# An Instance-Dependent Analysis for the Cooperative Multi-Player Multi-Armed Bandit

**Aldo Pacchiano** APACCHIANO@MICROSOFT.COM
*Microsoft Research, NYC*

**Peter Bartlett** PETER@BERKELEY.EDU
*University of California, Berkeley*

**Michael Jordan** JORDAN@CS.BERKELEY.EDU
*University of California, Berkeley*

**Editors:** Shipra Agrawal and Francesco Orabona

## Abstract

We study the problem of information sharing and cooperation in Multi-Player Multi-Armed bandits. We propose the first algorithm that achieves logarithmic regret for this problem when the collision reward is unknown. Our results are based on two innovations. First, we show that a simple modification to a successive elimination strategy can be used to allow the players to estimate their suboptimality gaps, up to constant factors, in the absence of collisions. Second, we leverage the first result to design a communication protocol that successfully uses the small reward of collisions to coordinate among players, while preserving meaningful instance-dependent logarithmic regret guarantees.

**Keywords:** List of keywords

## 1. Introduction

We consider the cooperative Multi-Player version of the Multi-Armed bandit problem. Generalizing the single-player case, the bandit instance is defined by mean rewards, $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_K) \in [0,1]^K$, all of which are unknown to each of $M$ players. There is a permutation $\sigma \in \mathbb{S}_K$ (unknown to the players) such that $\mu_{\sigma_1} \geq \mu_{\sigma_2} \cdots \geq \mu_{\sigma_K}$. At each time step $t = 1, \cdots, T$, each of the players $p \in [M]$ chooses an action $i_t^p \in [K]$ and observes the corresponding (random) reward. If two players pull the same arm, they receive a reward sampled from a distribution with unknown mean $\mu_{\text{collision}} \leq \mu_{\sigma_K}$. Our objective will be to design an algorithm with sublinear pseudo-regret:

$$\mathcal{R}_T = T \max_{\mathbf{a} \in \{0,1\}^K : \sum_{i=1}^K a(i) = M} \langle \mathbf{a}, \boldsymbol{\mu} \rangle - \sum_{t=1}^{T} \sum_{p=1}^{M} \mu_{i_t^p}.$$

As opposed to other work, such as Avner and Mannor (2014); Rosenski et al. (2016); Alatur et al. (2020), we do not make the assumption that collisions are announced to the players; rather, we simply assume that whenever two players select the same arm, they both observe a reward with mean $\mu_{\text{collision}}$. The players do not know for certain if there was a collision or not (cf. Boursier and Perchet, 2018; Shi et al., 2020; Bubeck and Budzinski, 2020). We will make use of the implicit information provided by collisions (a very low reward value) to design a communication protocol that will allow players to coordinate. We make the following boundedness assumption on the distribution of the reward signals observed by the players:

**Assumption 1** *All $K$ arms have bounded distributions with support in* $[0, 1]$.

Our main contribution is to design an elimination-based algorithm whose regret satisfies instance-dependent logarithmic bounds without assuming explicit knowledge of collision information. Our algorithm generalizes all previous approaches (for example those where collisions are announced Avner and Mannor (2014); Rosenski et al. (2016); Alatur et al. (2020) or even those where collisions are unannounced but their reward equals zero Huang et al. (2021)) to this problem since it does not require knowledge of the mean collision reward. We use the following notation to refer to the suboptimality gaps among the arms:

$$\Delta_{\sigma_i,\sigma_j} = \mu_{\sigma_i} - \mu_{\sigma_j},$$

where $i < j$. Our main result can be summarized as follows.

**Theorem 2 (simplified)** *There exists a strategy such that the regret is upper bounded by:*

$$\mathcal{R}_T \leq \widetilde{\mathcal{O}}\left(\frac{M(K-M)K^2\log(T)}{\Delta_{\sigma_M,\sigma_{M+1}}} + \mathbf{poly}(\log(T), K, M)\right),$$

*with probability at least* $1 - \min\left(\frac{1}{T}, \frac{K}{81}\right)$ *where* $\widetilde{\mathcal{O}}(\cdot)$ *hides factors logarithmic in* $M$ *and* $K$ *only and* $\mathbf{poly}(\log(T), K, M)$ *is linear in* $\log(T)$.

## 2. Previous Work

The Multi-Player bandit problem with bounded communication was first introduced in Lai et al. (2008); Liu and Zhao (2010); Anandkumar et al. (2011), and has been extensively studied since then under various assumptions on the communication patterns and the nature of the collisions (Avner and Mannor, 2014; Rosenski et al., 2016; Palicot, 2018; Lugosi and Mehrabian, 2018; Boursier and Perchet, 2018; Alatur et al., 2020; Bubeck et al., 2020b). Perhaps the first instance of a centralized version of the problem we study in this work appeared in Anantharam et al. (1987), where the problem of a single player selecting multiple arms simultaneously is considered.

The problem of Multi-Player, Multi-Armed bandits has commonly been motivated via its application to wireless communication and networking (Liu and Zhao, 2010; Rosenski et al., 2016); for example as a way to model the case where several users must access a wireless channel in a decentralized manner (Besson and Kaufmann, 2018).

We can classify the existing settings and algorithmic approaches to the Multi-Player Multi-Armed in broadly two categories. When collision information is available to the players and when it is not. In the first category, algorithms such as SIC-MMAB (Boursier and Perchet, 2018) or DPE1 (Wang et al., 2020) have been developed, the second of which achieves the same asymptotic regret as that obtained by an optimal centralized algorithm. Both of these algorithms crucially exploit the known collision information to establish communication between the players.

In the "No Sensing" setting where collision information is not readily available to the players, but instead players receive a diminished or zero reward, the problem of developing an optimal algorithm is substantially more challenging and has not been fully solved yet. Most importantly the three most prominent algorithms, SIC-MMAB2 (Boursier and Perchet, 2018), EC-SIC (Shi et al., 2020) and the algorithm of Lugosi and Mehrabian (2018) suffer from a variety of drawbacks.

SIC-MMAB2 satisfies a regret guarantee of order $\mathcal{O}\left(\sum_{i>M} \frac{M \log(T)}{\Delta_{\sigma_M,\sigma_i}} + \frac{MK^2}{\mu_{\sigma_K}} \log(T)\right)$. Unfortunately SIC-MMAB2 is suboptimal in two ways. First the algorithm requires knowledge of $\mu_{\sigma_K}$, and second its regret guarantee suffers an inverse dependence on $\mu_{\sigma_K}$, a quantity that may be astronomically large. Other algorithms for the No Sensing setting such as studied in Theorem 1.2 in Lugosi and Mehrabian (2018), and the ADAPTED SIC-MMAB algorithm from Boursier and Perchet (2018) suffer from the same limitations (see Table 1 in Boursier and Perchet (2018)). The more recent EC-SIC algorithm (Shi et al., 2020) improves the $M$ and $K$ dependence from the SIC-MMAB2 regret upper bound but still suffers from the substantial drawback of requiring knowledge of at least a lower bound to $\mu_{\sigma_K}$. Although EC-SIC achieves a better regret guarantee than SIC-MMAB2, it also requires knowledge of the gap $\Delta_{\sigma_M,\sigma_{M+1}}$ to be available to all players. Other algorithms such as the algorithm from Theorem 1.1 in Lugosi and Mehrabian (2018) suffer from more serious problems such as quadratic dependence on the inverse gap $\frac{1}{\Delta_{\sigma_M,\sigma_{M+1}}}$.

Some of these drawbacks have been addressed by recent work (Huang et al., 2021). Under the assumption the collision reward equals 0, the authors dispense with the assumptions of shared knowledge of both $\mu_{\sigma_K}$ and $\Delta_{\sigma_M,\sigma_{M+1}}$. Their collision communication protocol makes use of a test that is very much in the spirit of ours (see the subroutine CollisionTest in Section 4.2). Communication is achieved by finding large arms (that are up to a constant proportion the scale of the largest arm) and pulling them. The authors manage to obtain a logarithmic instance dependent regret guarantee scaling as $\mathcal{O}\left(\sum_{i>M} \frac{\log(T)}{\Delta_{\sigma_M,\sigma_i}} + MK^2 \log(T) + KM^2 \log\left(\frac{1}{\Delta_{\sigma_M,\sigma_{M+1}}}\right)^2\right)$. Unfortunately, their algorithm heavily depends on the assumption $\mu_{\text{collision}} = 0$.

In the present paper we avoid the aforementioned drawbacks and derive the first *truly* logarithmic problem-dependent guarantee for the No Sensing Multi-Player Multi-Armed bandit problem with unknown collision rewards. We leverage the implicit communication that exists when collisions occur, namely that the mean collision reward is small. We show that a simple modification of a successive elimination strategy can be used to allow the players to estimate the suboptimality gaps up to constant factors in the absence of collisions. Using this result we design a communication protocol that successfully leverages the small reward of collisions to coordinate among players, while at the same time preserves meaningful instance-dependent logarithmic regret guarantees.

A different setting for the Multi-Player Multi-Armed bandit problem is one in which the players are required to avoid all collisions. It was shown by Bubeck and Budzinski (2020) that one can obtain the optimal regret in this setting without any collisions at all. Their result was limited to two players and three actions. A more recent version of that result Bubeck et al. (2020a) shows that it is possible to achieve a regret scaling with $\sqrt{T}$ for the cooperative stochastic Multi-Player Multi-Armed bandit problem with a dependence of $K^{11}M$ in the number of arms $K$ and the number of players $M$. The algorithmic strategy proposed in Bubeck et al. (2020a) relies on a clever algorithm that, with high probability, avoids collisions altogether. More recent results Liu and Sellke (2022) suggest that achieving a logarithmic instance dependent rate is impossible in the absence of communication.

## 3. Assumptions and Notation

Our algorithm is based on the idea of exploiting a communication protocol that leverages collisions while maintaining favorable regret guarantees. In comparison to other work, we do not make the assumption that collisions are announced to the players; rather, we simply assume that whenever two players select the same arm, they both observe an i.i.d. reward with mean $\mu_{\text{collision}} \leq \mu_{\sigma_K}$.

Throughout the paper we will use the notation $t$ to index the rounds of play. In each round all $M$ players select an arm and collect a reward.

We denote by $N_i^p(t)$ the (random) number of pulls of arm $i$ by player $p$ up until and including round $t$. And let $\widehat{\mu}_i^p(t)$ be the empirical estimator maintained by player $p$ of the mean reward $\mu_i$ of arm $i$ at time $t$. This estimator consists of an average of $N_i^p(t)$ samples. Similarly for any $t, t'$ denote by $\widehat{\mu}_i^p(t:t')$ as the empirical mean estimator of arm $i$ computed by player $p$ during rounds $t$ to $t'$ (inclusive). Let $\delta \in (0,1)$ be a probability parameter. We will make use of the following confidence interval diameter function,

$$D : \mathbb{N} \to \mathbb{R}_+ \text{ such that } D(n) = \sqrt{\tfrac{2g(n)}{n}} \text{ where } g(n) = \log(4n^2MK/\delta).$$

As a simple consequence of Hoeffding's inequality, for any $p \in [M]$ and $i \in [K]$, with probability at least $1 - \frac{\delta}{MK}$, for all $t \in \mathbb{N}$ simultaneously, we have:

$$|\widehat{\mu}_i^p(t) - \mu_i| \leq D(N_i^p(t)). \tag{1}$$

**The Good Event $\mathcal{E}$.** We will denote the (at least) $1-\delta$ probability event that all confidence intervals from Equation 1 hold for all $p \in [M]$, all $i \in [K]$ and all $t \in \mathbb{N}$ simultaneously as $\mathcal{E}$.

**Round Robin Schedule.** Whenever we say the arms are pulled by the players in a Round Robin schedule, we mean that during the first round player $p$ will pull arm $p$, and in subsequent rounds player $p$ will pull the arm with an index one more than the one she pulled in the previous rounds, unless she has pulled arm $K$ in which case she will pull arm 1 the next round. Whenever all players are pulling arms according to a Round Robin schedule, they do not collide.

**Special Rounds.** We will refer to all rounds occurring right after a complete cycle of a Round Robin round as *special rounds*. At the beginning of time, before any arm is eliminated, this will occur exactly during rounds that are multiples of $K$.

We also make the following assumptions.

**Assumption 3 (Collisions)** *Whenever two players collide, both players get a reward sampled from a distribution with mean $\mu_{\text{collision}}$ such that $\mu_{\text{collision}} \leq \mu_{\sigma_K}$.*

Assumption 3 is not particularly limiting. Our main contribution is to design an algorithm that does not require the *identity* of colliding arms to be announced. Since we do not assume the collision reward to equal zero, the algorithm of Huang et al. (2021) is not applicable to our case. The techniques in Huang et al. (2021) rely on the overwhelming probability of seeing a nonzero after repeated sampling of a non-zero mean arm. This cannot be used in the setting where the collision reward may have a nonzero mean.

**Assumption 4 (Shared knowledge)** *All arms are labeled, and the labels are known by all players $p \in [M]$.*

Assumption 4 is mild in comparison with the shared randomness assumption of previous works such as Bubeck et al. (2020a) and Bubeck and Budzinski (2020). We also assume common knowledge of the problem-independent functions $f$, $B$ and $g$.

## 4. Algorithm Overview and Analysis

Our algorithm starts by having all players agree on a value $t_{\text{collision}-\text{test}}$ satisfying $t_{\text{collision}-\text{test}} = \Theta(\frac{\log(1/\delta)}{\Delta^2_{\text{collision}}})$ up to logarithmic factors where $\Delta_{\text{collision}} = \mu_{\sigma_1} - \mu_{\text{collision}}$ and with estimators $\widehat{\mu}^p_{\text{collision}}$. Achieving this requires the same set of methods used in the following part of the algorithm, thus we defer its explanation. Subsequently all players will pull arms in a Round Robin fashion, thus not incurring any collisions. During these rounds all players maintain confidence bands for the mean values of each of the arms $i \in [K]$. All players maintain a connectivity graph with node set $[K]$, such that for any $i, j \in [K]$, the edge $(i, j)$ is present if their confidence intervals overlap.

Consider the first *special* round $t^1_{\text{first}}$ when for player 1, the number of connected components of this graph is more than one. We define $t^p_{\text{first}}$ analogously for all other players $p \in [M]$. Let's refer to the connected component with the highest empirical means as $\mathcal{C}_{\text{top}}$ to its complement as $\mathcal{C}_{\text{bottom}} = [K]\backslash\mathcal{C}_{\text{top}}$ . Although this is not guaranteed, whenever this partition emerges the components $\mathcal{C}_{\text{bottom}}$ and $\mathcal{C}_{\text{top}}$ will be separated by a "consecutive" gap, $\min_{i\in\mathcal{C}_{\text{top}}} \mu_i - \max_{j\in\mathcal{C}_{\text{bottom}}} \mu_j$, roughly proportional to $\max_i \Delta_{\sigma_i,\sigma_{i+1}}$.

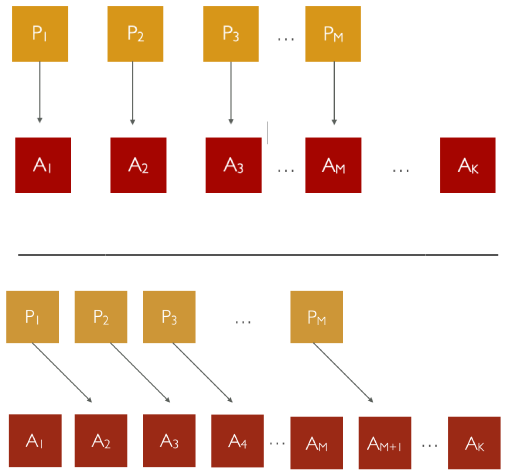


Figure 1: Illustration of the Round Robin Schedule. **Top:** Step 1. **Bottom:** Step 2.

By definition of the connectivity graph, when $\mathcal{E}$ holds all arms in $\mathcal{C}_{\text{top}}$ will have higher mean reward values than those of each of the arms in $\mathcal{C}_{\text{bottom}}$. Our algorithm is designed to ensure that $t^1_{\text{first}}$ does not occur too soon or too late. More specifically, we guarantee that up to logarithmic factors, $\frac{t^1_{\text{first}}}{K} \approx \frac{1}{\max_i \Delta^2_{\sigma_i,\sigma_{i+1}}}$,thus ensuring that the signals $t^p_{\text{first}}$ occur around the same time for all players. This is crucial to allow for successful communication between the players as this ensures player 1 start communicating after all remaining players are ready for listening. Once $t^1_{\text{first}}$ is triggered and player 1 has hold of a connectivity graph with at least two connected components, she will make use of the low reward experienced by collisions to transmit the partition $(\mathcal{C}_{\text{top}}, \mathcal{C}_{\text{bottom}})$ of the arm space to the other players. This can be achieved by communicating $\mathcal{C}_{\text{top}}$ to all players $p \neq 1$ using a bit string language agreed in advance. The main algorithmic challenge we face in designing this protocol is to ensure the collisions incurred during communication do not generate a substantial regret.

Once all other players have learned what arms belong to $\mathcal{C}_{\text{top}}$, they will RECURSE by playing the same strategy as before but now restricted to the two subproblems induced by the set of arms $\mathcal{C}_{\text{top}}$ and $\mathcal{C}_{\text{bottom}}$. If $|\mathcal{C}_{\text{top}}| < M$, the top indexed $|\mathcal{C}_{\text{top}}|$ players will forevermore play following a Round Robin schedule over $\mathcal{C}_{\text{top}}$ while the remaining $M - |\mathcal{C}_{\text{top}}|$ players restart the exploration strategy over the remaining arms. If instead $|\mathcal{C}_{\text{top}}| \geq M$ the $M$ players restart the exploration strategy over the set $\mathcal{C}_{\text{top}}$. RECURSE reduces the problem to a smaller Cooperative Multi-Player Multi-Armed bandit.

The players then re-index the arm labels[1], re-computes $t_{\text{collision}-\text{test}}$ (for simplicity of analysis) and starts playing in a Round Robin fashion within their assigned group, with the smallest indexed player in any group becoming the communicating player. Iterating over this procedure ensures the players converge to pull only the top $M$ arms.

Communicating $\mathcal{C}_{\text{top}}$ (and therefore $\mathcal{C}_{\text{bottom}}$) can be done by sending a bit string of size $K$, where the bits corresponding to the arms in $\mathcal{C}_{\text{top}}$ equal one and the remaining bits equal zero. Since the players have access to no signal other than the reward, and this can be modulated only when two players play the same arm, any communication between players needs to happen through collisions. With this in mind we design a communication mechanism that allows player 1 to transmit a bit string to all the other players with high probability.

While all players $p \in \{2, \cdots, K\}$ continue playing in a Round Robin fashion, player 1 will signal the start of the communication sequence at the second *special* round $t^1_{\text{comm1}}$ such that $\left\lfloor \frac{t^1_{\text{comm1}}/K}{g(t^1_{\text{comm1}}/K)} \right\rfloor$ is a power of nine and occurs after $t^1_{\text{first}}$. At this time player 1 will begin to pull arm $\widehat{\sigma}_1$—the arm with the largest empirical mean at time $t^1_{\text{comm1}}$—for a number of rounds equal to $Kt_{\text{collision}-\text{test}}$. All other players $p \in \{2, \cdots, M\}$ will begin listening for the start of the communication signal from player 1 at all *special* rounds[2] $t^p_{\text{listen}}$ such that $\left\lfloor \frac{t^p_{\text{listen}}/K}{g(t^p_{\text{listen}}/K)} \right\rfloor$ is a power of nine and occurs after $t^p_{\text{first}}$.

Having computed empirical estimators of the arm's rewards using data up to time $t^p_{\text{listen}}$, each of the players $p \in \{2, \cdots, M\}$ will have estimated the mean of $\widehat{\sigma}_1$ to sufficiently good accuracy to ensure that at the end of the next $Kt_{\text{collision}-\text{test}}$ rounds any of them can detect if there have been collisions with player 1 when pulling arm $\widehat{\sigma}_1$ during rounds $t^p_{\text{listen}} + 1$ to $t^p_{\text{listen}} + Kt_{\text{collision}-\text{test}}$. If player $p$ detects a small reward coming from an arm with a previously recorded high reward, it can conclude this arm is $\widehat{\sigma}_1$ and that player 1 has pulled it to signal the start of a communication round. If instead, none of the high reward arms record a substantial deviation in their collected reward during rounds $t^p_{\text{listen}} + 1$ to $t^p_{\text{listen}} + Kt_{\text{collision}-\text{test}}$, player $p$ can conclude player 1 has not started to communicate yet. By design our algorithm ensures that with high probability $t^p_{\text{listen}} \leq t^1_{\text{comm1}}$ and that after at most three trials $t^p_{\text{listen}} = t^1_{\text{comm1}}$. One of the main challenges in designing an algorithm with these properties is to ensure the listening players are able to listen for the incoming communication signal from the communicating player at the right time. The players $p \neq 1$ can only start listening for an incoming signal after they have collected enough samples to get an accurate enough estimator for $\widehat{\sigma}_1$. Since the time when this happens is a random variable we need to ensure both listening and communication protocols occur at times when all players have sufficiently accurate estimates. This is particularly challenging because the players are only aware of their own estimates.

Let's assume that $t^1_{\text{listen}} = t^1_{\text{comm1}}$. If player $p \in \{2, \cdots, M\}$ detects a communication signal from player 1 associated with arm $\widehat{\sigma}_1$, it will listen for the next $K^2 t_{\text{collision}-\text{test}}$ rounds (recall these "listening" players are still playing all arms in $[K]$ following a Round Robin schedule). Using the resulting $Kt_{\text{collision}-\text{test}}$ pulls of arm $\widehat{\sigma}_1$ player $p$ can decode the $K$ bit message sent by player 1. With the same test used to detect the start of communication, if the $i$-th block of $t_{\text{collision}-\text{test}}$ pulls of arm $\widehat{\sigma}_1$ has a low reward, player $p$ can conclude the bit value sent by player 1 is a one. If the average reward is large, player $p$ can conclude the bit value sent by player 1 is a zero.

1. For example the players assigned to $\mathcal{C}_1$ will re-index these arms so they are labeled 1 to $|\mathcal{C}_1|$ by switching the smallest label in $\mathcal{C}_1$ to a 1, the second smallest to a 2, and so on.
2. We impose the restriction that $t^p_{\text{listen}}$ is such that $\left\lfloor \frac{t^p_{\text{listen}}/K-1}{g(t^p_{\text{listen}}/K-1)} \right\rfloor$ is not a power of nine. Similarly for $t^1_{\text{comm1}}$.

The regret accrued by all players until the successful communication of $\mathcal{C}_{\text{top}}$ to all players can be decomposed in two parts: the Round Robin regret ($\text{RoundRobinRegret}([K])$) and the collision regret ($\text{CollisionRegret}([K])$).

Recall that $\frac{t^1_{\text{comm1}}}{K} \approx \frac{1}{\max_i \Delta^2_{\sigma_i,\sigma_{i+1}}}$ and observe that $\mu_{\sigma_1} - \mu_{\sigma_K} \leq K \max_i \Delta_{\sigma_i,\sigma_{i+1}}$. During a full Round Robin cycle over arms $\{1, \cdots, K\}$ regret is only incurred when arms in $\{\mu_{\sigma_i}\}_{i=M+1}^K$ are played. Each of these pulls may incur up to $K \max_i \Delta_{\sigma_i,\sigma_{i+1}}$ regret. Since there are $M(K-M)$ of these pulls, the $\text{RoundRobinRegret}([K])$ accrued by the $M$ players during the Round Robin plays[3] is of the order at most $\frac{KM(K-M)}{\max_i \Delta_{\sigma_i,\sigma_{i+1}}}$.

Now let's see what happens with the $\text{CollisionRegret}([K])$. Since $t_{\text{collision}-\text{test}} \approx \frac{\log(t/\delta)}{\Delta^2_{\text{collision}}}$ and the number of collisions experienced by player 1 during communication is upper bounded by $KMt_{\text{collision}-\text{test}}$ the $\text{CollisionRegret}([K])$ is of upper bounded by $\frac{KM\log(t^1_{\text{comm1}}/\delta)}{\Delta_{\text{collision}}}$ (notice this is of smaller order than $\text{RoundRobinRegret}([K])$ since $\Delta_{\text{collision}} \geq \max_i \Delta_{\sigma_i,\sigma_{i+1}}$ ).

The main challenges in our analysis are the following:

1. $t_{\text{collision}-\text{test}}$ and times $t^1_{\text{first}}$ and $t^1_{\text{comm1}}$ are random and thus unknown to players $2, \cdots, M$. We need to ensure that times $t^p_{\text{first}}$ occur at around the same time for all $p \in [M]$ in order to ensure players $p \in \{2, \cdots, M\}$ start "listening" for a potential communication start signal from player 1 at the right time. We do so by designing a mechanism that ensures $\frac{t^p_{\text{first}}}{g(t^p_{\text{first}})}$ is upper and lower bounded by a constant multiple of $\frac{1}{\max_i \Delta^2_{\sigma_i,\sigma_{i+1}}}$. This is the same mechanism we use in the subroutine dedicated to the estimation of $\Delta_{\text{collision}}$.

2. Since communication occurs via collisions, the regret incurred by player 1 (and any player colliding with it) may be linear in $\Delta_{\text{collision}}$ whenever these happen. We need to ensure the time needed for communicating and therefore the number of collisions involved is small, while still being sufficiently large to convey enough information.

3. As we have mentioned above, the start-communication signal is sent out by player 1 at a round such that $\left\lfloor \frac{t^1_{\text{comm1}}}{g(t^1_{\text{comm1}})} \right\rfloor$ is a power of nine. The reasoning behind this is to ensure the listening players $p \in \{2, \cdots, M\}$ are able to start listening at a recognizable time index. Since the times $t^p_{\text{first}}$ and $t^1_{\text{first}}$ are not equal, and all players $p \in \{2, \cdots, M\}$ only start listening after $t^p_{\text{first}}$ we need to ensure that $t^1_{\text{comm1}} \geq t^p_{\text{listen}}$ for all $p \in \{2, \cdots, M\}$.

We address each of these three challenges in the sections below.

### 4.1. Player 1 Communication Protocol

Let $I^p_i(t, \tilde{C}) = [\widehat{\mu}^p_i(t) - \tilde{C}D(N^p_i(t)), \widehat{\mu}^p_i(t) + \tilde{C}D(N^p_i(t))]$ be the $\tilde{C}-$blowup confidence interval for player $p$ at round $t$ around the mean reward of arm $i$. If $\tilde{C} \geq 1$, these confidence intervals are satisfied (i.e., $\mu_i \in I^p_i(t, \tilde{C})$) for all $t \in \mathbb{N}, i \in [K], p \in [M]$ whenever $\mathcal{E}$ holds, an event that happens with probability at least $1 - \delta$. We now introduce the empirical arm connectivity graph with blowup parameter $\tilde{C}$.

3. This is accounting for the regret collected during the time it takes for a single transmission of a partition $(\mathcal{C}_{\text{top}}, \mathcal{C}_{\text{bottom}})$. Since there could be up to $K-1$ such rounds, the algorithm's regret has an extra scaling with $K$ as in Theorem 2.

**Definition 5 ($\tilde{C}$-blowup Arm connectivity graph)** *Let $\tilde{C} \geq 1$. For each player we define the (random) $\tilde{C}-$blowup arm connectivity graph as $\mathcal{G}_t^p(\tilde{\mathcal{C}}) = ([K], E_t^p(\tilde{\mathcal{C}}))$ with node set $[K]$ and edge set $E_t^p(\tilde{\mathcal{C}})$ defined as:*

$$\{i,j\} \in E_t^p(\tilde{\mathcal{C}}), \quad \text{if } I_i^p(t,\tilde{C}) \cap I_j^p(t,\tilde{C}) \neq \emptyset.$$

Graph $\mathcal{G}_t^p(\tilde{\mathcal{C}})$ is a collection of connected components. The nodes $i, j \in \mathcal{G}_t^p(\tilde{\mathcal{C}})$ represent arms $i, j$ and are connected by an edge in $\mathcal{G}_t^p(\tilde{\mathcal{C}})$ if their $\tilde{\mathcal{C}}$-blowup confidence intervals overlap. If we identify each node $i$ of $\mathcal{G}_t^p(\tilde{\mathcal{C}})$ with the empirical mean $\widehat{\mu}_i^p(t)$, the graph has a natural geometric representation as a collection of intervals in $[0,1]$ with each connected interval in the collection representing a connected component of $\mathcal{G}_t^p(\tilde{\mathcal{C}})$. We say that two connected components of $\mathcal{G}_t^p(\tilde{\mathcal{C}})$ are *adjacent* if they are consecutive intervals in this geometric representation.

Let $\mathbf{conn}^p(t,\tilde{\mathcal{C}})$ be the number of connected components of $\mathcal{G}_t^p(\tilde{\mathcal{C}})$. Let $\{\mathcal{C}_j^p(t,\tilde{\mathcal{C}})\}_{j=1}^{\mathbf{conn}^p(t,\tilde{\mathcal{C}})}$ be the collection of connected components at time $t$, ordered by adjacency in the geometric representation of $\mathcal{G}_t^p(\tilde{\mathcal{C}})$, with the empirical mean values of $\mathcal{C}_1^p(t,\tilde{\mathcal{C}})$ being the connected component with the largest empirical mean values among all connected components in $\{\mathcal{C}_j^p(t,\tilde{\mathcal{C}})\}_{j=1}^{\mathbf{conn}^p(t,\tilde{\mathcal{C}})}$. This is the same as $\mathcal{C}_{\text{top}}$ in the previous discussion. Let's assume all players are playing using a Round Robin Schedule. Denote by $t_{\text{first}}^p$ to the first *special* round of player $p$ when $\mathbf{conn}^p(t_{\text{first}}^p,\tilde{\mathcal{C}}) > 1$. We start by showing that if we set $\tilde{C} = 10$, with high probability the condition $\mathbf{conn}^p\left(sK,\tilde{C}\right) \geq 2$ is triggered for all players $p \in [M]$ at a "special" Round Robin round $t_{\text{first}}^p$ (multiple of $K$) such that,

$$\frac{128}{\max_i \Delta^2_{\sigma_i,\sigma_{i+1}}} \leq \frac{N(t_{\text{first}}^p)}{g(N(t_{\text{first}}^p))} < \frac{1152}{\max_i \Delta^2_{\sigma_i,\sigma_{i+1}}}, \tag{2}$$

where $N(t_{\text{first}}^p) = N_i(t_{\text{first}}^p)$ for all $i \in [K]$ and therefore equal to $\frac{t_{\text{first}}^p}{K}$ (since $t_{\text{first}}^p$ is a special round it is a multiple of $K$). To simplify matters we will use the notation $s_{\text{first}}^p$ to denote the ratios $\frac{t_{\text{first}}^p}{K}$. We will use the same notational convention for all "named" rounds with subscripts such as $\text{first}, \text{comm}, \text{comm1}$,etc.. Let's start by showing that Equation 2 is a direct consequence of the following Lemma, setting $C = 10$. Recall that $D(n) = \sqrt{\frac{2g(n)}{n}}$.

**Lemma 6 (Confidence Bands)** *Let $\widehat{\mu}_{\sigma_i}(t)$ and $\widehat{\mu}_{\sigma_j}(t)$ be empirical estimators $\mu_{\sigma_i}$ and $\mu_{\sigma_j}$, each using $N(t)$ samples. Let $C > 3$ be a constant. If $t$ is the first special round such that*

$$\widehat{\mu}_{\sigma_i}(t) - \widehat{\mu}_{\sigma_j}(t) \geq CD(N(t)), \tag{3}$$

*then, whenever $\mathcal{E}$ holds, we have $\frac{\Delta_{\sigma_i,\sigma_j}}{2(C+2)} < D(N(t)) \leq \frac{\Delta_{\sigma_i,\sigma_j}}{C-2}$ and*

$$\frac{2(C-2)^2}{\Delta^2_{\sigma_i,\sigma_j}} \leq \frac{N(t)}{g(N(t))} < \frac{8(C+2)^2}{\Delta^2_{\sigma_i,\sigma_j}}. \tag{4}$$

The proof of Lemma 6 is in Appendix D.1. Equation 2 can be derived by simply plugging in $C = 10$. With the objective of ensuring all the times $t_{\text{first}}^p$ occur around the same time, let's now show that a simple function of $t_{\text{first}}^p$ is always around the vicinity of a power of 9. A simple number-theoretic implication of Equation 2 is there exists a unique power of nine in the interval $\left[\frac{128}{\max_i \Delta^2_{\sigma_i,\sigma_{i+1}}}, \frac{1152}{\max_i \Delta^2_{\sigma_i,\sigma_{i+1}}}\right)$ (see Lemma 9 in Appendix B.2).

For all $p \in [M]$ consider the first *special* round $t$ immediately following $t^p_{\text{first}}$ such that $\left\lfloor \frac{t/K}{g(t/K)} \right\rfloor$ is a power of nine. Call this round $t^p_{\text{comm}}$. Let $9^u$ be the unique (and player independent) power of nine in $\left[ \frac{128}{\max_i \Delta^2_{\sigma_i,\sigma_{i+1}}}, \frac{1152}{\max_i \Delta^2_{\sigma_i,\sigma_{i+1}}} \right)$. By definition $\left\lfloor \frac{t^p_{\text{comm}}/K}{g(t^p_{\text{comm}}/K)} \right\rfloor \in \{9^u, 9^{u+1}\}$ for all $p \in [M]$ and thus for all players $t^p_{\text{comm}} \in \left\{ \min_{t\in\mathbb{N}} \text{ s.t. } \left\lfloor \frac{t/K}{g(t/K)} \right\rfloor = 9^u, \min_{t\in\mathbb{N}} \text{ s.t. } \left\lfloor \frac{t/K}{g(t/K)} \right\rfloor = 9^{u+1} \right\}$ is one of two values provided $t$ is large enough so that $\frac{t/K}{g(t/K)}$ is an increasing function of $t$.

Instead of initiating communication at round $t^1_{\text{comm}}$, player 1 will wait until $t^1_{\text{comm1}}$ so that $t^1_{\text{comm1}} \in \left\{ \min_{t\in\mathbb{N}} \text{ s.t. } \left\lfloor \frac{t/K}{g(t/K)} \right\rfloor = 9^{u+1}, \min_{t\in\mathbb{N}} \text{ s.t. } \left\lfloor \frac{t/K}{g(t/K)} \right\rfloor = 9^{u+2} \right\}$. This is to ensure that no information-receiving players $p \in \{2, \cdots, M\}$ (all of which will start start listening for a communication signal either at round $\min_{t\in\mathbb{N}}$ s.t. $\left\lfloor \frac{t/K}{g(t/K)} \right\rfloor = 9^u$ or $\min_{t\in\mathbb{N}}$ s.t. $\left\lfloor \frac{t/K}{g(t/K)} \right\rfloor = 9^{u+1}$) will miss the start of player 1's message. The precise description of how player 1 waits for $t^1_{\text{first}}$ and $t^1_{\text{comm1}}$ (see Algorithm 3) and communicate (see Algorithm 4) can be found in Appendix B.2.

**The** ENCODE **function.** As we explained at the beginning of Section 4, communicating the composition of the connected component $\mathcal{C}^1_1(t^1_{\text{first}}, 10)$ can be done by sending a bit string of size $K$, where the bits corresponding to the arms in $\mathcal{C}^1_1(t^1_{\text{first}}, 10)$ equal one and the remaining bits equal zero. We will name this bitstring as $\text{ENCODE}(\mathcal{C}^1_1(t^1_{\text{first}}, 10))$.

In the following section we discuss how the communication protocol makes use of collisions to enable player 1 to transmit $\text{ENCODE}(\mathcal{C}^1_1(t^1_{\text{first}}, 10))$, and how it is that it is possible to do so while incurring a manageable regret.

### 4.2. Communication Analysis

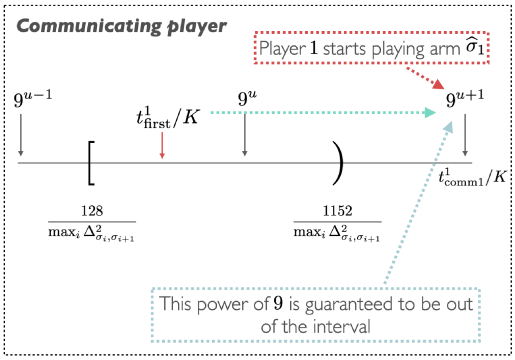


Figure 2: Illustration of the Communicating Player Strategy (up to log factors).

The main idea behind our communication algorithm is to make use of the small reward incurred by colliding players to enable communication between players. To gain some intuition for how our communication mechanism works, let's consider a simplified objective. Let's assume player 1 will be tasked with sending a bit $b \in \{0, 1\}$, and every other player $p \in \{2, \cdots, M\}$ will be tasked with figuring out the value of $b$. Just as in the preceding discussion imagine that initially the players pull arms $[K]$ following a Round Robin schedule, and that player 1 starts to communicate a bit at time $t^1_{\text{start}}$. Let's also assume in this discussion[4] that all players $p$ have knowledge of $t^1_{\text{start}}$. In Algorithm 3 time $t^1_{\text{start}}$ equals $t^1_{\text{comm1}}$. In the following section we will show that for all $p \in [M]$ with high probability $t^p_{\text{first}} \leq t^1_{\text{comm1}}$ therefore in what follows of this section we will assume $t^p_{\text{first}} \leq t^1_{\text{start}}$.

4. In the discussion that follows we will instantiate $t^1_{\text{start}}$ to different rounds including the rounds where each of the players $p \in \{2, \cdots, M\}$ use to listen for a "start of communication" signal from player 1 (denoted as $\{t^p_{\text{listen}}\}$) as well as the different message boundary times of $t^1_{\text{comm1}} + (j-1)Kt_{\text{collision-test}} + 1$ for all $j \in \{2, \cdots, |\alpha| + 1\}$ whenever player 1 is sending a message of size $\alpha$.

Let $\widehat{\sigma}_1$ be player 1's guess for the optimal arm at time $t^1_{\text{first}}$, i.e., the arm with the largest empirical mean. Let's consider the following communication algorithm. If $b = 1$, starting from round $t^1_{\text{start}} + 1$ and until round $t^1_{\text{start}} + Kt_{\text{collision}-\text{test}}$, player 1 will pull arm $\widehat{\sigma}_1$. If instead $b = 0$, player 1 will continue playing in a Round Robin fashion from round $t^1_{\text{start}} + 1$ until round $t^1_{\text{start}} + Kt_{\text{collision}-\text{test}}$, thus avoiding collisions. If $b = 1$ the average expected reward collected by players $p \in \{2, \cdots, M\}$ is $\mu_{\text{collision}}$ when pulling arm $\widehat{\sigma}_1$ during rounds $t^1_{\text{start}} + 1$ to $t^1_{\text{start}} + Kt_{\text{collision}-\text{test}}$. If $b = 0$ the average reward of pulling arm $\widehat{\sigma}_1$ during rounds $t^1_{\text{start}} + 1$ to $t^1_{\text{start}} + Kt_{\text{collision}-\text{test}}$ is $\mu_{\widehat{\sigma}_1}$. Since $t^1_{\text{start}} \geq t^1_{\text{first}}$ and $t^1_{\text{first}}$ satisfies the boundary conditions, Equation 2 implies that with high probability,

$$\frac{N^p_{\sigma_1}(t^1_{\text{start}})}{g(N^p_{\sigma_1}(t^1_{\text{start}}))} \geq \frac{128}{\max_i \Delta^2_{\sigma_i,\sigma_{i+1}}} \geq \frac{128}{\mu^2_{\sigma_1}}. \tag{5}$$

Since for all special rounds $t$ the number of pulls is $N^p_i(t) \approx t/K$, Equation 5 implies that with high probability at time $t^1_{\text{start}}$ player 1 has pulled each arm $\Omega\left(\frac{1}{\mu^2_{\sigma_1}}\right)$ times up to logarithmic factors. We will now show that the following three statements hold with high probability,

1. **Arm $\widehat{\sigma}_1$ has a large empirical mean for all players.**
   - Arm $\widehat{\sigma}_1 \in \{i \in [K] \text{ s.t. } \widehat{\mu}^p_i(t^1_{\text{start}}) - \widehat{\mu}^p_{\text{collision}} \geq \frac{1}{2}\max_{j\in[K]}\left(\widehat{\mu}^p_j(t^1_{\text{start}}) - \widehat{\mu}^p_{\text{collision}}\right)\}$ for all $p \in \{2, \cdots, M\}$ and $\widehat{\mu}^p_{\widehat{\sigma}_1}(t^1_{\text{start}}) - D(N_{\widehat{\sigma}_1}(t^1_{\text{start}})) \geq \frac{\mu_{\sigma_1}-\mu_{\text{collision}}}{2} + \mu_{\text{collision}}$.
2. **Arm $\widehat{\sigma}_1$ is comparable to $\sigma_1$.**
   - The witnesses $L^p_{\widehat{\sigma}_1}(t^1_{\text{start}})$ satisfy, $L^p_{\widehat{\sigma}_1}(t^1_{\text{start}}) \in \left[\frac{3\left(\mu_{\widehat{\sigma}_1}-\mu_{\text{collision}}\right)}{7}, \frac{4\left(\mu_{\widehat{\sigma}_1}-\mu_{\text{collision}}\right)}{7}\right] + \mu_{\text{collision}}$ for all $p \in \{2, \cdots, M\}$.
3. **When collisions are avoided the mean estimators $\widehat{\mu}^p_{\widehat{\sigma}_1}(t^1_{\text{start}} + 1 : +Kf(\widehat{\Delta}_{\text{collision}}, t^1_{\text{start}}))$ are far from $\mu_{\text{collision}}$.**
   - If $b = 1$, the estimators $\widehat{\mu}^p_{\widehat{\sigma}_1}(t^1_{\text{start}} + 1 : +Kt_{\text{collision}-\text{test}}) \leq L^p_{\widehat{\sigma}_1}$ for all $p \in [M]$.
   - If $b = 0$, the estimators $\widehat{\mu}^p_{\widehat{\sigma}_1}(t^1_{\text{start}} : t^1_{\text{start}} + Kt_{\text{collision}-\text{test}}) > L^p_{\widehat{\sigma}_1}$ for all $p \in [M]$.

If players $p \in \{2, \cdots, M\}$ have knowledge of $t^1_{\text{start}}$, all they need to do to decode $b$ is to test for all arms with a large empirical mean (as computed by these players up to time $t^1_{\text{start}}$), and compare these values with the empirical means computed during rounds $t^1_{\text{start}} + 1$ through $t^1_{\text{start}} + Kt_{\text{collision}-\text{test}}$, when potential collisions may have taken place. If the two estimators are vastly different, players $p \in \{2, \cdots, M\}$ can conclude that $b$ equals one. If instead these values are "similar," they can conclude that $b$ equals zero. We formalize this idea via the following CollisionTest algorithm. We use the notation $t^p_{\text{test}}$ to denote the rounds when each of the players $p \in \{2, \cdots, M\}$ starts probing for a communication signal from player 1. If $t^1_{\text{start}}$ is common knowledge, the CollisionTest algorithm can be instantiated by setting $t^p_{\text{test}} = t^1_{\text{start}}$.

**Algorithm 1** CollisionTest (player $p \neq 1$)

**Input** $t^p_{\text{test}}$, witnesses $\{L^p_i(t^p_{\text{test}})\}_{i\in[K]}$, empirical means $\{\widehat{\mu}^p_i(t^p_{\text{test}})\}_{i\in[K]}$, communication round empirical means $\{\widehat{\mu}^p_i(t^p_{\text{test1}}+1 : t^p_{\text{test1}}+Kt_{\text{collision-test}}\}_{i\in[K]}$

$\widehat{\mathbf{MaxArms}} \leftarrow \left\{ i \in [K] \text{ s.t. } \widehat{\mu}^p_i(t^p_{\text{test}}) - \widehat{\mu}^p_{\text{collision}} \geq \frac{1}{2}\max_{j\in[K]} \left(\widehat{\mu}^p_j(t^p_{\text{test}}) - \widehat{\mu}^p_{\text{collision}}\right)\right\}$

**if** $\exists i \in \widehat{\mathbf{MaxArms}}$ *s.t.* $\widehat{\mu}^p_i(t^p_{\text{test1}}+1 : t^p_{\text{test1}}+Kt_{\text{collision-test}} < L^p_i(t^p_{\text{test}})$ **then**

  **Return** 1

**end**

**Return** 0

Notice that in contrast with the discussion that preceded it CollisionTest allows the indices $t^p_{\text{test}}$ and $t^p_{\text{test1}}$ to be different but the length of the communication rounds remains $Kt_{\text{collision-test}}$. Our first result is to show this procedure allows players $p \in \{2, \cdots, M\}$ to recover $b$ with high probability.

**Lemma 7 (One Bit Recovery)** *Let* $A^p = \{L^p_i(t^1_{\text{start}})\}_{i\in[K]}$, $B^p = \{\widehat{\mu}^p_i(t^1_{\text{start}})\}_{i\in[K]}$, $C^p = \{\widehat{\mu}^p_i(t^1_{\text{start}}+1 : t^1_{\text{start}}+Kt_{\text{collision-test}})\}_{i\in[K]}$ *for all* $p$. *If the good event* $\mathcal{E}$ *holds then, with probability at least* $1-\frac{\delta}{4K^2}$, *all players* $p \in \{2, \cdots, M\}$ *will be able to recover exactly the bit transmitted by player* 1 *by calling* CollisionTest$(A^p, B^p, C^p)$.

The complete proof of Lemma 12 and an in-depth discussion on the CollisionTest concept can be found in Appendices B.3 and C.4. The logic behind why the CollisionTest works is as follows. At time $t^p_{\text{test}}$ (in our case equal to $t^1_{\text{comm1}}$ all players have access to a constant accuracy estimator for $\max_i \Delta_{\sigma_i,\sigma_{i+1}}$ (i.e. the empirical gap between the arms at the boundary of the two connected components of $\mathcal{C}^1_1(t^1_{\text{first}}, 10)$). Therefore they can estimate $\mu_{\sigma_1}$ up to a $\Delta_{\text{collision}}$-accuracy. By Hoeffding inequality testing at an accuracy of $c\Delta_{\text{collision}}$ the mean of an arm $\widehat{\sigma}_1$ satisfying $\mu_{\widehat{\sigma}_1} > \mu_{\text{collision}} + c'(\mu_{\sigma_1} - \mu_{\text{collision}})$ with $c' > c$ only requires $\widetilde{\mathcal{O}}\left(\frac{1}{\Delta^2_{\text{collision}}}\right)$ samples (up to log factors and each with regret at most $\Delta_{\text{collision}}$), thus incurring regret of only $\mathcal{O}(1/\Delta_{\text{collision}})$ (up to log factors).

### 4.3. The Listening Players

We have all the necessary pieces in place to spell out the details of the listening protocol used by players $\{2, \cdots, M\}$. The listening players shall wait until the first special round such that $\mathbf{conn}^p(sK, 10) \geq 2$. After this round has passed the players will start listening for an incoming signal from player 1 during subsequent special rounds $t = Ks$ such that $\left\lfloor \frac{s}{g(s)} \right\rfloor$ is a power of nine. The precise description of how to initialize the listening protocol (see Algorithm 6) can be found in Appendix B.4. If the player detects a signal, she will start listening for a size $K$ bit string using the DECODE function (see Algorithm 7 in Appendix B.4 for a detailed explanation). DECODE consists of $K$ consecutive CollisionTest calls.

Combining these communication and listening protocols for player 1 and players $\{2, \cdots, M\}$ respectively we can guarantee that with high probability the value of $t^p_{\text{listen}}$ passed down to the DECODE function satisfies $t^p_{\text{listen}} = t^1_{\text{comm1}}$ and that the listening players $\{2, \cdots, M\}$ will be able to decode the message handed down by player 1.

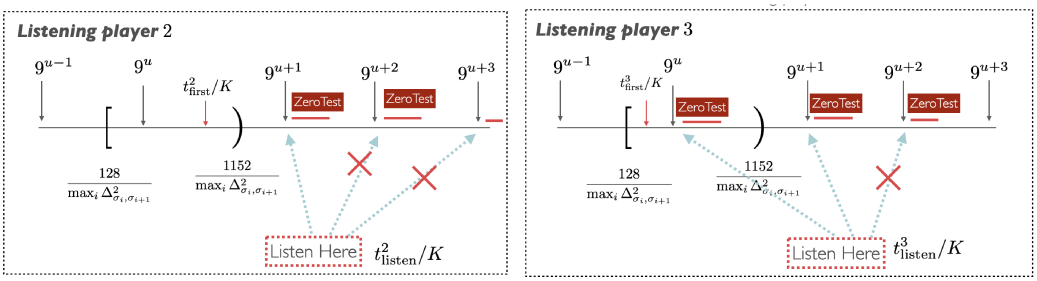


Figure 3: Illustration of the Listening Player Strategy (up to log factors) depending on what side $\frac{t^p_{\text{first}}/K}{g(t^p_{\text{first}}/K)}$ lies with respect to the power of nine in $\left[\frac{128}{\max_i \Delta^2_{\sigma_i,\sigma_{i+1}}}, \frac{1152}{\max_i \Delta^2_{\sigma_i,\sigma_{i+1}}}\right)$.

**Lemma 8 (Message Recovery)** *If $\mathcal{E}$ holds then with probability at least $1 - \frac{\delta}{K}$ for all $p \in [M]$ the value of $t^p_{\text{listen}}$ sent to the* DECODE *function of Algorithm 7 satisfies $t^p_{\text{listen}} = t^1_{\text{comm1}}$ and* DECODE *will recover the exact $K-$bit message sent by player 1.*

The proof of Lemma 8 is in Appendix B.4. The discussion is complemented by Appendix C.3.

### 4.4. Bounding the Regret

To finalize our regret analysis we first bound the regret incurred by the players during the first communication round and player 1 has communicated to all others the composition of $\mathcal{C}^1_1(t^1_{\text{first}}, 10)$. We denote this quantity as $\text{FirstPartitionRegret}([K],[M])$. We proceed to bound this quantity by splitting regret in two, the $\text{CollisionRegret}([K],[M])$ incurred by the players during communication induced by player 1 pulling arm $\widehat{\sigma}_1$ and the $\text{RoundRobinRegret}([K],[M])$ incurred by the players during the steps these followed a Round Robin schedule.

The length of each bit communication round is of order $t_{\text{collision}-\text{test}} \approx \frac{\log(1/\delta)}{\Delta^2_{\text{collision}}}$. During this communication round the number of collisions between player 1 and any other player is at most $(K+1) \times t_{\text{collision}-\text{test}}$ because the total number of bits required to communicate $\text{ENCODE}(\mathcal{C}^1_1(t^1_{\text{first}}, 10))$ equals $K+1$, one ON bit to signal the start of communication and $K$ to transmit $\mathcal{C}^1_1(t^1_{\text{first}}, 10)$. Since each collision incurs in at most regret $\Delta_{\text{collision}}$, $\text{CollisionRegret}([K],[M]) \leq \widetilde{\mathcal{O}}\left(\frac{KM \log(t^1_{\text{comm1}}/\delta)}{\Delta_{\text{collision}}}\right)$. The precise statement of this bound is in Corollary 20 in Section B.5.1.

To bound the $\text{RoundRobinRegret}([K],[M])$ note the regret during a single Round Robin round (i.e. $K$ steps of all $M$ players) is upper bounded by $M(K-M)K \max_i \Delta_{\sigma_i,\sigma_{i+1}}$. The total duration of the whole communication protocol for $\mathcal{C}^1_1(t^1_{\text{first}}, 10)$ is of the order of $\frac{\log(t^1_{\text{first}}/\delta)}{\max_i \Delta^2_{\sigma_i,\sigma_{i+1}}}$ and therefore $\text{RoundRobinRegret}([K],[M]) \leq \widetilde{\mathcal{O}}\left(\frac{M(K-M)K \log(t^1_{\text{first}}/\delta)}{\max_i \Delta_{\sigma_i,\sigma_{i+1}}}\right)$. The precise statement of this bound can be found in Corollary 21 in Section B.5.1. Combining these results and using $\Delta_{\text{collision}} \geq \max_i \Delta_{\sigma_i,\sigma_{i+1}}$ we conclude the regret to communicate $\mathcal{C}^1_1(t^1_{\text{first}}, 10)$ is $\widetilde{\mathcal{O}}\left(\frac{M(K-M)K \log(t^1_{\text{first}}/\delta)}{\max_i \Delta_{\sigma_i,\sigma_{i+1}}}\right)$.

We can now put together the whole algorithm and prove Theorem 2. As we explained at the beginning of Section 4, after communication of $\mathcal{C}^1_1(t^1_{\text{first}}, 10)$ and all players have learned what arms belong to $\mathcal{C}^1_1(t^1_{\text{first}}, 10)$, they will split themselves into one or two groups. If $|\mathcal{C}^1_1(t^1_{\text{first}}, 10)| \geq M$, from then on all players will pull arms from $\mathcal{C}^1_1(t^1_{\text{first}}, 10)$ exclusively. If $|\mathcal{C}^1_1(t^1_{\text{first}}, 10)| < M$, then

players 1 to $|\mathcal{C}_1^1(t^1_{\text{first}}, 10)|$ will pull arms from $\mathcal{C}_1^1(t^1_{\text{first}}, 10)$ following a Round Robin schedule thereafter (these players have identified a set of arms to exploit) while the remaining players will play arms from $[K] \backslash \mathcal{C}_1^1(t^1_{\text{first}}, 10)$ following the communication protocol we have outlined in the previous section but now restricted to a set of arms of size smaller than $M$. In both cases, the problem has been reduced to a smaller Cooperative Multi-Player Multi-Armed problem. The players then re-index the arm labels[5] and start playing following a Round Robin schedule within their assigned group and the smallest indexed player in any group will become the communicating player. We call the function that splits and iterates over smaller Cooperative Multi Armed Multi Player problems the RECURSE function. It is defined formally in Algorithm 8 in Appendix B.5.2.

The RECURSE function will converge to a steady state where the players are only pulling the top $M$ arms as soon as $\Delta_{\sigma_M, \sigma_{M+1}}$ is the basis of the communicated arm partition.

It can be shown the gaps recursed over are always in (roughly) decreasing order. Thus, each call to RECURSE will incur regret upper bounded by $\widetilde{\mathcal{O}}\left(\frac{M(K-M)K\log\left(t^1_{\text{first}}/\delta\right)}{\max_i \Delta_{\sigma_M, \sigma_{M+1}}}\right)$. Finally, since there can be at most $K-1$ calls to the RECURSE function, setting $\delta = \frac{1}{T}$ we conclude,

$$\mathcal{R}_T \leq \widetilde{\mathcal{O}}\left(\frac{M(K-M)K^2\log(T)}{\Delta_{\sigma_M, \sigma_{M+1}}}\right),$$

with probability at least $1-T$. This finalizes the proof of Theorem 2. The detailed explanation of each of these steps can be found in Appendices B.5 and C.3. The discussion on how to compute and agree on $t_{\text{collison}-\text{test}}$ for all players can also be found in Appendix C.3 and makes use of the same synchronization and communication ideas we have explained here. The regret incurred during that round is of order $\mathcal{O}\left(\frac{M(K-M)K^2\log(T)}{\Delta_{\text{collision}}}\right)$.

## 5. Conclusion

We have proposed a series of algorithms for the Multi-Player Multi-Armed bandit problem. We achieve a regret guarantee logarithmic in the number of rounds and inversely proportional to the sub-optimality gap. In contrast with previous work, we make no assumptions regarding the player's knowledge about the nature of the reward vector (such as assuming a known lower bound for the minimum reward value) and even the collision reward. This paper finally solves the no-sensing multi-player multi-armed bandit problem in its entirety when collisions are allowed. We believe the techniques we have introduced in this work (including for example the Bernstein ZeroTest for the zero collision reward setting discussed in Appendix B) can be used to prove sharp instance dependent guarantees in many decentralized bandit problems such as decentralized matching markets (see Liu et al. (2021)). We hope future research is also spent on simplifying our algorithmic implementations.

## References

Pragnya Alatur, Kfir Y Levy, and Andreas Krause. Multi-player bandits: The adversarial case. *Journal of Machine Learning Research*, 21:77, 2020.

5. For example the players assigned to $\mathcal{C}_1$ will re-index these arms so they are labeled 1 to $|\mathcal{C}_1|$ by switching the smallest label in $\mathcal{C}_1$ to a 1, the second smallest to a 2, and so on.

Animashree Anandkumar, Nithin Michael, Ao Kevin Tang, and Ananthram Swami. Distributed algorithms for learning and cognitive medium access with logarithmic regret. *IEEE Journal on Selected Areas in Communications*, 29(4):731–745, 2011.

Venkatachalam Anantharam, Pravin Varaiya, and Jean Walrand. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part i: Iid rewards. *IEEE Transactions on Automatic Control*, 32(11):968–976, 1987.

Orly Avner and Shie Mannor. Concurrent bandits and cognitive radio networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 66–81. Springer, 2014.

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

Etienne Boursier and Vianney Perchet. SIC-MMAB: Synchronisation involves communication in multiplayer multi-armed bandits. *arXiv preprint arXiv:1809.08151*, 2018.

Sébastien Bubeck and Thomas Budzinski. Coordination without communication: optimal regret in two players multi-armed bandits. In *Conference on Learning Theory*, pages 916–939. PMLR, 2020.

Sébastien Bubeck, Thomas Budzinski, and Mark Sellke. Cooperative and stochastic multi-player multi-armed bandit: Optimal regret with neither communication nor collisions. *arXiv preprint arXiv:2011.03896*, 2020a.

Sébastien Bubeck, Yuanzhi Li, Yuval Peres, and Mark Sellke. Non-stochastic multi-player multi-armed bandits: Optimal rate with collision information, sublinear without. In *Conference on Learning Theory*, pages 961–987. PMLR, 2020b.

Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Uniform, nonparametric, non-asymptotic confidence sequences. *arXiv preprint arXiv:1810.08240*, 2018.

Wei Huang, Richard Combes, and Cindy Trinh. Towards optimal algorithms for multi-player bandits without collision sensing information. *arXiv preprint arXiv:2103.13059*, 2021.

Lifeng Lai, Hai Jiang, and H Vincent Poor. Medium access in cognitive radio networks: A competitive multi-armed bandit framework. In *2008 42nd Asilomar Conference on Signals, Systems and Computers*, pages 98–102. IEEE, 2008.

Allen Liu and Mark Sellke. The pareto frontier of instance-dependent guarantees in multi-player multi-armed bandits with no communication. *arXiv preprint arXiv:2202.09653*, 2022.

Keqin Liu and Qing Zhao. Distributed learning in multi-armed bandit with multiple players. *IEEE Transactions on Signal Processing*, 58(11):5667–5681, 2010.

Lydia T Liu, Feng Ruan, Horia Mania, and Michael I Jordan. Bandit learning in decentralized matching markets. *Journal of Machine Learning Research*, 22(211):1–34, 2021.

Gábor Lugosi and Abbas Mehrabian. Multiplayer bandits without observing collision information. *arXiv preprint arXiv:1808.08416*, 2018.

Jacques Palicot. Multi-armed bandit learning in IoT networks: Learning helps even in non-stationary settings. In *Cognitive Radio Oriented Wireless Networks: Proceedings of the 12th International Conference*, volume 228, page 173. Springer, 2018.

Jonathan Rosenski, Ohad Shamir, and Liran Szlak. Multi-player bandits–a musical chairs approach. In *International Conference on Machine Learning*, pages 155–163. PMLR, 2016.

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

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

## Contents

## Appendix A. Guide to Appendix

The Appendix will be split in different sections. In Section B we include a much more detailed discussion of the different algorithmic components sketched out in the main paper with an emphasis in the zero collision reward case. In Section C we extend our results to a variety of settings more general than those presented in the main. The remaining Appendix sections contain missing proofs and supporting technical results.

## Appendix B. Detailed Discussion of the zero collision reward setting

In this section we will flesh out the details of the different algorithmic components to derive our instance dependent rates for the Multi-Player, Multi-Armed bandit problem with communication through collisions. Throughout the proofs we will make use of the following "Boundary conditions". In this section the following function will be useful,

- **Zero Collision Reward Length of a communication round.** $f : \mathbb{N} \to \mathbb{R}_+$ such that $f(n)$ is the first integer such that $\frac{f(n)}{B\left(f(n), \frac{\delta}{4K^2M}\right)} \geq 24\sqrt{\frac{n/K}{2g(n/K)}}$ where $B(n, \delta') = 2\log\log(2n) + \log\frac{5.2}{\delta'}$.

**Boundary Conditions** We will consider the following four boundary conditions.

1. Let $t_{\text{boundary1}}$ be such that $\frac{f(n)-1}{B(f(n)-1, \frac{\delta}{4K^2M})} \geq 1$ for all $n \geq t_{\text{boundary1}}$.
2. Let $s_{\text{boundary2}}$ be such that $\frac{\partial D(s)}{\partial s} \leq 0$ for all $s \geq s_{\text{boundary2}}$ ($s_{\text{boundary2}}$ can also be defined in terms of $g$ as $\frac{\partial g(s)}{\partial s} \geq 0$ for all $s \geq s_{\text{boundary2}}$).
3. Assume $\delta \leq \frac{1}{162}$ so that $4s^2_{\text{boundary2}}MKL/\delta \geq 162$.
4. Let $t_{\text{boundary3}}$ be such that $f(n) \leq n/K$ for all $n \geq t_{\text{boundary3}}$.

Let $t_{\text{firstBoundary}}$ be the first special round (i.e. $t_{\text{firstBoundary}}$ is a multiple of $K$) such that

$$t_{\text{firstBoundary}} \geq \max\left(t_{\text{boundary1}}, K s_{\text{boundary2}}, t_{\text{boundary3}}\right).$$

The analysis sketch for the regret rates achieved by our algorithm presented in Section 4 is only satisfied for timesteps larger than $t_{\text{firstBoundary}}$, these 'boundary conditions' appear in statements such as

"'By definition $\left\lfloor \frac{t^p_{\text{comm}}/K}{g(t^p_{\text{comm}}/K)} \right\rfloor \in \{9^u, 9^{u+1}\}$ for all $p \in [M]$ and thus for all players $t^p_{\text{comm}} \in \left\{\min_{t\in\mathbb{N}} \text{ s.t. } \left\lfloor \frac{t/K}{g(t/K)} \right\rfloor = 9^u, \min_{t\in\mathbb{N}} \text{ s.t. } \left\lfloor \frac{t/K}{g(t/K)} \right\rfloor = 9^{u+1}\right\}$ is one of two values provided $t$ is large enough so that $\frac{t/K}{g(t/K)}$ is an increasing function of $t$"'

from Section 4.1 (this statement here corresponds to $s_{\text{boundary2}}$. Throughout our detailed analysis of the regret rate of our algorithm presented in this section, we kept track of all the emerging boundary conditions and have compiled them in this list. We then show $t_{\text{firstBoundary}}$ is a function of $\log\left(\frac{1}{\delta}\right)$, $K$ and $M$ and that it depends polynomially on $K$ and $M$ and linearly on $\log(\frac{1}{\delta})$. See Section B.5 for a formal proof. The test used in the zero collision reward setting that forms the basis of our encoding and decoding strategy will be called the ZeroTest.

**Algorithm 2** Zero Test

**Input** Player $p \neq 1$, witnesses $\{L_i^p(t^1_{\text{start}})\}_{i\in[K],p\in[M]}$
**for** *$i$ such that* $\widehat{\mu}_i^p(t^1_{\text{start}}) \geq \frac{1}{2}\max_{j\in[K]} \widehat{\mu}_j^p(t^1_{\text{start}})$ **do**
  **if** $\widehat{\mu}_i^p(t^1_{\text{start}}+1 : t^1_{\text{start}} + Kf(t^1_{\text{start}})) < L_i^p(t^1_{\text{start}})$ **then**
    **Return** 1
  **end**
**end**
**Return** 0

### B.1. Analysis Desiderata

The main challenges in our analysis for the zero collision reward setting are the following:

1. Same as in item 1. of Section 4.
2. Since communication occurs via collisions, the regret incurred by player 1 (and any player colliding with it) may be linear whenever these happen. We need to ensure the time needed for communicating and therefore the number of collisions involved is small, while still being sufficiently large to convey enough information. Aided by Bernstein-style bounds we show the number of collisions needed to transmit information scales linearly with $\frac{1}{\mu_1} \leq \frac{1}{\max_i \Delta_{\sigma_i,\sigma_{i+1}}}$ and not quadratically.
3. Same as item 3. of Section 4.

### B.2. Detailed Discussion and Missing Supporting Results Player 1 Communication Protocol.

The following supporting Lemma allows us to show there exists a unique power of nine in the interval $\left[\frac{128}{\max_i \Delta^2_{\sigma_i,\sigma_{i+1}}}, \frac{1152}{\max_i \Delta^2_{\sigma_i,\sigma_{i+1}}}\right)$,

**Lemma 9** *Let $x \in \mathbb{R}$ be a positive real number and let $n \in \mathbb{N}$ be a natural number. There exists a unique power of $n$ in the interval $[x, nx)$.*

**Proof** Let $n^\alpha$ with $\alpha \in \mathbb{N}$ be the largest power of $n$ such that $n^\alpha < x$. Multiplying both sides of this inequality by $n$ we obtain $n^{\alpha+1} < nx$. Since by assumption $n^\alpha$ is the largest power of $n$ not in $[x, nx)$, it must be the case that $n^{\alpha+1}$ lies in $[x, nx)$. This proves there must be at least one power of $n$ in the interval $[x, nx)$. To show uniqueness again consider $x \leq n^{\alpha+1} < nx$ and multiply all parts of this inequality by $n$. We see that $nx \leq n^{\alpha+2}$, thus showing that $n^{\alpha+2} \notin [x, nx)$. ■

Algorithm 4 contains the protocol used by player 1 to broadcast an arbitrary bit message of size $\alpha$. The communicating player starts by sending a "ping." During rounds $t^1_{\text{comm1}} + 1$ to $t^1_{\text{comm1}} + Kf(t^1_{\text{comm1}})$ player 1 will pull arm $\widehat{\sigma}_1$ with the intention of transmitting a single ON bit to signal to the receiving players that an information transmission session has started. After this, the transmission of the bit string takes place. The total number of rounds necessary to transmit a bit string of size $\alpha$ is $(\alpha+1)Kf(t^1_{\text{comm1}})$. One bit to signal the start of the communication and $\alpha$ bits corresponding to the message.

**Algorithm 3** Prepare and Start Communication (Player 1)

**Input** Player 1
**Initialize** FLAG $\leftarrow$ NONE
**for** *special rounds* $s = 1, \cdots$ **do**
  Pull arm $K-1$ (player 1 has just finished a Round Robin round)
  **if** $\mathbf{conn}^1(sK, 10) \geq 2$ *and* FLAG $=$ NONE **then**
    $t^1_{\text{first}} \leftarrow sK$
    FLAG $\leftarrow$ FINDPOWER
  **end**
  **if** FLAG $=$ FINDPOWER, $\left\lfloor \frac{s}{g(s)} \right\rfloor = 9^w$ *for some* $w \in \mathbb{N}$ *and* $\left\lfloor \frac{s-1}{g(s-1)} \right\rfloor \neq 9^w$ **then**
    $t^1_{\text{comm}} \leftarrow Ks$ .
    FLAG $\leftarrow$ PRECOMM.
  **end**
  **else if** FLAG $=$ PRECOMM, $\left\lfloor \frac{s}{g(s)} \right\rfloor = 9^w$ *for some* $w \in \mathbb{N}$ *and* $\left\lfloor \frac{s-1}{g(s-1)} \right\rfloor \neq 9^w$ **then**
    $t^1_{\text{comm1}} \leftarrow Ks$
    Compute guess $\widehat{\sigma}_1 \in [K]$ for the maximal arm $\sigma_1$ :

$$\widehat{\sigma}_1 = \arg\max_{i \in [K]} \widehat{\mu}^1_i(t^1_{\text{comm1}}) - D(N^1_i(t^1_{\text{comm1}}))$$

    MESSAGE $\leftarrow$ ENCODE$(\mathcal{C}^1_1(t^1_{\text{first}}, 10))$
    Run COMMUNICATE$(t^1_{\text{comm1}}, \widehat{\sigma}_1, \text{MESSAGE})$ using Algorithm 4.
  **end**
**end**

**Properties of $t^1_{\text{comm1}}$.** By design, the $t^1_{\text{comm1}}$ index passed to the COMMUNICATE function is defined to be the *second* special round equal or larger to $t^1_{\text{first}}$ satisfying $t^1_{\text{comm1}} = Ks^1_{\text{comm1}}$ where $\left\lfloor \frac{s_{\text{comm1}^1}}{s^1_{\text{comm1}}} \right\rfloor = 9^w$ for some $w \in \mathbb{N}$. Similarly the $t^1_{\text{comm}}$ index is the *first* special round equal or larger to $t^1_{\text{first}}$ satisfying $t^1_{\text{comm}} = Ks^1_{\text{comm}}$ where $\lfloor \frac{s_{\text{comm1}^1}}{s^1_{\text{comm1}}} \rfloor = 9^w$ for some $w \in \mathbb{N}$. The following lemma addresses the question of how much larger can $t^1_{\text{comm1}}$ be than $t^1_{\text{first}}$.

**Lemma 10** *Let* $t^1_{\text{first}} = Ks^1_{\text{first}}$ *and* $t^1_{\text{comm1}} = Ks^1_{\text{comm1}}$. *If* $s^1_{\text{first}} \geq s_{\text{boundary2}}$ *and* $\delta \leq \frac{1}{162}$ *then* $s^1_{\text{comm1}} \leq 162 s^1_{\text{first}}$ *and*

$$\frac{s^1_{comm1}}{g(s^1_{comm1})} \leq \frac{162 s^1_{\text{first}}}{g(s^1_{\text{first}})}.$$

The proof of Lemma 10 can be found in Appendix D.2. We can leverage Lemma 10 to derive an explicit bound for $s^1_{\text{comm1}}$ that is satisfied whenever $\mathcal{E}$ holds.

**Lemma 11** *If* $\mathcal{E}$ *holds,* $s^1_{\text{first}} \geq s_{\text{boundary2}}$ *and* $\delta \leq \frac{1}{162}$ *then*

$$s^1_{\text{comm1}} \leq \frac{746496}{\max_i \Delta^2_{\sigma_i, \sigma_{i+1}}} \log\left( \frac{746496 MK}{\delta \max_i \Delta^2_{\sigma_i, \sigma_{i+1}}} \right)$$

The proof of Lemma 11 can be found in Appendix D.3.

In Algorithm 3 we detail player 1's steps to initialize the communication protocol.

**Algorithm 4** COMMUNICATE (Player 1)

**Input** Round number $t^1_{\text{comm1}}$, communicating arm $i_{\text{comm}}$, message $\mathbf{b} \in \{0,1\}^\alpha$, communication rounds length function $f : \mathbb{R} \to \mathbb{N}$

**for** $t = t^1_{\text{comm1}} + 1, \cdots, t^1_{\text{comm1}} + (\alpha+1)Kf(t^1_{\text{comm1}})$ **do**
  **if** $t \in [t^1_{\text{comm1}} + 1, \cdots, t^1_{\text{comm1}} + Kf(t^1_{\text{comm1}})]$ **then**
    **Start ping**. Play communicating arm $i_{\text{comm}}$.
  **end**
  **else if** $j \geq 2$ *and* $t \in [t^1_{\text{comm1}} + (j-1)Kf(t^1_{\text{comm1}}) + 1, \cdots, t^1_{\text{comm1}} + jKf(t^1_{\text{comm1}})]$ *and* $\mathbf{b}_{j-1} = 1$ **then**
    Play communicating arm $i_{\text{comm}}$.
  **end**
  **else if** $j \geq 2$ *and* $t \in [t^1_{\text{comm1}} + (j-1)Kf(t^1_{\text{comm1}}) + 1, \cdots, t^1_{\text{comm1}} + jKf(t^1_{\text{comm1}})]$ *and* $\mathbf{b}_{j-1} = 0$ **then**
    Play Round Robin arm $t - \lfloor \frac{t-1}{K} \rfloor K$.
  **end**
**end**

### B.3. Detailed Discussion and Missing Supporting Results for Communication Analysis

The main objective of this section is to present a proof of

**Lemma 12 (One Bit Recovery Zero Collision Reward)** *Let* $A^p = \{L^p_i(t^1_{\text{start}})\}_{i\in[K]}$, $B^p = \{\widehat{\mu}^p_i(t^1_{\text{start}})\}_{i\in[K]}$, $C^p = \{\widehat{\mu}^p_i(t^1_{\text{start}}+1 : t^1_{\text{start}} + Kf(t^1_{\text{start}}))\}_{i\in[K]}$ *for all* $p$. *If the good event* $\mathcal{E}$ *holds then, with probability at least* $1 - \frac{\delta}{4K^2}$, *all players* $p \in \{2, \cdots, M\}$ *will be able to recover exactly the bit transmitted by player* 1 *by calling* $\text{ZeroTest}(A^p, B^p, C^p)$.

thus establishing the reliability of the ZeroTest. We will now show that the following three statements hold with high probability,

1. **Arm $\widehat{\sigma}_1$ has a large empirical mean for all players.**
   - Arm $\widehat{\sigma}_1 \in \{i \in [K] \text{ s.t. } \widehat{\mu}^p_i(t^1_{\text{start}}) \geq \frac{1}{2}\max_{j\in[K]} \widehat{\mu}^p_j(t^1_{\text{start}})\}$ for all $p \in \{2, \cdots, M\}$ and $\widehat{\mu}^p_{\widehat{\sigma}_1}(t^1_{\text{start}}) - D(N_{\widehat{\sigma}_1}(t^1_{\text{start}})) \geq \frac{3\mu_{\sigma_1}}{4}$.
2. **Arm $\widehat{\sigma}_1$ is comparable to $\sigma_1$.**
   - The witnesses $L^p_{\widehat{\sigma}_1} = \frac{\widehat{\mu}^p_{\widehat{\sigma}_1}(t^1_{\text{start}}) - D(N_{\widehat{\sigma}_1}(t^1_{\text{start}}))}{2} \in [\frac{\mu_{\widehat{\sigma}_1}}{3}, \frac{\mu_{\widehat{\sigma}_1}}{2}]$ for all $p \in \{2, \cdots, M\}$.
3. **When collisions are avoided the empirical mean estimators $\widehat{\mu}^p_{\widehat{\sigma}_1}(t^1_{\text{start}} + 1 : +Kf(t^1_{\text{start}}))$ are far from zero.**

- If $b = 1$, the estimators $\widehat{\mu}^p_{\widehat{\sigma}_1}(t^1_{\text{start}} + 1 : +Kf(t^1_{\text{start}})) \leq L^p_{\widehat{\sigma}_1}$ for all $p \in [M]$.
- If $b = 0$, the estimators $\widehat{\mu}^p_{\widehat{\sigma}_1}(t^1_{\text{start}} : t^1_{\text{start}} + Kf(t^1_{\text{start}})) > L^p_{\widehat{\sigma}_1}$ for all $p \in [M]$.

We start by presenting an 'easy-to-read' proof sketch, followed by an in-depth discussion of the more nuaced aspects of the proof.

**Proof** [sketch] Let's assume that $\mathcal{E}$ holds. We start by showing that because $t^1_{\text{start}} \geq \frac{128}{\mu^2_{\sigma_1}} g(t^1_{\text{start}}/K)$ (see Equation 5) for all $p \in [M]$ the lower confidence bound estimator around the optimal arm $\sigma_1$ computed at time $t^1_{\text{start}}$ is at least a constant proportion of its magnitude:

$$\widehat{\mu}^p_{\sigma_1}(t^1_{\text{start}}) - D(N^p_{\sigma_1}(t^1_{\text{start}})) \geq \frac{3\mu_{\sigma_1}}{4}. \tag{6}$$

See Lemma 13 for a proof of Equation 6. Since $\widehat{\sigma}_1 = \arg\max_{i\in[K]} \widehat{\mu}_i(t^1_{\text{start}}) - D(N^p_{\sigma_1}(t^1_{\text{start}}))$ and $\mathcal{E}$ holds, we conclude that $\mu_{\widehat{\sigma}_1} \geq \frac{3\mu_{\sigma_1}}{4}$. In other words, the true mean of $\widehat{\sigma}_1$ is at least a constant $(3/4)$ multiple of the mean reward of the maximal arm. This observation can be used to show that whenever the good event holds:

A) Arm $\widehat{\sigma}_1$ is always in the set of arms inspected during the ZeroTest; i.e.,

$$\widehat{\mu}_{\widehat{\sigma}_1}(t^1_{\text{start}}) \geq \frac{1}{2} \max_{j\in[K]} \mu_{j\in[K]} \widehat{\mu}^p_j(t^1_{\text{start}}).$$

B) The witnesses $L^p_{\widehat{\sigma}_1}(t^1_{\text{start}}) \in \left[\frac{\mu_{\widehat{\sigma}_1}}{3}, \frac{\mu_{\widehat{\sigma}_1}}{2}\right]$

Both A) and B) hold whenever $\mathcal{E}$ holds, and thus occur with probability at least $1 - \delta$. See Lemma 14 for a proof of this claim. The fundamental property of the witnesses $L^p_{\widehat{\sigma}_1}(t^1_{\text{start}})$ is that they are neither close to zero nor close to $\mu_{\widehat{\sigma}_1}$. Since $\mu_{\widehat{\sigma}_1} \geq \frac{3\mu_{\sigma_1}}{4}$, the witnesses are bounded away from zero and from $\mu_{\widehat{\sigma}_1}$ by a factor of at least $\frac{\mu_{\sigma_1}}{4}$.

The remainder of this proof sketch is based on the discussion and results from Appendix B.3.1. Let $\mathbb{P}_X$ be a distribution with support over $[0, 1]$ and mean $\mu_X$ equal to either the null (always zero) distribution or to the reward distribution of arm $\widehat{\sigma}_1$. We consider and answer the following question. Provided we have knowledge of the witnesses $\{L^p_{\widehat{\sigma}_1}(t^1_{\text{comm1}})\}_{p\in\{2,\cdots,M\}}$, how many i.i.d. samples from $\mathbb{P}_X$ are needed to be able to distinguish with probability $1 - \delta$ what type of distribution the samples come from (either the zero distribution or the reward distribution of arm $\widehat{\sigma}_1$).

Since the witnesses are all in the interval $\left[\frac{\mu_{\widehat{\sigma}_1}}{3}, \frac{\mu_{\widehat{\sigma}_1}}{2}\right]$, it follows that $\mu_{\widehat{\sigma}_1} - L^p_{\widehat{\sigma}_1} \geq \frac{\mu_{\widehat{\sigma}_1}}{2} \geq \frac{3\mu_{\sigma_1}}{8}$. This implies that in order to distinguish between these distributions it is enough to estimate $\mu_X$ up to accuracy $\mathcal{O}(\mu_{\sigma_1})$.

As a consequence of Assumption 1 we see the variance of the reward distribution of arm $\widehat{\sigma}_1$ is upper bounded by $\mu_{\widehat{\sigma}_1}(1 - \mu_{\widehat{\sigma}_1}) \leq \mu_{\sigma_1}$ and therefore the variance of $\mathbb{P}_X$ is also upper bounded by $\mu_{\sigma_1}$. Using this variance bound along with a Uniform Empirical Bernstein bound (see Lemmas 35 and 36 in Appendix E) we can show that with probability at least $1 - \delta$ it is enough for player $p$ to look at the empirical average of $N_p$ samples from $\mathbb{P}_X$, where $N_p$ is such that $\frac{N_p}{B(N_p, \frac{\delta}{4K^2M})} \geq \frac{48}{L^p_{\widehat{\sigma}_1}}$.

We require the length of the communication rounds to be player independent. Since all witnesses are in $\left[\frac{\mu_{\widehat{\sigma}_1}}{3}, \frac{\mu_{\widehat{\sigma}_1}}{2}\right]$ and $t^1_{\text{start}}$ can be related to $\mu_{\sigma_1}$ via Equation 5, we instead define $N$ (common to all players $p$) to be a function of $t^1_{\text{start}}$. We show that instead of $N_p$ we can use a player-independent value

$N = f(t^1_{\text{start}})$ such that $f(t^1_{\text{start}})$ is the first integer such that $\frac{f(t^1_{\text{start}})}{B(f(t^1_{\text{start}}), \frac{\delta}{4K^2M})} \geq 24\sqrt{\frac{t^1_{\text{start}}/K}{2g(t^1_{\text{start}}/K)}}$ also achieves this goal. This definition of $f$ gracefully extends to all integers $t$.

■

The following discussion will be devoted to present a detailed proof of Lemma 12.

**Detailed Proof of Lemma 12** Algorithm 5 is a simplified version of the full communication protocol which we introduced in Section 4. First, we assume the algorithm takes as input a known time index $t^1_{\text{start}}$, a bit $b \in \{0, 1\}$ that player 1 wishes to communicate to all other players and the communication rounds length function $f : \mathbb{R} \to \mathbb{N}$ that will determine the length of the communication rounds as a function of $t$. The algorithm works as follows. Initially all players in $\{1, \cdots, M\}$ play arms 1 through $K$ in a Round Robin fashion until time $t^1_{\text{start}}$, which is assumed to be a special round such that for all $i, j \in [K]$ and $p, p' \in [M]$, the counts $N^p_i(t^1_{\text{start}}) = N^{p'}_j(t^1_{\text{start}})$ (this means $t^1_{\text{start}}$ must be a multiple of $K$). After this time has elapsed each player $p$ has access to empirical estimators $\widehat{\mu}^p_i(t^1_{\text{start}})$ of $\mu_i$ for all $i \in [K]$. At this time, player 1 (the communicating player) computes a guess for the maximal arm $\widehat{\sigma}_1 \in [K]$. This will be the arm that player 1 will use to transmit bit $b$. The remaining $f(t^1_{\text{start}})$ rounds, which we call the communication rounds (where $f$ is a function known by all players $p \in [M]$) are used by player 1 to transmit $b$ and by the other players $p \neq 1$ to receive it.

If $b = 1$, player 1 will keep playing in a Round Robin fashion during the communication rounds, while if $b = 0$, player 1 will instead play arm $\widehat{\sigma}_1$. When transmitting a single bit, the length of the communication round equals $Kf(t^1_{\text{start}})$ rounds. At the end of the communication rounds, all players (except possibly player 1 if transmitting $b = 0$) will have played each arm $f(t^1_{\text{start}})$ times.

Using these $f(t^1_{\text{start}})$ samples per arm the players $p \in \{2, \cdots, M\}$ will conduct a test with the objective of verifying if there was any arm whose reward estimator $\widehat{\mu}(t^1_{\text{start}} + 1 : t^1_{\text{start}} + Kf(t^1_{\text{start}}))$ was substantially lower than the average values $\widehat{\mu}^p_i(t^1_{\text{start}})$ computed up to time $t^1_{\text{start}}$. If this is the case, they can safely conclude player 1 was pulling the same arm throughout the communication rounds $[t^1_{\text{start}} + 1, \cdots, t^1_{\text{start}} + Kf(t^1_{\text{start}})]$, and therefore incurring in collisions with player 1, while if they do not detect any substantial difference in their estimators, they can conclude that player 1 continued to play in a Round Robin fashion. We explain this procedure in more detail in Algorithms 5 and 2.

When $b = 1$ player 1 plays arm $\widehat{\sigma}_1$ during all rounds from $t = t^1_{\text{start}} + 1$ to $t = t^1_{\text{start}} + Kf(t^1_{\text{start}})$ while the remaining players are playing in a Round Robin fashion. Collisions will occur anytime a player distinct from 1 attempts to pull arm $\widehat{\sigma}_1$. At this moment, both players will receive a reward of zero. In case $b = 0$, no collisions will occur during rounds $t = t^1_{\text{start}} + 1, \cdots, t^1_{\text{start}} + Kf(t^1_{\text{start}})$ and the reward each player receives from pulling arm $\widehat{\sigma}_1$ should have a mean value of $\mu_{\widehat{\sigma}_1}$. In order for player $p' \neq 1$ to discern if player 1 is transmitting a one or a zero, there are two challenges. First, since the optimal arm index estimator $\widehat{\sigma}_1$ is random, none of the players $p' \neq 1$ knows the precise identity of arm $\widehat{\sigma}_1$. Second, none of the players knows the exact value of the true mean $\mu_{\widehat{\sigma}_1}$ and therefore, discerning between samples of the null (always zeros) distribution $\mathbb{P}_{\text{null}}$ and $\mathbb{P}_{\widehat{\sigma}_1}$ may prove challenging. We address these issues below.

Algorithm 5 below contains the detailed description of the simplified Communication Protocol.

**Algorithm 5** One Bit Communication Protocol Appendix

**Input** Players $[M]$, arms $[K]$, bit to communicate $b \in \{0,1\}$, communication round start $t^1_{\text{start}}$, communication rounds length function $f : \mathbb{R} \to \mathbb{N}$.

**for** $t = 1, \cdots, t^1_{\text{start}}$ **do**
  All players $p \in [M]$ play arms $[K]$ in a Round Robin fashion.
  **for** $p \in [M]$ **do**
    Play following a Round Robin schedule.
  **end**
**end**

Player 1 computes a guess $\widehat{\sigma}_1 \in [K]$ for the maximal arm $\sigma_1$ :

$$\widehat{\sigma}_1 = \arg\max_{i \in [K]} \widehat{\mu}^1_i(t^1_{\text{start}}) - D(N^1_i(t^1_{\text{start}}))$$

**for** $t = t^1_{\text{start}} + 1, \cdots, t^1_{\text{start}} + Kf(t^1_{\text{start}})$ **do**
  **if** $b = 1$ **then**
    Player 1 plays arm $\widehat{\sigma}_1$.
  **end**
  **else**
    Player 1 plays arms $[K]$ following a Round Robin schedule.
  **end**
  All players $p \neq 1$ continue playing arms $[K]$ following a Round Robin schedule.
**end**

In order to recover the value of the transmitted bit $b$ at the end of round $t^1_{\text{start}} + Kf(t^1_{\text{start}})$ all players $p \neq 1$ will compare $\widehat{\mu}^p_i(t^1_{\text{start}} + 1 : t^1_{\text{start}} + Kf(t^1_{\text{start}}))$ with the values $\widehat{\mu}^p_i(t^1_{\text{start}})$. If player $p \neq 1$ detects $\widehat{\mu}^p_i(t^1_{\text{start}} + 1 : t^1_{\text{start}} + Kf(t^1_{\text{start}}))$ to be substantially lower than $\widehat{\mu}^p_i(t^1_{\text{start}})$ for any arm $i$, then it can conclude there have been collisions and that $b = 1$, otherwise it will conclude that $b = 0$. We turn these intuitions into a precise mechanism in the discussion that follows.

For all $p \in [M]$ and all $i \in [K]$ define the lower bound 'witnesses' as,

$$L^p_i(t) := \frac{\widehat{\mu}^p_i(t) - D(N^p_i(t))}{2}$$

We now consider the following test to be executed by all players $p \in [M]$ with the data collected during rounds $t_0 + 1, \cdots, t_0 + Kf(t_0)$ and designed to decode bit $b$ transmitted by player 1. Let $\widehat{\mu}^p_i(t^1_{\text{start}} + 1 : t^1_{\text{start}} + Kf(t^1_{\text{start}}))$ be the empirical mean of arm $i$ computed by player $p$ and using only samples from $t = t^1_{\text{start}} + 1, \cdots, t^1_{\text{start}} + Kf(t^1_{\text{start}})$. Since all players $p \neq 1$ are playing the arms in $[K]$ following a Round Robin schedule this estimator will consist of $f(t^1_{\text{start}})$ samples for each arm $i \in [K]$. The following test will be used by all players $p \in [M]$ such that $p \neq 1$ to decode $b$,

Recall that $t^1_{\text{start}}$ is assumed to satisfy $t^1_{\text{start}} \geq t^1_{\text{first}}$. Since $t^1_{\text{start}}$ is assumed to be a special round, for all players $p, p' \in [M]$ and arms $i, j \in [K]$ the number of pulls satisfies $N^p_i(t^1_{\text{start}}) =$

$N_j^{p'}(t^1_{\text{start}}) = \frac{t^1_{\text{start}}}{K}$ and for the maximal arm $\sigma_1$,

$$\frac{2}{D^2(N^p_{\sigma_1}(t^1_{\text{start}}))} = \frac{N^p_{\sigma_1}(t^1_{\text{start}})}{g(N^p_{\sigma_1}(t^1_{\text{start}}))} \geq \frac{128}{\mu^2_{\sigma_1}} \tag{7}$$

For the remainder of this subsection and for all $p \in [M]$ and $i \in [K]$ we denote $N_i^p(t^1_{\text{start}})$ as $N(t^1_{\text{start}})$ to refer to $\frac{t^1_{\text{start}}}{K}$, the number of times each arm was played by any player $p \in [M]$ up to time $t^1_{\text{start}}$. In Lemma 13 we see that whenever $t^1_{\text{start}}$ satisfies Equation 7, the lower confidence bound estimators around the optimal arm $\widehat{\mu}^p_{\sigma_1}(t^1_{\text{start}}) - D(N^p_{\sigma_1}(t^1_{\text{start}}))$ are at least a constant fraction of the value of $\mu_{\sigma_1}$.

**Lemma 13** *If $t^1_{\text{start}}$ satisfies Equation 7 and $\mathcal{E}$ holds, the lower confidence bound estimator around the optimal arm $\sigma_1$ is at least a constant proportion of its magnitude*

$$\widehat{\mu}^p_{\sigma_1}(t^1_{\text{start}}) - D(N^p_{\sigma_1}(t^1_{\text{start}})) \geq \frac{3\mu_{\sigma_1}}{4}.$$

**Proof** If $\mathcal{E}$ holds, $\widehat{\mu}^p_{\sigma_1}(t^1_{\text{start}}) \in [\mu_{\sigma_1} - D(N^p_{\sigma_1}(t^1_{\text{start}})), \mu_{\sigma_1} + D(N^p_{\sigma_1}(t^1_{\text{start}}))]$ and therefore $\widehat{\mu}_{\sigma_1}(t^1_{\text{start}}) - D(N^p_{\sigma_1}(t^1_{\text{start}})) \geq \mu_{\sigma_1} - 2D(N^p_{\sigma_1}(t^1_{\text{start}}))$. Since $t^1_{\text{start}}$ satisfies Equation 7,

$$D(N^p_{\sigma_1}(t^1_{\text{start}})) \leq \frac{\mu_{\sigma_1}}{8}$$

and we conclude that $\widehat{\mu}^p_{\sigma_1}(t^1_{\text{start}}) - 2D(N^p_{\sigma_1}(t^1_{\text{start}})) \geq \frac{3\mu_{\sigma_1}}{4}$. ■

We can use the result of Lemma 13 to show that whenever $\mathcal{E}$ is true and Equation 7 holds for $t^1_{\text{start}}$ the witnesses of $\widehat{\sigma}_1$ are both upper and lower bounded by constant multiples of the true mean $\mu_{\widehat{\sigma}_1}$.

**Lemma 14** *Whenever $\mathcal{E}$ holds and Equation 7 holds for $t^1_{\text{start}}$, all witnesses of arm $\widehat{\sigma}_1$ for all $p \in [M]$ satisfy*

$$L^p_{\widehat{\sigma}_1}(t^1_{\text{start}}) \in \left[\frac{\mu_{\widehat{\sigma}_1}}{3}, \frac{\mu_{\widehat{\sigma}_1}}{2}\right] \tag{8}$$

*Furthermore $\mu_{\widehat{\sigma}_1} \geq \frac{3\mu_{\sigma_1}}{4}$ and $\widehat{\mu}_{\widehat{\sigma}_1}(t^1_{\text{start}}) \geq \frac{1}{2}\max_{j\in[K]} \widehat{\mu}^p_j(t^1_{\text{start}})$.*

**Proof** Recall that $\widehat{\sigma}_1 = \arg\max_{i\in[K]} \widehat{\mu}^1_i(t^1_{\text{start}}) - D(N^1_i(t^1_{\text{start}}))$. It follows that

$$\widehat{\mu}^1_{\widehat{\sigma}_1}(t^1_{\text{start}}) - D(N^1_{\widehat{\sigma}_1}(t^1_{\text{start}})) \geq \widehat{\mu}^1_{\sigma_1}(t^1_{\text{start}}) - D(N^1_{\sigma_1}(t^1_{\text{start}})) \overset{(i)}{\geq} \frac{3\mu_{\sigma_1}}{4} \overset{(ii)}{\geq} \frac{3\mu_{\widehat{\sigma}_1}}{4},$$

where inequality $(i)$ holds by Lemma 13 and inequality $(ii)$ holds by definition of $\mu_{\sigma_1}$. Furthermore, notice that since $\mathcal{E}$ holds,

$$\widehat{\mu}^p_{\widehat{\sigma}_1}(t^1_{\text{start}}) - D(N^p_{\widehat{\sigma}_1}(t^1_{\text{start}})) \leq \mu_{\widehat{\sigma}_1}. \tag{9}$$

for all $p \in [M]$.

Combining these two inequalities together we see that $\mu_{\widehat{\sigma}_1} \geq \frac{3\mu_{\sigma_1}}{4}$. Recall that $t^1_{\text{start}}$ is a special round. By definition we have $N_i^p(t^1_{\text{start}}) = N_j^{p'}(t^1_{\text{start}})$ for all $i, j \in [K]$ and all $p, p' \in [M]$, and $t^1_{\text{start}} \geq t^1_{\text{first}}$ therefore since $\mathcal{E}$ holds Equation 7 must be satisfied for all players. Equation 7

implies that $D(N^p_{\widehat{\sigma}_1}(t^1_{\text{start}})) \leq \frac{\mu_{\sigma_1}}{8}$ and therefore that $D(N^p_{\widehat{\sigma}_1}(t^1_{\text{start}})) \leq \frac{\mu_{\widehat{\sigma}_1}}{6}$. Combining these two observations we see that for all $p \in [M]$,

$$\widehat{\mu}^p_{\widehat{\sigma}_1}(t^1_{\text{start}}) - D(N^p_{\widehat{\sigma}_1}(t^1_{\text{start}})) \overset{(i)}{\geq} \mu_{\widehat{\sigma}_1} - 2D(N^p_{\widehat{\sigma}_1}(t^1_{\text{start}})) \overset{(ii)}{\geq} \frac{2\mu_{\widehat{\sigma}_1}}{3} \tag{10}$$

Inequality $(i)$ is satisfied because $\mathcal{E}$ holds, and $(ii)$ is a consequence of the observation that $D(N^p_{\widehat{\sigma}_1}(t^1_{\text{start}})) \leq \frac{\mu_{\widehat{\sigma}_1}}{6}$. Combining Equations 9 and 10 we conclude that

$$\frac{\mu_{\widehat{\sigma}_1(t^1_{\text{start}})}}{3} \leq L^p_{\widehat{\sigma}_1}(t^1_{\text{start}}) = \frac{\widehat{\mu}_{\widehat{\sigma}_1}(t^1_{\text{start}}) - D(N^p_{\widehat{\sigma}_1}(t^1_{\text{start}}))}{2} \leq \frac{\mu_{\widehat{\sigma}_1(t^1_{\text{start}})}}{2}.$$

For the second part, the following sequence of inequalities holds:

$$\widehat{\mu}^p_{\widehat{\sigma}_1}(t^1_{\text{start}}) \overset{(i)}{\geq} \mu_{\widehat{\sigma}_1} - D(N^p_{\widehat{\sigma}_1}(t^1_{\text{start}})) \overset{(ii)}{\geq} \frac{3\mu_{\sigma_1}}{4} - D(N^p_{\widehat{\sigma}_1}(t^1_{\text{start}})) \overset{(iii)}{\geq} \frac{5\mu_{\sigma_1}}{8}.$$

Inequality $(i)$ is satisfied because Equation 10 is true whenever $\mathcal{E}$ holds. Inequality $(ii)$ is a consequence of $\mu_{\widehat{\sigma}_1} \geq \frac{3\mu_{\sigma_1}}{4}$ and inequality $(iii)$ holds because $D(N^p_{\widehat{\sigma}_1}(t^1_{\text{start}})) \leq \frac{\mu_{\sigma_1}}{8}$. The later implies that

$$\frac{9}{8}\mu_{\sigma_1} \geq \mu_{\sigma_1} + D(N^p_{\widehat{\sigma}_1}(t^1_{\text{start}})) \geq \max_{j\in[K]} \widehat{\mu}^p_j(t^1_{\text{start}}),$$

where the inequality on the RHS holds because $\mathcal{E}$ is satisfied and $\mu_{\sigma_1}$ is the maximal arm. We conclude that

$$\widehat{\mu}^p_{\widehat{\sigma}_1}(t^1_{\text{start}}) \geq \frac{5\max_{j\in[K]} \widehat{\mu}^p_j(t^1_{\text{start}})}{9}.$$

Since $5/9 > 1/2$ the result follows.

■

Among other things Lemma 14 implies that whenever $\mathcal{E}$ holds, arm $\widehat{\sigma}_1$ (the *random* arm used by player 1 to communicate) is among the arms inspected by all players $p \neq 1$ during the Zero Test in Algorithm 2. In section B.3.1 we show how to choose $f$ in order to ensure that Algorithm 2 allows all players $p \neq 1$ to decode $b$ with high probability. This is the same functional form that is highlighted at the start of Section 3.

### B.3.1. DETAILED ANALYSIS OF THE ZERO TEST

We will take a small step back and consider a simplified version of the bit communication protocol which will be useful in analyzing the Zero Test of Algorithm 2. Let $X$ be a random variable with support in $[0, 1]$ and mean $\mu_X$. Assume $\mu_X > L$ for $L$ known. Let $Z_1, \cdots, Z_N$ be $N$ i.i.d. samples from either $\mathbb{P}_X$ or the null distribution $\mathbb{P}_{\text{null}}$ (all $Z_i = 0$) and let $\widehat{\mu}_Z = \frac{1}{N}\sum_{i=1}^N Z_i$. The problem we consider is the following:

*How many i.i.d. samples are required to determine with high probability from which of the two distributions ($\mathbb{P}_X$ or $\mathbb{P}_{\text{null}}$) do the samples of $\{Z_i\}_{i=1}^N$ come from?*

We will analyze the following simplified version of the zero test,

$$\text{If } \widehat{\mu}_Z \geq L, \text{ then output } \mathbb{P}_X \text{ else output } \mathbb{P}_{\text{null}} \tag{Zero-Test-Simple}$$

We use the fact that for any random variable $\mathbb{P}_X$ with support in $[0,1]$ and mean $\mu_X$ the variance can be upper bounded by $\mu_X(1-\mu_X)$ (see Lemma 34 in Appendix E) in conjunction with a Uniform Empirical Bernstein bound (see Lemmas 35 and 36 in Appendix E) to show the following bound on the required number of samples $N$,

**Lemma 15** *Let $\delta' \in (0,1)$. If $X$ is a random variable with support in $[0,1]$, distribution $\mathbb{P}_X$ and mean $\mu_X$ satisfying $L \leq \mu_X$, then with probability at least $1-\delta'$ for all $N$ such that $\frac{N}{B(N,\delta')} \geq \max\left(\frac{2}{\mu_X - L}, \frac{16\min(\mu_X, 1-\mu_X)}{(\mu_X - L)^2}\right)$ we have,*

$$\widehat{\mu}_X \geq L,$$

*where $B(n,\delta') = 2\log\log(2n) + \log\frac{5.2}{\delta'}$.*

**Proof**

Let $\alpha = \min(\mu_X, 1-\mu_X)$. A simple use of Lemma 36 implies that with probability at least $1-\delta'$ for all $n \in \mathbb{N}$:

$$\widehat{\mu}_X \geq \mu_X - 2\sqrt{\frac{\alpha B(n,\delta')}{n}} - \frac{B(n,\delta')}{n}.$$

The LHS of this inequality attains a value of at least $L$ whenever:

$$\mu_X - L \geq 2\sqrt{\frac{\alpha B(n,\delta')}{n}} + \frac{B(n,\delta')}{n}.$$

We finalize the proof by noting that for all $n$ such that $\frac{n}{B(n,\delta')} \geq \max\left(\frac{2}{\mu_X - L}, \frac{16\min(\mu_X, 1-\mu_X)}{(\mu_X - L)^2}\right)$ we have that $\frac{\mu_X - L}{2} \geq 2\sqrt{\frac{\alpha B(n,\delta')}{n}}$ and $\frac{\mu_X - L}{2} \geq \frac{B(n,\delta')}{n}$. ■

The bound in Lemma 15 implies that when the sampling distribution for $Z$ equals $\mathbb{P}_X$, the empirical mean $\widehat{\mu}_Z$ will be larger than $L$ with high probability provided the number of test samples $\{Z_i\}_{i=1}^N$ is large enough. Observe that $N$ has an inverse dependence on the gap $\mu_X - L$. We will be applying this result to the case where although $L \neq \mu_X$ it is of the order of $\mu_X$.

**Lemma 16** *Let $\delta' \in (0,1)$. Assume that $\frac{\mu_X}{2} \geq L \geq \frac{\mu_X}{3}$ and let be $N$ be an integer such that $\frac{N}{B(N,\delta')} \geq \frac{48}{L}$, then Zero-Test-Simple succeeds with probability at least $1-\delta'$.*

**Proof** This follows from Lemma 15 by noting that in this case $\frac{1}{\mu_X - L} \leq \frac{1}{L}$ and $\frac{16\min(\mu_X, 1-\mu_X)}{(\mu_X - L)^2} \leq \frac{48}{L}$. ■

Lemma 16 is an instantiation of the results of Lemma 15 when $\mu_X - L$ is of the order of $\mu_X$. This result says that up to logarithmic factors, it is enough for $N \approx \frac{1}{\mu_X}$ for the empirical estimator $\widehat{\mu}_Z$ to be at least a constant fraction of the true mean $\mu_X$.

We can now apply these results to the Communication Protocol in Algorithm 5 and the Zero Test in Algorithm 2. Recall that by definition of $t^1_{\text{start}}$ we have $t^1_{\text{start}} \geq t^1_{\text{first}}\ N_i^p(t^1_{\text{start}}) = N_j^{p'}(t^1_{\text{start}}) = N(t^1_{\text{start}})$ for all $i,j \in [K]$ and all $p,p' \in [M]$.

**Lemma 17** *Let $\delta \in (0,1)$ and $t^1_{\text{start}}$ satisfy Equation 7. If $f(t^1_{\text{start}})$ is an integer such that $\frac{f(t^1_{\text{start}})}{B(f(t^1_{\text{start}}),\frac{\delta}{4K^2M})} \geq 24\sqrt{\frac{t^1_{\text{start}}/K}{2g(t^1_{\text{start}}/K)}}$ then whenever the good event $\mathcal{E}$ holds, at the end of the Communication protocol in Algorithm 5 all players $p \in [M]$ with $p \neq 1$ will be able to recover exactly the bit transmitted by player 1 via the Zero Test of Algorithm 2 with probability at least $1 - \frac{\delta}{4K^2}$.*

**Proof** Let's assume $\mathcal{E}$ holds. Equation 7 implies that $\mu_{\sigma_1} \geq 8\sqrt{\frac{g(N(t^1_{\text{start}}))}{N(t^1_{\text{start}})}}$ Lemma 13 tells us that $\mu_{\widehat{\sigma}_1} \geq \frac{3\mu_{\sigma_1}}{4}$ and therefore $\mu_{\widehat{\sigma}_1} \geq 6\sqrt{\frac{g(N(t^1_{\text{start}}))}{N(t^1_{\text{start}})}}$. Furthermore, by Equation 8 of Lemma 14, $L^p_{\widehat{\sigma}_1}(t^1_{\text{start}}) \in \left[\frac{\mu_{\widehat{\sigma}_1}}{3}, \frac{\mu_{\widehat{\sigma}_1}}{2}\right]$ and therefore $L^p_{\widehat{\sigma}_1} \geq 2\sqrt{\frac{g(N(t^1_{\text{start}}))}{N(t^1_{\text{start}})}}$. We can then conclude that $\frac{48}{L^p_{\widehat{\sigma}_1}} \leq 24\sqrt{\frac{N(t^1_{\text{start}})}{g(N(t^1_{\text{start}}))}}$ for all $p \in [M]$.

Recall that $N(t^1_{\text{start}}) = \frac{t^1_{\text{start}}}{K}$ (a known function of $t^1_{\text{start}}$). If we define $f(t^1_{\text{start}})$ to be the any integer[6] such that $\frac{f(t^1_{\text{start}})}{B(f(t^1_{\text{start}}),\frac{\delta}{4K^2M})} \geq 24\sqrt{\frac{N(t^1_{\text{start}})}{g(N(t^1_{\text{start}}))}} = 24\sqrt{\frac{t^1_{\text{start}}/K}{g(t^1_{\text{start}}/K)}}$, the conditions of Lemma 16 are satisfied since $\frac{f(t^1_{\text{start}})}{B(f(t^1_{\text{start}}),\frac{\delta}{4K^2M})} \geq \frac{48}{L^p_{\widehat{\sigma}_1}}$ and $L^p_{\widehat{\sigma}_1}(t^1_{\text{start}}) \in \left[\frac{\mu_{\widehat{\sigma}_1}}{3}, \frac{\mu_{\widehat{\sigma}_1}}{2}\right]$. We can conclude that the Zero Tests performed by each of the players $p \in [M]$ are successful in recovering player 1's transmission over the pulls of arm $\widehat{\sigma}_1$ with probability at least $1 - \frac{\delta}{4K^2M}$ each. A union bound over the $M$ players yields the result. ■

This finalizes the formal proof of Lemma 12.

### B.4. Detailed Discussion and Missing Supporting Results for The Listening Players

In this section we present the full versions of the algorithms used by the listening players to 1) prepare and start listening (see Algorithm 6) and 2) decode player 1's message (see Algorithm 7). We also present the proof of Lemma 8, which we restate for readability.

6. For example the first one that satisfies this bound.

**Algorithm 6** Prepare and Start Listening ($p \in \{2, \cdots, M\}$)

**Input** Players $p \in \{2, \cdots, K\}$
**Initialize** FLAG $\leftarrow$ NONE
**for** special rounds $s = 1, \cdots$ **do**
  Pull arm $K - p$ (Round Robin schedule) **if** $\mathbf{conn}^p(sK, 10) \geq 2$ **then**
    FLAG $\leftarrow$ FINDPOWER
  **end**
  **if** FLAG $=$ FINDPOWER, $\left\lfloor \frac{s}{g(s)} \right\rfloor = 9^w$ *for some* $w \in \mathbb{N}$ *and* $\left\lfloor \frac{s-1}{g(s-1)} \right\rfloor \neq 9^w$ **then**
    $t^p_{\text{listen}} \leftarrow Ks$
    $A^p \leftarrow \{L^p_i(t^p_{\text{listen}})\}_{i\in[K]}$, $B^p \leftarrow \{\widehat{\mu}^p_i(t^p_{\text{listen}})\}_{i\in[K]}$
    Start listening for a communication start signal from player 1
    FLAG $\leftarrow$ LISTENCOMM1
  **end**
  **else if** FLAG $=$ LISTENCOMM1 *and* $t = t^p_{\text{listen}} + Kf(t^p_{\text{listen}})$ **then**
    **if** $\text{ZeroTest}(A^p, B^p, C^p) = 0$ **then**
      $C^p \leftarrow \{\widehat{\mu}^p_i(t^p_{\text{listen}} + 1 : t^p_{\text{listen}} + Kf(t^p_{\text{listen}}))\}_{i\in[K]}$
      The algorithm did not detect a communication start signal:
      FLAG $\leftarrow$ FINDPOWER
    **end**
    **else**
      $\mathbf{DecodedMessage} \leftarrow$ DECODE($t^p_{\text{listen}}, A^p, B^p, C^p$, message size $= K$) using Algorithm 7.
    **end**
  **end**
**end**

**Algorithm 7** DECODE

**Input** Round number $t^p_{\text{listen}}$, witnesses $\{L^p_i(t^p_{\text{listen}})\}_{i\in[K]}$, empirical means $\{\widehat{\mu}^p_i(t^p_{\text{listen}})\}_{i\in[K]}$, message size $\alpha$
$A^p \leftarrow \{L^p_i(t^p_{\text{listen}})\}_{i\in[K]}$, $B^p \leftarrow \{\widehat{\mu}^p_i(t^p_{\text{listen}})\}_{i\in[K]}$
**for** $t = t^p_{\text{listen}} + 1, \cdots, t^p_{\text{listen}} + \alpha Kf(t^p_{\text{listen}})$ **do**
  **if** $t = t^p_{\text{listen}} + jKf(t_{\text{listen}})$ **then**
    $C^p \leftarrow \{\widehat{\mu}^p_i(t^p_{\text{listen}} + (j-1)Kf(t^p_{\text{listen}}) + 1 : t^p_{\text{listen}} + jKf(t^p_{\text{listen}}))\}_{i\in[K]}$
    $\mathbf{DecodedMessage}_j \leftarrow \text{ZeroTest}(A^p, B^p, C^p)$
  **end**
  Return $\mathbf{DecodedMessage}$.
**end**

**Lemma 18 (Message Recovery)** *If $\mathcal{E}$ holds then with probability at least $1-\frac{\delta}{K}$ for all $p\in[M]$ the value of $t^p_{\text{listen}}$ sent to the* DECODE *function of Algorithm 7 satisfies $t^p_{\text{listen}}=t^1_{\text{comm1}}$ and* DECODE *will recover the exact $K-$bit message sent by player 1.*

**Proof** Let $9^u$ be the unique power of nine in the interval $\left[\frac{128}{\max_i \Delta^2_{\sigma_i,\sigma_{i+1}}},\frac{1152}{\max_i \Delta^2_{\sigma_i,\sigma_{i+1}}}\right)$. Recall that whenever the good event $\mathcal{E}$ holds by design the first $t^p_{\text{listen}}$ of Algorithm 6 satisfies

$$t^p_{\text{listen}}\in\left\{\min_{t\in\mathbb{N}}\text{ s.t. }\left\lfloor\frac{t/K}{g(t/K)}\right\rfloor=9^u,\min_{t\in\mathbb{N}}\text{ s.t. }\left\lfloor\frac{t/K}{g(t/K)}\right\rfloor=9^{u+1}\right\}.$$

Similarly recall that whenever the good event $\mathcal{E}$ holds

$$t^1_{\text{comm1}}\in\left\{\min_{t\in\mathbb{N}}\text{ s.t. }\left\lfloor\frac{t/K}{g(t/K)}\right\rfloor=9^{u+1},\min_{t\in\mathbb{N}}\text{ s.t. }\left\lfloor\frac{t/K}{g(t/K)}\right\rfloor=9^{u+2}\right\}.$$

This means that each player $p\in\{2,\cdots,M\}$ may require at most three invocations to the ZeroTest function to detect the $b=1$ bit that player 1 will use to signal the start of the communication sequence. Applying Lemma 12 with $t^1_{\text{start}}$ equals the different $t^p_{\text{listen}}$ guesses of the listening players and over the $K$ bits transmitted by player 1, we see that a union bound over at most $K+3$ uses of Lemma 12 are required. Since $K+3\leq 4K$ the result follows. ■

### B.5. Detailed Discussion and Missing Supporting Results for Bounding Regret

In this section we present the missing detailed proofs of Section 4.4 in the main. The discussion is divided in two parts. First we analyze the $\text{FirstPartitionRegret}([K],[M])$ derived from a single communication and listening round required to transmit $\mathcal{C}^1_1(t^1_{\text{first}},10)$ (see Section B.5.1). The next section B.5.2 deals with the RECURSE function and with assembling the final algorithm.

#### B.5.1. BOUNDING THE FIRST PARTITION REGRET

We have now the necessary ingredients to characterize the regret of the strategy to communicate the composition of $\mathcal{C}^1_1(t^1_{\text{first}},10)$ from player 1 to all other players ($\text{FirstPartitionRegret}([K],[M])$) Observe that regret is generated only when collisions occur. During the communication interaction between player 1 and any other single player $p\in\{2,\cdots,M\}$, the number of collisions is upper bounded by $(K+1)f(t^1_{\text{comm1}})$. Thus the total CollisionRegret experienced by the $M$ players during the first communication round (when player 1 informs players $\{2,\cdots,M\}$ of the partition resulting from the first time $\mathcal{G}^1_{t^1_{\text{first}}}(10)$ has more than one connected component) is upper bounded by $M(K+1)f(t^1_{\text{comm1}})$. Let's prove an upper bound for $f(t^1_{\text{comm1}})$.

**Lemma 19** *If $\mathcal{E}$ holds, $t_{\text{comm1}}\geq\max(t_{\text{boundary1}},t_{\text{boundary3}})$, $s^1_{\text{first}}\geq s_{\text{boundary2}}$ and $\delta\leq\frac{1}{162}$ then,*

$$f(t^1_{\text{comm1}})\leq\frac{20736B\left(\frac{186624}{\max_i\Delta^2_{\sigma_i,\sigma_{i+1}}},\frac{\delta}{4K^2M}\right)}{\max_i\Delta_{\sigma_i,\sigma_{i+1}}}.\tag{11}$$

The proof of Lemma 19 can be found in Appendix D.4. We can now combine Lemmas 8 and 19 to bound the regret incurred by Algorithms 3, 4, 6 and 7 during the transmission of the partition message $\text{ENCODE}(\mathcal{C}_1^1(t^1_{\text{first}}, 10))$.

**Corollary 20 (First Partition Collision Regret)** *If $\mathcal{E}$ holds and $\delta \leq \frac{1}{162}$ then with probability at least $1 - \frac{\delta}{K}$ the total collision regret (the regret generated by collisions occurring during the communication rounds used to communicate the composition of $\mathcal{C}_1^1(t^1_{\text{first}}, 10)$ satisfies,*

$$\text{CollisionRegret}([K],[M]) \leq \frac{20736(K+1)MB\left(\frac{186624}{\max_i \Delta^2_{\sigma_i,\sigma_{i+1}}}, \frac{\delta}{4K^2M}\right)}{\max_i \Delta_{\sigma_i,\sigma_{i+1}}} + \tilde{c}(\delta, K, M),$$

*where $\tilde{c}(\delta, K, M)$ is a logarithmic problem independent cost resulting from the regret incurred*[7] *before the boundary conditions $t_{\text{comm1}} \geq \max\left(t_{\text{boundary1}}, t_{\text{boundary3}}\right)$, $s^1_{\text{first}} \geq s_{\text{boundary2}}$ hold*[8].

We can also get a bound for the $\text{RoundRobinRegret}$. By Lemma 11 whenever $\mathcal{E}$ holds, $s^1_{\text{comm1}} \leq \frac{746496}{\max_i \Delta^2_{\sigma_i,\sigma_{i+1}}} \log\left(\frac{746496MK}{\delta \max_i \Delta^2_{\sigma_i,\sigma_{i+1}}}\right)$. During a single Round Robin cycle regret is only incurred when arms in $\{\mu_{\sigma_i}\}_{i=M+1}^K$ are played. A full cycle consists of $KM$ arm pulls, all players pull each arm once. Out of these $KM$ pulls the $M^2$ pulls of arms $\mu_{\sigma_1}, \cdots, \mu_{\sigma_M}$ incurr in no regret. The remaining $(K-M)M$ pulls incur in a regret of

$$(K-M)\left(\sum_{i=1}^M \mu_{\sigma_i}\right) - M\left(\sum_{i=M+1}^K \mu_{\sigma_i}\right)$$

We can further upper bound this quantity as follows,

$$\begin{aligned}(K-M)\left(\sum_{i=1}^M \mu_{\sigma_i}\right) - M\left(\sum_{i=M+1}^K \mu_{\sigma_i}\right) &= \sum_{i=1}^M \sum_{j=M+1}^K \Delta_{\sigma_i,\sigma_j} \\ &\leq M(K-M)K \max_i \Delta_{\sigma_i,\sigma_{i+1}}\end{aligned}$$

Where we have used the bound $\Delta_{\sigma_{i_1},\sigma_{i_2}} \leq K \max_i \Delta_{\sigma_i,\sigma_{i+1}}$ for all $i_1 < i_2$. Thus the $\text{RoundRobinRegret}$ incurred by the algorithm in the rounds preceding active communication can be upper bounded by

$$M(K-M)K \max_i \Delta_{\sigma_i,\sigma_{i+1}} s^1_{\text{comm1}}.$$

Recall that during the communication rounds, all players $p \in \{2, \cdots, M\}$ are still using a Round Robin schedule. This goes on for $(K+1)Kf(t^1_{\text{comm1}})$ rounds after $t^1_{\text{comm1}}$, thus completing a total

7. A slightly more careful algorithm that uses an estimator of $\mu_{\hat{\sigma}_1}$ as the input to determine the length of the one bit communication rounds yields a regret bound of the form $\text{CollisionRegret}([K],[M]) = \mathcal{O}\left(\frac{KM\log(t/\delta)}{\mu_{\sigma_1}}\right)$. Since this quantity would be dominated by the $\text{RoundRobinRegret}([K],[M])$ it wouldn't change the final result.
8. We will provide a bound for this quantity in the following section.

of $(K+1)f(t^1_{\text{comm1}})$ Round Robin cycles. Using Equation 11 from Lemma 19 we can upper bound the Round Robin regret incurred during these rounds as

$$M(K-M)K \max_i \Delta_{\sigma_i,\sigma_{i+1}} \times (K+1)f(t^1_{\text{comm1}}) \leq 20736(K+1)M(K-M)KB\left(\frac{186624}{\max_i \Delta^2_{\sigma_i,\sigma_{i+1}}}, \frac{\delta}{4K^2M}\right).$$

These observations imply the following upper bound for RoundRobinRegret,

**Corollary 21 (First Partition Round Robin Regret)** *If $\mathcal{E}$ holds and $\delta \leq \frac{1}{162}$ then with probability at least $1-\frac{\delta}{K}$ the Round Robin regret is bounded as follows:*

$$\text{RoundRobinRegret}([K],[M]) \leq \frac{746496M(K-M)K}{\max_i \Delta_{\sigma_i,\sigma_{i+1}}} \log\left(\frac{746496MK}{\delta \max_i \Delta^2_{\sigma_i,\sigma_{i+1}}}\right) 20736(K+1)M(K-M)KB\left(\frac{186624}{\max_i \Delta^2_{\sigma_i,\sigma_{i+1}}}, \frac{\delta}{4K^2M}\right) + \tilde{c}(\delta,K,M),$$

*where $\tilde{c}(\delta,K,M)$ is a logarithmic problem independent cost resulting from the regret incurred before the boundary conditions $t_{\text{comm1}} \geq \max(t_{\text{boundary1}}, t_{\text{boundary3}})$, $s^1_{\text{first}} \geq s_{\text{boundary2}}$ hold and is the same as in Corollary 20.*

Combining Corollaries 20 and 21 we can infer that if $\mathcal{E}$ holds and $\delta \leq \frac{1}{162}$ then with probability at least $1-\frac{\delta}{K}$ the total regret incurred up to time $t^1_{\text{comm1}} + (K+1)f(t^1_{\text{comm1}})$, when all players are aware of the composition of $\mathcal{C}^1_1(t^1_{\text{first}}, 10)$ is upper bounded by

$$\text{FirstPartitionRegret}([K],[M]) \leq \frac{746496M(K-M)K}{\max_i \Delta_{\sigma_i,\sigma_{i+1}}} \log\left(\frac{746496MK}{\delta \max_i \Delta^2_{\sigma_i,\sigma_{i+1}}}\right) + 20736(K+1)M(K-M)KB\left(\frac{186624}{\max_i \Delta^2_{\sigma_i,\sigma_{i+1}}}, \frac{\delta}{4K^2M}\right) + \frac{20736(K+1)MB\left(\frac{186624}{\max_i \Delta^2_{\sigma_i,\sigma_{i+1}}}, \frac{\delta}{4K^2M}\right)}{\max_i \Delta_{\sigma_i,\sigma_{i+1}}} + \tilde{c}(\delta,K,M), \tag{12}$$

Where the term $\tilde{c}(\delta,K,M)$ captures a crude linear upper bound on the regret collected before the boundary conditions hold true. We can also invoque the results in Lemma 11 and 19 to bound on the total number of rounds needed until all players are aware of the composition of $\mathcal{C}^1_1(t^1_{\text{first}}, 10)$,

$$\text{Runtime}([K],[M]) \leq \frac{746496K}{\max_i \Delta^2_{\sigma_i,\sigma_{i+1}}} \log\left(\frac{746496MK}{\delta \max_i \Delta^2_{\sigma_i,\sigma_{i+1}}}\right) + \frac{20736(K+1)KB\left(\frac{186624}{\max_i \Delta^2_{\sigma_i,\sigma_{i+1}}}, \frac{\delta}{4K^2M}\right)}{\max_i \Delta_{\sigma_i,\sigma_{i+1}}}.$$

**Bounding** $\tilde{c}(\delta, K, M)$ The cost of satisfying the boundary conditions bound is not additive between the RoundRobinRegret and the CollisionRegret components of FirstPartititionRegret. In order to deal with these we introduce a slight modification to the communication and listening protocols of Algorithms 3, 4, 6 and 7 by modifying the definition of $t^p_{\text{first}}$ for all $p \in [M]$. Instead we use $\tilde{t}^p_{\text{first}}$ times defined as $\tilde{t}^p_{\text{first}} = \max(t^p_{\text{first}}, t_{\text{firstBoundary}})$. It is easy to see this will not affect the regret too much. If $t^p_{\text{first}} \geq t_{\text{firstBoundary}}$ for all $p \in [M]$, or $t_{\text{firstBoundary}} \geq t^p_{\text{first}}$ for some $p \in [M]$ but not for all, the analysis will remain unchanged. If instead $t_{\text{firstBoundary}} > t^p_{\text{first}}$ for all $p \in [M]$, all the player's $\tilde{t}^p_{\text{first}} = t_{\text{firstBoundary}}$. This definition induces that of $\tilde{s}^p_{\text{first}}$ $\forall p \in [M]$, $\tilde{t}^1_{\text{comm1}}$ and $\tilde{s}^1_{\text{comm1}}$. As a consequence of Lemma 11 we see that $\tilde{s}^1_{\text{comm1}} \leq 162\tilde{s}^1_{\text{first}} = 162\frac{t_{\text{firstBoundary}}}{K}$. Notice that by definition of the $4-$th boundary condition $f(\tilde{t}^1_{\text{comm1}}) \leq \frac{\tilde{t}^1_{\text{comm1}}/K}{g(\tilde{t}^1_{\text{comm1}}/K)} \leq \tilde{t}^1_{\text{comm1}}/K \leq 162t_{\text{firstBoundary}}/K$.

The protocol thus ensures communicating the composition of $\mathcal{C}^p_1(t^1_{\text{first}}, 10)$ (notice that we are still transmitting the composition of $\mathcal{C}^p_1(t^1_{\text{first}}, 10)$ and not $\mathcal{C}^p_1(\tilde{t}^1_{\text{first}}, 10)$ ) can be achieved while incurring regret of at most,

$$M(K-M)K \max_i \Delta_{\sigma_i,\sigma_{i+1}} \left(\tilde{s}_{\text{comm1}} + (K+1)f(\tilde{t}^1_{\text{comm1}})\right) + M(K+1)f(\tilde{t}^1_{\text{comm1}}).$$

We define $\tilde{c}(\delta, K, M)$ to be a problem independent upper bound of this quantity.

$$\begin{aligned}\tilde{c}(\delta, K, M) &= M(K-M)K\left(\frac{162t_{\text{firstBoundary}}}{K} + (K+1)\frac{162t_{\text{firstBoundary}}}{K}\right.\\ &\quad \left. + M(K+1)\frac{162t_{\text{firstBoundary}}}{K}\right)\\ &= \mathbf{poly}\left(\log\left(\frac{1}{\delta}\right), K, M\right)\end{aligned}$$

And where the dependence on $\log\left(\frac{1}{\delta}\right)$ is linear.

### B.5.2. ANALYZING THE RECURSE FUNCTION

In order to implement the recursion strategy described at the start of Section 4 and at the end of Section 4.4, when faced with a smaller Cooperative Multi-Player Multi-Armed problem the players will restart their empirical mean estimators from scratch. In Appendix C.5 we describe a warm-start strategy that allows the players to start their empirical mean estimators using a constant proportion of the samples that have been gathered so far. The two strategies have the same performance up to constant factors.

**Algorithm 8** RECURSE

**Input** players $p \in \{1, \cdots, M\}$, arm indices $\{1, \cdots, K\}$

Run Algorithms 3, 4, 6, and 7 to communicate $\mathcal{C}_1^1(t^1_{\text{first}}, 10)$ to all players.

**if** $|\mathcal{C}_1^1(t^1_{\text{first}}, 10)| > M$ **then**
  RECURSE on $\mathcal{C}_1^1(t^1_{\text{first}}, 10)$ with players $[M]$.
**end**
**else**
  Run RoundRobin on $\mathcal{C}_1^1(t^1_{\text{first}}, 10)$ with players $\{1, \cdots, |\mathcal{C}_1^1(t^1_{\text{first}}, 10)|\}$.
  RECURSE on $[K] \backslash \mathcal{C}_1^1(t^1_{\text{first}}, 10)$ with players $[M] \backslash \{1, \cdots, |\mathcal{C}_1^1(t^1_{\text{first}}, 10)|\}$.
**end**

To analyze the regret guarantees of Algorithm 8 let's start by noting that in case $|\mathcal{C}_1^1(t^1_{\text{first}}, 10)| \leq < M$, the subset of players $\{1, \cdots, |\mathcal{C}_1^1(t^1_{\text{first}}, 10)|\}$ that has been assigned to RoundRobin over the subset $\mathcal{C}_1^1(t^1_{\text{first}}, 10)$ will not incur in any more regret. It is therefore only necessary to bound the regret incurred by the algorithm during each of its successive calls to the RECURSE subroutine.

The main difficulty we face in deriving an upper bound for the regret of RECURSE is that a successful execution of the communication protocol may not imply the gap between the arms in $\mathcal{C}_1^1(t^1_{\text{first}}, 10)$ and those in $[K] \backslash \mathcal{C}_1^1(t^1_{\text{first}}, 10)$ equals $\max_i \Delta_{\sigma_i, \sigma_{i+1}}$. Nevertheless we can show they cannot be more than a constant multiple fraction apart,

**Lemma 22** *In the event the communication protocol succeeds in transmitting the composition of* $\mathcal{C}_1^1(t^1_{\text{first}}, 10)$ *to all players* $p \in [P]$. *If* $\tilde{\Delta} = \min_{\sigma_i \in \mathcal{C}_1^1(t^1_{\text{first}}, 10)} \mu_{\sigma_i} - \max_{\sigma_j \in [K] \backslash \mathcal{C}_1^1(t^1_{\text{first}}, 10)} \mu_{\sigma_j}$. *Then,*

$$\max_i \Delta_{\sigma_i, \sigma_{i+1}} \leq 3\tilde{\Delta}.$$

**Proof** If the communication protocol succeeded, then the sandwich property of Equation 4 holds for all $\sigma_i, \sigma_j$ and therefore,

$$\frac{s^1_{\text{first}}}{g(s^1_{\text{first}})} \in \left[\frac{128}{\tilde{\Delta}^2}, \frac{1152}{\tilde{\Delta}^2}\right]$$

Since the 'trigger' time $\bar{s}^1_{\text{first}}$ for $\max_i \Delta_{\sigma_i, \sigma_{i+1}}$ satisfies,

$$\frac{\bar{s}_{\text{first}}}{g(\bar{s}_{\text{first}})} \in \left[\frac{128}{\max_i \Delta^2_{\sigma_i, \sigma_{i+1}}}, \frac{1152}{\max_i \Delta^2_{\sigma_i, \sigma_{i+1}}}\right]$$

And by definition $s^1_{\text{first}} \leq \bar{s}^1_{\text{first}}$, we can conclude that $\frac{128}{\tilde{\Delta}^2} \leq \frac{1152}{\max_i \Delta^2_{\sigma_i, \sigma_{i+1}}}$ and therefore,

$$\max_i \Delta_{\sigma_i, \sigma_{i+1}} \leq 3\tilde{\Delta}.$$

■

Let's consider the set of consecutive gaps $\{\Delta_{\sigma_i, \sigma_{i+1}}\}_{i=1}^{K-1}$ and assume their ordering to be $\bar{\Delta}_1 \geq \cdots \geq \bar{\Delta}_{K-1}$ where $\bar{\Delta}_i = \Delta_{\sigma_{\ell(i)}, \sigma_{\ell(i)+1}}$ for some bijective mapping $\ell : [K-1] \to [K-1]$. The inverse mapping $\ell^{-1}(i)$ satisfies, $\Delta_{\sigma_i, \sigma_{i+1}} = \bar{\Delta}_{\ell^{-1}(i)}$.

For all $i \in [K]$ denote by $J_{\text{up}}(i)$ to the index,

$$J_{\text{up}}(i) = \arg\max\{j \text{ s.t. } 3\bar{\Delta}_j \geq \bar{\Delta}_i\}.$$

By definition $J_{\text{up}}(i) \geq i$. As an immediate consequence of Lemma 22, the first sub-problem the RECURSE algorithm will solve (in case the communication protocol was successful) will break the arm set through one of the gaps in $\{\Delta_{\sigma_{\ell(i)},\sigma_{\ell(i)+1}}, i \leq J_{\text{up}}(1)\}$. Furthermore Lemma 22 also implies the RECURSE algorithm will break the $\bar{\Delta}_1 = \Delta_{\sigma_{\ell(1))},\sigma_{\ell(1)+1}}$ gap in at most $J_{\text{up}}(1)$ recursive calls. In fact the same argument holds for all $i \in [K-1]$. The RECURSE algorithm will break the $\bar{\Delta}_i = \Delta_{\sigma_{\ell(i)},\sigma_{\ell(i)+1}}$ gap in at most $J_{\text{up}}(i)$ recursive calls.

After the RECURSE algorithm has successfully broken the $\Delta_{\sigma_M,\sigma_{M+1}}$ gap, the players will cease to experience any regret. As a result of the previous discussion this will happen in at most $J_{\text{up}}(\ell^{-1}(M))$ recursive calls. We are ready to bound the regret of the RECURSE Algorithm.

For any $\Delta > 0$ we define the partition regret function for gap $\Delta$, number of arms $\bar{K}$ and number of players $\bar{M}$ as,

$$\begin{aligned}\text{PartitionRegret}(\Delta, \bar{K}, \bar{M}, \delta) = \frac{746496\bar{M}(\bar{K}-\bar{M})\bar{K}}{\Delta}\log\left(\frac{746496\bar{M}\bar{K}}{\delta\Delta^2}\right) + \\ 20736(\bar{K}^2+\bar{K})\bar{M}(\bar{K}-\bar{M})\, B\left(\frac{186624}{\Delta^2}, \frac{\delta}{4\bar{K}^2\bar{M}}\right) + \\ \frac{20736(\bar{K}+1)\bar{M}B\left(\frac{186624}{\Delta^2}, \frac{\delta}{4\bar{K}^2\bar{M}}\right)}{\Delta} + \tilde{c}(\delta, \bar{K}, \bar{M})\end{aligned}$$

This is the parametric form of the upper bound in Equation 12. Recall that for any sub-problem of $\bar{K}$ arms the communication protocol succeeds with probability at least $\underbrace{1-\delta}_{\text{satisfying } \mathcal{E}} - \frac{\delta}{K}$ (see the discussion surrounding Equation 12). This concludes the proof of one of our main results.

**Theorem 23** *If $\delta \leq \frac{1}{162}$ with probability $1-\delta$ $\left(J_{\text{up}}(\ell^{-1}(M)) + \sum_{i=1}^{J_{\text{up}}(\ell^{-1}(M))} \frac{1}{i}\right)$ the regret of Algorithm 8 satisfies*

$$\text{RegretRECURSE}([K],[M]) \leq \sum_{i=1}^{J_{\text{up}}(\ell^{-1}(M))} \text{PartitionRegret}(\bar{\Delta}_i, \ell(i), M, \delta)$$

Using the definition $\delta = \frac{\xi}{2K}$ and the fact that PartitionRegret is monotonic w.r.t. $\frac{1}{\Delta}$, $\bar{K}$ and M as well as the inequalities $\ell(i) \leq K$ and $J_{\text{up}}(\ell^{-1}(M)) \leq M$ and $3\Delta_{J_{\text{up}}(\ell^{-1}(M))} \geq$ we can conclude the following,

**Corollary 24** *If $\frac{\xi}{2K} \leq \frac{1}{162}$ then with probability at least $1-\xi$ the regret of Algorithm 8 satisfies*

$$\begin{aligned}\text{RegretRECURSE}([K],[M]) &\leq K \cdot \text{PartitionRegret}\left(\frac{\Delta_{\sigma_M,\sigma_{M+1}}}{3}, K, M, \frac{\xi}{2K}\right) \\ &= \frac{3 \times 746496 M(K-M)K^2}{\Delta_{\sigma_M,\sigma_{M+1}}} \log\left(\frac{18 \times 746496 M K^2}{\xi \Delta^2_{\sigma_M,\sigma_{M+1}}}\right) + \\ &\quad 20736(K^3+K^2)M(K-M)B\left(\frac{9\times 186624}{\Delta^2_{\sigma_M,\sigma_{M+1}}}, \frac{\xi}{8K^3M}\right) + \\ &\quad \frac{20736(K^2+K)MB\left(\frac{9\times 186624}{\Delta^2_{\sigma_M,\sigma_{M+1}}}, \frac{\xi}{8K^3M}\right)}{\Delta_{\sigma_M,\sigma_{M+1}}} + K\tilde{c}\left(\frac{\xi}{2K}, K, M\right)\end{aligned}$$

Note that both $g(n)$ and $B(n,\delta)$ are of the order of $\widetilde{\mathcal{O}}(\log(n/\delta))$ where $\widetilde{\mathcal{O}}(\cdot)$ hides logarithmic factors in $K$ and $M$ only. By setting $\xi = \min\left(\frac{1}{T}, \frac{K}{81}\right)$ we can easily turn the results of Corollary 24 into the following Corollary that corresponds to the statement of Theorem 2,

**Corollary 25 (Main-Simplified)** *There exists a strategy such that the regret is upper bounded by:*

$$\mathcal{R}_T \leq \widetilde{\mathcal{O}}\left(\frac{M(K-M)K^2\log(T)}{\Delta_{\sigma_M,\sigma_{M+1}}} + \textbf{poly}(\log(T), K, M)\right),$$

*with probability at least $1-\min\left(\frac{1}{T}, \frac{K}{81}\right)$ where $\widetilde{\mathcal{O}}(\cdot)$ hides factors logarithmic in $M$ and $K$* ***only***[9].

Our results also imply anytime guarantees,

**Corollary 26 (Main-Simplified Anytime)** *Let $\delta \in (0,1)$. There exists a strategy such that the regret is upper bounded by:*

$$\mathcal{R}_t \leq \widetilde{\mathcal{O}}\left(\frac{M(K-M)K^2\log(t/\delta)}{\Delta_{\sigma_M,\sigma_{M+1}}} + \textbf{poly}(\log(t/\delta), K, M)\right),$$

*with probability at least $1-\delta$ for all $t \in \mathbb{N}$ and where $\widetilde{\mathcal{O}}(\cdot)$ hides factors logarithmic in $M$ and $K$* ***only***.

## Appendix C. Sharpening of the Zero Collision Reward Setting

In this section we describe a couple of Extensions of our main results.

9. More careful analysis may be possible that could ameliorate the dependence on the number of arms by at least one factor of $K$. This may be achieved by not upper bounding the RoundRobinRegret and PartitionRegret in each partition by those of the partition defined by $\Delta_{\sigma_M,\sigma_{M+1}}$. We leave this sharpening for future work.

### C.1. Problem independent Collision Regret

Recall the the logic behind why the ZeroTest works. At time $t^p_{\text{test}}$ (in our case equal to $t^1_{\text{comm1}}$ all players have access to a constant accuracy estimator for $\max_i \Delta_{\sigma_i,\sigma_{i+1}}$ (i.e. the empirical gap between the arms at the boundary of the two connected components of $\mathcal{C}^1_1(t^1_{\text{first}}, 10)$). Therefore they can estimate $\mu_{\sigma_1}$ up to a constant accuracy. By Freedman's inequality, testing at an accuracy of $c\mu_{\sigma_1}$ the mean of an arm $\widehat{\sigma}_1$ satisfying $\mu_{\widehat{\sigma}_1} > c'\mu_{\sigma_1}$ with $c' > c$ only requires $\widetilde{\mathcal{O}}\left(\frac{1}{\mu_{\sigma_1}}\right)$ samples (up to log factors) since the variance of $\widehat{\sigma}_1$ is upper bounded by $\mu_{\sigma_1}$.

Thus, it is enough for the ZeroTest to succeed to use $\widetilde{\mathcal{O}}\left(\frac{\text{poly}(K,M)\log(1/\delta)}{\mu_{\sigma_1}}\right)$ collisions to transmit each bit. Since each collision incurs in regret of at most $\mu_{\sigma_1}$, this implies the CollisionRegret can be upper bounded by a problem independent term of the form $\mathcal{O}(\text{poly}(K, M)\log(T))$.

This can be achieved by substituting the communication length function $f$ with an agreed estimator $t_{\text{communication}-\text{length}}$ of order $\approx 1/\mu_{\sigma_1}$. This can be agreed upon by all players using the same initial procedure as in the $\mu_{\text{collision}} > 0$ case, which would yield an estimator of $1/\mu^2_{\sigma_1}$ since $\Delta_{\text{collision}} = \mu_{\sigma_1}$ in the zero collision reward setting.

### C.2. Unknown number of players

Our algorithms work in the setting where each player has a known player index but does not know how many players may exist with a larger index. The RECURSE algorithm needs to be slightly modified. Instead of running RoundRobin on $\mathcal{C}^1_1(t^1_{\text{first}}, 10)$ if $|\mathcal{C}^1_1(t^1_{\text{first}}, 10)| \leq M$, the players will simply RECURSE and run the full communication protocol on the two sub-problems induced by $\mathcal{C}^1_1(t^1_{\text{first}}, 10)$ and $[K]\backslash\mathcal{C}^1_1(t^1_{\text{first}}, 10)$. Since all players are aware of their index, upon receiving $\mathcal{C}^1_1(t^1_{\text{first}}, 10)$ each of them can determine what sub-problem it is meant to play after a call to RECURSE, either $\mathcal{C}^1_1(t^1_{\text{first}}, 10)$ or the $[K]\backslash\mathcal{C}^1_1(t^1_{\text{first}}, 10)$. Within each of the sub-problems each player is perfectly capable of inferring if it should be the communicating player or not. The techniques we have used to derive the regret guarantees of Corollary 25 can be used to derive the same instance dependent logarithmic regret rate for this slightly more complex algorithm.

### C.3. Unknown lower bound for the collision reward

Our algorithms work in the setting where the collision reward is a random variable with mean $\mu_{\text{collision}} \in [0, 1]$ satisfying the condition $\mu_{\text{collision}} \leq \mu_{\sigma_K}$ and unknown to the learners in advance. Our algorithm will consist of an initial phase aimed at discovering an estimator for $\mu_{\sigma_1} - \mu_{\text{collision}}$. This phase, not present in the vanilla version of the algorithm discussed in the previous sections will be executed at the start of any RECURSE subroutine. The objective of this discovery phase is to ensure player 1 can convey the identity of a round value $t_{\text{collision}-\text{test}}$ to all remaining players $p \in \{2, \cdots, M\}$ that will be used to determine the length of the communication rounds in the second phase of the algorithm. During the second phase of the algorithm, the players will engage in a similar interaction as that described by Algorithms 4, 6 and 7 where instead of using $f$ to infer the length of the communication rounds, the players will use $t_{\text{collision}-\text{test}}$. Transmitting the identify of $t_{\text{collision}-\text{test}}$ from player 1 to all players $p \in \{2, \cdots, M\}$ is achieved via a bastardized version of the ZeroTest we outline below.

1. All players go through a modified version of the RoundRobin schedule that we'll call CollisionRoundRobin. Instead of cycling in batches of $K$ rounds, they will use a cycle

length of size $K+M$. Let's see how the very first such cycle works. All the remaining ones are a repetition of this basic structure. During the first $K$ rounds all players cycle through the $K$ arms in $[K]$ following the usual $\mathrm{RoundRobin}$ schedule. From round $K+1$ to $K+M$, players $\{2,\cdots,M\}$ will continue pursuing a traditional $\mathrm{RoundRobin}$ schedule while player 1 will instead pull arm $M$. All players $p\in[M]$ will build their empirical estimators of $\mu_1,\cdots,\mu_K$ using only the samples collected during the first $K$ rounds of each $K+M$ $\mathrm{CollisionRoundRobin}$ cycle. To estimate $\mu_{\mathrm{collision}}$ player 1 will use the samples collected at time $K+1$ of each $\mathrm{CollisionRoundRobin}$ cycle while players $p\in\{2,\cdots,M\}$ will use the samples collected at times $K+M-p+1$. All other samples will be discarded. If the number of players is unknown, we could extend the length of a $\mathrm{CollisionRoundRobin}$ cycle to be of size $2K$ instead.

2. Define $\widehat{\sigma}_1^p(t)=\arg\max_{i\in[K]}\widehat{\mu}_k^p(t)$ be player $p$'s guess for the largest arm during round $t$. Define as $\widehat{\mu}^p_{\mathrm{collision}}(t)$ to be the empirical estimator of $\mu_{\mathrm{collision}}$ by player $p\in[M]$ at time $t$. Let $t^p_{\mathrm{first-collision}}$ be the first special round $t$ such that when $I^p_{\widehat{\sigma}_1^p(t)}(t,10)\cap I^p_{\mathrm{collision}}(t,10)=\emptyset$. The same logic used to prove Equations 2 and 5 can be used to show that whenever $\mathcal{E}$ holds and for all $p\in[M]$,

$$\frac{128}{\left(\mu_{\sigma_1}-\mu_{\mathrm{collision}}\right)^2}\leq\frac{N^p_{\sigma_1}(t^p_{\mathrm{first-collision}})}{g(N^p_{\sigma_1}(t^p_{\mathrm{first-collision}})}\leq\frac{1152}{\left(\mu_{\sigma_1}-\mu_{\mathrm{collision}}\right)^2}\tag{13}$$

3. Define $t^1_{\mathrm{comm1-collision}}$ and $t^1_{\mathrm{comm-collision}}$ as functions of $t^1_{\mathrm{first-collision}}$ the same way as $t^1_{\mathrm{comm1}}$ and $t^1_{\mathrm{comm}}$ are functions of $t^1_{\mathrm{first}}$ in the previous sections and set the length of the communicating sequence starting at $t^1_{\mathrm{comm1-collision}}$ to be of size $t^1_{\mathrm{comm1-collision}}$ instead of $Kf(t^1_{\mathrm{comm1}})$. The listening protocol for players $p\in\{2,\cdots M\}$ remains unchanged except for the communication rounds length $f$. Player 1 will communicate a bit by pulling arm $\widehat{\sigma}^1_1=\widehat{\sigma}^1_1(t^1_{\mathrm{first-collision}})$ from round $t^1_{\mathrm{comm1-collision}}+1$ to round $2t^1_{\mathrm{comm1-collision}}$. The listening protocol for players $p\in\{2,\cdots,M\}$ remains mostly unchanged. All players $p\in\{2,\cdots,M\}$ will compute a set of large empirical reward arms $\widehat{\mathbf{MaxArms}}^p$ defined as ,

$$\begin{aligned}\widehat{\mathbf{MaxArms}}^p\leftarrow\Big\{&i\in[K]\text{ s.t.}\\&\widehat{\mu}^p_i(t^p_{\mathrm{listen-collision}})-\widehat{\mu}^p_{\mathrm{collision}}(t^p_{\mathrm{listen-collision}})\geq\\&\frac{1}{2}\max_{j\in[K]}\left(\widehat{\mu}^p_j(t^p_{\mathrm{listen-collision}})-\widehat{\mu}^p_{\mathrm{collision}}(t^p_{\mathrm{listen-collision}})\right)\Big\}.\end{aligned}$$

The players will then compute witness estimators $L^p_i(t^p_{\mathrm{listen-collision}})$ for all arms in $i\in\widehat{\mathbf{MaxArms}}^p$ defined as,

$$L^p_i(t^p_{\mathrm{listen-collision}})=\frac{\widehat{\mu}^p_i(t^p_{\mathrm{listen-collision}})-\widehat{\mu}^p_{\mathrm{collision}}(t^p_{\mathrm{listen-collision}})}{2}+\widehat{\mu}^p_{\mathrm{collision}}(t^p_{\mathrm{listen-collision}}).$$

4. A bit $b$ with value 1 is communicated by player 1 to make sure all players $p\in\{2,\cdots,M\}$ learn $t^1_{\mathrm{comm1-collision}}$. During the listening protocol all players $p\in\{2,\cdots,M\}$ will collect samples from $t^p_{\mathrm{listen-collision}}+1$ to $2t^p_{\mathrm{listen-collision}}$ following a $\mathrm{CollisionRoundRobin}$ schedule. These samples will be used by players $p\in\{2,\cdots,M\}$ as input to the $\mathrm{CollisionTest}$

to figure if player 1 has been pulling arm $\widehat{\sigma}_1^1$. Since $t^1_{\text{comm1}-\text{collision}} \geq 2t^1_{\text{comm}-\text{collision}} \geq 4t^1_{\text{first}-\text{collision}}$ the different $t^p_{\text{listen}-\text{collision}}$ times do not overlap with the sample collection for the CollisionZeroTest:

$$\begin{aligned}&\text{If } \exists i \in \widehat{\mathbf{MaxArms}}^p \text{ s.t. } \widehat{\mu}_i^p(t^p_{\text{listen}-\text{collision}}+1 : 2t^p_{\text{listen}-\text{collision}}) < L_i^p(t^p_{\text{listen}-\text{collision}}) :\\ &\quad \text{Return } b = 1\\ &\text{Else :}\\ &\quad \text{Return } b = 0\end{aligned} \tag{14}$$

If $\mathcal{E}$ holds, $t^1_{\text{comm1}-\text{collision}}$ equals the first $t^p_{\text{listen}-\text{collision}}$ that returns $b = 1$ for all $p \in [M]$. Once this signal has been received all players have shared knowledge of the value of $t^1_{\text{comm1}-\text{collision}}$.

The regret incurred during this phase of the algorithm is at most $2(\mu_{\sigma_1} - \mu_{\text{collision}})\, t^p_{\text{comm1}-\text{collision}}$. Since $t^p_{\text{comm1}-\text{collision}}$ satisfies Equation 13 the same argument as in Corollary 21 implies - ignoring any polynomial factors of $K$ and $M$ - the regret is upper bounded by $\mathcal{O}\left(\frac{\log(1/\delta)}{\mu_{\sigma_1}-\mu_{\text{collision}}}\right)$. Once having established shared knowledge of $t^1_{\text{comm1}-\text{collision}}$ among all players, the second phase of the algorithm starts. During this phase the player all players are to restart their empirical mean estimators for all arms although all listening players will keep a copy of their last witness values $L_i^p(t^p_{\text{comm1}-\text{collision}})$ for all arms in $i \in \widehat{\mathbf{MaxArms}}^p$. The second phase of the algorithm bears more resemblance with Algorithms 4, 6 and 7. It is designed to transmit the composition of $\mathcal{C}_1^1(t^1_{\text{first}}, 10)$. The players will start pulling arms following a traditional $\text{RoundRobin}-$schedule and follow the exact same logic as Algorithms 4 and 6 including the definitions of $t^p_{\text{first}}$, $t^1_{\text{comm}}$, $t^1_{\text{comm1}}$ and $t^p_{\text{listen}}$. The only difference is that instead of using $f$ to decide the length of the communication rounds, each bit is to be transmitted by player 1 using the same protocol described above where empirical estimators of the rewards of arms $i \in \widehat{\mathbf{MaxArms}}^p$ are compared with the witness values $L_i^p(t^p_{\text{comm1}-\text{collision}})$. Thus each bit transmission costs -ignoring polynomial factors of $K$ and $M$- at most $\mathcal{O}\left(\frac{\log(1/\delta)}{\mu_{\sigma_1}-\mu_{\text{collision}}}\right)$ regret. By the same argument as in the previous sections the lead-up to communicating $\mathcal{C}_1^1(t^1_{\text{first}}, 10)$ incurrs in regret of order $\mathcal{O}\left(\frac{\log(1/\delta)}{\max_i \Delta_{\sigma_i,\sigma_{i+1}}}\right)$. Since $\frac{1}{\mu_{\sigma_1}-\mu_{\text{collision}}} \leq \frac{1}{\max_i \Delta_{\sigma_i,\sigma_{i+1}}}$, we conclude the total regret -ignoring polynomial factors in $K$ and $M$ - up to the time all players have knowledge of $\mathcal{C}_1^1(t^1_{\text{first}}, 10)$ is upper bounded by $\mathcal{O}\left(\frac{\log(1/\delta)}{\max_i \Delta_{\sigma_i,\sigma_{i+1}}}\right)$. The polynomial factors in $K$ and $M$ remain the same as in the zero collision reward setting. We flesh out this strategy in more detail below.

### C.4. Detailed Analysis of Unknown Collision Reward

In this section we explore the setting where the collision reward is a random variable with a mean of $\mu_{\text{collision}}$ that is unknown to the learner. We assume that $\mu_{\text{collision}} \leq \mu_{\sigma_K}$. We focus on showing the CollisionTest works as intended. We will follow the analysis of the communication protocol in Section 4.2. Let $t^1_{\text{start}-\text{collision}}$ be the known start of the communication sequence.

Since $\max_i \Delta_{\sigma_i,\sigma_{i+1}} \leq \Delta_{\sigma_1,\sigma_K} \leq \mu_{\sigma_1} - \mu_{\text{collision}}$, the same logic used to prove Equations 2 and 5 can be utilized to prove that whenever $\mathcal{E}$ holds, in Phase 1 we have that

$$\frac{N^p_{\sigma_1}(t^1_{\text{first}-\text{collision}})}{g(N^p_{\sigma_1}(t^1_{\text{first}-\text{collision}}))} \geq \frac{128}{(\mu_{\sigma_1} - \mu_{\text{collision}})^2} \tag{15}$$

and in Phase 2 of the protocol,

$$\frac{N^p_{\sigma_1}(t^1_{\text{first}})}{g(N^p_{\sigma_1}(t^1_{\text{first}}))} \geq \frac{128}{\Delta^2_{\sigma_1,\sigma_K}} \geq \frac{128}{(\mu_{\sigma_1} - \mu_{\text{collision}})^2} \tag{16}$$

thus during Phase 1 and Phase 2,

$$\frac{N^p_{\sigma_1}(t^1_{\text{start}})}{g(N^p_{\sigma_1}(t^1_{\text{start}}))} \geq \frac{128}{(\mu_{\sigma_1} - \mu_{\text{collision}})^2} \tag{17}$$

Let's call $\widehat{\sigma}_1$ to player 1's guess for the largest arm at time $t^1_{\text{comm1}-\text{collision}}$. Equation 17 can be used to show that a similar but stronger set of properties as those presented at the beginning of Section 4.2 holds with high probability,

**1. If $t^1_{\text{start}} = t^1_{\text{comm1}-\text{collision}}$ then arm $\widehat{\sigma}_1$ has a large empirical mean for all players.** Arm $\widehat{\sigma}_1 \in \{i \in [K] \text{ s.t. } \widehat{\mu}^p_i(t^1_{\text{start}}) - \widehat{\mu}^p_{\text{collision}}(t^1_{\text{start}}) \geq \frac{1}{2}\max_{j\in[K]}\left(\widehat{\mu}^p_j(t^1_{\text{start}}) - \widehat{\mu}^p_{\text{collision}}(t^1_{\text{start}})\right)\}$ for all $p \in \{2, \cdots, M\}$ and $\widehat{\mu}^p_{\widehat{\sigma}_1}(t^1_{\text{start}}) - D(N_{\widehat{\sigma}_1}(t^1_{\text{start}})) \geq \frac{\mu_{\sigma_1} - \mu_{\text{collision}}}{2} + \mu_{\text{collision}}$.

- To see a formal proof of this statement refer to Lemma 27 in Appendix C.4. Set $\tilde{C} = 10$. This result makes sure that $\widehat{\sigma}_1 \in \widehat{\mathbf{MaxArms}}^p$ with high probability. As a side product of this result it is possible to show that $\mu_{\widehat{\sigma}_1} - \mu_{\text{collision}} \geq \frac{3}{4}(\mu_{\sigma_1} - \mu_{\text{collision}})$. In other words, the gap $\mu_{\widehat{\sigma}_1} - \mu_{\text{collision}}$ is at least a constant multiple of the gap $\mu_{\sigma_1} - \mu_{\text{collision}}$.

**2. Arm $\widehat{\sigma}_1$ is comparable to $\sigma_1$.** If $t^1_{\text{start}} = t^1_{\text{comm1}-\text{collision}}$ the witnesses $L^p_{\widehat{\sigma}_1}(t^1_{\text{start}})$ satisfy, $L^p_{\widehat{\sigma}_1}(t^1_{\text{start}}) \in \left[\frac{3(\mu_{\widehat{\sigma}_1} - \mu_{\text{collision}})}{7}, \frac{4(\mu_{\widehat{\sigma}_1} - \mu_{\text{collision}})}{7}\right] + \mu_{\text{collision}}$ for all $p \in \{2, \cdots, M\}$.

- To see a formal proof of this statement refer to Lemma 28 in Appendix C.4. Set $\tilde{C} = 10$. This results guarantees we can compute a 'witness' value that is a constant multiple of $\mu_{\sigma_1} - \mu_{\text{collision}}$ away from $\mu_{\sigma_1}$ and $\mu_{\text{collision}}$. Indeed it is easy to see $\mu_{\widehat{\sigma}_1} - \mu_{\text{collision}} \geq \frac{3}{4}(\mu_{\sigma_1} - \mu_{\text{collision}})$ (see Equation 21) implies $L^p_{\widehat{\sigma}_1}(t^1_{\text{start}}) \in \left[\frac{9(\mu_{\sigma_1} - \mu_{\text{collision}})}{28}, \frac{16(\mu_{\sigma_1} - \mu_{\text{collision}})}{28}\right] + \mu_{\text{collision}}$ for all $p \in \{2, \cdots, M\}$.

**3. When collisions are avoided with high probability the empirical mean estimators $\widehat{\mu}^p_{\widehat{\sigma}_1}(t^1_{\text{start}} + 1 : +2t^1_{\text{start}})$ are far from $\mu_{\text{collision}}$.** If $\mathcal{E}$ holds and

- $b = 1$, the estimators $\widehat{\mu}^p_{\widehat{\sigma}_1}(t^1_{\text{start}} + 1 : 2t^1_{\text{start}}) \leq L^p_{\widehat{\sigma}_1}$ for all $p \in [M]$.
- $b = 0$, the estimators $\widehat{\mu}^p_{\widehat{\sigma}_1}(t^1_{\text{start}} + 1 : 2t^1_{\text{start}}) > L^p_{\widehat{\sigma}_1}$ for all $p \in [M]$.

To prove the third item, notice that in case $b = 1$ is to be transmitted $\widehat{\mu}^p_{\widehat{\sigma}_1}(t^1_{\text{start}} + 1 : 2t^1_{\text{start}}) \leq \mu_{\text{collision}} + D(N_{\widehat{\sigma}_1}(t^1_{\text{start}})) \leq \mu_{\text{collision}} + \frac{\mu_{\widehat{\sigma}_1} - \mu_{\text{collision}}}{r-2}$ where $r = 2(\tilde{C} - 2)$ (see Equation 22 for a proof). Therefore $\widehat{\mu}^p_{\widehat{\sigma}_1}(t^1_{\text{start}} + 1 : 2t^1_{\text{start}}) \leq \mu_{\text{collision}} + \frac{\mu_{\widehat{\sigma}_1} - \mu_{\text{collision}}}{14} < \mu_{\text{collision}} + \frac{9(\mu_{\widehat{\sigma}_1} - \mu_{\text{collision}})}{28}$. Similarly in case $b = 0$, $\widehat{\mu}^p_{\widehat{\sigma}_1}(t^1_{\text{start}} + 1 : 2t^1_{\text{start}}) \geq \mu_{\text{collision}} - \frac{\mu_{\widehat{\sigma}_1} - \mu_{\text{collision}}}{14} > \mu_{\text{collision}} + \frac{16(\mu_{\widehat{\sigma}_1} - \mu_{\text{collision}})}{28}$.

Lemmas 27 and 28 comprise the bulk of the necessary steps to adapt our results where $\mu_{\text{collision}} = 0$ to the more general setting where $\mu_{\text{collision}} > 0$ and therefore these finalize the proof of Lemma 7. Let's assume the constant defining the $\tilde{C}-$blowup connectivity graph that defines time $t^1_{\text{first}}$ equals $\tilde{C}$. From Lemma 6 we can conclude that,

$$\frac{N^p_{\sigma_1}(t^1_{\text{first}})}{g(N^p_{\sigma_1}(t^1_{\text{first}})} \geq \frac{2(\tilde{C}-2)^2}{(\mu_{\sigma_1}-\mu_{\text{collision}})^2}. \tag{18}$$

And since $t^1_{\text{start}} \geq t^1_{\text{first}}$ and $t^1_{\text{first}} \geq Ks_{\text{boundary2}}$, the ratio $\frac{N^p_{\sigma_1}(t)}{g(N^p_{\sigma_1}(t))}$ is non decreasing for all $t \geq t^1_{\text{first}}$. Thus,

$$\frac{N^p_{\sigma_1}(t^1_{\text{start}})}{g(N^p_{\sigma_1}(t^1_{\text{start}})} \geq \frac{2(\tilde{C}-2)^2}{(\mu_{\sigma_1}-\mu_{\text{collision}})^2}. \tag{19}$$

**Lemma 27** *If $\mathcal{E}$ holds and $\tilde{C} \geq 9$,*

$$\widehat{\sigma}_1 \in \left\{ i \in [K] \text{ s.t. } \widehat{\mu}^p_i(t^1_{\text{start}}) - \widehat{\mu}^p_{\text{collision}}(t^1_{\text{start}}) \geq \frac{1}{2}\max_{j\in[K]}\left(\widehat{\mu}^p_j(t^1_{\text{start}}) - \widehat{\mu}^p_{\text{collision}}(t^1_{\text{start}})\right)\right\}$$

*for all $p \in \{2,\cdots,M\}$ and $\widehat{\mu}^p_{\widehat{\sigma}_1}(t^1_{\text{start}}) - D(N_{\widehat{\sigma}_1}(t^1_{\text{start}})) \geq \left(1-\frac{4}{r}\right)(\mu_{\sigma_1}-\mu_{\text{collision}})+\mu_{\text{collision}}$.*

**Proof**

By Equations 18 and 19, similar to the proof of Lemma 12 we conclude that $D(N_{\widehat{\sigma}_1}(t^1_{\text{start}})) = \sqrt{\frac{g(N^p_{\sigma_1}(t^1_{\text{start}}))}{2N^p_{\sigma_1}(t^1_{\text{start}})}} \leq \frac{\mu_{\sigma_1}-\mu_{\text{collision}}}{r}$ where $r = 2(\tilde{C}-2)$ if $\mathcal{E}$ holds, $\widehat{\mu}^p_i(t^1_{\text{start}}) \in [\mu_i - \frac{\mu_{\sigma_1}-\mu_{\text{collision}}}{r}, \mu_i + \frac{\mu_{\sigma_1}-\mu_{\text{collision}}}{r}]$ for all $i \in [K]\cup\{\text{collision}\}$ and $\widehat{\mu}^1_i(t^1_{\text{first}}) \in [\mu_i - \frac{\mu_{\sigma_1}-\mu_{\text{collision}}}{r}, \mu_i + \frac{\mu_{\sigma_1}-\mu_{\text{collision}}}{r}]$ for all $i \in [K]\cup\{\text{collision}\}$. These facts in conjunction with the definition of $\widehat{\sigma}_1$ imply,

$$\mu_{\sigma_1} - \frac{\mu_{\sigma_1}-\mu_{\text{collision}}}{r} \leq \widehat{\mu}^1_{\sigma_1}(t^1_{\text{first}}) \leq \widehat{\mu}^1_{\widehat{\sigma}_1}(t_{\text{first}^1}) \leq \mu_{\widehat{\sigma}_1} + \frac{\mu_{\sigma_1}-\mu_{\text{collision}}}{r}. \tag{20}$$

and therefore $\mu_{\sigma_1} \leq \mu_{\widehat{\sigma}_1} + \frac{2(\mu_{\sigma_1}-\mu_{\text{collision}})}{r}$. Thus,

$$\begin{aligned}
\widehat{\mu}^p_{\widehat{\sigma}_1}\left(t^1_{\text{start}}\right) - \widehat{\mu}^p_{\text{collision}}\left(t^1_{\text{start}}\right) &\geq \mu_{\widehat{\sigma}_1} - \frac{\mu_{\sigma_1}-\mu_{\text{collision}}}{r} - \frac{\mu_{\sigma_1}-\mu_{\text{collision}}}{r} - \mu_{\text{collision}} \\
&\geq \mu_{\sigma_1} - \frac{2(\mu_{\sigma_1}-\mu_{\text{collision}})}{r} - \frac{2(\mu_{\sigma_1}-\mu_{\text{collision}})}{r} - \mu_{\text{collision}} \\
&= (\mu_{\sigma_1}-\mu_{\text{collision}})\left(1-\frac{4}{r}\right)
\end{aligned}$$

Similarly for any $j \in [K]$,

$$\begin{aligned}
\widehat{\mu}^p_j(t^1_{\text{start}}) - \widehat{\mu}^p_{\text{collision}}(t^1_{\text{start}}) &\leq \mu_{\sigma_1} + \frac{\mu_{\sigma_1}-\mu_{\text{collision}}}{r} - \mu_{\text{collision}} + \frac{\mu_{\sigma_1}-\mu_{\text{collision}}}{r} \\
&\leq \mu_{\sigma_1} + \frac{2(\mu_{\sigma_1}-\mu_{\text{collision}})}{r} - \mu_{\text{collision}} \\
&\leq (\mu_{\sigma_1}-\mu_{\text{collision}})\left(1+\frac{2}{r}\right)
\end{aligned}$$

Since $\tilde{C} \geq 9$, $r \geq 14$, it follows that $\left(\frac{1}{2}+r\right) \leq \left(1-\frac{4}{r}\right)$ and therefore for any $j \in [K]$

$$\begin{aligned}\frac{\widehat{\mu}_j^p(t_{\text{start}}^1) - \widehat{\mu}_{\text{collision}}^p(t_{\text{start}}^1)}{2} &\leq (\mu_{\sigma_1} - \mu_{\text{collision}})\left(\frac{1}{2}+\frac{1}{r}\right) \leq (\mu_{\sigma_1} - \mu_{\text{collision}})\left(1-\frac{4}{r}\right) \\ &\leq \widehat{\mu}_{\widehat{\sigma}_1}^p\left(t_{\text{start}}^1\right) - \widehat{\mu}_{\text{collision}}^p\left(t_{\text{start}}^1\right)\end{aligned}$$

Thus implying the first result. The second statement is a result of the inequality,

$$\begin{aligned}\widehat{\mu}_{\widehat{\sigma}_1}^p(t_{\text{start}}^1) - D(N_{\widehat{\sigma}_1}(t_{\text{start}}^1)) &\geq \mu_{\widehat{\sigma}_1} - 2D(N_{\widehat{\sigma}_1}(t_{\text{start}}^1)) \\ &\geq \mu_{\sigma_1} - \frac{2(\mu_{\sigma_1} - \mu_{\text{collision}})}{r} - \frac{2(\mu_{\sigma_1} - \mu_{\text{collision}})}{r} \\ &= \mu_{\text{collision}} + (\mu_{\sigma_1} - \mu_{\text{collision}})\left(1-\frac{4}{r}\right)\end{aligned}$$

The result follows. ■

By Equation 20 in the proof of Lemma 27 we infer that:

$$\mu_{\widehat{\sigma}_1} \leq \mu_{\sigma_1} \leq \mu_{\widehat{\sigma}_1} + \frac{2\left(\mu_{\sigma_1} - \mu_{\text{collision}}\right)}{r}$$

Therefore the gap between $\mu_{\widehat{\sigma}_1}$ and $\mu_{\text{collision}}$ is lower bounded by,

$$\mu_{\widehat{\sigma}_1} - \mu_{\widehat{\sigma}_K^p} \geq (\mu_{\sigma_1} - \mu_{\text{collision}})\left(1-\frac{2}{r}\right). \tag{21}$$

Most importantly Equation 21 shows the gap $\mu_{\widehat{\sigma}_1} - \mu_{\text{collision}}$ is at least a constant multiple of the gap $\mu_{\sigma_1} - \mu_{\text{collision}}$.

**Lemma 28**

*The witnesses* $L_{\widehat{\sigma}_1}^p\left(t_{\text{start}}^1\right) = \frac{\widehat{\mu}_{\widehat{\sigma}_1}^p(t_{\text{start}}^1) - \widehat{\mu}_{\text{collision}}^p(t_{\text{start}}^1)}{2} + \widehat{\mu}_{\text{collision}}^p(t_{\text{start}}^1)$ *satisfy,*

$$L_{\widehat{\sigma}_1}^p\left(t_{\text{start}}^1\right) \in \mu_{\text{collision}} + \left[\left(\frac{1}{2}-\frac{2}{r-2}\right)\left(\mu_{\widehat{\sigma}_1} - \mu_{\text{collision}}\right), \left(\frac{1}{2}+\frac{2}{r-2}\right)\left(\mu_{\widehat{\sigma}_1} - \mu_{\text{collision}}\right)\right]$$

*for all* $p \in \{2, \cdots, M\}$ *where* $r = 2(\tilde{C}-2)$.

**Proof** By Equation 19, similar to the discussion above, $D(N_{\widehat{\sigma}_1}(t_{\text{start}}^1)) \leq \frac{\mu_{\sigma_1} - \mu_{\text{collision}}}{r}$ (for $r = 2(\tilde{C}-2)$) if $\mathcal{E}$ holds and therefore $\mu_{\sigma_1} \leq \mu_{\widehat{\sigma}_1} + \frac{2(\mu_{\sigma_1} - \mu_{\text{collision}})}{r}$. Hence,

$$\mu_{\sigma_1} - \mu_{\text{collision}} \leq \mu_{\widehat{\sigma}_1} + \frac{2(\mu_{\sigma_1} - \mu_{\text{collision}})}{r} - \mu_{\text{collision}}.$$

Thus implying,

$$\mu_{\sigma_1} - \mu_{\text{collision}} \leq \frac{r}{r-2}\left(\mu_{\widehat{\sigma}_1} - \mu_{\text{collision}}\right)$$

and therefore that

$$D(N_{\widehat{\sigma}_1}(t^1_{\text{start}})) \leq \frac{\mu_{\widehat{\sigma}_1} - \mu_{\text{collision}}}{r-2} \tag{22}$$

then,

$$\begin{aligned}\widehat{\mu}^p_{\widehat{\sigma}_1}(t^1_{\text{start}}) - \widehat{\mu}^p_{\text{collision}}(t^1_{\text{start}}) &\leq \mu_{\widehat{\sigma}_1} - \mu_{\text{collision}} + \frac{2(\mu_{\widehat{\sigma}_1} - \mu_{\text{collision}})}{r-2} \\ &= \left(1 + \frac{2}{r-2}\right)(\mu_{\widehat{\sigma}_1} - \mu_{\text{collision}})\end{aligned}$$

Similarly,

$$\begin{aligned}\widehat{\mu}^p_{\widehat{\sigma}_1}(t^1_{\text{start}}) - \widehat{\mu}^p_{\text{collision}}(t^1_{\text{start}}) &\geq \mu_{\widehat{\sigma}_1} - \mu_{\text{collision}} - \frac{2(\mu_{\widehat{\sigma}_1} - \mu_{\text{collision}})}{r-2} \\ &= \left(1 - \frac{2}{r-2}\right)(\mu_{\widehat{\sigma}_1} - \mu_{\text{collision}})\end{aligned}$$

Since $\mu_{\text{collision}} - \frac{\mu_{\widehat{\sigma}_1} - \mu_{\text{collision}}}{r-2} \leq \widehat{\mu}^p_{\text{collision}}(t^1_{\text{start}}) \leq \mu_{\text{collision}} + \frac{\mu_{\widehat{\sigma}_1} - \mu_{\text{collision}}}{r-2}$,

$$\begin{aligned}\mu_{\text{collision}} + \left(\frac{1}{2} - \frac{2}{r-2}\right)(\mu_{\widehat{\sigma}_1} - \mu_{\text{collision}}) &\leq \widehat{\mu}^p_{\text{collision}}(t^1_{\text{start}}) + \frac{\widehat{\mu}^p_{\widehat{\sigma}_1}(t^1_{\text{start}}) - \widehat{\mu}^p_{\text{collision}}(t^1_{\text{start}})}{2} \\ &\leq \mu_{\text{collision}} + \left(\frac{1}{2} + \frac{2}{r-2}\right)(\mu_{\widehat{\sigma}_1} - \mu_{\text{collision}})\end{aligned}$$

■

Finally the following 'inverted' version of Lemma 15 will prove useful,

**Lemma 29** *Let $\delta' \in (0,1)$. If $X$ is a random variable with support in $[0,1]$, distribution $\mathbb{P}_X$ and mean $\mu_X$ satisfying $\mu_X \leq U$, then with probability at least $1-\delta'$ for all $N$ such that $\frac{N}{B(N,\delta')} \geq \max\left(\frac{2}{U-\mu_X}, \frac{16\min(\mu_X, 1-\mu_X)}{(U-\mu_X)^2}\right)$ we have,*

$$\widehat{\mu}_X \leq U,$$

*where $B(n,\delta') = 2\log\log(2n) + \log\frac{5.2}{\delta'}$.*

**Proof**

Let $\alpha = \min(\mu_X, 1-\mu_X)$. A simple use of the reversed version of Lemma 36 implies that with probability at least $1-\delta'$ for all $n \in \mathbb{N}$:

$$\mu_X + 2\sqrt{\frac{\alpha B(n,\delta')}{n}} - \frac{B(n,\delta')}{n} \geq \widehat{\mu}_X.$$

The LHS of this inequality attains a value of at most $U$ whenever:

$$U - \mu_X \geq 2\sqrt{\frac{\alpha B(n,\delta')}{n}} - \frac{B(n,\delta')}{n}.$$

We finalize the proof by noting that for all $n$ such that $\frac{n}{B(n,\delta')} \geq \max\left(\frac{2}{U-\mu_X}, \frac{16\min(\mu_X,1-\mu_X)}{(U-\mu_X)^2}\right)$ we have that $\frac{U-\mu_X}{2} \geq 2\sqrt{\frac{\alpha B(n,\delta')}{n}}$ and $\frac{U-\mu_X}{2} \geq \frac{B(n,\delta')}{n}$. ■

### C.5. Complex Restart Strategy

Here we describe a way to reuse some of the collected samples so far and warm start the estimators. Let's assume $t^1_{\text{comm1}} = \min_{t\in\mathbb{N}} \text{ s.t. } \left\lfloor \frac{t/K}{g(t/K)} \right\rfloor = 9^{\tilde{w}}$ for some $\tilde{u}$ and define

$$t_{\text{restart}} = \min_{t\in\mathbb{N}} \text{ s.t. } \left\lfloor \frac{t/K}{g(t/K)} \right\rfloor = 9^{\tilde{w}-3}.$$

Let's see that whenever $\mathcal{E}$ holds $t^p_{\text{first}} > t_{\text{restart}}$ for all $p \in [M]$. Similar to the proof of Lemma 8 let's denote by $9^u$ the unique power of nine in the interval $\left[\frac{128}{\max_i \Delta^2_{\sigma_i,\sigma_{i+1}}}, \frac{1152}{\max_i \Delta^2_{\sigma_i,\sigma_{i+1}}}\right)$. Recall that whenever the good event $\mathcal{E}$ holds $\frac{s^p_{\text{first}}}{g(s^p_{\text{first}})} \in \left[\frac{128}{\max_i \Delta^2_{\sigma_i,\sigma_{i+1}}}, \frac{1152}{\max_i \Delta^2_{\sigma_i,\sigma_{i+1}}}\right)$ for all $p \in [M]$ and

$$t^1_{\text{comm1}} \in \left\{ \min_{t\in\mathbb{N}} \text{ s.t. } \left\lfloor \frac{t/K}{g(t/K)} \right\rfloor = 9^{u+1}, \min_{t\in\mathbb{N}} \text{ s.t. } \left\lfloor \frac{t/K}{g(t/K)} \right\rfloor = 9^{u+2} \right\}.$$

Since by definition $\tilde{w} - 3 < u$ it must be the case that $t_{\text{restart}} < t^p_{\text{first}}$ for all $p \in [M]$.

When jumping into these smaller problems, all players will warm-start their empirical reward estimators at $\{\hat{\mu}^p_i(t_{\text{restart}})\}_{i\in[K],p\in[M]}$, throwing away all the information gathered during the communication rounds.

Each player will now re-index time to suit the sub-problem it has landed on by throwing away the data corresponding to historical rounds where samples not belonging to the connected component assigned to her were collected from $t = 1$ to $t_{\text{restart}}$. This procedure ensures each sub-problem is at a state where there is no player for which the condition $\mathbf{conn}^p(sK, 5) \geq 2$ has been triggered, while ensuring a substantial proportion of the data collected so far can be reused.

Since the proportion of samples that can be reused using this strategy is constant, no substantial speedup can be gained from following this strategy.

## Appendix D. Missing Proofs

In this section we present the proofs of all those lemmas for which having the proof present in the main or the Appendix discussion section would have hindered the flow of the text.

### D.1. Proof of Lemma 6

We restate Lemma 6 for the reader's convenience.

**Lemma 30 (Confidence Bands)** *Let $\widehat{\mu}_{\sigma_i}(t)$ and $\widehat{\mu}_{\sigma_j}(t)$ be empirical estimators $\mu_{\sigma_i}$ and $\mu_{\sigma_j}$, each using $N(t)$ samples. Let $C > 3$ be a constant. If $t$ is the first special round such that*

$$\widehat{\mu}_{\sigma_i}(t) - \widehat{\mu}_{\sigma_j}(t) \geq CD(N(t)), \tag{3}$$

*then, whenever $\mathcal{E}$ holds, we have $\frac{\Delta_{\sigma_i,\sigma_j}}{2(C+2)} < D(N(t)) \leq \frac{\Delta_{\sigma_i,\sigma_j}}{C-2}$ and*

$$\frac{2(C-2)^2}{\Delta^2_{\sigma_i,\sigma_j}} \leq \frac{N(t)}{g(N(t))} < \frac{8(C+2)^2}{\Delta^2_{\sigma_i,\sigma_j}}. \tag{4}$$

**Proof** Notice that whenever $\mathcal{E}$ holds:

$$\widehat{\mu}_{\sigma_i}(t) - \widehat{\mu}_{\sigma_j}(t) \geq \Delta_{\sigma_i,\sigma_j} - 2D(N(t)).$$

Hence whenever $D(N(t)) \leq \frac{\Delta_{\sigma_i,\sigma_j}}{C+2}$,

$$\widehat{\mu}_{\sigma_i}(t) - \widehat{\mu}_{\sigma_j}(t) \geq \Delta_{\sigma_i,\sigma_j} - 2D(N(t)) \geq CD(N(t)),$$

and therefore condition 3 will trigger. This implies that special round $t-1$, being one before condition 3 is ever triggered must satisfy $D(N(t-1)) > \frac{\Delta_{\sigma_i,\sigma_j}}{C+2}$. Since $\delta < \frac{1}{2}$ Lemma 38 ensures that $D(N(t-1)) < 2D(N(t))$ and therefore that

$$D(N(t)) > \frac{\Delta_{\sigma_i,\sigma_j}}{2(C+2)}. \tag{23}$$

Similarly note that whenever $\mathcal{E}$ holds

$$\widehat{\mu}_{\sigma_i}(t) - \widehat{\mu}_{\sigma_j}(t) - 2D(N(t)) \leq \mu_{\sigma_i} + D(N(t)) - \mu_{\sigma_j} + D(N(t)) - 2D(N(t)) \leq \Delta_{\sigma_i,\sigma_j}.$$

and therefore if the condition $\widehat{\mu}_{\sigma_i}(t) - \widehat{\mu}_{\sigma_j}(t) \geq CD(N(t))$ was true, then,

$$(C-2)D(N(t)) \leq \widehat{\mu}_{\sigma_i}(t) - \widehat{\mu}_{\sigma_j}(t) - 2D(N(t)) \leq \Delta_{\sigma_i,\sigma_j}$$

Therefore,

$$D(N(t)) \leq \frac{\Delta_{\sigma_i,\sigma_j}}{C-2}. \tag{24}$$

We now turn our attention to lower and upper bounding $\widehat{\Delta}_{\sigma_i,\sigma_j}$. Since $\widehat{\Delta}_{\sigma_i,\sigma_j} = \widehat{\mu}_{\sigma_i}(t) - \widehat{\mu}_{\sigma_j}(t) - 2D(N(t))$ we can conclude that $\widehat{\Delta}_{\sigma_i,\sigma_j} \leq \Delta_{\sigma_i,\sigma_j}$. We use Equation 24 to produce a lower bound,

$$\widehat{\Delta}_{\sigma_i,\sigma_j} = \widehat{\mu}_{\sigma_i}(t) - \widehat{\mu}_{\sigma_j}(t) - 2D(N(t)) \geq \Delta_{\sigma_i,\sigma_j} - 4D(N(t)) \geq \frac{C-3}{C-2}\Delta_{\sigma_i,\sigma_j}.$$

Plugging in the definition of $D(N(t))$ and using the lower and upper bounds of equations 23 and 24 yields:

$$\frac{2(C-2)^2\log(4(N(t))^2MK/\delta)}{\Delta^2_{\sigma_i,\sigma_j}} \leq N(t) \leq \frac{8(C+2)^2\log(4(N(t))^2MK/\delta)}{\Delta^2_{\sigma_i,\sigma_j}}.$$

■

### D.2. Proof of Lemma 10

We restate Lemma 10 for readability.

**Lemma 31** *Let* $t^1_{\text{first}} = Ks^1_{\text{first}}$ *and* $t^1_{\text{comm1}} = Ks^1_{\text{comm1}}$. *If* $s^1_{\text{first}} \geq s_{\text{boundary2}}$ *and* $\delta \leq \frac{1}{162}$ *then* $s^1_{\text{comm1}} \leq 162s^1_{\text{first}}$ *and*

$$\frac{s^1_{comm1}}{g(s^1_{comm1})} \leq \frac{162s^1_{\text{first}}}{g(s^1_{\text{first}})}.$$

**Proof** Notice that by definition there is a $u \in \mathbb{N}$ such that,

$$9^{u-1} < \left\lfloor \frac{s^1_{\text{first}}}{g(s^1_{\text{first}})} \right\rfloor \leq 9^u = \left\lfloor \frac{s^1_{\text{comm}}}{g(s^1_{\text{comm}})} \right\rfloor < 9^{u+1} = \left\lfloor \frac{s^1_{\text{comm1}}}{g(s^1_{\text{comm1}})} \right\rfloor \tag{25}$$

Since by assumption $\delta \leq \frac{1}{162}$ it is easy to see that $4\left(s^1_{\text{first}}\right)^2 MK/\delta \geq 162$, it follows that $g(162s^1_{\text{first}}) \leq 2g(s^1_{\text{first}})$. Thus by inequality 25,

$$\frac{s^1_{\text{comm1}}}{g(s^1_{\text{comm1}})} \leq 9^{u+1} + 81 = (9^{u-1}+1)\cdot 81 \leq 81\frac{s^1_{\text{first}}}{g(s^1_{\text{first}})} \leq \frac{162s^1_{\text{first}}}{g(162s^1_{\text{first}})}$$

Recall that by definition $s^1_{comm1}$ is the first integer such that $\left\lfloor \frac{s^1_{\text{comm1}}}{g(s^1_{\text{comm1}})} \right\rfloor = 9^{u+1}$. Since $D(s)$ is decreasing for all $s \geq s_{\text{boundary2}}$ and $s^1_{\text{first}}$ is assumed to be at least $s_{\text{boundary2}}$, we can conclude that $162s^1_{\text{first}} \geq s^1_{\text{comm1}}$. The first result follows. Since $g(s)$ is an increasing function we conclude that

$$\frac{162s^1_{\text{first}}}{g(162s^1_{\text{first}})} \leq \frac{162s^1_{\text{first}}}{g(s^1_{\text{first}})}$$

The second result follows. ■

### D.3. Proof of Lemma 11

We restate Lemma 11 for readability.

**Lemma 32** *If $\mathcal{E}$ holds, $s^1_{\text{first}} \geq s_{\text{boundary2}}$ and $\delta \leq \frac{1}{162}$ then*

$$s^1_{\text{comm1}} \leq \frac{746496}{\max_i \Delta^2_{\sigma_i,\sigma_{i+1}}} \log\left(\frac{746496MK}{\delta \max_i \Delta^2_{\sigma_i,\sigma_{i+1}}}\right)$$

**Proof** As a consequence of Lemma 10, we see that $\frac{s^1_{comm1}}{g(s^1_{comm1})} \leq \frac{162s^1_{\text{first}}}{g(s^1_{\text{first}})}$. If $\mathcal{E}$ holds, Equation 2 implies that $\frac{s^1_{comm1}}{g(s^1_{comm1})} \leq \frac{186624}{\max_i \Delta^2_{\sigma_i,\sigma_{i+1}}}$. Let $h(n) = \frac{n}{g(n)}$. Notice that $h'(n) = \frac{\log(4MKn/\delta)-1}{\log^2(4MKn/\delta)}$. Since $\delta < \frac{1}{162}$ we conclude that $h'(n) > 0$ for all $n \geq 1$. Since by definition both $s^1_{\text{comm1}}$ and $\frac{186624}{\max_i \Delta^2_{\sigma_i,\sigma_{i+1}}}$ are both at least 1, and $\frac{4MK}{\delta} \times \frac{186624}{\max_i \Delta^2_{\sigma_i,\sigma_{i+1}}} \geq 4$ for all, a simple use of Lemma 40 where $x = s^1_{\text{comm1}}, c = 4MK/\delta$ and $b = \frac{186624}{\max_i \Delta^2_{\sigma_i,\sigma_{i+1}}}$ implies that,

$$s^1_{\text{comm1}} \leq \frac{746496}{\max_i \Delta^2_{\sigma_i,\sigma_{i+1}}} \log\left(\frac{746496MK}{\delta \max_i \Delta^2_{\sigma_i,\sigma_{i+1}}}\right).$$

■

### D.4. Proof of Lemma 19

We restate Lemma 19 for readability.

**Lemma 33** *If $\mathcal{E}$ holds, $t_{\text{comm1}} \geq \max\left(t_{\text{boundary1}}, t_{\text{boundary3}}\right)$, $s^1_{\text{first}} \geq s_{\text{boundary2}}$ and $\delta \leq \frac{1}{162}$ then,*

$$f(t^1_{\text{comm1}}) \leq \frac{20736B\left(\frac{186624}{\max_i \Delta^2_{\sigma_i,\sigma_{i+1}}}, \frac{\delta}{4K^2M}\right)}{\max_i \Delta_{\sigma_i,\sigma_{i+1}}}. \tag{11}$$

**Proof**

$$\begin{aligned}
f(t^1_{\text{comm1}}) &\overset{(i)}{\leq} 48B\left(f(t^1_{\text{comm1}}), \frac{\delta}{4K^2M}\right)\sqrt{\frac{t^1_{\text{comm1}}/K}{2g(t^1_{\text{comm1}}/K)}} \\
&\overset{(ii)}{\leq} 48B\left(f(t^1_{\text{comm1}}), \frac{\delta}{4K^2M}\right)\sqrt{\frac{162t^1_{\text{first}}/K}{2g(t^1_{\text{first}}/K)}} \\
&\overset{(iii)}{\leq} 48B\left(f(t^1_{\text{comm1}}), \frac{\delta}{4K^2M}\right)\frac{\sqrt{1152 * 162}}{\max_i \Delta_{\sigma_i,\sigma_{i+1}}} \\
&= \frac{20736B\left(f(t^1_{\text{comm1}}), \frac{\delta}{4K^2M}\right)}{\max_i \Delta_{\sigma_i,\sigma_{i+1}}},
\end{aligned}$$

where inequality $(i)$ follows from elementary properties of $f(\cdot)$ (see Lemma 39 in Appendix F.2) along with the assumption $t^1_{\text{comm1}} \geq t_{\text{boundary1}}$, and $(ii)$ follows from Lemma 10 along with the assumptions $s^1_{\text{first}} \geq s_{\text{boundary2}}$ and $\delta \leq \frac{1}{162}$. Inequality $(iii)$ follows because $\mathcal{E}$ holds, $N(t^1_{\text{first}}) = t^1_{\text{first}}/K$ and Equation 2 implies,

$$\frac{N(t^1_{\text{first}})}{g(N(t^1_{\text{first}}))} = \frac{t^1_{\text{first}}/K}{g(t^1_{\text{first}}/K)} < \frac{1152}{\max_i \Delta^2_{\sigma_i,\sigma_{i+1}}}.$$

Finally since $t^1_{\text{comm1}} \geq t_{\text{boundary3}}$, and $B(n, \frac{\delta}{4K^2M})$ is an increasing function of $n$,

$$B(f(t^1_{\text{comm1}}), \frac{\delta}{4K^2M}) \leq B\left(s^1_{\text{comm1}}, \frac{\delta}{4K^2M}\right)$$

Finally, following the same argument from $(ii)$ and $(iii)$ above by applying Lemma 10 to the ratio $\frac{s^1_{\text{comm1}}}{g(s^1_{\text{comm1}})} \leq \frac{162s^1_{\text{first}}}{g(s^1_{\text{first}})}$ and using the fact that $\mathcal{E}$ holds (and therefore Equation 2) and that $B(n, \frac{\delta}{4K^2M})$ is increasing in $n$, we can conclude that

$$\begin{aligned}
B\left(s^1_{\text{comm1}}, \frac{\delta}{4K^2M}\right) &\leq B\left(\frac{186624}{\max_i \Delta^2_{\sigma_i,\sigma_{i+1}}} g(s^1_{\text{comm1}}), \frac{\delta}{4K^2M}\right) \\
&\leq B\left(\frac{186624}{\max_i \Delta^2_{\sigma_i,\sigma_{i+1}}} \log\left(\frac{MK186624}{\delta \max_i \Delta^2_{\sigma_i,\sigma_{i+1}}}\right), \frac{\delta}{4K^2M}\right)
\end{aligned}$$

■

## Appendix E. The Zero Test - Supporting Lemmas

In order to answer this question we will make use of the following Lemmas:

**Lemma 34** *If $X$ is a random variable with support in $[0,1]$ with mean $\mu_X$ then:* $\mathrm{Var}(X) \leq \mu_X(1-\mu_X)$.

**Proof** By definition $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mu_X^2$. Since $X \in [0,1]$ then $\mathbb{E}[X^2] \leq \mathbb{E}[X]$. The result follows. ■

**Lemma 35** *[Uniform empirical Bernstein bound] In the terminology of Howard et al. (2018), let $S_n = \sum_{i=1}^n Y_i$ be a sub-$\psi_P$ process with parameter $c > 0$ and variance process $W_n$. Then with probability at least $1-\delta'$ for all $n \in \mathbb{N}$.*

$$S_n \leq 1.44\sqrt{(W_n \vee m)\left(1.4\log\log\left(2\left(\frac{W_n}{m}\vee 1\right)\right) + \log\frac{5.2}{\delta'}\right)} + 0.41c\left(1.4\log\log\left(2\left(\frac{W_n}{m}\vee 1\right)\right) + \log\frac{5.2}{\delta'}\right),$$

*where $m > 0$ is arbitrary but fixed.*

**Proof** Setting $s = 1.4$ and $\eta = 2$ in the polynomial stitched boundary in Equation (10) of Howard et al. (2018) shows that $u_{c,\delta'}(v)$ is a sub-$\psi_G$ boundary for constant $c$ and level $\delta$ where

$$u_{c,\delta'}(v) = 1.44\sqrt{(v\vee 1)\left(1.4\log\log\left(2(v\vee 1)\right) + \log\frac{5.2}{\delta'}\right)} + 1.21c\left(1.4\log\log\left(2(v\vee 1)\right) + \log\frac{5.2}{\delta'}\right).$$

By the boundary conversions in Table 1 in Howard et al. (2018) $u_{c/3,\delta'}$ is also a sub-$\psi_P$ boundary for constant $c$ and level $\delta'$. The desired bound then follows from Theorem 1 by Howard et al. (2018). ■

We now apply the results of Lemma 35 to a random variable $X$ satisfying the assumptions of Lemma 34:

**Lemma 36** *Let $\delta' \in (0,1)$. If $X$ is a random variable with support in $[0,1]$ with mean $\mu_X$ and law $\mathbb{P}_X$ and let $\{X_i\}_{i=1}^\infty$ be i.i.d. samples from $\mathbb{P}_X$, then with probability at least $1-\delta'$ and for all $n \in \mathbb{N}$ simultaneously:*

$$\mu_X - 2\sqrt{\frac{\min(\mu_X, 1-\mu_X)B(n,\delta')}{n}} - \frac{B(n,\delta')}{n} \leq \frac{1}{n}\sum_{i=1}^n X_i,$$

*where $B(n,\delta') = 2\log\log(2n) + \log\frac{5.2}{\delta'}$.*

**Proof** Consider the martingale difference sequence $Y_i = X_i - \mu_X$. The process $S_n = \sum_{i=1}^n Y_i$ with variance process $W_n = n\mathrm{Var}(X)$ satisfies the sub-$\psi_P$ condition of Howard et al. (2018) with constant $c = 1$ (see Bennett case in Table 3 of Howard et al. (2018)). By Lemma 35 the bound:

$$S_n \leq 1.44\sqrt{(W_t \vee m)\left(1.4\log\log\left(2\left(\frac{W_t}{m}\vee 1\right)\right) + \log\frac{5.2}{\delta'}\right)} + 0.41c\left(1.4\log\log\left(2\left(\frac{W_t}{m}\vee 1\right)\right) + \log\frac{5.2}{\delta'}\right)$$

holds for all $n \in \mathbb{N}$ with probability at least $1-\delta'$. Observe that as a consequence of Lemma 34, the variance process $W_n$ satisfies $W_n \leq n\min(\mu_X, 1-\mu_X)$. If we set $m = t\mu_X$, we can futher upper bound the RHS as:

$$S_n \leq 1.44\sqrt{n\min(\mu_X,1-\mu_X)\left(1.4\log\log(2n) + \log\frac{5.2}{\delta'}\right)} + 0.41\left(1.4\log\log(2n) + \log\frac{5.2}{\delta'}\right).$$

The result follows. ■

## Appendix F. Ancillary Technical Lemmas

### F.1. Properties of $D(\cdot)$

**Lemma 37** *The function $D : \mathbb{R} \to \mathbb{R}$ defined as $D(\ell) = \sqrt{\frac{2g(\ell)}{\ell}}$ for $g(\ell) = \log\left(4\ell^2 MK/\delta\right)$ is increasing for $\ell \geq 1$ whenever $\delta < \frac{1}{2}$.*

**Proof** Let $c' = 4MK/\delta$ and consider the function $h(\ell) = \frac{\log(c'\ell^2)}{\ell}$. The derivative of $h$ equals

$$h'(\ell) = \frac{2 - \log(c'\ell^2)}{\ell^2}.$$

Therefore $h'(\ell) \leq 0$ iff $2 \leq \log(c'\ell^2)$, which holds iff $\exp(2) \leq c'\ell^2$. As long as $\delta < \frac{1}{2}$, the constant $c' > \exp(2)$ which implies the result. ■

**Lemma 38** *For any $\ell \geq 1$, and whenever $\delta < \frac{1}{2}$ the function $D(\cdot)$ doesn't decrease too fast:*

$$2D(\ell+1) > D(\ell)$$

**Proof** Observe that $\log(4\ell^2 ML/\delta) \leq \log(4(\ell+1)^2 ML/\delta)$ since $\log(\cdot)$ is an increasing function. Similarly for all $\ell \geq 1$ we have that $\sqrt{\frac{1}{\ell}} \leq \sqrt{\frac{2}{\ell+1}} < 2\sqrt{\frac{1}{\ell+1}}$. Therefore:

$$D(\ell) = \sqrt{\frac{2\log(4\ell^2 ML/\delta)}{\ell}} < 2\sqrt{\frac{2\log(4(\ell+1)^2 ML/\delta)}{\ell+1}} = 2D(\ell+1).$$

■

### F.2. Properties of $f(\cdot)$

Let's tart by showing that $f(n)$ can be upper-bounded.

**Lemma 39** *If* $\frac{f(n)-1}{B\left(f(n)-1,\frac{\delta}{4K^2M}\right)} \geq 1$ *then*

$$\frac{f(n)}{B(f(n),\frac{\delta}{4K^2M})} \leq 48\sqrt{\frac{n/K}{2Kg(n/K)}}$$

**Proof** By definition of $f(n)$,

$$\frac{f(n)-1}{B(f(n)-1,\frac{\delta}{4K^2M})} < 24\sqrt{\frac{n/K}{2Kg(n/K)}} \leq \frac{f(n)}{B(f(n),\frac{\delta}{4K^2M})}$$

Since $B(f(n),\frac{\delta}{4K^2M}) \geq 1$ and $B(f(n)-1,\frac{\delta}{4K^2M}) \leq B(f(n),\frac{\delta}{4K^2M})$,

$$\frac{f(n)}{B\left(f(n),\frac{\delta}{4K^2M}\right)} - \frac{f(n)-1}{B\left(f(n)-1,\frac{\delta}{4K^2M}\right)} \leq \frac{f(n)}{B\left(f(n),\frac{\delta}{4K^2M}\right)} - \frac{f(n)-1}{B\left(f(n),\frac{\delta}{4K^2M}\right)} \leq 1$$

We can conclude that

$$\frac{f(n)}{B\left(f(n),\frac{\delta}{4K^2M}\right)} \leq \frac{f(n)-1}{B\left(f(n)-1,\frac{\delta}{4K^2M}\right)} + 1 \overset{(i)}{\leq} 2\frac{f(n)-1}{B\left(f(n)-1,\frac{\delta}{4K^2M}\right)} \leq 48\sqrt{\frac{n/K}{2Kg(n/K)}}.$$

Inequality $(i)$ holds because by assumption $\frac{f(n)-1}{B\left(f(n)-1,\frac{\delta}{4K^2M}\right)} \geq 1$. ■

### F.3. Miscellaneous

The following lemma will prove useful in upper bounding $s^1_{\text{comm1}}$.

**Lemma 40** *Let* $h(x) = \frac{x}{\log(cx)}$ *for* $c > 0$. *Let* $x_0$ *be the first positive real number such that*[10] *for all* $x' \geq x_0$, $h'(x) \geq 0$. *Let* $x, b \geq x_0$ *such that* $h(x) = \frac{x}{\log(cx)} \leq b$ *and* $cb \geq 4$ *then*

$$x \leq 4b\log(cb)$$

**Proof** We shall show the desired result by the way of contradiction. Let $x' \geq x_0$ be such that $x' > 4b\log(cb)$. The following inequalities hold

$$\begin{aligned}\frac{x'}{\log(cx')} &\overset{(i)}{\geq} \frac{4b\log(cb)}{\log(4cb\log(cb))} \\ &= \frac{4b\log(cb)}{\log(cb)+\log(4)+\log(\log(cb))} \end{aligned} \tag{26}$$

10. It is easy to see that $h'(x) = \frac{\log(cx)-1}{\log^2(cx)}$ so that $x_0 = \frac{e}{c}$.

Inequality $(i)$ holds because we have assumed $cb \geq 4 > 3$ and therefore $\log(cb) \geq 1$, $b \geq x_0$ and therefore that $4b\log(cb) \geq x_0$. Since $cb \geq 4 > 1$, it follows that $\log(cb) \geq \log(\log(cb))$. We can upper bound the denominator of 26 by $3\log(cb)$ thus,

$$\frac{x'}{\log(cx')} \geq \frac{4b\log(cb)}{3\log(cb)} > b \tag{27}$$

Since $x$ is assumed to satisfy $\frac{x}{\log(cx)} \leq b$ this concludes the proof. ■