An Attackability Perspective on No-Sensing Adversarial Multi-player Multi-armed Bandits

Chengshuai Shi and Cong Shen
University of Virginia
Charlottesville, VA 22904, USA
{cs7ync, cong}@virginia.edu

2021 IEEE International Symposium on Information Theory (ISIT) | 978-1-5386-8209-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ISIT45174.2021.9517726

***Abstract*—In this work, we study the no-sensing adversarial multi-player multi-armed bandits problem. A new dimension of hardness, called *attackability*, is introduced, which is orthogonal to the hardness of multiple players. All adversaries can be categorized based on the attackability and we introduce *Adversary-Adaptive Collision-Communication (A2C2)*, a family of algorithms with forced-collision communications among players. Information-theoretic tools of the Z-channel model, error-correction/detection coding, and randomized communication are utilized to address the challenge of implicit communication without collision information in an adversarial environment. Theoretical analysis proves that asymptotic attackability-dependent sublinear regrets can be achieved, which do not have an exponential dependence on the number of players and as a result reveal a fundamental tradeoff between the two dimensions of hardness in this problem.**

I. Introduction

The multi-player version of the multi-armed bandits (MP-MAB) problem, in which players simultaneously play the bandit game and interact with each other through arm collisions, has received increasing interest in recent years [2]–[6]. This model is largely motivated by practical applications such as cognitive radio [7]–[10] and wireless caching [11], [12], where user interactions must be taken into account along with the bandit game.

Depending on how rewards are generated, the MP-MAB game can be either stochastic or adversarial. While most of the existing works focus on the stochastic setting, the (oblivious) adversarial problem is considerably harder. The need of fighting the adversary while interacting with other players introduces significant difficulties to the problem. 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 [2], [3], [8], [13], [14]. Recently, [4] proposes to purposely instigate collisions as a way to communicate 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].

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. A widely-recognized more difficult MP-MAB problem is the no-sensing scenario, in which players cannot access the collision indicators. Recently, some progress has been made on the stochastic no-sensing problem [5], [10], [17], but the more difficult setting of decentralized no-sensing adversarial MP-MAB remains open. To the best of our knowledge, [6] is the only work that achieves a sublinear regret of $O(T^{1-\frac{1}{2M}})$, where M is the number of players and T is the time horizon. We note that this exponential dependence on M reveals a particular *dimension of hardness* (multiple players) in the no-sensing adversarial MP-MAB problem.

This work makes progress in the no-sensing adversarial MP-MAB problem by addressing the challenges in incorporating implicit communications. A new dimension of hardness is revealed: *attackability* of the adversary, that is orthogonal to the multi-player dimension of hardness [6]. Under both attackability-aware and attackability-unaware settings, we develop a suite of *Adversary-Adaptive Collision-Communication (A2C2)* algorithms. All of the A2C2 algorithms utilize some (common) new elements, such as the information-theoretical Z-channel model and error-correction coding, to design a communication protocol that can effectively counter the adversary with a non-dominant communication regret. 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. The A2C2 algorithms are proved to achieve attackability-dependent sublinear regrets asymptotically, without an exponential dependence on the number of players as in [6].

We may view A2C2 of this paper and the method of [6] as operating at different regimes in the two-dimensional hardness space (multi-player versus attackability). Philosophically speaking, this result shows that one can trade off the multi-player dimension of hardness with the attackability dimension of hardness. A comparison of the regret bounds is given in Table I, including a preview of the main results of this work.

II. Related Work

Collision-sensing MP-MAB. Initial approaches for collision-sensing MP-MAB adopt single-player algorithms with various collision-avoidance protocols [2], [3], [8], [13],

A full version of this paper [1] is accepted to be published at IEEE Journal on Selected Areas in Information Theory. The work was supported in part by the US National Science Foundation (NSF) under Grant CNS-2002902 and ECCS-2029978, and a Commonwealth Cyber Initiative (CCI) cybersecurity research collaboration grant.

TABLE I
REGRET BOUNDS OF ADVERSARIAL MP-MAB ALGORITHMS

Model/Reference	Asymptotic Bound
Centralized, Optimal Regret [18]	$\Theta(\sqrt{MKT})$
Collision Sensing [14]	$\tilde{O}(K^M M^2 T^{\frac{3}{4}})$
Collision Sensing [19]	$\tilde{O}(M^{\frac{4}{3}} K^{\frac{1}{3}} T^{\frac{2}{3}})$
Collision Sensing, $M = 2$ [6]	$\tilde{O}(K^2\sqrt{T})$
No Sensing [6]	$\tilde{O}(MK^{\frac{3}{2}} T^{1-\frac{1}{2M}})$
No Sensing, α-aware (this work)	$\tilde{O}(M^{\frac{4}{3}} K^{\frac{1}{3}} T^{\frac{2+\alpha+\epsilon}{3}})$
No Sensing, β-aware (this work)	$\tilde{O}(M^2 K^{\frac{2}{3}} T^{\max\{\frac{1+\beta}{2}, \frac{2}{3}\}})$
No Sensing, α-unaware (this work)	$\tilde{O}(M^{\frac{4}{3}} K^{\frac{1}{3}} T^{\frac{5+\alpha+\epsilon}{6}})$
No Sensing, β-unaware (this work)	$\tilde{O}(M^2 K^{\frac{1}{3}} T^{\max\{\frac{2+\beta+\epsilon}{3}, \frac{3}{4}\}})$

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

including EXP3 for adversarial MP-MAB with a regret of $O(T^{\frac{3}{4}})$ [14]. The idea of implicit communication with forced collisions is introduced and developed by [4], [20]–[22]. The theoretical analysis in [4] shows, for the first time, that the regret of decentralized stochastic MP-MAB can approach the centralized lower bound [15]. For the adversarial environment, a regret of $O(T^{\frac{2}{3}})$ is achieved in [19] by invoking forced collisions to let players coordinately perform an EXP3 algorithm. The two-players performance has been improved to $O(\sqrt{T\log(T)})$ in [6].

No-sensing MP-MAB. Early attempts to incorporate implicit communication in the no-sensing stochastic setting are discussed in [4], [10]. For the more difficult case of no-sensing adversarial MP-MAB, progress is very limited. To the best of our knowledge, [6] is the only work studying this problem. The idea is to design a collision-avoidance approach by reserving "safe" arms for players, which results in a regret of $O(T^{1-\frac{1}{2M}})$ that has an exponential dependency on M.

III. PROBLEM FORMULATION

A. The no-sensing adversarial MP-MAB problem

In the decentralized no-sensing adversarial MP-MAB model, there are K arms and $1 < M \leq K$ players. The arms are labeled 1 to K and the players 1 to M, respectively. At each time step $t \in [T]$, each player $m \in [M]$ individually chooses and pulls arm $\pi_m(t)$. Simultaneously, an adversary assigns loss $l_k(t)$ for each arm $k \in [K]$. The true loss $l_k(t)$ is assumed to be player-independent. 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 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

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

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

If the players have access to both $s_{\pi_m(t)}(t)$ and $\eta_{\pi_m(t)}(t)$, it is a collision-sensing problem. 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. The lack of information on the collision indicators complicates the MP-MAB problem in general [4], [10], and this challenge is more significant for the adversarial setting [6]. Note that if $l_k(t) \neq 1$, $\forall k, t$, the no-sensing setting is equivalent to collision-sensing.

For this model, the notion of regret can be generalized w.r.t. the best allocation of players to arms as follows [19]:

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

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

As shown in [6], one cannot obtain any non-trivial regret guarantees facing an adaptive adversary. This work thus focuses on the oblivious adversary who chooses the loss sequence independently of the actions of players.

B. Attackabilities of the adversary

To explore the idea of forced-collision communications, the overall horizon T is divided into the exploration and communication phases [4], [19]. 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 different phases:

- **Exploration phase.** Assuming that the preceding communication phase is successful, no negative influence occurs when the adversary assigns loss 1 in the exploration phase, since the regret is measured by the gap to the optimal choice.
- **Communication phase.** The loss-1 attack in a communication phase may lead to communication errors, which jeopardize the essential coordination among players and lead to a potentially *linear* regret due to collisions in the subsequent exploration phase, as illustrated in Fig. 1. The worst-case scenario can have all-one loss sequences for all communication phases, which prevents any information sharing.

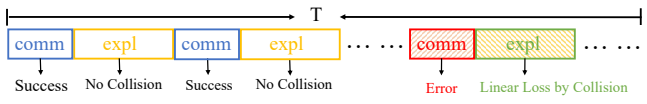

Fig. 1. Illustration of communication and exploration phases.

Previous studies of the stochastic no-sensing MP-MAB [4], [10] show that any policy with collision-communications has a dependency on the environment's ability to "attack" such communications. In the stochastic settings, such ability is characterized by a lower bound $\mu_{\min}$ such that $0 < \mu_{\min} \leq \min_{k\in[K]} \mu_k$, where μ_k is the expectation of arm k's rewards.

Analogous to the role of $\mu_{\min}$ [4], [10], we propose a new concept 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. More specifically, we define two types of attackabilties: the **local attackability** and the **global attackability**, which provide two different ways to *categorize* all the adversaries.

First, 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, and is defined as follows.

Definition 1. *For a 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.*

Corollary 1. *For any given adversary, there exists $\alpha \in [0,1]$ such that her local attackability satisfies $W(T) \leq O\left(T^{\alpha}\right)$.*

Any possible adversary can be characterized by the local attackability parameter α in Corollary 1, and 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 a 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 loss 1's that are assigned by the adversary on arm k throughout the T time slots.*

Corollary 2. *For any given adversary, there exists $\beta \in [0,1]$ such that her global attackability satisfies $V(T) \leq O\left(T^{\beta}\right)$.*

Similar to local attackability, the adversaries share the same parameter β can be viewed as in the same category. However, note that since the local attackability parameter α in Corollary 1 does not provide any bound on the overall attack, it is more stringent than the global attack parameter β in Corollary 2.

It is important to note 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 must be in one category. In addition, such categorization does not need to be aware by the players. In this sense, we do not impose more assumptions than [6]. 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.

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 [19], [22]. Player 1 (leader) determines arm assignments and transmits them to the remaining players (followers) in the communication phases. Then, in the following exploration phases, all the players keep sampling the assigned arms.

A. Exploration Phase

Assuming explicit communications are allowed, the problem is similar to the adversarial multi-play problem, where the leader (the centralized agent) assigns M arms $A(t) = \{A_1(t), ..., A_M(t)\}$ at each time step t to the followers, i.e., arm $A_m(t)$ for player m. As commonly adopted in the multi-play setting [23], we can view each subset of M distinct arms $\{A_1, ..., A_M\}$ is viewed as a single meta-arm A and the set of all meta-arms $\mathcal{K}$ is defined as: $\mathcal{K} := \{\{A_1, ..., A_M\} \subseteq [K] | A_m \neq A_n \text{ for any } m \neq n\}$.

An arm assignment policy that builds on [19] is adopted in this paper. The key idea is to perform a centralized EXP3 algorithm but with only the leader's observations, which is designed to facilitate the generalization to the decentralized setting. Specifically, after exploration at time $t-1$, using her own observation $\hat{l}_{A_1(t-1)}(t-1)$ from arm $A_1(t-1)$, the leader updates an unbiased loss estimator as $\forall k \in [K], \tilde{l}_k(t-1) = \frac{M\hat{l}_{A_1(t-1)}(t-1)}{\sum_{A:k\in A\in\mathcal{K}} P_A(t-1)} \mathbb{1}\{k = A_1(t-1)\}$, where $P_A(t-1)$ is the probability that meta-arm A is chosen at time $t-1$.

For time t, the decision is based on the cumulative loss estimator $\tilde{L}_A(t)$ for each meta-arm $A \in \mathcal{K}$ as $\tilde{L}_A(t) = \sum_{\upsilon=1}^{t-1} \sum_{k\in A} \tilde{l}_k(\upsilon)$. Then, the EXP3 algorithm [24] is applied to the meta-arms, so that each meta-arm $A \in \mathcal{K}$ is sampled with a probability $P_A(t) \propto \exp(-\eta \tilde{L}_A(t))$ as the exploration meta-arm $A(t)$. The loss estimator $\{\tilde{l}_k(t)\}_{k\in[K]}$ is then again updated and the same procedures iterate for time $t+1$.

Lastly, 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 and the update happens only after each exploration phase. The leader also uses her samples of losses observed during the entire exploration phase as the feedback to assign arms for the next phase.

B. Communication Phase

In the communication phases, arm assignments are transmitted from the leader to the followers with the forced collisions mechanism [4], [10]. Every player is first assigned a unique communication arm corresponding to her index, e.g., 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. In the more challenging no-sensing setting, without information on the collision indicator, loss-1 attacks from the adversary may cause communication errors and incur a linear regret as shown in Fig. 1. The no-sensing settings are discussed in the following sections.

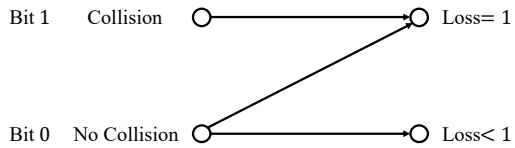

Fig. 2. The Z-channel model for forced-collision communications.

V. Local Attackability: α-(un)aware A2C2

In this section, we first assume that the local attackability parameter α is known to the players, i.e., α-aware A2C2, and introduce some important ideas. Then, we discuss the case with unknown α, i.e., α-unaware A2C2.

A. α-aware A2C2

Two information-theoretic concepts play important roles in the design of α-aware A2C2 and subsequent algorithms.

Z-channel model. There exists an asymmetry in collision-communication of no-sensing MP-MAB: bit 1 (collision) is *always* received correctly, while bit 0 (no collision) can be potentially corrupted by an attack of loss 1. Thus, the adversary attack corresponds to a Z-channel model [25] as shown in Fig. 2: she can attack bit 0 but not bit 1. Compared with the stochastic no-sensing MP-MAB [10], the key difference here is that a *fixed* crossover probability does not exist.

Error-correction code with long blocklength. With the Z-channel model, the key idea to overcome the full attackability is to "overpower" the adversary via coding that has sufficient error-correction capabilities [26]. To facilitate the regret analysis, we choose to use the simple repetition code [27].[1] At the encoder, each information bit is repeated to a string of length $h(T,\alpha+\epsilon) = \Theta(T^{\alpha+\epsilon}) = \omega(T^{\alpha})$, where $\epsilon > 0$ is an arbitrarily small constant.[2] 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 and thus provides successful communications *asymptotically*.

B. α-unaware A2C2

With no information of α, the main difficulty lies in how to prepare for the worst case without incurring a linear loss. In addition to the key features in α-aware A2C2, several new ideas are needed in the α-unaware A2C2 algorithm. A sketch of α-unaware A2C2 is presented in Algorithm 1.[3]

Estimation of α. To effectively protect communications, we propose to adaptively estimate α in an escalation fashion. The estimated value $\alpha' \in [0,1]$ starts with 0, and increases with a step size of ϵ upon each communication failure, where

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

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

[3]For compactness, Algorithm 1 is written to be applicable to all M players, but note that these players are operating in a decentralized fashion.

Algorithm 1 Sketch of α-unaware A2C2 Algorithm

1: **for** phase p **do**
Leader:
2: Sample one meta-arm $A(p)$ w.r.t. $P_A(p)$
3: Send assigned arms in $A(p)$ to followers with the error-detection repetition code ▷ *Communication*
4: Iterate "Uplink-Downlink" for $N(\xi)$ times and decide to update estimation α' or not ▷ *Synchronization*
5: Pull the assigned arm for τ time steps and update loss estimators with the collected rewards ▷ *Exploration*
Follower m:
6: Receive and decode the assigned arm, and check if a communication error happens ▷ *Communication*
7: Iterate "Uplink-Downlink" for $N(\xi)$ times and decide to update estimation α' or not ▷ *Synchronization*
8: Pull the assigned arm for τ time steps ▷ *Exploration*
9: **end for**

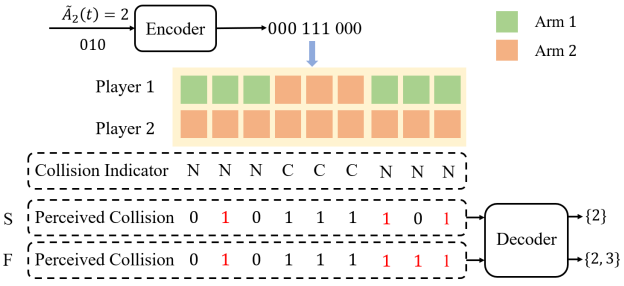

Fig. 3. Example of one communication phase and error-detection coding with $M = 2$, $K = 3$ and $h(T,\alpha') = 3$.

$\epsilon > 0$ is an arbitrarily small positive constant. Once reaching $\alpha' = v\epsilon > \alpha$, we have $h(T,\alpha') = \omega(W(T))$, which means there will be no more communication errors (asymptotically).

Error-detection repetition code. To provide information of communication failure for the aforementioned escalation mechanism, the constant weight code [28], which is an error-detection code for the Z-channel, is utilized. To facilitate discussions, a specific kind of constant weight code is adopted. As shown in Fig. 3, while transmitting arm index k, the leader represents it by a bit sequence of length K where 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 codewords share the weight of $h(T,\alpha')$. Upon receiving, the entire bit string is divided into K blocks and the decoder outcome is the (possibly multiple) indices of the blocks that have all-ones.

Thanks to the property of the Z-channel, the source index k is always decoded correctly. Thus, if the decoder outputs more than one index, there must be a communication error (example 'F' in Fig. 3), meaning the estimation should be updated. Otherwise, the decoder outputs only one index, which indicates the communication is successful (example 'S' in Fig. 3) and it is sufficient to maintain the current estimation.

Synchronization with randomized length. Error-detection repetition code allows each follower to decide whether the α-estimation needs to be updated. However, such decisions may vary *across* players, which calls for communication for syn-

chronization. We note 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 estimation, while the others do not.

To solve this problem, we 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 (referred to as 'uplink'). The followers send bit 1 representing error and bit 0 representing no error with the same error-correction repetition code of length $h(T, \alpha')$. As long as one follower has communication errors, the leader can correctly receive bit 1 (signal for updating). This '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 (referred to as 'downlink'). If bit 1 is sent, all the followers can receive it correctly. However, in the case of bit 0, uneven attacks may happen and thus the synchronization may fail.

To keep synchronization successful with a high probability, the 'uplink-downlink' procedure keeps iterating for a *random* number of rounds $N(\xi)$, which is 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. This procedure with a carefully chosen ξ is crucial in maintaining a sub-linear regret.

VI. Global Attackability: β-(un)aware A2C2

In the discussion of Section V, although the local attackability is bounded, the adversary can attack *arbitrarily many* times. However, with the global attackability, the adversary has an overall budget for attacking rather than a one-time budget. Thus, it is an overkill to prevent the global attackability in *every* communication phase. More efficient algorithms, β-(un)aware A2C2, are proposed with a holistic consideration for the total budget of the adversary. Details of β-(un)aware A2C2 can be found in [1], and we only highlight the key design philosophies here. As illustrated in Fig. 1, attacks in one communication phase can only cause a one-time linear loss in the next (finite-length) exploration phase, 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, it is sufficient to adopt a less powerful error-correction/detection code and a less frequent synchronization scheme, as long as the loss caused by the global attackability does not dominate the regret.

VII. Performance Analysis

This section provides the theoretical analysis for all proposed A2C2 algorithms. Detailed proofs can be found in [1].

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* $\mathbb{E}[R(T)] \leq O(M^{\frac{4}{3}} K^{\frac{1}{3}} \log(K)^{\frac{2}{3}} T^{\frac{2+\alpha+\epsilon}{3}})$, *where ϵ is an arbitrarily small positive constant.*

Theorem 2 (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* $\mathbb{E}[R(T)] \leq O(M^{\frac{4}{3}} K^{\frac{1}{3}} \log(K)^{\frac{1}{3}} T^{\frac{5+\alpha+\epsilon}{6}})$, *where $\epsilon > 0$ is an arbitrarily small positive constant.*

Noting that for $\alpha = 0$, the known regret in collision-sensing setting $O(T^{\frac{2}{3}})$ [19] is recovered by α-aware A2C2. When $\alpha = 1$, i.e., $W(T) = \Omega(T)$, the adversary can asymptotically attack all time slots, and thus our regrets become $O(T)$. Also, 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 2.

We also note that with the properly chosen parameters in [1], the expected regret of β-aware and β-unaware A2C2 algorithms are of order $O(M^2 K^{\frac{2}{3}} \log(K)^{\frac{1}{3}} T^{\max\{\frac{1+\beta}{2}, \frac{2}{3}\}})$and $O(M^2 K^{\frac{1}{3}} \log(K)^{\frac{1}{3}} T^{\max\{\frac{2+\beta+\epsilon}{3}, \frac{3}{4}\}})$, respectively.

Compared with the regret of $O(MK^{\frac{3}{2}} T^{1-\frac{1}{2M}})$ in [6], 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 [6], 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.

VIII. Conclusions

This work made progress in the no-sensing adversarial MP-MAB problem by incorporating implicit communications. 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, with several new tools from information theory and communication theory. Theoretical analysis showed that the proposed algorithms have attackability-dependent regrets, which eliminated 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 of attackability that compliments the hardness associated with the number of players.

REFERENCES

[1] C. Shi and C. Shen, "On no-sensing adversarial multi-player multi-armed bandits with collision communications," *IEEE Journal on Selected Areas in Information Theory*, 2021.

[2] K. Liu and Q. Zhao, "Distributed learning in multi-armed bandit with multiple players," *IEEE Trans. Signal Processing*, vol. 58, no. 11, pp. 5667–5681, 2010.

[3] L. Besson and E. Kaufmann, "Multi-player bandits revisited," in *Algorithmic Learning Theory*, 2018, pp. 56–92.

[4] E. Boursier and V. Perchet, "SIC-MMAB: synchronisation involves communication in multiplayer multi-armed bandits," in *Advances in Neural Information Processing Systems*, 2019, pp. 12 071–12 080.

[5] G. Lugosi and A. Mehrabian, "Multiplayer bandits without observing collision information," *arXiv preprint arXiv:1808.08416*, 2018.

[6] S. Bubeck, Y. Li, Y. Peres, and M. Sellke, "Non-stochastic multi-player multi-armed bandits: Optimal rate with collision information, sublinear without," in *Conference on Learning Theory*, 2020, pp. 961–987.

[7] 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, 2011.

[8] O. Avner and S. Mannor, "Concurrent bandits and cognitive radio networks," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2014, pp. 66–81.

[9] C. Gan, R. Zhou, J. Yang, and C. Shen, "Cost-aware cascading bandits," *IEEE Trans. Signal Processing*, vol. 68, pp. 3692–3706, June 2020.

[10] C. Shi, W. Xiong, C. Shen, and J. Yang, "Decentralized multi-player multi-armed bandits with no collision information," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 1519–1528.

[11] 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, April 2020.

[12] X. Xu and M. Tao, "Decentralized multi-agent multi-armed bandit learning with calibration for multi-cell caching," *IEEE Trans. Commun.*, pp. 1–1, 2020.

[13] J. Rosenski, O. Shamir, and L. Szlak, "Multi-player bandits–a musical chairs approach," in *International Conference on Machine Learning*, 2016, pp. 155–163.

[14] M. Bande and V. V. Veeravalli, "Adversarial multi-user bandits for uncoordinated spectrum access," in *IEEE International Conference on Acoustics, Speech and Signal Processing (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: Iid rewards," *IEEE Trans. Autom. Control*, vol. 32, no. 11, pp. 968–976, 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 *Proceedings of the 32nd International Conference on Machine Learning*, 2015, pp. 1152–1161.

[17] S. Bubeck and T. Budzinski, "Coordination without communication: optimal regret in two players multi-armed bandits," in *Conference on Learning Theory*. PMLR, 2020, pp. 916–939.

[18] J.-Y. Audibert, S. Bubeck, and G. Lugosi, "Regret in online combinatorial optimization," *Mathematics of Operations Research*, vol. 39, no. 1, pp. 31–45, 2013.

[19] P. Alatur, K. Y. Levy, and A. Krause, "Multi-player bandits: The adversarial case," *Journal of Machine Learning Research*, vol. 21, 2020.

[20] H. Tibrewal, S. Patchala, M. K. Hanawal, and S. J. Darak, "Distributed learning and optimal assignment in multiplayer heterogeneous networks," in *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE, 2019, pp. 1693–1701.

[21] P.-A. Wang, A. Proutiere, K. Ariu, Y. Jedra, and A. Russo, "Optimal algorithms for multiplayer multi-armed bandits," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 4120–4129.

[22] E. Boursier, E. Kaufmann, A. Mehrabian, and V. Perchet, "A practical algorithm for multiplayer bandits when arm means vary among players," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 1211–1221.

[23] T. Uchiya, A. Nakamura, and M. Kudo, "Algorithms for adversarial bandit problems with multiple plays," in *International Conference on Algorithmic Learning Theory*. Springer, 2010, pp. 375–389.

[24] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The non-stochastic multiarmed bandit problem," *SIAM Journal on Computing*, vol. 32, no. 1, pp. 48–77, 2002.

[25] L. G. Tallini, S. Al-Bassam, and B. Bose, "On the capacity and codes for the Z-channel," in *Proceedings of the IEEE International Symposium on Information Theory*, 2002, p. 422.

[26] S. Lin and D. J. Costello, *Error Control Coding*, 2nd ed. USA: Prentice-Hall, Inc., 2004.

[27] P.-N. Chen, H.-Y. Lin, and S. M. Moser, "Optimal ultrasmall block-codes for binary discrete memoryless channels," *IEEE Trans. Info. Theory*, vol. 59, no. 11, pp. 7346–7378, 2013.

[28] J. M. Borden, "Optimal asymmetric error detecting codes," *Information and Control*, vol. 53, no. 1-2, pp. 66–73, 1982.