace the originally generated losses. Fig. 6 illustrates the performance of α-unaware and β-aware A2C2s against the other two benchmarks. It can be observed that α-unaware and β-aware A2C2s outperform the baselines, and β-aware A2C2 has the best performance, which corroborates the regret analysis.

The second loss sequence is generated by shifting the mean values of the stochastic loss distribution at a changing point, which is similar to [12]. Specifically, before $T' = 4 \times 10^5$, the loss expectations are $[a_k]_{k=1}^K = [0.2, 0.2, 0.2, 0.2, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4]$, where arms $1-4$ are better than the others, and the loss for each arm k is sampled from the uniform distribution on $[a_k - 0.15, a_k + 0.15]$. Then, after T', some shifts happen to the loss expectations as $a_1 = a_4 = 0.8$ and $a_5 = a_6 = 0.2$, thus arms $2, 3, 5, 6$ now generate lower losses. Then, contiguous loss-1 sequences of length 50 are similarly spread into the whole loss sequences. The results are reported in Fig. 7. It can be observed that

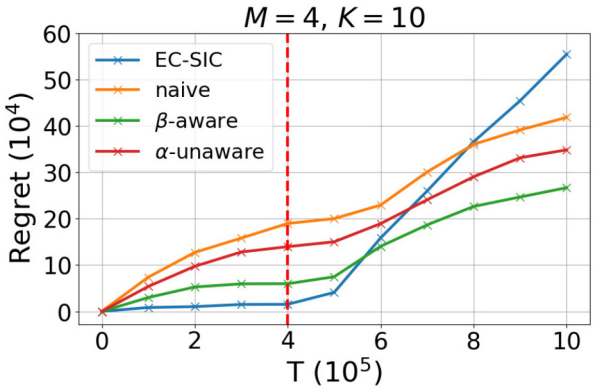


Fig. 7. Regret comparison where the loss generations change at $T' = 4 \times 10^5$ (labeled with the dotted red line).

while EC-SIC has the best performance at the beginning, its performance quickly degrades after the changing point T' since it has already converged to the originally good arms, some of which (arm 1 and 4) however become sub-optimal afterwards. Although the performance of the two A2C2 algorithms also fluctuate after the changing point, they quickly adjust to the new environment and re-gain a sublinear regret behavior.

X. Conclusion

This work made progress in the no-sensing adversarial MP-MAB problem by incorporating implicit communication. We have introduced the concept of *attackability* to categorize all possible adversaries from either a local view or a global view, and designed *Adversary-Adaptive Collision-Communication* (*A2C2*), a family of algorithms that can handle known or unknown attackabilities. The algorithmic contributions mainly came from several new tools in information theory and communication theory, such as Z-channel model, error-correction coding, (novel) error-detection repetition code, and uplink-downlink communication with randomized length. Theoretical analyses showed that the proposed algorithms have attackability-dependent regrets, which eliminate the exponential dependence on the number of players in the state-of-the-art no-sensing adversarial MP-MAB research, and revealed a new dimension of hardness, i.e., attackability, that compliments the hardness associated with the number of players.

Appendix A
Special Cases: $M = 1$ and Collision-Sensing

The regrets from communications, synchronization errors, and explorations after failed communications do not exist for $M = 1$ (single player); our result recovers $O(T^{\frac{1}{2}})$ in this special case. The reason that the asymptotic regret in Theorems 1 to 4 cannot recover this special case is that we focus on the scaling behavior, and thus the multiplicative factor $M - 1$ is simplified to M in the ensuing analysis.

We further note that if the collisions can be perceived by players, i.e., collision-sensing, there is no need for error-correction/detecting coding, attackability estimation and synchronization. In this case, the proposed algorithms recover the collision-sensing algorithm in [12].

Appendix B
The Centralized Algorithm

For the completeness of this work, we first provide a regret analysis of the algorithm described in Section IV-A, which is based on [12].

Theorem 5: The exploration regret of the centralized algorithm with $\eta = \sqrt{\log \binom{K}{M}/TKM}$ described in Section IV-A has a regret upper bound of:

$$\mathbb{E}[R_c(T)] \leq 2M\sqrt{K \log(K) T}. \tag{2}$$

Proof: First, we show that $\tilde{l}_A(t) = \sum_{k \in A} \tilde{l}_k(t)$ is an unbiased estimation of the true loss from meta-arm A at time t, i.e., $l_A(t) = \sum_{k \in A} l_k(t)$. Denoting $P(t) = \{P_A(t)\}_{A \in \mathcal{K}}$, we first show that for any arm $k \in [K]$, $\tilde{l}_k(t)$ is an unbiased estimation of $l_k(t)$ as

$$\begin{aligned}
\mathbb{E}\left[\tilde{l}_k(t)|P(t)\right] &\overset{(i)}{=} M \frac{l_k(t)}{\sum_{A:k\in A\in\mathcal{K}} P_A(t)} \mathbb{P}(k = A_1(t)) \\
&= M \frac{l_k(t)}{\sum_{A:k\in A\in\mathcal{K}} P_A(t)} \mathbb{P}(k \in A(t)) \\
&\quad \times \mathbb{P}(k = A_1(t)|k \in A(t)) \\
&\overset{(ii)}{=} M \frac{l_k(t)}{\sum_{A:k\in A\in\mathcal{K}} P_A(t)} \sum_{A:k\in A\in\mathcal{K}} P_A(t) \frac{1}{M} \\
&= l_k(t),
\end{aligned}$$

where equations (i) and (ii) are from the definitions of $\tilde{l}_k(t)$ and $P_A(t)$, respectively. From the law of total expectation, we can derive $\mathbb{E}[\tilde{l}_k(t)] = \mathbb{E}[\mathbb{E}[\tilde{l}_k(t)|P(t)]] = l_k(t)$. Finally, with $\tilde{l}_A(t) = \sum_{k \in A} \tilde{l}_k(t)$, by the linearity of expectation, $\tilde{l}_A(t)$ is an unbiased estimation of $l_A(t)$.

With the standard EXP3 regret guarantees, the centralized regret [12], [33] is bounded as:

$$\mathbb{E}[R_c(T)] \leq \eta \sum_{t=1}^{T} \sum_{A \in \mathcal{K}} \mathbb{E}\left[P_A(t) \mathbb{E}\left[\left(\tilde{l}_A(t)\right)^2 |P(t)\right]\right] + \frac{\log \binom{K}{M}}{\eta}.$$

The term $\mathbb{E}[(\tilde{l}_A(t))^2|P(t)]$ can be simplified as:

$$\begin{aligned}
\mathbb{E}\left[(\tilde{l}_A(t))^2|P(t)\right] &= \mathbb{E}\left[\left(\sum_{k\in A} \tilde{l}_k(t)\right)^2 |P(t)\right] \\
&= \sum_{j,k\in A} \mathbb{E}\left[\tilde{l}_k(t)\tilde{l}_j(t)|P(t)\right] \\
&\overset{(i)}{=} \sum_{k\in A} \mathbb{E}\left[\left(\tilde{l}_k(t)\right)^2 |P(t)\right] \\
&= \sum_{k\in A} \left(\frac{M l_k(t)}{\sum_{B:k\in B\in\mathcal{K}} P_B(t)}\right)^2 \mathbb{P}(k = A_1(t)) \\
&= \sum_{k\in A} \left(\frac{M l_k(t)}{\sum_{B:k\in B\in\mathcal{K}} P_B(t)}\right)^2 \mathbb{P}(k \in A(t)) \\
&\quad \cdot \mathbb{P}(k = A_1(t)|k \in A(t)) \\
&= M \sum_{k\in A} \frac{(l_k(t))^2}{\sum_{B:k\in B\in\mathcal{K}} P_B(t)},
\end{aligned}$$

where equation (i) is because $\tilde{l}_k(t) \neq 0$ holds for at most one arm. With this result, we get:

$$\begin{aligned}
\mathbb{E}[R_c(T)] &\leq \eta \sum_{t=1}^{T} \sum_{A\in\mathcal{K}} \mathbb{E}\left[P_A(t) M \sum_{k\in A} \frac{(l_k(t))^2}{\sum_{B:k\in B\in\mathcal{K}} P_B(t)} \right] + \frac{\log\binom{K}{M}}{\eta} \\
&= M\eta \sum_{t=1}^{T} \mathbb{E}\left[\sum_{A\in\mathcal{K}} P_A(t) \sum_{k\in A} \frac{(l_k(t))^2}{\sum_{B:k\in B\in\mathcal{K}} P(t)} \right] + \frac{\log\binom{K}{M}}{\eta} \\
&= M\eta \sum_{t=1}^{T} \mathbb{E}\left[\sum_{k\in[K]} \frac{(l_k(t))^2 \sum_{A:k\in A\in\mathcal{K}} P_A(t)}{\sum_{B:k\in B\in\mathcal{K}} P_B(t)} \right] + \frac{\log\binom{K}{M}}{\eta} \\
&= M\eta \sum_{t=1}^{T} \mathbb{E}\left[\sum_{k\in[K]} (l_k(t))^2 \right] + \frac{\log\binom{K}{M}}{\eta} \\
&\overset{(i)}{\leq} MKT\eta + \frac{\log\binom{K}{M}}{\eta} \\
&\overset{(ii)}{=} 2\sqrt{MKT \log\binom{K}{M}} \\
&\leq 2M\sqrt{KT\log(K)},
\end{aligned}$$

where inequality (i) is the result of $l_k(t) \leq 1$, and equation (ii) is from $\eta = \sqrt{\log\binom{K}{M}/MKT}$. ∎

The following result establishes a regret bound for the blocked version of the centralized algorithm, which is important for the ensuing analysis in the decentralized setting.

Theorem 6 [12], [35]: Let Π be a bandit algorithm with an expected regret upper bound of $R(T)$. Then the blocked version of Π with a block size τ has a regret upper bound of $\tau R(T/\tau) + \tau$.

In the multi-player case, the last term τ, which represents the additional regret when T is not divisible by τ, is converted into $M\tau$.

Appendix C
α-Aware

The overall regret of α-aware A2C2 can be decomposed as: $R_1(T) = R_1^{expl}(T) + R_1^{comm}(T)$, with $R_1^{expl}(T)$ and $R_1^{comm}(T)$ referring to the exploration and communication regret, respectively.

The local attackability $W(T)$ satisfies $W(T) \leq O(T^\alpha)$. Intuitively, the length of repetition code in α-aware A2C2 is $h(T, \alpha+\epsilon) = \Theta(T^{\alpha+\epsilon})$. It is implied that $h(T,\alpha+\epsilon) = \omega(W(T))$, which means asymptotically no successful attack can possibly happen and communication phases are guaranteed to be successful. With no communication phase in the centralized algorithm, the upper bound in Eqn. (2) can also serve as an upper bound for the exploration regret of α-aware A2C2 after successful communications with Theorem 6. Thus, the exploration regret $R_1^{expl}(T)$ can be bounded in the following lemma.

Lemma 8: The exploration regret of α-aware A2C2 satisfies:

$$\mathbb{E}\left[R_1^{expl}(T)\right] \leq O\left(M\sqrt{K\log(K)T\tau}\right).$$

Proof: Since $h(T,\alpha+\epsilon) = \Theta(T^{\alpha+\epsilon})$, there exist $m_1 > 0, m_2 > 0$, and T_0 such that, $\forall T > T_0, m_1 T^{\alpha+\epsilon} \leq h(T,\alpha+\epsilon) \leq m_2 T^{\alpha+\epsilon}$. Furthermore, since $W(T) = O(T^\alpha)$, $\exists n > 0$ and T_1 such that $\forall T > T_1, |W(T)| \leq nT^\alpha$. Thus, $\exists T^* = \max\{T_0, T_1, (n/m_1)^{\frac{1}{\epsilon}}\}$, $h(T,\alpha+\epsilon) \geq m_1 T^{\alpha+\epsilon} > nT^\alpha \geq W(T)$, $\forall T > T^*$, which means communications are guaranteed to be successful for $T > T^*$. Thus, all explorations are collision-free. Denote $T_e = T - T_c$, where T_e is the overall exploration time and T_c is the overall communication time. With Theorem 6 and Eqn. (2), we have that $\forall T > T^*$,

$$\begin{aligned}
\mathbb{E}\left[R_1^{expl}(T)\right] &\leq \tau R_c(T_e/\tau) + M\tau \\
&= 2M\sqrt{K\log(K)T_e\tau} + M\tau \\
&\leq 3M\sqrt{K\log(K)T\tau},
\end{aligned}$$

and thus $\mathbb{E}[R_1^{expl}(T)] \leq O(M\sqrt{K\log(K)T\tau})$. ∎

There are at most $\lceil \frac{T}{\tau} \rceil$ rounds communication, while each round contains $(M-1)\lceil \log_2(K) \rceil h(T,\alpha+\epsilon)$ time slots. Thus, the communication regret can be bounded as follows.

Lemma 9: The communication regret of α-aware A2C2 satisfies:

$$\begin{aligned}
\mathbb{E}\left[R_1^{comm}(T)\right] &\leq M(M-1)\left\lceil \log_2(K) \right\rceil \lceil T/\tau \rceil h(T,\alpha+\epsilon) \\
&\leq O\left(M^2 \log(K) T^{1+\alpha+\epsilon}/\tau\right).
\end{aligned}$$

With Lemmas 8 and 9, Theorem 1 can be proven via solving the optimization problem of

$$\min_{\tau\in\mathbb{N}} \max\left\{ O\left(M\sqrt{K\log(K)T\tau}\right), O\left(M^2\log(K)\frac{T^{1+\alpha+\epsilon}}{\tau}\right) \right\}$$

which leads to $\tau = \lceil M^{\frac{2}{3}} K^{-\frac{1}{3}} \log(K)^{\frac{1}{3}} T^{\frac{1+2\alpha+2\epsilon}{3}} \rceil$.

Appendix D
β-Aware

The same error detection code in Section VI-A is also used in the β-aware A2C2 algorithm. However, it is not used for the update of estimations, but rather only to maintain an unbiased loss estimation. The β-aware A2C2 algorithm is presented in Algorithms 5 (leader) and 6 (follower).

The overall regret of β-aware A2C2 can be decomposed as $R_2(T) = R_2^{expl}(T) + R_2^{err}(T) + R_2^{comm}(T)$, where $R_2^{expl}(T)$ and $R_2^{err}(T)$ refer to the exploration regret after successful and failed communications, respectively, and $R_2^{comm}(T)$ characterizes the communication regret. With Theorem 6 and Eqn. (2), $R_2^{expl}(T)$ can be upper bounded, as stated in the following lemma.

Lemma 10: The exploration regret of β-aware A2C2 after successful communication satisfies:

$$\mathbb{E}\left[R_2^{expl}(T)\right] \leq O\left(M\sqrt{K\log(K)T\tau}\right).$$

Algorithm 5 β-Aware A2C2: Leader

Input: M, K, T

Initialize: $\tau \leftarrow \lceil K^{\frac{1}{3}}\log(K)^{-\frac{1}{3}}T^{\max\{\beta,\frac{1}{3}\}}\rceil$; $\eta \leftarrow \sqrt{\log\binom{K}{M}\tau/MKT}$; $\nu \leftarrow \max\left\{\frac{3\beta-1}{2}, 0\right\}$; $F \leftarrow 0$

for $p = 1, 2, \ldots$ **do**

  $\forall A \in \mathcal{K}, \tilde{L}_A(p) \leftarrow \sum_{\nu=1}^{p-1}\sum_{k\in A}\tilde{l}_k(\nu)$

  $\forall A \in \mathcal{K}, P_A(p) \leftarrow \frac{e^{-\eta\tilde{L}_A(p)}}{\sum_{J\in\mathcal{K}}e^{-\eta\tilde{L}_J(p)}}$ ▷ Loss estimator

  Choose $A(p) = \{A_1(p), \ldots, A_M(p)\}$ with $P_A(p)$

  Randomly permute $A(p)$ into $\tilde{A}(p)$

▷ **Communication Phase:**

  $\forall m \in [M]$, $\text{msg}_m \leftarrow \text{eEncoder}(\tilde{A}_m(p), k(T,\nu))$

  $\forall m \in [M]$, $\text{Send}(m, \text{msg}_m)$ ▷ Send Assignment

  $\text{msg}_F \leftarrow \text{Receive}(k(T,\nu))$

  $F \leftarrow \text{rDecoder}(\text{msg}_F, k(T,\nu))$

▷ **Exploration Phase:**

  Stay on arm $\tilde{A}_1(p)$ for τ time steps ▷ Exploration

  if $F = 0$ **then**

    Record cumulative loss $\hat{l}_{\tilde{A}_1(p)}(p)$

    $\forall k \in [K]$, set $\tilde{l}_k(p) \leftarrow \frac{M}{\tau}\frac{\hat{l}_{\tilde{A}_1(p)}(p)\mathbb{1}\{\tilde{A}_1(p)=k\}}{\sum_{A:k\in A\in\mathcal{K}}P_A(p)}$

  else $\forall k \in [K], \tilde{l}_k(p) \leftarrow 0$

  end if

end for

Algorithm 6 β-Aware A2C2: Follower

Input: M, K, T, index m

Initialize: $\tau \leftarrow \lceil K^{\frac{1}{3}}\log(K)^{-\frac{1}{3}}T^{\max\{\beta,\frac{1}{3}\}}\rceil$; $\eta \leftarrow \sqrt{\log\binom{K}{M}\tau/MKT}$; $\nu \leftarrow \max\left\{\frac{3\beta-1}{2}, 0\right\}$; $F \leftarrow 0$; $\mathcal{S} \leftarrow \emptyset$

for $p = 1, 2, \ldots$ **do**

▷ **Communication Phase:**

  $\text{msg}_m \leftarrow \text{Receive}(k(T,\nu))$

  $\mathcal{S} \leftarrow \text{eDecoder}(\text{msg}_m, k(T,\nu))$ ▷ Receive Assignment

  Randomly choose in $\mathcal{S}$ for $\tilde{A}_m(p)$

  $F \leftarrow \mathbb{1}\{|\mathcal{S}| > 1\}$

  $\text{msg}_F \leftarrow \text{rEncoder}(k(T,\nu))$

  $\text{Send}(1, \text{msg}_F)$ ▷ Feedbak Communication Error

▷ **Exploration Phase:**

  Stay on arm $\tilde{A}_m(p)$ for τ time steps ▷ Exploration

end for

Proof: Denote $T_s = T - T_c - T_f$, where T_f is the overall exploration time after failed communications and T_c is the overall communication time. Since $\forall t \in T_s$, all explorations are collision-free, and the same combination of Theorem 6 and Eqn. (2) leads to the following upper bound:

$$\begin{aligned}\mathbb{E}\Big[R_2^{expl}(T)\Big] &\le \tau R_c(T_s/\tau) + M\tau \\ &= 2M\sqrt{K\log(K)T_s\tau} + M\tau \\ &\le 3M\sqrt{K\log(K)T\tau},\end{aligned}$$

which means $\mathbb{E}[R_2^{expl}(T)] \le O(M\sqrt{K\log(K)T\tau})$. ∎

The overall number of loss ones on one arm, i.e., the global attackability, is $V(T) \le O(T^\beta)$. To succeed in attacking one communication phase, at least $k(T,\nu) = \Theta(T^\nu)$ of loss ones are required. Thus, the overall number of successful attacks is no more than $\frac{MV(T)}{k(T,\nu)} \le O(MT^{\beta-\nu})$ times . The second term $R_2^{err}(T)$ in the overall regret can be bounded as follows.

Lemma 11: The exploration regret caused by failed communications of β-aware A2C2 satisfies:

$$\mathbb{E}\big[R_2^{err}(T)\big] \le O\Big(M^2T^{\beta-\nu}\tau\Big).$$

Proof: Since $k(T,\nu) = \Theta(T^\nu)$, there exist $m_1 > 0, m_2 > 0$ and T_0 such that, $\forall T > T_0, m_1T^\nu \le k(T,\nu) \le m_2T^\nu$. Since $V(T) = O(T^\beta)$, $\exists n > 0$ and T_1 such that, $\forall T > T_1, |V(T)| \le nT^\beta$. Thus, $\exists T^* = \max\{T_0, T_1\}$, $\frac{V(T)}{k(T,\nu)} \le \frac{n}{m_1}T^{\beta-\nu}$, $\forall T > T^*$. With a loss of at most $M\tau$ caused by each successful attack, we have that $\forall T > T^*$,

$$\mathbb{E}\big[R_2^{err}(T)\big] = M\frac{V(T)}{k(T,\nu)}\cdot M\tau \le M^2\frac{n}{m_1}T^{\beta-\nu}\tau.$$

Thus, $\mathbb{E}[R_2^{err}(T)] \le O(M^2T^{\beta-\nu}\tau)$. ∎

With at most $\lceil T/\tau\rceil$ rounds of communications consisting of $M-1$ times of arm assignment and one time error report, the communication regret can be bounded as follows.

Lemma 12: The communication regret of β-aware A2C2 satisfies:

$$\begin{aligned}\mathbb{E}\big[R_2^{comm}(T)\big] &\le M(M-1)K\lceil T/\tau\rceil k(T,\nu) + M\lceil T/\tau\rceil k(T,\nu) \\ &\le O\Big(M^2KT^{1+\nu}/\tau\Big).\end{aligned}$$

With Lemmas 10 to 12, Theorem 2 can be proven via solving the following optimization problem:

$$\min_{\tau\in\mathbb{N},\nu\ge0}\max\Big\{O\Big(M\sqrt{K\log(K)T\tau}\Big), O\Big(M^2T^{\beta-\nu}\tau\Big), O\Big(M^2KT^{1+\nu}/\tau\Big)\Big\}$$

which leads to $\tau = \lceil K^{\frac{1}{3}}\log(K)^{-\frac{1}{3}}T^{\max\{\beta,\frac{1}{3}\}}\rceil$ and $\nu = \max\{\frac{3\beta-1}{2}, 0\}$.

APPENDIX E
α-UNAWARE

Since multiple estimations are required before successful communications can be guaranteed, $R_3^{expl}(T)$ consists of not only the exploration regret $R_3^{expl}(T|\zeta = 1)$ after successful communication phases, but also explorations regret $R_3^{expl}(T|\zeta = 0)$ after failed communications, which is different from $R_1^{expl}(T)$ in α-aware A2C2. The estimation $\alpha' = j\epsilon$ is denoted as α_j for an integer j for simplicity. The overall exploration time steps when α_j is used as the α-estimation is denoted as $T_{e,j}$, and the corresponding τ and ξ with α_j are denoted as τ_j and ξ_j, respectively.

With the update, $\exists\nu > 0$, $\delta \in [0,\epsilon)$, $\alpha \le \alpha + \delta \le \alpha_\nu = \nu\epsilon \le \alpha + \epsilon$. Since $h(T,\alpha_\nu) = \Theta(T^{\nu\epsilon})$, $\exists\kappa_1 > 0, \kappa_2 > 0$ and T_0 such that, $\forall T > T_0$, $\kappa_1T^{\nu\epsilon} \le h(T,\alpha_\nu) \le \kappa_2T^{\nu\epsilon}$. Since $W(T) = O(T^\alpha)$, $\exists n > 0$ and T_1 such that, $\forall T > T_1, |W(T)| \le nT^\alpha$. It implies that $W(T) \le nT^\alpha < \kappa_1T^{\alpha+\delta} \le h(T,\alpha_\nu)$, $\forall T > T^* = \max\{T_0, T_1, (n/\kappa_1)^{\frac{1}{\delta}}\}$, which means $\forall T > T^*$, the repetition code can overpower the local attacks, and thus successful communications are guaranteed with α_ν. For a given adversary sequence, $\forall T > T^*$, the update is assumed to stop at $\alpha_w = w\epsilon$ with $w \le \nu$.

Proof of Lemma 1: Denote $T_{s,j}$ and $T_s = \sum_{j=0}^{w}T_{s,j} \le \sum_{j=0}^{w}T_{e,j} \le T$ as the length of exploration with successful

preceding communications under estimation α_j and in total, respectively. The exploration under each estimation α_j with successful preceding communications can be viewed alone as a bandit game with horizon $T_{s,j}$ and block τ_j. Thus, by applying Theorem 6 to Eqn. (2), $\forall T > T^*$, the first term in the exploration regret can be bounded as

$$\begin{aligned}\mathbb{E}\Big[R_3^{expl}(T|\zeta=1)\Big] &= \sum_{j=0}^{w}\mathbb{E}\Big[R_3^{expl}\big(T|\zeta=1,\alpha'=\alpha_j\big)\Big]\\ &\le \sum_{j=0}^{w}\Big(2M\sqrt{K\log(K)T_{s,j}\tau_j}+M\tau_j\Big)\\ &\overset{(i)}{\le} 3(w+1)M\sqrt{K\log(K)T_s\tau_w}\\ &\le 3\sqrt{2}(v+1)M^{\frac{4}{3}}K^{\frac{1}{3}}\log(K)^{\frac{1}{3}}T^{\frac{5+\alpha_v}{6}},\end{aligned}$$

where inequality (i) is from that $T_{s,j} \le T_s$ and τ_j is monotonically increasing with j. Thus, $\mathbb{E}[R_3^{expl}(T)|\zeta = 1] \le O(M^{\frac{4}{3}}K^{\frac{1}{3}}\log(K)^{\frac{1}{3}}T^{\frac{5+\alpha+\epsilon}{6}})$.

The second term is caused by the potential collisions due to failed communications with an underestimated attackability, and $\forall T > T^*$, for the adversary given above, it takes w trials to eliminate communication errors, and the trial with estimation α_j leads to a regret of $M\tau_j$. Thus, the second term can be bounded as:

$$\begin{aligned}\mathbb{E}\Big[R_3^{expl}(T|\zeta=0)\Big] &= \sum_{j=0}^{w-1}\mathbb{E}\Big[R_3^{expl}\big(T|\zeta=0,\alpha'=\alpha_j\big)\Big]\\ &\le \sum_{j=0}^{w-1}M\tau_j\\ &= \sum_{j=0}^{w-1}M\left\lceil M^{\frac{2}{3}}K^{-\frac{1}{3}}\log(K)^{-\frac{1}{3}}T^{\frac{2+\alpha_j}{3}}\right\rceil\\ &\le 2M^{\frac{5}{3}}K^{-\frac{1}{3}}\log(K)^{-\frac{1}{3}}T^{\frac{2+\alpha_v}{3}}.\end{aligned}$$

Since $\alpha_v \le 1$, the first term dominates the second term as $\frac{2+\alpha_v}{3} \le \frac{5+\alpha_v}{6}$. Thus, the overall exploration regret can be bounded as $\mathbb{E}[R_3^{expl}(T)] \le O(M^{\frac{4}{3}}K^{\frac{1}{3}}\log(K)^{\frac{1}{3}}T^{\frac{5+\alpha+\epsilon}{6}})$.

Proof of Lemma 2: The communication regret consists of that for arm assignment and synchronization under different estimations. The arm assignment in each round consists of $(M-1)Kh(T,\alpha_j)$ time slots while the synchronization has an average length of $Mh(T,\alpha_j)\frac{\lceil T^{\xi_j}\rceil}{2}$. For the given adversary in the proof above, $\forall T > T^*$, the overall communication regret can be bounded as:

$$\begin{aligned}\mathbb{E}\big[R_3^{comm}(T)\big] \le \sum_{j=0}^{w}&\left\lceil\frac{T_{e,j}}{\tau_j}\right\rceil\\ &\times\left(M(M-1)Kh(T,\alpha_j)+M^2h(T,\alpha_j)\frac{\lceil T^{\xi_j}\rceil}{2}\right)\end{aligned}$$

Considering $h(T,\alpha_j) = \Theta(T^{\alpha_j})$, $\forall j = 0,1,\ldots,v$, $\exists f_{j,1} > 0, f_{j,2} > 0$ and $T_{j,0}$ such that, $\forall T > T_{j,0}$, $f_{j,1}T^{\alpha_j} \le h(T,\alpha_j) \le f_{j,2}T^{\alpha_j}$. Thus, $\forall j = 0,1,\ldots,v$, $\exists f_1 \le \min_j\{f_{j,1}\}, \exists f_2 \ge \max_j\{f_{j,2}\}$ and $T_x > \max_j\{T_{j,0}\}$ such that $\forall T > T_x$, $f_1T^{\alpha_j} \le h(T,\alpha_j) \le f_2T^{\alpha_j}$. Based on this, $\forall T > \max\{T^*, T_x\}$, the communication regret can be bounded as:

$$\begin{aligned}\mathbb{E}\big[R_3^{comm}(T)\big] &\le 2\sum_{j=0}^{w}\frac{T_{e,j}}{\tau_j}\Big(M^2Kf_2T^{\alpha_j}+M^2f_2T^{\alpha_j}T^{\xi_j}\Big)\\ &= 2\sum_{j=0}^{w}T_{e,j}M^{\frac{4}{3}}K^{\frac{4}{3}}\log(K)^{\frac{1}{3}}f_2T^{\frac{\alpha_j-1}{3}}\\ &\quad+2\sum_{j=0}^{w}T_{e,j}M^{\frac{4}{3}}K^{\frac{1}{3}}\log(K)^{\frac{1}{3}}f_2T^{\frac{\alpha_j-1}{6}}\\ &\le 2M^{\frac{4}{3}}K^{\frac{4}{3}}\log(K)^{\frac{1}{3}}f_2T^{\frac{2+\alpha_v}{3}}\\ &\quad+2M^{\frac{4}{3}}K^{\frac{1}{3}}\log(K)^{\frac{1}{3}}f_2T^{\frac{5+\alpha_v}{6}}.\end{aligned}$$

With $\alpha_v < 1$, we have $\mathbb{E}[R_3^{comm}(T)] \le O(M^{\frac{4}{3}}K^{\frac{1}{3}}\log(K)^{\frac{1}{3}}T^{\frac{5+\alpha+\epsilon}{6}})$.

Proof of Lemma 3: Before completing the estimation of α', each communication for synchronization has a failure probability of $\frac{1}{\lceil T^{\xi}\rceil}$, which in the worst case has a linear regret MT. With a union bound for all communication phases with estimation less than α_w, we have:

$$\begin{aligned}\mathbb{E}\big[R_3^{sync}(T)\big] &\le \sum_{j=0}^{w-1}\left\lceil\frac{T_{e,j}}{\tau_j}\right\rceil\frac{1}{\lceil T^{\xi_j}\rceil}MT\\ &\le 2\sum_{j=0}^{w-1}\frac{T_{e,j}}{\tau_j}\frac{1}{T^{\xi_j}}MT\\ &\le 2M^{\frac{1}{3}}K^{\frac{1}{3}}\log(K)^{\frac{1}{3}}T^{\frac{5+\alpha+\epsilon}{6}}.\end{aligned}$$

Thus, Lemma 3 can be obtained as $\mathbb{E}[R_3^{sync}(T)] \le O(M^{\frac{1}{3}}K^{\frac{1}{3}}\log(K)^{\frac{1}{3}}T^{\frac{5+\alpha+\epsilon}{6}})$.

APPENDIX F
β-UNAWARE

The β-unaware algorithm for the leader and followers are presented in Algorithms 7 and 8, respectively. The following proofs focus on $\beta \ge \frac{1}{4}$, and the case of $\beta \le \frac{1}{4}$ can be obtained as a special case where the estimation β' is kept as $\frac{1}{4}$. Under estimation $\beta' = \frac{1}{4} + j\epsilon$, denoted as β_j for simplicity, similar notations of $T_{e,j}$, τ_j, ξ_j and ν_j are applied referring to the overall time that the current estimation holds and the corresponding parameters.

With the update, $\exists \nu > 0, \delta \in (0,\epsilon], \beta+\epsilon \ge \beta' = \beta_v > \beta+\delta \ge \beta$. Since $k(T,\beta_v) = \Theta(T^{\frac{1}{4}+\nu\epsilon})$, $\exists \kappa_1 > 0$, $\kappa_2 > 0$, and T_0 such that, $\forall T > T_0$, $\kappa_1T^{\beta+\delta} \le \kappa_1T^{\frac{1}{4}+\nu\epsilon} \le k(T,\beta_v) \le \kappa_2T^{\frac{1}{4}+\nu\epsilon} \le \kappa_2T^{\beta+\epsilon}$. With $V(T) = O(T^{\beta})$, $\exists n > 0$, T_1 so that $\forall T > T_1$, $|V(T)| \le nT^{\beta}$. We then have $V(T) \le nT^{\beta} < \kappa_1T^{\beta+\delta} \le k(T,\beta_v)$, $\forall T > T^* = \max\{T_0, T_1, (n/\kappa_1)^{\frac{1}{\delta}}\}$, which means updating stops with β_v. For a given adversary, $\forall T > T^*$, the estimation is assumed to be completed as $\beta' = \beta_w = \frac{1}{4} + w\epsilon$, where $w \le v$.

Proof of Lemma 4: Denote $T_s = \sum_{j=0}^{w}T_{s,j} \le T$ as the overall length of the exploration phases after successful communications. The term $R_4^{expl}(T)$ now consists of only explorations after successful communications, which is again

Algorithm 7 β-Unaware A2C2: Leader

Input: M, K, T

Initialize: $\beta' \leftarrow \frac{1}{4}$; communication flag: $F_1, F_2 \leftarrow 0$; round counter: $R \leftarrow 0$; collision counter $C \leftarrow 0$
for $p = 1, 2, \ldots$ **do**
  $\tau \leftarrow \lceil K^{-\frac{1}{3}} \log(K)^{-\frac{1}{3}} T^{\frac{1+2\beta'}{3}} \rceil$
  $\eta \leftarrow \sqrt{\log\binom{K}{M}\tau/MKT}$; $\xi \leftarrow \frac{1+2\beta'}{3}$; $\nu' \leftarrow \frac{4\beta'-1}{3}$
  $F_1 \leftarrow 0$; $R \leftarrow R+1$;
  $\forall A \in \mathcal{K}, \tilde{L}_A(p) \leftarrow \sum_{\nu=1}^{p-1} \sum_{k \in A} \tilde{l}_k(\nu)$
  $\forall A \in \mathcal{K}, P_A(p) \leftarrow \frac{e^{-\eta \tilde{L}_A(p)}}{\sum_{J \in \mathcal{K}} e^{-\eta \tilde{L}_J(p)}}$
  Choose $A(p) = \{A_1(p), \ldots, A_M(p)\}$ with $P_A(p)$
  Randomly permute $A(p)$ into $\tilde{A}(p)$
  ▷ **Communication Phase:**
  $\forall m \in [M]$, $\text{msg}_m \leftarrow \text{eEncoder}(\tilde{A}_m(p), k(T, \nu'))$
  $\forall m \in [M]$, $\text{Send}(m, \text{msg}_m)$ ▷ Send Assignment
  $\text{msg}_{F_1} \leftarrow \text{Receive}(k(T, \nu'))$
  $F_1 \leftarrow \text{rDecoder}(\text{msg}_{F_1}, k(T, \nu'))$; ▷ Uplink
  $C \leftarrow C + F_1 k(T, \nu')$ ▷ Count Attack
  if $R \geq \left\lceil \frac{T^{\beta'}}{k(T,\nu')} \right\rceil$ **then**
    $F_2 \leftarrow \mathbb{1}\{C \geq \lceil T^{\beta'} \rceil\}$ ▷ Update Point
    for $q = 1, 2, \ldots, N(\xi)$ **do**
      $\text{msg}_{F_2} \leftarrow \text{rEncoder}\big(F_2, k(T, \beta')\big)$
      $\forall m \in [M]$, $\text{Send}(m, \text{msg}_{F_2})$ ▷ Downlink
      $\text{msg}_{F_2} \leftarrow \text{Receive}(k(T, \beta'))$
      $F_2 \leftarrow \text{rDecoder}(\text{msg}_{F_2}, k(T, \beta'))$ ▷ Uplink
    end for
    $R \leftarrow 0$; $\beta' \leftarrow \beta' + F_2 \epsilon$ ▷ Update
  end if
  ▷ **Exploration Phase:**
  Stay on arm $\tilde{A}_1(p)$ for τ time steps
  if $F_1 = 0$ **then**
    Record cumulative loss $\hat{l}_{\tilde{A}_1(p)}(p)$
    $\forall k \in [K]$, set $\tilde{l}_k(p) \leftarrow \frac{M}{\tau} \frac{\hat{l}_{\tilde{A}_1(p)}(p) \mathbb{1}\{\tilde{A}_1(p)=j\}}{\sum_{A: k \in A \in \mathcal{K}} P_A(p)}$
  else $\forall k \in [K]$, $\tilde{l}_k(p) \leftarrow 0$
  end if
end for

Algorithm 8 β-Unaware A2C2: Follower

Input: M, K, T, index m

Initialize: $\beta' \leftarrow \frac{1}{4}$; communication flag: $F_1, F_2 \leftarrow 0$; round counter: $R \leftarrow 0$; collision counter $C \leftarrow 0$
for $p = 1, 2, \ldots$ **do**
  $\tau \leftarrow \lceil K^{-\frac{1}{3}} \log(K)^{-\frac{1}{3}} T^{\frac{1+2\beta'}{3}} \rceil$
  $\eta \leftarrow \sqrt{\log\binom{K}{M}\tau/MKT}$; $\xi \leftarrow \frac{2-2\beta'}{3}$; $\nu' \leftarrow \frac{4\beta'-1}{3}$
  $F_1, F_2 \leftarrow 0$; $R \leftarrow R+1$
  ▷ **Communication Phase:**
  $\text{msg}_m \leftarrow \text{Receive}(k(T, \nu'))$
  $\mathcal{S} \leftarrow \text{eDecoder}(\text{msg}_m, k(T, \nu'))$ ▷ Receive Assignment
  Randomly choose in $\mathcal{S}$ for $\tilde{A}_m(p)$
  $F_1 \leftarrow \mathbb{1}\{|\mathcal{S}| > 1\}$ ▷ Communication Error
  $C \leftarrow C + (|S| - 1)k(T, \nu')$ ▷ Count Attack
  $\text{msg}_{F_1} \leftarrow \text{rEncoder}(F_1, k(T, \nu'))$
  $\text{Send}(1, \text{msg}_{F_1})$ ▷ Uplink
  if $R \geq \left\lceil \frac{T^{\beta'}}{k(T,\nu')} \right\rceil$ **then**
    $F_2 \leftarrow \mathbb{1}\{C \geq \lceil T^{\beta'} \rceil\}$
    for $q = 1, 2, \ldots, N(\xi)$ **do** ▷ Update Point
      $\text{msg}_{F_2} \leftarrow \text{Receive}(k(T, \beta'))$ ▷ Downlink
      $F_2 \leftarrow \max\{F_2, \text{rDecoder}(\text{msg}_{F_2}, k(T, \beta'))\}$
      $\text{msg}_{F_2} \leftarrow \text{rEncoder}(F_2, k(T, \beta'))$
      $\text{Send}(1, \text{msg}_{F_2})$ ▷ Uplink
    end for
    $R \leftarrow 0$; $\beta' \leftarrow \beta' + F_2 \epsilon$; $F_2 \leftarrow 0$ ▷ Update
  end if
  ▷ **Exploration Phase:**
  Stay on arm $\tilde{A}_m(p)$ for τ time steps
end for

bounded by Theorem 6 and Eqn. (2) as:

$$\begin{aligned}
\mathbb{E}\Big[R_4^{expl}(T)\Big] &= \sum_{j=0}^{w} \mathbb{E}\Big[R_4^{expl}\big(T|\beta' = \beta_j\big)\Big] \\
&\leq \sum_{j=0}^{w} \Big(2M\sqrt{K\log(K)T_{s,j}\tau_j} + M\tau_j\Big) \\
&\leq 3(v+1)M\sqrt{K\log(K)T_s\tau_v} \\
&\leq 3(v+1)MK^{\frac{1}{3}}\log(K)^{\frac{1}{3}}T^{\frac{2+\beta_v}{3}}.
\end{aligned}$$

Thus, we have $\mathbb{E}[R_4^{expl}(T)] \leq O(MK^{\frac{1}{3}}\log(K)^{\frac{1}{3}}T^{\frac{2+\beta+\epsilon}{3}})$.

Proof of Lemma 7: With the estimated β_j, the adversary at most attacks $D_j = 2\lceil \frac{T^{\beta_j}}{k(T,\nu_j)} \rceil$ rounds on each arm before updating, and each successful attack leads to a loss of at most $M\tau_j$. With $k(T, \nu_j) = \Theta(T^{\nu_j}) = \Theta(T^{\frac{4\beta_j-1}{3}})$, $\forall j = 0, 1, \ldots, w$, $\exists h_{j,1} > 0, \exists h_{j,2} > 0, \exists T_{m,y}, \forall T > T_{j,0}$, $h_{j,1}T^{\frac{4\beta_j-1}{3}} \leq k(T, \nu_j) \leq h_{j,2}T^{\frac{4\beta_j-1}{3}}$. Thus, $\forall j = 0, 1, \ldots, v$, $\exists h_1 \leq \min_j\{h_{j,1}\}$, $\exists h_2 \geq \max_j\{h_{j,2}\}$, $\exists T_y \geq \max_j\{T_{j,0}\}$, $\forall T > T_y$, $h_1T^{\frac{4\beta_j-1}{3}} \leq k(T, \nu_j) \leq h_2T^{\frac{4\beta_j-1}{3}}$. It then follows that

$$\begin{aligned}
\mathbb{E}\big[R_4^{err}(T)\big] &\leq \sum_{j=0}^{w} MD_jM\tau_j \\
&\leq \sum_{j=0}^{w} \frac{4T^{\beta_j}}{k\big(T, \nu_j\big)}M^2\tau_j \\
&\leq \sum_{j=0}^{w} \frac{4}{h_1}M^2K^{\frac{1}{3}}\log(K)^{-\frac{1}{3}}T^{\frac{2+\beta_j}{3}} \\
&\leq (v+1)\frac{4}{h_1}M^2K^{\frac{1}{3}}\log(K)^{-\frac{1}{3}}T^{\frac{3+\beta_v}{3}},
\end{aligned}$$

$\forall T > \max\{T^*, T_y\}$, which means $\mathbb{E}[R_4^{err}(T)] \leq O(M^2K^{\frac{1}{3}}\log(K)^{-\frac{1}{3}}T^{\frac{2+\beta+\epsilon}{3}})$.

Proof of Lemma 5: Communications consist of three parts: those for arm assignment, communication error report, and synchronization. With the estimated β_j, the arm assignment and communication error report happen at most $\lceil T_{e,j}/\tau_j \rceil$ rounds while each round lasts $(M-1)Kk(T, \nu_j) + k(T, \nu_j)$ time slots. On the other hand, under the same β_j, the synchronization happens at most $\lceil \frac{T_{e,j}}{\tau_j} \rceil \frac{k(T,\nu_j)}{T^{\beta_j}}$ rounds while each round lasts $Mk(T, \nu_j)\frac{\lceil T^{\xi_j} \rceil}{2}$ time slots on average. Thus, it can be get that

$$\begin{aligned}
\mathbb{E}\big[R_4^{comm}(T)\big] = &\sum_{j=0}^{w} \left\lceil \frac{T_{e,j}}{\tau_j} \right\rceil (M-1)\big[MKk\big(T, \nu_j\big) + k\big(T, \nu_j\big)\big] \\
&+ \sum_{j=0}^{w} \left\lceil \frac{T_{e,j}}{\tau_j} \right\rceil \frac{M^2k^2\big(T, \nu_j\big)}{T^{\beta_j}} \frac{\lceil T^{\xi_j} \rceil}{2}
\end{aligned}$$

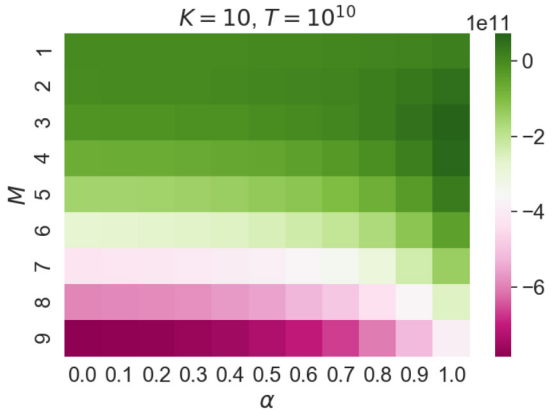


Fig. 8. Regret differences between α-unaware A2C2 and [5].

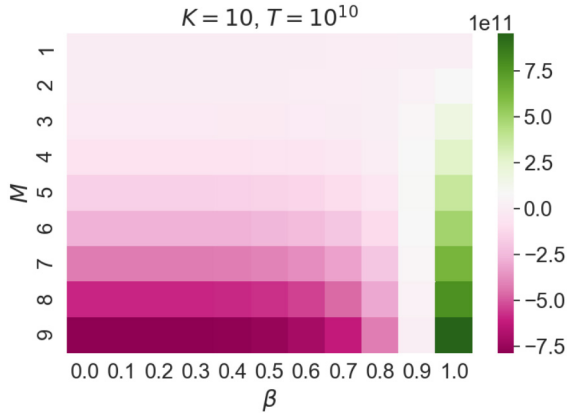


Fig. 9. Regret differences between β-unaware A2C2 and [5].

$$\leq 2\sum_{j=0}^{w} h_2 M^2 K^{\frac{4}{3}} \log(K)^{\frac{1}{3}} T_{e,j} T^{\frac{2\beta_j-2}{3}} + 2\sum_{j=0}^{w} (h_2)^2 M^2 K^{\frac{1}{3}} \log(K)^{\frac{1}{3}} T_{e,j} T^{\frac{\beta_j-1}{3}} \leq 2h_2 M^2 K^{\frac{4}{3}} \log(K)^{\frac{1}{3}} T^{\frac{1+2\beta_v}{3}} + 2(h_2)^2 M^2 K^{\frac{1}{3}} \log(K)^{\frac{1}{3}} T^{\frac{2+\beta_v}{3}}.$$

$\forall T > T^*$. Since $\beta_v \leq 1$, $\frac{2+\beta_v}{3} \geq \frac{1+2\beta_v}{3}$, $\mathbb{E}[R_4^{comm}(T)] \leq O(M^2 K^{\frac{1}{3}} \log(K)^{\frac{1}{3}} T^{\frac{2+\beta+\epsilon}{3}})$.

Proof of Lemma 6: With the estimated β_j, the synchronization happens every $\lceil \frac{T^{\beta_j}}{k(T,v_j)} \rceil$ iterations of exploration and communication, and each synchronization has a failure probability $\frac{1}{\lceil T^{\xi_j} \rceil}$, which leads to a worst-case linear regret of MT in the following time slots:

$$\mathbb{E}\left[R_4^{sync}(T)\right] \leq \sum_{j=0}^{w} \left\lceil \frac{T_{e,j}}{\tau_j} \right\rceil \frac{k(T,v_j)}{T^{\beta_j}} \frac{1}{\lceil T^{\xi_j} \rceil} MT \leq 2h_2 \sum_{j=0}^{w} \frac{T_{e,j}}{\tau_j} MT^{1+v_j-\beta_j-\xi_j} \leq 2h_2 \sum_{j=0}^{w} MK^{\frac{1}{3}} \log(K)^{\frac{1}{3}} T_{e,j} T^{\frac{\beta_j-1}{3}} \leq 2h_2 MK^{\frac{1}{3}} \log(K)^{\frac{1}{3}} T^{\frac{2+\beta_w}{3}} \leq 2h_2 MK^{\frac{1}{3}} \log(K)^{\frac{1}{3}} T^{\frac{2+\beta+\epsilon}{3}}.$$

Thus, we have $\mathbb{E}[R_4^{sync}(T)] \leq O(MK^{\frac{1}{3}} \log(K)^{\frac{1}{3}} T^{\frac{2+\beta+\epsilon}{3}})$.

APPENDIX G
ADDITIONAL NUMERICAL ILLUSTRATIONS

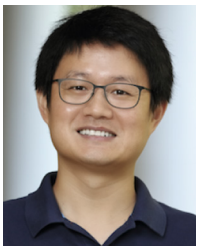
In addition to Figs. 1 and 2 in Section I, a more detailed numerical illustration is given here to compare A2C2 with the algorithm in [5]. In Fig. 8, the regret differences between α-unaware and [5] with different number of participating players (i.e., M) and attackabilities (i.e., α) are illustrated. We see that the advantage of α-unaware A2C2 is more pronounced in the region of large M and small α. A similar observation can be made from Fig. 9, which shows the regret differences between β-unaware A2C2 and [5].

REFERENCES

[1] K. Liu and Q. Zhao, "Distributed learning in multi-armed bandit with multiple players," *IEEE Trans. Signal Process.*, vol. 58, no. 11, pp. 5667–5681, Nov. 2010.
[2] L. Besson and E. Kaufmann, "Multi-player bandits revisited," in *Proc. Algorithmic Learn. Theory*, 2018, pp. 56–92.
[3] E. Boursier and V. Perchet, "SIC-MMAB: Synchronisation involves communication in multiplayer multi-armed bandits," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 12071–12080.
[4] G. Lugosi and A. Mehrabian, "Multiplayer bandits without observing collision information," 2018. [Online]. Available: arXiv:1808.08416.
[5] S. Bubeck, Y. Li, Y. Peres, and M. Sellke, "Non-stochastic multi-player multi-armed bandits: Optimal rate with collision information, sublinear without," in *Proc. Conf. Learn. Theory*, 2020, pp. 961–987.
[6] A. Anandkumar, N. Michael, A. K. Tang, and A. Swami, "Distributed algorithms for learning and cognitive medium access with logarithmic regret," *IEEE J. Select. Areas Commun.*, vol. 29, no. 4, pp. 731–745, Apr. 2011.
[7] O. Avner and S. Mannor, "Concurrent bandits and cognitive radio networks," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Disc. Databases*, 2014, pp. 66–81.
[8] C. Gan, R. Zhou, J. Yang, and C. Shen, "Cost-aware cascading bandits," *IEEE Trans. Signal Process.*, vol. 68, no. 6, pp. 3692–3706, Jun. 2020.
[9] C. Shi, W. Xiong, C. Shen, and J. Yang, "Decentralized multi-player multi-armed bandits with no collision information," in *Proc. Int. Conf. Artif. Intell. Stat.*, 2020, pp. 1519–1528.
[10] X. Xu, M. Tao, and C. Shen, "Collaborative multi-agent multi-armed bandit learning for small-cell caching," *IEEE Trans. Wireless Commun.*, vol. 19, no. 4, pp. 2570–2585, Apr. 2020.
[11] C. Shen, "Universal best arm identification," *IEEE Trans. Signal Process.*, vol. 67, no. 17, pp. 4464–4478, Sep. 2019.
[12] P. Alatur, K. Y. Levy, and A. Krause, "Multi-player bandits: The adversarial case," *J. Mach. Learn. Res.*, vol. 21, no. 77, pp. 1–23, 2020.
[13] J. Rosenski, O. Shamir, and L. Szlak, "Multi-player bandits—A musical chairs approach," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 155–163.
[14] M. Bande and V. V. Veeravalli, "Adversarial multi-user bandits for uncoordinated spectrum access," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2019, pp. 4514–4518.
[15] V. Anantharam, P. Varaiya, and J. Walrand, "Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays—Part I: I.I.D rewards," *IEEE Trans. Autom. Control*, vol. 32, no. 11, pp. 968–976, Nov. 1987.
[16] J. Komiyama, J. Honda, and H. Nakagawa, "Optimal regret analysis of Thompson sampling in stochastic multi-armed bandit problem with multiple plays," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 1152–1161.
[17] P.-A. Wang, A. Proutiere, K. Ariu, Y. Jedra, and A. Russo, "Optimal algorithms for multiplayer multi-armed bandits," in *Proc. Int. Conf. Artif. Intell. Stat.*, 2020, pp. 4120–4129.
[18] H. Tibrewal, S. Patchala, M. K. Hanawal, and S. J. Darak, "Multiplayer multi-armed bandits for optimal assignment in heterogeneous networks," 2019. [Online]. Available: arXiv:1901.03868.
[19] A. Magesh and V. V. Veeravalli, "Multi-user MABs with user dependent rewards for uncoordinated spectrum access," in *Proc. 53rd Asilomar Conf. Signals Syst. Comput.*, 2019, pp. 969–972.

[20] E. Boursier, E. Kaufmann, A. Mehrabian, and V. Perchet, "A practical algorithm for multiplayer bandits when arm means vary among players," in *Proc. Int. Conf. Artif. Intell. Stat.*, 2020, pp. 1211–1221.
[21] S. Bubeck and T. Budzinski, "Coordination without communication: Optimal regret in two players multi-armed bandits," in *Proc. Conf. Learn. Theory*, 2020, pp. 916–939.
[22] J.-Y. Audibert, S. Bubeck, and G. Lugosi, "Regret in online combinatorial optimization," *Math. Oper. Res.*, vol. 39, no. 1, pp. 31–45, 2013.
[23] A. Garivier and O. Cappé, "The KL-UCB algorithm for bounded stochastic bandits and beyond," in *Proc. 24th Annu. Conf. Learn. Theory*, 2011, pp. 359–376.
[24] P. Landgren, V. Srivastava, and N. E. Leonard, "On distributed cooperative decision-making in multiarmed bandits," in *Proc. Eur. Control Conf. (ECC)*, 2016, pp. 243–248.
[25] S. Shahrampour, A. Rakhlin, and A. Jadbabaie, "Multi-armed bandits in multi-agent networks," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2017, pp. 2786–2790.
[26] Y. Wang, J. Hu, X. Chen, and L. Wang, "Distributed bandit learning: Near-optimal regret with efficient communication," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–31.
[27] B. Awerbuch and R. Kleinberg, "Competitive collaborative learning," *J. Comput. Syst. Sci.*, vol. 74, no. 8, pp. 1271–1288, 2008.
[28] N. Cesa-Bianchi, C. Gentile, Y. Mansour, and A. Minora, "Delay and cooperation in nonstochastic bandits," in *Proc. Conf. Learn. Theory*, 2016, pp. 605–622.
[29] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.
[30] K. Xu, M. Gerla, and S. Bae, "How effective is the IEEE 802.11 RTS/CTS handshake in ad hoc networks," in *Proc. IEEE Global Telecommun. Conf.*, vol. 1, Nov. 2002, pp. 72–76.
[31] R. Bonnefoi, L. Besson, C. Moy, E. Kaufmann, and J. Palicot, "Multi-armed bandit learning in IoT networks: Learning helps even in non-stationary settings," in *Proc. Int. Conf. Cogn. Radio Oriented Wireless Netw.*, 2017, pp. 173–185.
[32] T. Uchiya, A. Nakamura, and M. Kudo, "Algorithms for adversarial bandit problems with multiple plays," in *Proc. Int. Conf. Algorithmic Learn. Theory*, 2010, pp. 375–389.
[33] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The non-stochastic multiarmed bandit problem," *SIAM J. Comput.*, vol. 32, no. 1, pp. 48–77, Jan. 2003.
[34] A. Kulesza and B. Taskar, "Determinantal point processes for machine learning," *Found. Trends Mach. Learn.*, vol. 5, nos. 2–3, pp. 123–286, 2012.
[35] R. Arora, O. Dekel, and A. Tewari, "Online bandit learning against an adaptive adversary: From regret to policy regret," in *Proc. 29th Int. Conf. Mach. Learn.*, 2012, pp. 1747–1754.
[36] L. G. Tallini, S. Al-Bassam, and B. Bose, "On the capacity and codes for the Z-channel," in *Proc. IEEE Int. Symp. Inf. Theory*, 2002, p. 422.
[37] S. Lin and D. J. Costello, *Error Control Coding*, 2nd ed. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 2004.
[38] P.-N. Chen, H.-Y. Lin, and S. M. Moser, "Optimal ultrasmall block-codes for binary discrete memoryless channels," *IEEE Trans. Inf. Theory*, vol. 59, no. 11, pp. 7346–7378, Nov. 2013.
[39] J. M. Borden, "Optimal asymmetric error detecting codes," *Inf. Control*, vol. 53, nos. 1–2, pp. 66–73, 1982.

Chengshuai Shi (Graduate Student Member, IEEE) received the B.E. degree in electrical engineering from the School of the Gifted Young, University of Science and Technology of China in 2019. He is currently pursuing the Ph.D. degree with the Charles L. Brown Department of Electrical and Computer Engineering, University of Virginia. His current research focuses are on multiarmed bandits, federated learning, and reinforcement learning.



Cong Shen (Senior Member, IEEE) received the B.S. and M.S. degrees from the Department of Electronic Engineering, Tsinghua University, China, in 2002 and 2004, respectively, and the Ph.D. degree in electrical engineering from the University of California at Los Angeles in 2009. He is currently an Assistant Professor with the Charles L. Brown Department of Electrical and Computer Engineering, University of Virginia. He also has extensive industry experience, having worked for Qualcomm, SpiderCloud Wireless, Silvus Technologies, and Xsense.ai, in various full time and consulting roles. His general research interests are in the area of communications, wireless networks, and machine learning. He currently serves as an Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS and IEEE WIRELESS COMMUNICATIONS LETTERS.