Collaborative Exploration and Exploitation in massively Multi-Player Bandits

Hiba Dakdouk, Raphaël Féraud, Romain Laroche, Nadège Varsier, Patrick Maillé

Collaborative Exploration and Exploitation in massively Multi-Player Bandits

Hiba Dakdouk
Orange Labs

Raphaël Féraud
Orange Labs

Romain Laroche
Microsoft Research

Nadège Varsier
Orange Labs

Patrick Maillé
IMT Atlantique

October 8, 2021

Abstract

In this paper, we propose an approach to optimize the performance of Internet of Things (IoT) networks. We formulate the optimization problem as a massive multi-player multi-armed bandit problem, where the devices are the players and the radio channels are the arms, with collisions possibly preventing message reception. For handling a realistic IoT network, we do not assume that *sensing information is available* (i.e. that the collision are observed) or that *the number of players is smaller than the number of arms*. As the optimization problem is intractable, we propose two greedy policies: the first one focusing on the number of successful communications, while the second one also takes into account fairness between players. In order to implement an approximation of the targeted policies, we propose an *explore-then-exploit* approach, and establish a regret lower bound in $\Omega\left(T^{2/3}\left(\frac{\log T}{N}+K^{3/2}\right)\right)$. For estimating the mean reward of arms, we propose a decentralized exploration algorithm with controlled information exchanges between players. Then we state that the regret of the estimated target policy is optimal with respect to the time horizon T. Finally, we provide some experimental evidences that the proposed algorithms outperform several baselines.

1 Introduction

Optimization of IoT wireless network performance In Internet of Things (IoT) networks, a large number of devices is connected to the Internet through wireless gateways. The communication protocols used in IoT, such as LoRa, allow to evaluate a binary reward for each transmission (success or not) since each uplink transmission is followed by a downlink time windows where the node listens to the gateway to receive the acknowledgement of the uplink transmission. The frequency of sending messages through the gateway depends on the application (healthcare, security, smart cities, marketing, gaming, home automation...). Moreover, for several real-time applications, the device has to send a packet when an unknown and uncontrolled event occurs. For instance, the device can interact with its environment in real-time, to get a green light when there is no car at a crossroad, an ad when the user is in front of a shop, a ticket when getting on the bus... Which is why in the following we assume that each player has a probability of sending a message at each time step.

We consider a large number N of devices communicating with a unique gateway on a limited number K of orthogonal channels ($N \geq K$), using an acknowledgement protocol slotted in time. At each time slot, each device n has a probability p_n to send a message to the gateway. When two or more devices send a message to the gateway at the same time slot through the same channel, an *internal collision* occurs. Hence, due to local interference the gateway does not receive any of the messages from the colliding devices and does not send the acknowledgements to the devices. Moreover, collisions with other types of networks may occur, so that even if only one player sends a message in one channel at a given time slot, the message may not be received by the gateway. These *external collisions* make the probabilities of successful transmission (and hence the channels' qualities) different for each channel. When an *internal or external collision* occurs, we consider that the message is simply lost. We do not consider collisions when the gateway sends acknowledgements, since these downlink collisions necessitate

that at least two acknowledgements are sent from the gateway at the same time to devices which are located at the same place. This is impossible if there is a unique gateway using a protocol slotted in time (as we assume), and unlikely if there are more gateways. Another important feature of the studied problem is that only the devices can estimate the quality of channels, since the gateway cannot know that messages have been sent by some devices if a collision occurs. As a consequence, the estimation of the channel quality has to be done in a decentralized way.

To set this scenario into the framework of multi-player multi-armed bandits, each device is considered as a player, a communication channel is considered as an arm, and the reward corresponds to the reception or not of the acknowledge from the gateway.

Related Work The *decentralized multi-player multi-armed bandits* have been studied for opportunistic spectrum access in [Liu and Zhao, 2010, Anandkumar et al., 2010, Avner and Mannor, 2014, Nayyar et al., 2015]. In opportunistic spectrum access, primary users have a strict priority over secondary users, which are allowed *sensing* a channel before sending a packet in order to check that it is free. The objective of those works is to avoid collisions between concurrent secondary users, that share the same channels, while choosing the best channels, i.e., with the highest probabilities to be free of primary users. This line of work makes the assumption that there are less players than channels, that the collisions with other players are observed, and uses orthogonalization techniques to avoid collisions. In [Rosenski et al., 2016], the authors propose to use collisions to estimate in a first phase the number of players and the value of arms and then a Musical Chair approach to allocate each player on a different N-best arm. In [Hanawal and Darak, 2018], the authors improve this approach by reducing the first phase to the estimation of the value of arms and then use a trekking approach to allocate each player on a different N-best arm without the knowledge of the number of players. In [Boursier and Perchet, 2019], the authors propose a communication protocol based on controlled collisions that achieves almost the same performance as a centralized algorithm. In [Wang et al., 2020], the authors improve this result by electing a leader that explores the arms and allocates other players on different estimated N-best arms. The leader communicates to other players the list of estimated N-best arms when it changes using the same communication protocol as in [Boursier and Perchet, 2019]. This algorithm is asymptotically optimal. An interesting extension of the problem setting was proposed in [Boursier et al., 2020] for handling the case where the mean rewards of arms are not the same for each player. However, this line of works makes the assumption that *sensing* information is available and the number of players is small ($N \leq K$), which are both unrealistic assumptions for IoT networks.

In [Boursier and Perchet, 2019], the authors propose an adaptation of their algorithm to the case where *sensing* is not allowed, that preserves the logarithmic behavior with respect to the time horizon. This approach has been improved in [Shi et al., 2020] for the case where sensing is not allowed thanks to the use of *Z-channel coding*, quantization of transmitted statistics and a tree structured communication, where a leader gathers the statistics and then decides for all players which is the best set of arms. In [Lugosi and Mehrabian, 2018], the authors define the multi-player stochastic multi-armed bandit as an anti-coordination game, where the goal is to quickly reach an approximate Nash equilibrium. Finally, in [Bubeck et al., 2019] the difficult case of non-stochastic multi-player multi-armed bandits is addressed. In all those works, the number of players is assumed to be below the number of channels, which is not realistic for IoT networks.

Motivated by IoT networks, in [Bonnefoi et al., 2018, Besson and Kaufmann, 2018] the authors propose a new problem setting where *sensing* is not allowed, the number of players is larger than the number of channels, and the players asynchronously play: each player has the same probability to send a packet at each time slot. The authors show experimentally that *selfish UCB*, which consists in each player independently playing *UCB* [Auer et al., 2002a], works surprisingly well. This experimental result has been confirmed in the case of LoRa networks using stochastic and non-stochastic multi-armed bandits [Kerkouche et al., 2018] or in the case of IEEE 802.15.4 time-slotted channel hopping protocol [Dakdouk et al., 2018]. Despite its good experimental performance, this algorithm has no theoretical guarantees, and it has been shown that *selfish UCB* can fail badly on some cases [Besson and Kaufmann, 2018]. With a similar problem setting but with different probabilities to send packets the authors in [Dakdouk et al., 2020] propose a cooperative algorithm that aims to find a set of optimal arms while minimizing the number of plays. However that work does not optimize the number of optimal arms to find, and the exploitation policy followed by the players is uniform, which is clearly sub-optimal. Those lacks are overcome with our proposed policies.

Finally, the optimization problem we propose to solve is related to Aloha protocol [Bertsekas et al.,

1992], where each player n transmits a packet with a probability p_n in a slot. For instance in [Wang and Kar, 2004], the authors formulate the decentralized throughput maximization problem in an Aloha network with a single channel in a way that is close to our optimization problem. However that work considers a single channel, and the decision variable is the sending probability p_n rather than the choice of the channel. If the probabilities of sending a message are optimized, then the application constraints of IoT (frequency of sending messages or real-time messages) cannot be respected. In [Cohen et al., 2013], the authors propose a best-response algorithm which solves the throughput maximization problem for multi-channels Aloha protocol. They notably show that the best-response algorithm converges to a Nash Equilibrium in a finite time. However they consider that the channel capacities and the strategies of other players are known, and that each player has the same probability to send a message at each slot, which is unrealistic and restrictive for IoT networks.

Contributions and paper organization In this paper, with the aim of optimizing transmissions in IoT networks, we study the extension of the problem proposed in [Bonnefoi et al., 2018], where each player has a different probability to send a packet at a each time slot [Dakdouk et al., 2020]. We propose an *explore-then-exploit* approach, where a decentralized exploration algorithm outputs an estimation of the parameters, and then a target policy, which can be computed by the gateway, is used during the exploitation phase.

The remainder of this paper is organized as follows. In section 2, we formalize the objective of optimizing the successful transmissions. We show in Theorem 1, that there exists a deterministic policy (an assignment of devices over channels) that is optimal. Then we propose two deterministic target greedy policies: `DORG` (decreasing-order-reward-greedy) optimizes the number of successful transmissions, while `DOFG` (decreasing-order-fair-greedy) guarantees some fairness between players, which is defined as the ratio between the lowest successful transmission rate and the highest successful transmission rate obtained over the players. In Theorem 2, we show that `DORG` is an optimal policy in the setting proposed in [Bonnefoi et al., 2018] (when $\forall n$, $p_n = p$), while in Theorem 3 we show that `DOFG` is fair.

In section 3, we state a first regret lower bound in $\Omega\left(T^{2/3}\left(\log T/N + K^{3/2}\right)\right)$, which holds for any *explore-then-exploit* policy. This notably shows that the problem is much more difficult than standard multi-armed bandits for which the regret lower bound is in $\Omega(\sqrt{KT})$ [Bubeck and Cesa-Bianchi, 2012].

Since the collisions can only be observed by devices, in section 4 we propose a decentralized exploration algorithm that outputs with high probability an approximation of the mean rewards of arms, i.e., the channel qualities. We allows players to collaborate by exchanging messages using the IoT protocol. We provide a deep analysis of the proposed algorithms. Theorem 6 states an upper bound of the number of time steps needed to output a controlled approximation of the arms that is near optimal in comparison to the lower bound of K biased coin estimations in $\Omega\left(K/\epsilon^2 \log 1/\delta\right)$ [Anthony and Bartlett, 1999]. Theorem 5 states an upper bound of the communication cost in $O(N)$, which is order optimal. Then, in the setting proposed in [Bonnefoi et al., 2018], Theorem 7 states that when using `DORG`, the proposed algorithm benefits from a regret upper bound that is optimal with respect to T, and near optimal with respect to K. Finally, Theorem 8 shows that when using approximation of mean rewards, `DOFG` is still fair.

In section 5, we compare our approach with state-of-the-art methods on a large set of synthetic problems. We show that when using `DORG` the proposed algorithm outperforms the baselines in terms of successful communication rate, and when using `DOFG` it outperforms them in terms of fairness between players. The reader will find additional experiments in appendix B and the proofs in appendix C.

2 Optimizing transmissions in IoT networks

2.1 Problem Formulation

Let $[N]$ be a set of N players, such that at each time slot t each player $n \in [N]$ has a constant probability p_n to send a message, such that $1 > p > p_n > 0$, where p is the duty cycle that is imposed to the IoT network in order to share the free bandwidth with other users. Without loss of generality, in the following we assume that: $p_1 \geq ... \geq p_N$. At each time slot t, the set $\mathcal{N}_t$ of players sending messages is selected by N independent Bernoulli samples: $\mathcal{N}_t := \{n \in [N] \text{ such that } a_n = 1, \text{ with } a_n \sim \mathcal{B}(p_n)\}$. Let $[K]$ be the set of K arms. The transmission of a message is successful if it does not collide with other messages. The random variable representing an external collision on arm k is denoted by

$E^k \sim \mathcal{B}(\theta^k)$ (equals 0 if collision, 1 otherwise). Similarly, internal collisions between the controlled players are represented by the random variables $(I^k)_{k\in[K]}$ (equals 0 if collision, 1 otherwise) and depend on the implemented policy. After playing arm k, player n observes the binary outcome $Y^{k_n} = E^{k_n} I^{k_n}$, i.e., knows whether a collision occurred or not (through an acknowledgement message) but cannot distinguish external and internal collisions.

We assume that the number of players is known by the gateway, which is realistic in IoT protocols, and that the gateway sends this information to each player at the beginning of the game. Considering that the probability of sending a message depends mainly on the type of devices, we assume that each player knows its own probability of sending a message p_n. We recall that no assumption is made on the number N of players, which means that the number of players may be higher that the number K of arms.

We will call a *policy* a (possibly randomized) way for players to select the channel to use for their next transmission. Formally, a policy π will be a vector of probability distributions over the set of arms: $\pi = (\pi_1, ..., \pi_N)$, with $\pi_n = (\pi_n^1, ..., \pi_n^K)$, where π_n^k denotes the probability that player n chooses arm k for sending a message. We denote by $\mu_{n,\boldsymbol{\theta}}^k(\pi)$ the expected reward in model $\boldsymbol{\theta} = \{\theta^1, ..., \theta^K\}$ of playing arm k while the other players follow policy π. It is the probability that no external collision occurs times the probability that no internal collision occurs, or mathematically

$$\mu_{n,\boldsymbol{\theta}}^k(\pi) = \theta^k \prod_{n'=1, n'\neq n}^{N} (1 - p_{n'}\pi_{n'}^k). \tag{1}$$

Equation (1) shows the difficulty of the studied problem: the mean reward of an arm for a given player depends on the probabilities of the other players to send a message and on the policies they follow. The aggregated average reward in model $\boldsymbol{\theta} = \{\theta^1, ..., \theta^K\}$ per time slot over all players $\mu_{\boldsymbol{\theta}}(\pi)$ is:

$$\mu_{\boldsymbol{\theta}}(\pi) = \sum_{k=1}^{K} \theta^k \sum_{n=1}^{N} p_n.\pi_n^k \prod_{n'\in[N]\setminus\{n\}} (1 - p_{n'}\pi_{n'}^k). \tag{2}$$

This performance metric corresponds to the expected number of successful uplink transmissions per time slot in the IoT network. This optimization problem with respect to π has a solution, since the objective function is continuous and the set of decision variables is compact. But the problem itself is not convex, hence classical convex optimization methods cannot be applied. While suspected, proving the NP-hardness of the problem remains an open question. Still, Theorem 1 states that at least one solution is a deterministic policy.

Theorem 1. *There exists a policy maximizing the overall network utility (equation* (2)*) that is deterministic.*

Based on Theorem 1, from now on, we only consider deterministic policies (each player sticks to an arm that it will play for sending messages) and write k_n the arm assigned to player n. The expected reward per time slot in model $\boldsymbol{\theta} = \{\theta^1, ..., \theta^K\}$ of a deterministic policy π can then be written as:

$$\begin{aligned} \mu(\pi) &= \sum_{n=1}^{N} p_n \theta^{k_n} \prod_{n'\neq n,\ \text{s.t.}\ k_{n'}=k_n} (1 - p_{n'}) \qquad (3) \\ &= \sum_{k=1}^{K} \theta^k \underbrace{\prod_{n\in[N],\ \text{s.t.}\ k_n=k} (1 - p_n)}_{z^k} \underbrace{\sum_{n\in[N],\ \text{s.t.}\ k_n=k} \frac{p_n}{1 - p_n}}_{\ell^k} \end{aligned}$$

where z^k is the probability that all players assigned to arm k do not send messages, and ℓ^k is the sum of the activation odds for all players assigned to arm k.

2.2 Reward greedy algorithm

In this section, we propose a greedy algorithm that aims to maximize the network utility (equation (2)).

Lemma 1. *For a deterministic policy π, let $\mu(\pi[n])$ denote the expected reward when only players $1, ..., n$ are playing (all players $n' > n$ are deactivated). Then we have the recursive expression:*

Algorithm 1 Reward Greedy
(`DORG` if players are sorted in p_n decreasing order)

Inputs: $[K]$, $[N]$, $\{\theta^k\}_{k\in[K]}$, $\{p_n\}_{n\in[N]}$
Output: π
Init: per-arm inactivity probabilities: $z^k = 1$.
Init: per-arm activation odds sums: $\ell^k = 0$.

for $n = 1$ to N **do**
  Set $k_n \in \operatorname{argmax}_{k\in[K]} \theta^k z^k (1 - \ell^k)$.
  Update $z^{k_n} \leftarrow z^{k_n}(1 - p_n)$.
  Update $\ell^{k_n} \leftarrow \ell^{k_n} + \frac{p_n}{1-p_n}$.
  Set $\pi_n^{k_n} = 1$, and $\forall k \neq k_n, \pi_n^k = 0$.
end for

Algorithm 2 Fairness Greedy
(`DOFG` if players are sorted in p_n decreasing order)

Inputs: $[K]$, $[N]$, $\{\theta^k\}_{k\in[K]}$, $\{p_n\}_{n\in[N]}$
Output: π
Init: per-arm inactivity probabilities: $z^k = 1$.

for $n = 1$ to N **do**
  Let $k_n \in \operatorname{argmax}_{k\in[K]} \theta^k z^k$
  Update $z^{k_n} \leftarrow z^{k_n}(1 - p_n)$
  Set $\pi_n^{k_n} = 1$, and $\forall k \neq k_n, \pi_n^k = 0$.
end for

$$\mu_{\boldsymbol{\theta}}(\pi[n]) = \mu_{\boldsymbol{\theta}}(\pi[n-1]) + p_n \theta^{k_n} \left(1 - \ell^{k_n}_{[n-1]}\right) z^{k_n}_{[n-1]},$$

where $z^k_{[n]}$ is the probability that arm k is not used by any of the first n players, and $\ell^k_{[n]}$ is the sum of activation odds of the n first players for arm k.

Lemma 1 reveals a recurrence relation over n of the expected total reward. Under the assumption that the problem parameters are known, Lemma 1 paves the way to the definition of `DORG`, decreasing-order-reward-greedy (Algorithm 1), a recursive algorithm that assigns player n to arm k_n such that the right-hand term of the recursive equation in Lemma 1 is maximized. The result is highly dependent on the order in which the players are added to the pool, but the following theorem suggests the algorithm can lead to an actual optimum.

Theorem 2. *If $\sum_{n\in[N]} \frac{p_n}{1-p_n} \leq K + 1$, then there exists an ordering over players $\sigma^* : [N] \to [N]$ such that Algorithm 1 returns an optimal policy.*

Remark 1. *When $\forall n, p_n = p$ [Bonnefoi et al., 2018] Theorem 2 states that `DORG` returns the optimal policy. The precondition of Theorem 2 clearly holds in IoT networks, where the duty cycle p is commonly set to less than* 0.01.

2.3 Fairness greedy algorithm

Theorem 1 states that the resource assignment of an optimal deterministic policy is a Pareto optimum: as the network utility is maximum, if a user increases its own utility (equation (1)) another user has necessarily to decrease its utility (due to equation (2)). Notice that a Pareto optimum does not provide any guarantee about the *fairness* of the resource allocation among players. In this section, we design a policy to ensure fairness among players, for which we will use the definition below.

Definition 1 (α-fairness)**.** *A policy π is said to be α-fair if* $\frac{\min_{n\in[N]} \sum_{k=1}^K \mu^k_{n,\boldsymbol{\theta}}(\pi)}{\max_{n\in[N]} \sum_{k=1}^K \mu^k_{n,\boldsymbol{\theta}}(\pi)} \geq \alpha$.

Building a fair policy can be done by balancing the load with respect to the mean rewards of arms. The fair greedy algorithm (see Algorithm 2) assigns sequentially each player to the arm that

maximizes the reward of the arm times the probability of no internal collision. The player scheduling also plays an important role and we prove a lower bound on the fairness of Algorithm 2, when players are sorted in decreasing order of p_n. In that case we coin this algorithm `DOFG`, which stands for decreasing-order-fair-greedy.

Theorem 3. *`DOFG` generates α-fair policies, with $\alpha \geq 1 - \max_{n\in[N]} p_n$.*

Theorem 3 implies that when the probability of sending messages of the most frequent player is not high, which is the case in IoT networks, `DOFG` is a fair policy.

3 An Explore-Then-Exploit approach

The choice of the policy depends on the metric to be maximized: for maximizing network utility, `DORG` policy (Algorithm 1) should be used, while to guarantee some fairness among players, `DOFG` policy (Algorithm 2) is to be used. However both policies necessitate the model $\boldsymbol{\theta}$, which is unknown: hence exploration is necessary. To maximize the objective metric, we propose an *explore-then-exploit* approach: an exploration algorithm shares the probabilities of sending messages of players and outputs an ϵ-approximation of the model $\boldsymbol{\theta}$ with high probability for a sufficiently small ϵ, and then a target policy is used during the exploitation phase.

In IoT networks, due to the use of batteries by devices, the energy consumption and hence the number of sent messages should be minimal. That is why in comparison to an *explore-and-exploit* approach, which can adapt the sampling of arms according to their estimated mean rewards, the advantage of *explore-then-exploit* approach is that in the case of decentralized algorithms, the number of communications between players can be reduced to its minimum: N for sharing the probabilities of sending messages and N for sharing the estimation of arms.

Definition 2 (Regret)**.** *Let π_t be a policy generated at time t by an algorithm, and $\mu_{\boldsymbol{\theta}}(\pi_t)$ be its value in model $\boldsymbol{\theta} = \{\theta^1, ..., \theta^K\}$, we define the expected regret with respect to the optimal policy $\pi^*_{\boldsymbol{\theta}}$ as $E[R(T)] = \sum_{t=1}^{T}(\mu_{\boldsymbol{\theta}}(\pi^*_{\boldsymbol{\theta}}) - \mu_{\boldsymbol{\theta}}(\pi_t))$.*

Definition 3 (ϵ-approximation)**.** *$\hat{\theta}^k$ is said to be an ϵ-approximation of arm k, if the difference between it and θ^k is less than $\epsilon : |\theta^k - \hat{\theta}^k| \leq \epsilon$.*

Theorem 4. *When $\epsilon = K/\sqrt[3]{T}$, there exists a model $\boldsymbol{\theta} = \{\theta^1, ..., \theta^k\}$ and a distribution of players $p_1, ..., p_N$ such that the expected regret with respect to the deterministic optimal policy $\pi^*_{\boldsymbol{\theta}}$ of any exploration algorithm that outputs an ϵ-approximation of each arm θ^k with probability at least $1-1/T$ and which is followed by the optimal policy using the estimated model is at least: $\Omega\left(T^{2/3}\left(\frac{\log T}{N} + K^{3/2}\right)\right)$.*

Theorem 4 reveals that the studied problem is much more difficult than the multi-armed bandit problem, for which the regret lower bound of Explore-Then-Exploit algorithms is in $\Omega(\sqrt{KT})$ [Bubeck and Cesa-Bianchi, 2012].

4 Collaborative Exploration in Multi-Players Bandits

4.1 Principle

The gateway could not perform the exploration since when an *internal or external* collision occurs, the gateway cannot know that messages have been sent. That is why, we propose a decentralized algorithm for exploring the mean rewards of arms, which is performed with the messages that the devices have to send. For computing the exploration policy on each player, the probabilities of sending a messages have to be shared at the beginning of the exploration phase. In order to reduce the exploration time needed to find an ϵ-approximation of each arm, each player is responsible of a predefined number of samples t^*_n for each arm according to its probability of sending a message, so that all players would finish their estimations almost at the same time. At end of the exploration phase, each player sends its ϵ-approximation of each arm to other players through the gateway. Then, the target policy can be computed in a centralized way (by the gateway) or separately within each node.

Algorithm 3 Collaborative Exploration in Multi-Player Multi-Armed Bandits

Inputs: $[K]$, $[N]$, $\epsilon \in [0,1]$, $\delta \in (0,1)$ **Output:** $\hat{\theta} = \{\hat{\theta}^k, \forall k \in [K]\}$

Init: $t = 0$, $\forall n \in [N], \forall k \in [K], t_n^* = \infty$, $ack2_n = 0$, $\forall (n_1, n_2) \in [N]^2$ $ack1_{n_1,n_2} = 0$

1: **repeat**
2: $\mathcal{N}_t := \{n \in [N], a_n \sim \mathcal{B}(p_n), a_n = 1\}$
3: **for** $n \in \mathcal{N}_t$ **do**
4: $k_n \sim \mathcal{U}(1, K)$
5: $Y_n^{k_n}(t_n^{k_n}) := I_n^{k_n} E^{k_n}$
6: $\hat{\mu}_n^{k_n}(\pi_u) := \sum_{t=1}^{t=t_n^{k_n}} Y_n^{k_n}(t)/t_n^{k_n}$
7: $t_n^{k_n} := t_n^{k_n} + 1$
8: **if** $ack1_{n,n} = 0$ **then**
9: $ack1_{n,n} = send(p_n)$
10: **else**
11: **if** $\sum_i^N ack1_{n,i} = N$ **then**
12: $\forall n, t_n^* := \dfrac{p_n \log(2K/\delta)}{2(\epsilon \rho_n^k(\pi_u))^2 \sum_{i=1}^N p_i}$
13: **end if**
14: **if** $\forall k, t_n^k \geq t_n^*$ **and** $ack2_n = 0$ **then**
15: $ack2_n = \text{send}(\hat{\theta}_n^1, t_n^1, ..., \hat{\theta}_n^K, t_n^K)$
16: **end if**
17: **end if**
18: **end for**
19: $t = t + 1$
20: **until** $\exists \mathcal{N}' \subset \mathcal{N}, \begin{cases} \forall k \sum\limits_{n \in \mathcal{N}'} t_n^k \geq T = \sum\limits_{n \in \mathcal{N}} t_n^* \\ \sum\limits_{n \in \mathcal{N}'} ack2_n = |\mathcal{N}'| \end{cases}$
21: all players calculate $\hat{\theta}^k := \dfrac{\sum_{n \in \mathcal{N}'} \hat{\theta}_n^k t_n^k}{\sum_{n \in \mathcal{N}'} t_n^k}$

4.2 Description of the algorithm

The sampling strategy used in *collaborative exploration* is the *Uniform Policy* π_u: $\forall n, \forall k, \pi_n^k = \frac{1}{K}$. Then, player n can estimate the mean reward of arms using:

$$\hat{\theta}_n^k = \frac{\hat{\mu}_n^k(\pi_u)}{\rho_n^k(\pi_u)}, \text{ where} \tag{4}$$

$$\rho_n^k(\pi_u) = \prod_{n'=1, n' \neq n}^{N} (1 - p_{n'} \pi_{n'}^k) = \prod_{n'=1, n' \neq n}^{N} (1 - p_{n'}/K)$$

The function send(s), used in Algorithm 3, means that message s is broadcast to other players through the gateway on a channel chosen uniformly over K. The function send(s) returns 1 if an acknowledgement is received from the gateway or 0 else. When player n receives the probabilities of sending messages of all other players (Algorithm 3 line 11), it computes the required number of samples of each arm t_n^* according to Lemma 2. When player n has sampled t_n^* times each arm k, it sends their estimations $\hat{\theta}_n^k$ and t_n^k to other players (Algorithm 3 line 15). $\hat{\theta}_n^k$ is computed according to equation (4). The exploration phase ends when the arms have been sampled enough by a subset of players (Algorithm 3 line 20). Finally, the players compute the global estimations of arms by combining the received local ones (Algorithm 3 line 21).

Lemma 2. *By using Algorithm 3, in order to obtain with a probability* $1 - \delta$ *an* ϵ*-approximation of the mean rewards of arms, player* n *needs to sample each arm at least*

$$t_n^* = \left\lceil \frac{p_n \log(2K/\delta)}{2\epsilon^2 (\prod_{n' \neq N}(1 - p_{n'}/K))^2 \sum_{i=1}^N p_i} \right\rceil \textit{ times.}$$

4.3 Analysis of the algorithm

As in IoT network the payload of each transmission can contain 255 bytes [Augustin et al., 2016], we control the number of sent messages rather than the size of messages.

Theorem 5. ***Communication Cost*** *The communication cost of Algorithm 3 is with probability* $1-\delta$ *less than:* $C(2N)$ *messages, where* $C(m) = m\left\lceil \frac{\log \delta/m}{\log\left(1-\sum_{k=1}^{K}\frac{\rho_n(\pi_u)}{K}\theta^k\right)} + 1 \right\rceil$.

Theorem 5 states an upper bound of the number of messages sent by the N players for sharing the probabilities of sending messages, and for sharing their estimations at the end of the exploration phase. The communication cost is in the order of $2N$ messages, which is order optimal since for computing an optimal policy, any algorithm needs to share at least the player probabilities. Notice that an exploration algorithm, which elects a leader for performing a forced exploration, has a communication cost at least in the order of $N + \Omega(K/\epsilon^2 \log 1/\delta)$, which is sub-optimal in the general case.

Theorem 6. *With a probability at least* $1-\delta$*, Algorithm 3 stops while finding the* ϵ*-approximations of model* $\boldsymbol{\theta} = \{\theta^1, ..., \theta^K\}$ *at:*

$$t^* \leq \frac{K \log 2K/\delta}{2\epsilon^2 (1-p_1/K)^{2N-2} \sum_{i=1}^{N} p_i} + \frac{K}{p_N}\left(\sqrt{\frac{1}{2}\log\frac{NK}{\delta}} + C(2)\right),$$

where $p_N = \min_{n\in[N]} p_n$*,* $p_1 = \max_{n\in[N]} p_n$*, and* $C(2)$ *is the needed number of sent messages to successfully send* 2 *messages.*

Theorem 6 states an upper bound on the number of time slots needed by all players to finish their estimations of the mean rewards of the arms and to share them. The left term in $O\left(K/\epsilon^2 \log K/\delta\right)$ is the dominating term of the upper bound of the sample complexity. It is near optimal in comparison to the lower bound of K biased coin estimations in $\Omega\left(K/\epsilon^2 \log 1/\delta\right)$ [Anthony and Bartlett, 1999]. The right term of the upper bound in $O\left(K/p_N\sqrt{\log NK/\delta}\right)$ mainly depends of the least frequent player. This is due to the fact that, in worst case, before stopping Algorithm 3 has to wait that the least frequent player has sent its estimation of the arms.

Theorem 7. *when* $\delta = 1/T, \epsilon = K/\sqrt[3]{T}$*,* $\forall n \in [N], p_n = p$*, the expected cumulative regret with respect to the optimal policy* $\pi^*_{\boldsymbol{\theta}}$ *of Algorithm 3 followed by the policy* $\pi^*_{\hat{\theta}}$ *is upper bounded by:*

$$E[R(T)] \leq T^{2/3}\left(2K^2 + \frac{\log 2KT}{2(1-p/K)^{2N-2}Np}\right) + \frac{K^2}{p}\left(\sqrt{\frac{1}{2}\log NKT} + C(2)\right) + K$$

Theorem 7 shows that in the setting proposed by [Bonnefoi et al., 2018], the regret of Algorithm 3 followed by `DORG` is in $O\left(T^{2/3}\left((\log KT)/(1-p/K)^N N + K^2\right)\right)$, which is optimal with respect to T, and near optimal with respect to K (see Theorem 4). However concerning N, there is a gap with the lower bound. This gap is due to the lower bound that is built on a particular class of problems, where the optimal policy can be evaluated and produces no collision between players.

Theorem 8. *Applying Algorithm 3 followed by* `DOFG` *(Algorithm 2) on* $\hat{\boldsymbol{\theta}}$ *returns an* α*-fair policy in the true model* $\boldsymbol{\theta}$*, with*

$$\alpha \geq 1 - p_1 - \frac{2K\epsilon}{\max_{n\in[N]} \frac{\hat{\theta}^{k_n} z^{k_n}}{1-p_n}}.$$

Theorem 8 implies that, using ϵ-approximations of arms, `DOFG` still has the same fairness guarantee minus a term mainly depending on ϵ.

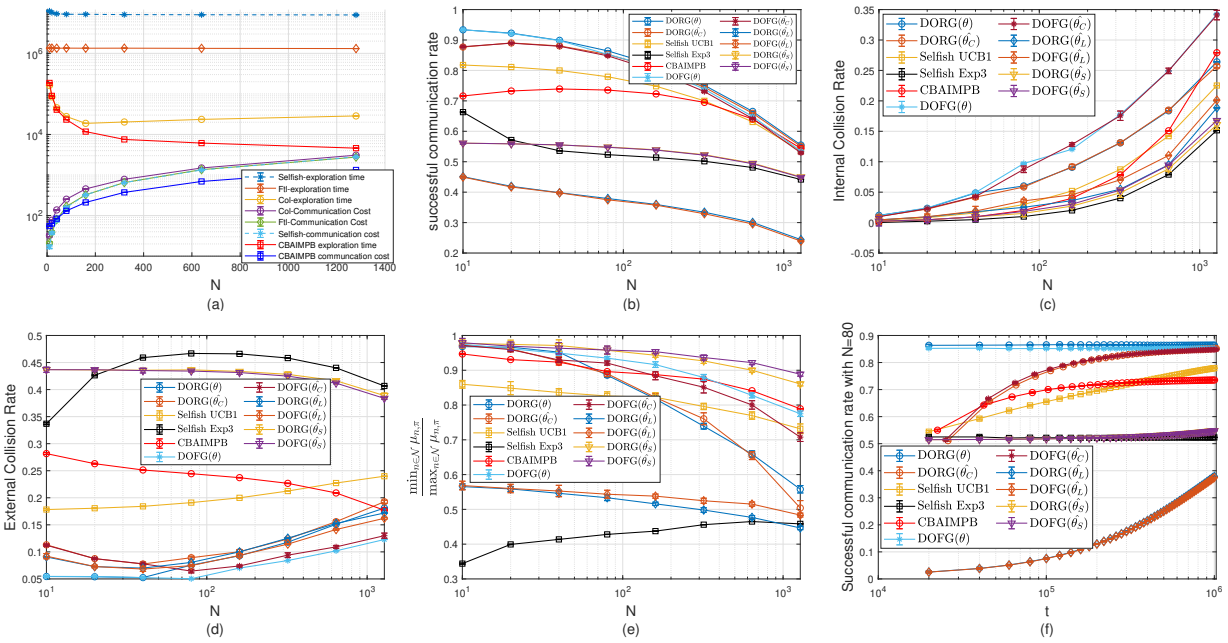

Figure 1: (a) exploration phase, (b) successful communication rate, (c) internal collision rate, (d) external collision rate, (e) fairness, (f) successful communication rate versus time. The successful communication and collision rates are cumulative over time. $\hat{\theta}_C$ when *collaborative exploration* is used, $\hat{\theta}_S$ when *selfish exploration* is used, and $\hat{\theta}_L$ when *follow-the-leader exploration* is used.

5 Simulation and Results

In order to illustrate and complete the analysis of the aforementioned algorithms, we first compare the performance of *collaborative exploration* (Algorithm 3) with *selfish exploration*, where each player explores selfishly, and with *follow-the-leader exploration (FtL)*, where only the most frequent player explores. Then we compare *collaborative exploration* followed by $\texttt{DORG}(\hat{\boldsymbol{\theta}})$ and $\texttt{DOFG}(\hat{\boldsymbol{\theta}})$, with *selfish UCB* [Bonnefoi et al., 2018] and *selfish EXP3* [Auer et al., 2002b], which respectively consist in independently playing *UCB* and *EXP3* on each player, and with CBAIMPB [Dakdouk et al., 2020], where the players find $(\epsilon', m)-$ optimal arms and exploit them uniformly with $m = 5, \epsilon' = 0.2$. We run both algorithms with various values of N, and $K = 10$, such that $\forall k, \theta^k \sim \mathcal{U}(0, 1)$. The distribution of players is uniform and the upper bound of the distribution is chosen such that the internal collision rate does not exceed 0.15 when the number of players reaches 1300 and play the arms uniformly, so $\forall n, p_n \sim \mathcal{U}(3.10^{-4}, 2.2.10^{-3})$. $\delta = 0.05$, $\epsilon = 0.1$. The curves are averaged over 10 trials and run on 10^6 time steps.

In figure 1a, we observe that the exploration time of *collaborative exploration* is two orders of magnitude less than *follow-the-leader exploration* and three orders of magnitude less than *selfish exploration* but one order of magnitude more than *CBAIMPB*, which stops exploration when it finds the best arms. Concerning the communication cost, we observe that the communication cost of the four exploration policies are close. In particular, the communication cost of *collaborative exploration* is two time less than the upper bound stated in Theorem 5, which is in the order of $2N$. This is due to the fact that the stopping condition of Algorithm 3 does not imply that all players have been sampled enough, but that the arms have been sampled enough. As a consequence, all the estimations of all players do not need to be shared, but only those of players that have finished their estimations.

The performance differences of the exploration policies affect the whole performance of $\texttt{DORG}(\hat{\boldsymbol{\theta}})$ and $\texttt{DOFG}(\hat{\boldsymbol{\theta}})$, which consist of the exploration algorithm followed by the corresponding exploitation phase. That is why in figures 1b and 1f, the successful communication rate when using *selfish exploration* and *follow-the-leader exploration* are dramatically lesser than the one of *collaborative exploration*. In figures 1b and 1f, $\texttt{DOFG}(\boldsymbol{\theta})$ is slightly outperformed in terms of successful communication rate by $\texttt{DORG}(\boldsymbol{\theta})$. $\texttt{DORG}(\hat{\boldsymbol{\theta}})$ and $\texttt{DOFG}(\hat{\boldsymbol{\theta}})$ exhibit the same behavior, and we can notice that $\texttt{DORG}(\hat{\boldsymbol{\theta}})$ and $\texttt{DOFG}(\hat{\boldsymbol{\theta}})$ clearly outperform *selfish UCB1*, *selfish Exp3* and *CBAIMPB* , and tend to perform as well as $\texttt{DORG}(\boldsymbol{\theta})$ and

DOFG($\boldsymbol{\theta}$) as N increases (figure 2b). This improvement is due to their low external collision rate (figure 2d) thanks to playing more the best arms, while because of playing more the best arms their internal collision rate is higher (figure 2c). Finally, while *Selfish Exp3* is theoretically better suited for our problem setting, it is clearly outperformed by *Selfish UCB*.

Concerning fairness, DOFG($\hat{\boldsymbol{\theta}}$) clearly outperforms *selfish UCB1*, *selfish Exp3* and DORG($\hat{\boldsymbol{\theta}}$), while DORG($\hat{\boldsymbol{\theta}}$) is outperformed by them when N is high (Figure 1e). *CBAIMPB* offers a high fairness between players due to the uniform selection of the arms by all players during both exploration and exploitation phases. The use of *selfish exploration* leads to high fairness level due to its very long uniform exploration phase, in contrast to *follow-the-leader exploration* that suffers of very low fairness level due to the fact that during the exploration time, only the leader can send messages. The observed fairness of DOFG($\boldsymbol{\theta}$) in figure 1e differs from the theoretical one (Theorem 3). This is due to the fact that the mean rewards of players are observed on a finite number of time slots (10^6). As time passes the observed fairness tends to the theoretical fairness (Appendix B.2).

6 Conclusion

With the aim of optimizing transmission in IoT networks we have proposed an *explore-then-exploit* approach. Despite the fact that the optimal policy cannot be evaluated in the general case, we have stated a first regret lower bound for this problem. We have proposed two target policies DORG and DOFG that are efficient with any number of players, and can handle internal and external collision without *sensing*. We have shown that our algorithm when using DORG is optimal with respect to T and near optimal with respect to K in the setting proposed in [Bonnefoi et al., 2018] (when $\forall n, p_n = p$), and that when using DOFG it is fair. Our experiments confirm the good behavior of *selfish UCB* and *CBAIMPB*, but show that both are outperformed in terms of network successful communication rate by DORG($\hat{\boldsymbol{\theta}}$) and DOFG($\hat{\boldsymbol{\theta}}$), and in terms of fairness by DOFG($\hat{\boldsymbol{\theta}}$). This work can be extended in many directions: studying *explore-and-exploit* approach for the proposed problem, handling an evolving number of active players, handling more general non-stationary environments handling players with different mean rewards of arms... Finally, showing the NP-Hardness of the optimization problem stated in equation (2) is an open problem.

References

[Anandkumar et al., 2010] Anandkumar, A., Michael, N., and Tang, A. (2010). Opportunistic spectrum access with multiple users: Learning under competition. In 2010 Proceedings IEEE INFOCOM.

[Anthony and Bartlett, 1999] Anthony, M. and Bartlett, P. L. (1999). Neural Network Learning: Theoretical Foundations. Cambridge University Press, USA, 1st edition.

[Auer et al., 2002a] Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002a). Finite-time analysis of the multiarmed bandit problem. Machine learning, 47(2-3):235–256.

[Auer et al., 2002b] Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (2002b). The non-stochastic multiarmed bandit problem. SIAM journal on computing, 32(1):48–77.

[Augustin et al., 2016] Augustin, A., Yi, J., Clausen, T., and Townsley, W. M. (2016). A study of lora: Long range & low power networks for the internet of things. Sensors, 16(9):1466.

[Avner and Mannor, 2014] Avner, O. and Mannor, S. (2014). Concurrent bandits and cognitive radio networks. In ECML PKDD, Berlin, Heidelberg. Springer-Verlag.

[Bertsekas et al., 1992] Bertsekas, D. P., Gallager, R. G., and Humblet, P. (1992). Data networks, volume 2. Prentice-Hall International New Jersey.

[Besson and Kaufmann, 2018] Besson, L. and Kaufmann, E. (2018). Multi-player bandits revisited. In Proceedings of Algorithmic Learning Theory, volume 83, pages 56–92.

[Bonnefoi et al., 2018] Bonnefoi, R., Besson, L., Moy, C., Kaufmann, E., and Palicot, J. (2018). Multi-armed bandit learning in iot networks: Learning helps even in non-stationary settings. In Marques, P., Radwan, A., Mumtaz, S., Noguet, D., Rodriguez, J., and Gundlach, M., editors, Cognitive Radio Oriented Wireless Networks, pages 173–185. Springer International Publishing.

[Boursier and Perchet, 2019] Boursier, E. and Perchet, V. (2019). Sic-mmab: Synchronisation involves communication in multiplayer multi-armed bandits. In Advances in Neural Information Processing Systems 32, pages 12048–12057.

[Boursier et al., 2020] Boursier, E., Perchet, V., Kaufmann, E., and Mehrabian, A. (2020). A practical algorithm for multiplayer bandits when arm means vary among players. In AISTATS.

[Bubeck and Cesa-Bianchi, 2012] Bubeck, S. and Cesa-Bianchi, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. arXiv preprint arXiv:1204.5721.

[Bubeck et al., 2019] Bubeck, S., Li, Y., Peres, Y., and Sellke, M. (2019). Non-stochastic multi-player multi-armed bandits: Optimal rate with collision information, sublinear without.

[Cohen et al., 2013] Cohen, K., Leshem, A., and Zehavi, E. (2013). Game theoretic aspects of the multi-channel aloha protocol in cognitive radio networks. IEEE Journal on Selected Areas in Communications, 31(11):2276–2288.

[Dakdouk et al., 2020] Dakdouk, H., Féraud, R., Varsier, N., and Maillé, P. (2020). Collaborative exploration in stochastic multi-player bandits. In Asian Conference on Machine Learning, pages 193–208. PMLR.

[Dakdouk et al., 2018] Dakdouk, H., Tarazona, E., Alami, R., Féraud, R., Papadopoulos, G. Z., and Maillé, P. (2018). Reinforcement learning techniques for optimized channel hopping in ieee 802.15.4-tsch networks. In Proceedings of the 21st ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems, MSWIM '18.

[Hanawal and Darak, 2018] Hanawal, M. K. and Darak, S. J. (2018). Multi-player bandits: A trekking approach. arXiv preprint arXiv:1809.06040.

[Kerkouche et al., 2018] Kerkouche, R., Alami, R., Féraud, R., Varsier, N., and Maillé, P. (2018). Node-based optimization of lora transmissions with multi-armed bandit algorithms. In 2018 25th International Conference on Telecommunications (ICT).

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

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

[Nayyar et al., 2015] Nayyar, N., Katathil, D., and Jain, R. (2015). On regret-optimal learning in decentralized multi-player multi-armed bandits. In EEE Transactions on Control of Network Systems.

[Rosenski et al., 2016] Rosenski, J., Shamir, O., and Szlak, L. (2016). Multi-player bandits – a musical chairs approach. In ICML.

[Shi et al., 2020] Shi, C., Xiong, W., Shen, C., and Yang, J. (2020). Decentralized multi-player multi-armed bandits with no collision information. In International Conference on Artificial Intelligence and Statistics, pages 1519–1528. PMLR.

[Wang et al., 2020] Wang, P.-A., Proutiere, A., Ariu, K., Jedra, Y., and Russo, A. (2020). Optimal algorithms for multiplayer multi-armed bandits. In International Conference on Artificial Intelligence and Statistics, pages 4120–4129. PMLR.

[Wang and Kar, 2004] Wang, X. and Kar, K. (2004). Distributed algorithms for max-min fair rate allocation in aloha networks. In Proceedings of the 42nd Annual Allerton Conference. Citeseer.

A Additional experiments

B Additional experiments

B.1 Preliminary experiments

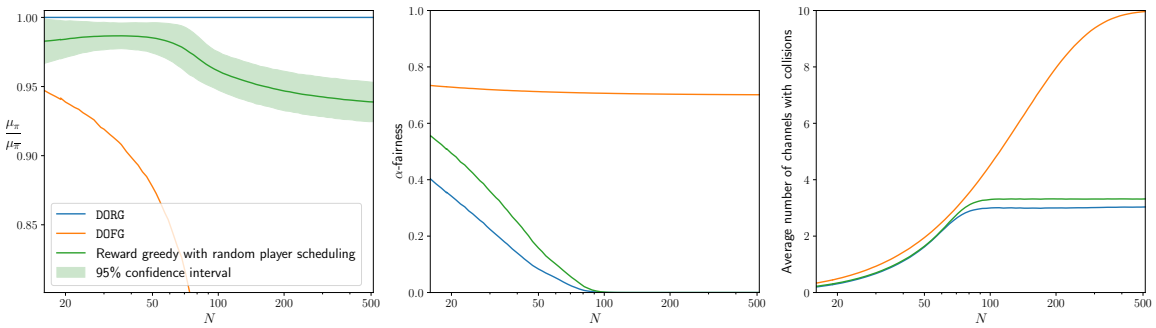

Figure 2: With number of arm $K = 10$ fixed, and for N values (ranging from 16 to 512 on a log scale), the algorithms have been compared in terms of expected reward ratio with `DORG` (left), $\bar{\pi}$ denotes the policy to be compared with `DORG`, α-fairness (center), and number of channels with internal collision (right).

In this section, we perform the following experiment: the problem parameters are sampled as follows: $\forall n \in [N], p_n \sim \mathcal{U}(0, 0.3)$[1] and $\forall k \in [K], \theta^k \sim \mathcal{U}(0, 1)$. Figure 2 compares the performance of `DORG`, `DOFG`, and Reward Greedy (Algorithm 1) with random ordering, where each point is the average of 10,000 runs. Figure 2 (left) reveals that sorting the players in decreasing order is a good policy. However, it has to be noted that the difference between `DORG` and a random ordering is much thinner when p_n are smaller, as expected in a real setting. We also notice that `DOFG` expected reward loss, as compared to `DORG`, is below 20% until $N \approx 75$. Figure 2 (center) illustrates the result of Theorem 3, and indicates that the fairness lower bound is tight. It also shows that, while `DOFG` only loses 20% rewards when $N \approx 75$ as compared to `DORG`, its fairness is approximately 30 times larger.

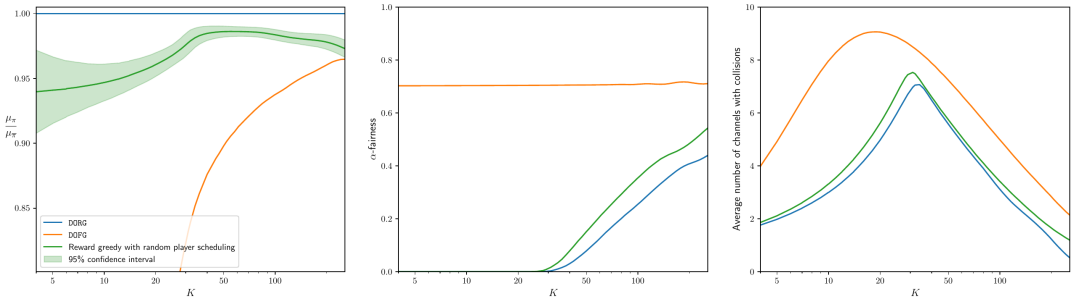

Figure 3: `DORG` and `DOFG`: experiments with N fixed and K varying. With number of players $N = 200$ fixed, and for K values (ranging from 4 to 256 on a log scale), the algorithms have been compared in terms of expected reward ratio with `DORG` (above), α-fairness (middle), and number of collided channels (below). Each point has been obtained from 10,000 problems where the parameters are sampled as follows: $\forall n \in [N], p_n \sim \mathcal{U}(0, 0.3)$ and $\forall k \in [K], \theta^k \sim \mathcal{U}(0, 1)$.

Further, on figure 2 (right), we notice that the expected number of channels with collisions stops increasing as N grows around $N = 100$. It is the moment when the channels get completely saturated.

[1]Such high values for p_n are used to graphically observe the expected properties. Experiments with realistic p_n values ($p_n < 0.01$) may be found in Appendix A.

$N = 100$ coincides with the point where the fairness gets to 0 on figure 2 (center). We explain this phenomenon as follows: each channel k fills up, up to the point when $\ell^k > 1$. When all the channels reached this point, adding new players to the network actually decreases the expected reward, and `DORG`'s strategy condemns the arms with the lowest θ^k and use them as a garbage bin for new players. These channels get so crowded that there is a collision on it with a very high probability, in order to keep the other channels functionally unspoiled. In comparison, to guarantee fairness `DOFG` does not throw away players on a bin channel.

Similar figures with $N = 200$ and 100 K values ranging from 4 to 256 on a log scale are available below.

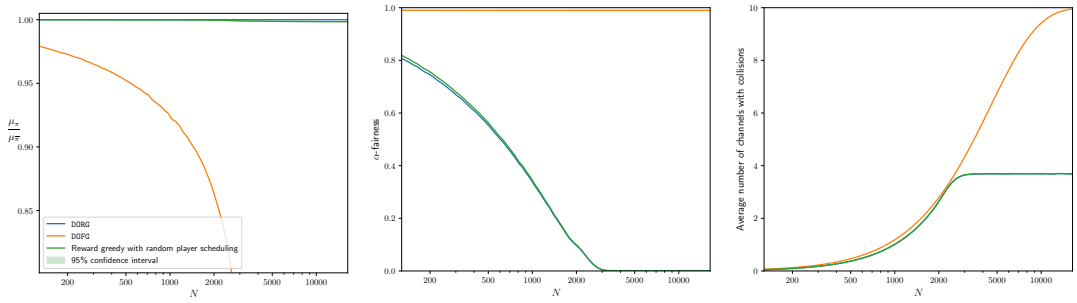

Figure 4: `DORG` and `DOFG`: experiments with K fixed and N varying, and smaller p_n. With number of players $K = 10$ fixed, and for N values (ranging from 128 to 16384 on a log scale), the algorithms have been compared in terms of expected reward ratio with `DORG` (above), α-fairness (middle), and number of collided channels (below). Each point has been obtained from 10,000 problems where the parameters are sampled as follows: $\forall n \in [N], p_n \sim \mathcal{U}(0, 0.01)$ and $\forall k \in [K], \theta^k \sim \mathcal{U}(0, 1)$.

B.2 Fairness

Figure 5 shows the progress of the fairness level achieved by `DOFG`(θ) policy as time passes. The experimental settings are the same as those in section 5. The black plot corresponds to the theoretical fairness level proved in Theorem 3. In order to reach the theoretical fairness level, the observed mean rewards of all players have to reach their expected values. Due to the low probabilities of sending messages of the players, this would take a long time. As shown by figure 5, the observed fairness tends to the theoretical fairness in 10^8 times steps for 10 players.

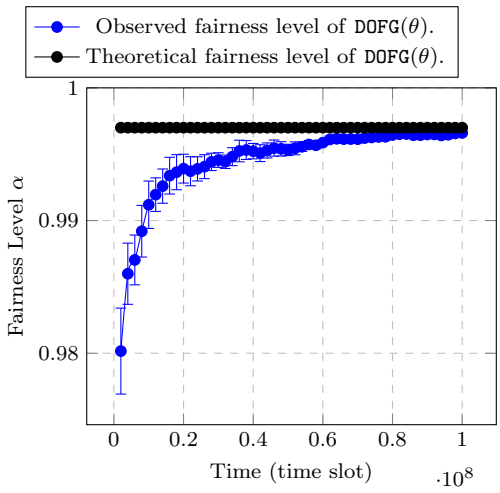

Figure 5: Fairness level achieved by DOFG(θ) as a function of time with 10 players.

C Proofs

C.1 Notations

For the sake of ease the reading of proofs, we provide below the notations.

notation	meaning
N	number of players.
$[N]$	set of players.
p_n	probability that player n sends a message.
K	number of arms.
$[K]$	set of arms.
θ_k	mean reward of arms k.
$\boldsymbol{\theta}$	model $\boldsymbol{\theta} = (\theta_1, ..., \theta_K)$.
$\hat{\theta}_k$	estimated mean reward of arms k.
$\hat{\boldsymbol{\theta}}$	estimated model $\hat{\theta} = (\hat{\theta}_1, ..., \hat{\theta}_K)$.
ϵ	approximation term.
δ	probability of failure.
π_n^k	probability that player n chooses arm k.
π_n	policy of player n, $\pi_n = (\pi_n^1, ..., \pi_n^K)$.
π	policy of players, $\pi = (\pi_1, ..., \pi_n)$.
π_u	uniform policy.
$\pi^\dagger$	decreasing order fair greedy policy generated by Algorithm 2.
$\pi^*_{\boldsymbol{\theta}}$	optimal policy in model $\boldsymbol{\theta}$, which is deterministic, when it is clear in the context, we use π^*.
$\mu_{\boldsymbol{\theta}}(\pi)$	mean reward in model $\boldsymbol{\theta}$ of the policy π, when it is clear in the context, we use $\mu(\pi)$.
	For a stochastic policy: $\mu_{\boldsymbol{\theta}}(\pi) = \sum_{k=1}^K \theta^k \sum_{n=1}^N p_n \pi_n^k \prod_{n' \neq n} (1 - p_{n'} \pi_{n'}^k)$.
	For a deterministic policy $\mu_{\boldsymbol{\theta}}(\pi) = \sum_{k=1}^K \theta^k z^k l^k$.
z^k	probability that arm k is not used by any other players, $z^k = \prod_{n' \in [N], k_n = k} (1 - p_n)$.
l^k	sum of activation odds on arm k of other players, $l^k = \sum_{n' \in [N], k_n = k} \frac{p_n}{1 - p_n}$.
k_n	arm assigned to player n.
$\pi[n]$	policy π when players $n' > n$ do not play.
$z^k[n]$	probability that arm k is not used by any of the first n players.
$l^k[n]$	sum of activation odds of the n first players for arm k.
$\rho_n^k(\pi)$	probability that no other players have chosen arm k using policy π.

C.2 Proof of Theorem 1

There exists an optimal policy which is deterministic.

Proof. We may write the global objective as:

$$\mu(\pi) = \sum_{k=1}^{K} \underbrace{\theta^k}_{\text{mean reward of arm k}} \sum_{n=1}^{N} \underbrace{p_n.\pi_n^k}_{\text{probability that player n chooses arm k}} \underbrace{\prod_{n'=1, n' \neq n}^{N} (1 - p_{n'}.\pi_{n'}^k)}_{\text{probability that no collision occurs}} \tag{5}$$

Let us assume that $\pi^* = \{\pi_n\}_{n \in [N]}$ is optimal. Let us fix all player policies but player n's. Then, we notice that $\mu(\pi)$ is linear (see (5)) in each $\pi_n^k, k = 1, ..., K$, meaning that the maximum is achieved for any $k_n^* \in \text{argmax}_{k \in [K]} \frac{\partial \mu(\pi)}{\partial \pi_n^k}$, and therefore the optimal policy may have been chosen so that π_n is deterministic: $\pi_n^{k^*} = 1$ and $\forall k \neq k^*, \pi_n^k = 0$. The same reasoning can be repeated for the other players, so that there exists an optimal policy that is deterministic. □

C.3 Proof of Lemma 1

For a deterministic policy π, let $\mu(\pi[n])$ denote the aggregated expected reward when only the players $1, ..., n$ are playing (all players $n' > n$ are deactivated). Then we have the recursive expression

$$\mu(\pi[n]) = \mu(\pi[n-1]) + p_n \theta^{k_n} \left(1 - \ell_{[n-1]}^{k_n}\right) z_{[n-1]}^{k_n},$$

where $z_{[n]}^k$ is the probability that arm k is not used by any of the first n players, and $\ell_{[n]}^k$ is the sum of activation odds of the n first players for arm k.

Proof. We have:

$$\begin{aligned}
\mu(\pi[n]) &= \mu(\pi[n-1]) + \mu(\pi[n]) - \mu(\pi[n-1]) \\
&= \mu(\pi[n-1]) + \sum_{k \in [K]} \theta^k z_{[n]}^k \ell_{[n]}^k - \sum_{k \in [K]} \theta^k z_{[n-1]}^k \ell_{[n-1]}^k \\
&= \mu(\pi[n-1]) + \theta^{k_n} z_{[n]}^{k_n} \ell_{[n]}^{k_n} - \theta^{k_n} z_{[n-1]}^{k_n} \ell_{[n-1]}^{k_n} \qquad (6) \\
&= \mu(\pi[n-1]) + \theta^{k_n} \left(z_{[n]}^{k_n} \ell_{[n]}^{k_n} - z_{[n-1]}^{k_n} \ell_{[n-1]}^{k_n}\right) \\
&= \mu(\pi[n-1]) + \theta^{k_n} \left((1 - p_n) z_{[n-1]}^{k_n} \left(\ell_{[n-1]}^{k_n} + \frac{p_n}{1 - p_n}\right) - z_{[n-1]}^{k_n} \ell_{[n-1]}^{k_n}\right) \\
&= \mu(\pi[n-1]) + \theta^{k_n} \left(-p_n z_{[n-1]}^{k_n} \ell_{[n-1]}^{k_n} + p_n z_{[n-1]}^{k_n}\right) \\
&= \mu(\pi[n-1]) + p_n \theta^{k_n} z_{[n-1]}^{k_n} \left(1 - \ell_{[n-1]}^{k_n}\right),
\end{aligned}$$

where the line (6) comes from the fact that $z_{[n]}^k = z_{[n-1]}^k$ and $\ell_{[n]}^k = \ell_{[n-1]}^k$ for all $k \neq k_n$. □

C.4 Proof of Theorem 2

Lemma 3. *As long as $\ell_{n-1}^k \leq 2$, the reward-greedy criterion for Algorithm 1 decreases as we add a new player n:*

$$z_{[n]}^k \left(1 - \ell_{[n]}^k\right) \leq z_{[n-1]}^k \left(1 - \ell_{[n-1]}^k\right). \tag{7}$$

Proof. We look at the difference:

$$\forall k \neq k_n, \quad z^k_{[n]}\left(1-\ell^k_{[n]}\right) - z^k_{[n-1]}\left(1-\ell^k_{[n-1]}\right) = 0 \tag{8}$$

$$\begin{aligned} z^{k_n}_{[n]}\left(1-\ell^{k_n}_{[n]}\right) - z^{k_n}_{[n-1]}\left(1-\ell^{k_n}_{[n-1]}\right) &= (1-p_n)z^{k_n}_{[n-1]}\left(1-\ell^{k_n}_{[n-1]}-\frac{p_n}{1-p_n}\right) \\ &\quad - z^{k_n}_{[n-1]}\left(1-\ell^{k_n}_{[n-1]}\right) && (9) \\ &= (1-p_n)z^{k_n}_{[n-1]}\left(1-\ell^{k_n}_{[n-1]}\right) - p_n z^{k_n}_{[n-1]} \\ &\quad - z^{k_n}_{[n-1]}\left(1-\ell^{k_n}_{[n-1]}\right) && (10) \\ &= -p_n z^{k_n}_{[n-1]}\left(1-\ell^{k_n}_{[n-1]}\right) - p_n z^{k_n}_{[n-1]} && (11) \\ &= -p_n z^{k_n}_{[n-1]}\left(2-\ell^{k_n}_{[n-1]}\right) && (12) \end{aligned}$$

Since p_n and $z^{k_n}_{[n-1]}$ are always positive, we may conclude.

□

Theorem 2: *If $\sum_{n\in[N]}\frac{p_n}{1-p_n} \leq K+1$, then, there exists an ordering over players $\sigma^* : [N] \to [N]$ such that Algorithm 1 returns an optimal policy.*

Proof. The proof makes use of Lemma 3 which states that, as long as $\ell^k_{n-1} \leq 2$, the reward-greedy criterion for Algorithm 1 decreases as we add a new player n.

We prove below that this Lemma applies for all picked arms if $\sum_{n\in[N]}\frac{p_n}{1-p_n} \leq K+1$. By *reductio ad absurdum*, we assume that $\sum_{n\in[N]}\frac{p_n}{1-p_n} \leq K+1$ and that there exists some arm k and some player ordering σ (not necessarily σ^*) such that $\pi^*(\sigma(N)) = k$ and $\ell^k_{\sigma([N-1])} > 2$, where π^* is an optimal policy and $\sigma([N-1])$ denotes the $N-1$ first indexes in the σ reordering. Then, there must exist an arm k' for which $\ell^{k'}_{\sigma([N-1])} < 1$, otherwise we would have $\sum_{n\in[N]}\frac{p_n}{1-p_n} > \sum_{n\in[N-1]}\frac{p_{\sigma(n)}}{1-p_{\sigma(n)}} > K+1$. It means that, for k', the reward-greedy criterion $z^{k'}_{\sigma([N-1])}\left(1-\ell^{k'}_{\sigma([N-1])}\right)$ is positive, and therefore larger than that of k: $z^k_{\sigma([N-1])}\left(1-\ell^k_{\sigma([N-1])}\right)$, which is negative. As Lemma 1 states that the reward-greedy criterion is incrementally optimal, it means that k' would have been a strictly better arm for player $\sigma(N)$, which contradicts the assumption that π^* is optimal.

Let an optimal policy π^* be given, and let us construct the player ordering σ^* such that Algorithm 1 applied on the σ^* ordering returns π^*.

Algorithm 4 Reconstruction of a player ordering that allows Algorithm 1 to return π^*

Inputs: $[K]$, $[N]$, $\{\theta^k\}_{k\in[K]}$, $\{p_n\}_{n\in[N]}$, π^*
Output: σ^* such that Algorithm 1 returns π^*
Init: per-arm inactivity probabilities: $z^k = 1$.
Init: per-arm activation odds sums: $\ell^k = 0$.
Init: Set of players remaining to be assigned: $\mathcal{N} = [N]$.

1: **for** $n = 1$ to N **do**
2: Let $\sigma^*(n)$ be an element of $\mathcal{N}$ such that $\pi^*(\sigma^*(n)) \in \text{argmax}_{k\in[K]}\,\theta^k z^k(1-\ell^k)$.
3: Update $\mathcal{N} \leftarrow \mathcal{N} - \{\sigma^*(n)\}$.
4: Update $z^{k_n} \leftarrow z^{k_n}\left(1-p_{\sigma^*(n)}\right)$.
5: Update $\ell^{k_n} \leftarrow \ell^{k_n} + \frac{p_{\sigma^*(n)}}{1-p_{\sigma^*(n)}}$.
6: **end for**

It is direct to understand that Algorithm 1 applied on a σ^* player ordering would retrieve π^*. Indeed, Algorithm 4 makes it so the players are ordered to be incrementally optimal. The last piece of the proof is to check the existence of a player $\sigma^*(n)$ assigned to a reward-greedy arm on line 2.

Again by *reductio ad absurdum*, we assume that there is no remaining player that π^* assigned to a reward-greedy arm k^*. Then, it means that until the last selection, this arm will not be picked and another arm k will be picked instead. We showed at the beginning of the proof that the reward-greedy

criterion is only decreasing as the arms are being selected, and that the reward-greedy criterion of an arm not being selected, such as k^*, is constant. So it means that $\pi^*(\sigma^*(N))$ should be k^*, hence, the contradiction.

We may therefore conclude the proof by stating that Algorithm 4 will never fail to construct σ^* and that Algorithm 1 applied to the σ^* player ordering will return π^*. □

C.5 Proof of Theorem 3

DOFG generates α-fair policies, with

$$\alpha \geq 1 - \max_{n \in [N]} p_n. \tag{13}$$

Proof. For every arm, we have the following equality:

$$\mu_n(\pi^\dagger) = \theta^{k_n} \prod_{n' \neq n, \text{ s.t. } k_{n'} = k_n} (1 - p_{n'}) = \frac{\theta^{k_n} z^{k_n}}{1 - p_n}. \tag{14}$$

We prove now that $\min_{n \in [N]} \mu_n(\pi^\dagger) = \mu_N(\pi^\dagger)$. We proceed by induction. The base case is direct for $N = 1$. Now, we prove the induction step by assuming that it is true for N and prove it for $N + 1$. We have to distinguish two cases whether k_N equals k_{N+1} or not.

Case $k_N = k_{N+1}$, then from Equation 14, we have $\mu_{N+1}(\pi^\dagger) = \frac{1-p_N}{1-p_{N+1}} \mu_N(\pi^\dagger)$. Since we know by construction that $p_{N+1} \leq p_N$, we may conclude that $\mu_{N+1}(\pi^\dagger) \leq \mu_N(\pi^\dagger)$.

Case $k_N \leq k_{N+1}$, then stating that $\mu_{N+1}(\pi^\dagger) > \mu_N(\pi^\dagger)$ would imply that k_N was not optimally selecting the arm at the previous step, which brings a contradiction.

Let us assume without loss of generality that player N has been assigned to arm K. Since $\pi^\dagger_N$ has been chosen so that to maximize $\theta^k z^k$ at iteration N, it means that:

$$\min_{n \in [N]} \mu_n(\pi^\dagger) = \mu_N(\pi^\dagger) \geq \max_{k \in [K]} \theta^k z^k. \tag{15}$$

We also know that:

$$\max_{n \in [N]} \mu_n(\pi^\dagger) = \max_{n \in [N]} \frac{\theta^{k_n} z^{k_n}}{1 - p_n} \tag{16}$$

$$\leq \frac{\max_{k \in [K]} \theta^k z^k}{1 - \max_{n \in [N]} p_n} \tag{17}$$

$$\leq \frac{1}{1 - p_1} \min_{n \in [N]} \mu_n(\pi^\dagger), \tag{18}$$

which concludes the demonstration. □

C.6 Proof of Lemma 2

By using Algorithm 3, in order to obtain with a probability $1 - \delta$ an ϵ-approximation of the mean rewards of arms, player n needs to sample each arm at least

$$t_n^* = \left\lceil \frac{p_n \log(2K/\delta)}{2\epsilon^2 (\prod_{n' \neq N} (1 - p_{n'}/K))^2 \sum_{i=1}^N p_i} \right\rceil \text{ times.}$$

Proof. Due to equations 1 and 4, for a given probability of failure $\delta \in [0, 1]$, and a given approximation factor ϵ, $\forall n \in [N]$, $\forall k \in [K]$ we have:

$$P(|\mu^k - \hat{\mu}_n^k| \geq \epsilon) \leq \frac{\delta}{K} \iff P(|\theta^k - \hat{\theta}_n^k| \geq \epsilon'_n) \leq \frac{\delta}{K}, \tag{19}$$

where $\epsilon'_n = \epsilon . \prod_{n' \neq n} (1 - p_{n'}/K)$.

Applying Hoeffding's inequality:

$$P(|\theta_n^k - \hat{\theta}_n^k| \geq \epsilon'_n) \leq 2e^{-2t_n^k {\epsilon'_n}^2}. \tag{20}$$

Therefore for obtaining an ϵ-approximation of arm k on player n with a probability $1 - \frac{\delta}{K}$:

$$t_n^k \geq \frac{\log(2K/\delta)}{2\epsilon_n'^2} \iff t_n^k \geq \frac{\log(2K/\delta)}{2\epsilon^2(\prod_{n' \neq n}(1 - p_{n'}/K))^2} \geq t^\dagger = \frac{\log(2K/\delta)}{2\epsilon^2(\prod_{n' \neq N}(1 - p_{n'}/K))^2}$$

Now, as Algorithm 3 shares the estimations of the N players for finding ϵ-approximation of arm k with high probability, we need $\sum_{n=1}^N t_n^* = t^\dagger$ samples.

Hence, if each player samples arm k at least $t_n^* \geq \lceil \frac{p_n \log(2K/\delta)}{2\epsilon^2(\prod_{n' \neq N}(1 - p_{n'}/K))^2 \sum_{i=1}^N p_i} \rceil$ times, an ϵ-approximation of arm θ^k is obtained with a probability $1 - \frac{\delta}{K}$.

□

C.7 Proof of Theorem 4

When $\epsilon = K/\sqrt[3]{T}$, there exists a model $\boldsymbol{\theta} = \{\theta^1, ..., \theta^k\}$ and a distribution of players $p_1, ..., p_N$ such that the expected regret with respect to the deterministic optimal policy $\pi^*_{\boldsymbol{\theta}}$ of any exploration algorithm that outputs an ϵ-approximation of each arm θ^k with probability at least $1 - 1/T$ and which is followed by the optimal policy using the estimated model is at least:

$$E[R(T)] \geq \Omega\left(T^{2/3}\left(\frac{\log T}{N} + K^{3/2}\right)\right).$$

In the following we show that a lower bound holds for a class of models $\boldsymbol{\theta}$ and distribution of players $p_1, ..., p_N$. For the sake of simplifying notations, we assume in the following that:

- $\theta^1 > \theta^2, ..., \theta^{K-1} > \theta^K$,
- $p_1 > p_2, ..., p_{N-1} > p_N$.

Choice of a class of problems. The most difficult point for evaluating a regret lower bound is that in the general case the optimal policy, which maximizes the mean reward (see equation (3)), is unknown. For handling this point we choose a particular class of problems, where $N = K$. Then, we assume that the distribution of arms is such that $\forall k \in [K - 1]$,

$$\theta^k = \theta^{k+1} + \epsilon, \tag{21}$$

$$\frac{\epsilon}{2p_k} \leq \theta^k \leq \frac{\epsilon}{2p_{k+1}}. \tag{22}$$

The optimal policy. When $\frac{\epsilon}{2p_k} \leq \theta^k$ (equation (22)), then superposing players on any arm provides less reward than spreading players on the arms. Indeed, let Δ_s be the gap between the mean reward of two players $k_1, k_2, k_1 < k_2$ assigned on different arms, and the mean reward of two players assigned on the same arm:

$$\Delta_s = p_{k_1}\theta^{k_1} + p_{k_2}\theta^{k_2} - p_{k_1}\theta^{k_1}(1 - p_{k_2}) - p_{k_2}\theta^{k_1}(1 - p_{k_1}), \tag{23}$$

$$= p_{k_2}(\theta^{k_2} - \theta^{k_1}) + 2p_{k_1}p_{k_2}\theta^{k_1}, \tag{24}$$

$$= -p_{k_2}\epsilon + 2p_{k_1}p_{k_2}\theta^{k_1} \geq 0. \tag{25}$$

Hence, when equation (22) holds, the optimal assignment of players over arms is:

$$(p_1, \theta^1), (p_2, \theta^2), ..., (p_K, \theta^K). \tag{26}$$

Regret decomposition. Let T be the time horizon. Let π_E^* be the optimal (in term of sample complexity) exploration policy that outputs an ϵ-approximation with high probability of $\boldsymbol{\theta}$, i.e. each arm θ^k, and $\pi_{\boldsymbol{\theta}}^*$ be the optimal policy. We consider the time t^*, where the optimal exploration algorithm π_E^* outputs exactly an ϵ-approximation of model $\boldsymbol{\theta}$. Then, the expected cumulative regret with respect to the deterministic policy $\pi_{\boldsymbol{\theta}}^*$ is expressed as:

$$E[R(T)] = t^*(\mu_{\boldsymbol{\theta}}(\pi_{\boldsymbol{\theta}}^*) - \mu_{\boldsymbol{\theta}}(\pi_E^*)) + (T - t^*)(\mu_{\boldsymbol{\theta}}(\pi_{\boldsymbol{\theta}}^*) - \mu_{\boldsymbol{\theta}}(\pi_{\hat{\boldsymbol{\theta}}}^*)), \tag{27}$$

where $\mu_{\boldsymbol{\theta}}(\pi_{\hat{\boldsymbol{\theta}}}^*)$ denotes the mean reward in the model $\boldsymbol{\theta}$ of the optimal policy using the estimated model $\hat{\boldsymbol{\theta}}$.

Lower bound of the right term. The right term equation (27) is the instantaneous regret of the estimated optimal policy $\pi_{\hat{\boldsymbol{\theta}}}^*$. For stating a lower bound on this term, we lower bound it by the expected number of mistakes in assignment times the minimal gap between the optimal policy and the estimated optimal policy when a mistake is done:

$$\mu_{\boldsymbol{\theta}}(\pi_{\boldsymbol{\theta}}^*) - \mu_{\boldsymbol{\theta}}(\pi_{\hat{\boldsymbol{\theta}}}^*) \geq E\left[\left| k \in [K], \hat{\theta}^{k+1} > \hat{\theta}^k \right|\right] \min_{k \in [K], \hat{\theta}^{k+1} > \hat{\theta}^k} \left(\mu_{\boldsymbol{\theta}}(\pi_{\boldsymbol{\theta}}^*) - \mu_{\boldsymbol{\theta}}(\pi_{\hat{\boldsymbol{\theta}}}^*)\right), \tag{28}$$

A mistake in the ranking of two arms $k_1, k_2, k_1 < k_2$ can lead to two different mistakes in assignment of players over arms: superposition of players or inversion of players Let $\Delta_{s,i}$ be the gap between the mean reward of the optimal policy (see equation (26)) with a superposition of two players and the mean reward of the optimal policy with an inversion of the two players $k_1, k_2, k_1 < k_2$:

$$\begin{aligned}
\Delta_{s,i} &= p_{k_1}\theta^{k_1}(1 - p_{k_2}) + p_{k_2}\theta^{k_1}(1 - p_{k_1}) - p_{k_1}\theta^{k_2} - p_{k_2}\theta^{k_1}, && (29)\\
&= p_{k_1}\theta^{k_1} - p_{k_2}p_{k_2}\theta^{k_1} - p_{k_1}p_{k_2}\theta^{k_1} - p_{k_1}\theta^{k_2}, && (30)\\
&= p_{k_1}(\theta^{k_1} - 2p_{k_2}\theta^{k_1} - \theta^{k_2}), && (31)\\
&= p_{k_1}(\theta^{k_1}(1 - 2p_{k_2}) - \theta^{k_1} + \epsilon), && (32)\\
&= p_{k_1}(\epsilon - 2p_{k_2}\theta^{k_1}) \geq 0, && (33)
\end{aligned}$$

where equation (32) is due to equation (21) and the last inequality is due to equation (22).

Hence, the minimum gap of any policy with the optimal policy is obtained by inverting the assignments of two players:

$$\begin{aligned}
\min_{k \in [K], \hat{\theta}^{k+1} > \hat{\theta}^k} \left(\mu_{\boldsymbol{\theta}}(\pi_{\boldsymbol{\theta}}^*) - \mu_{\boldsymbol{\theta}}(\pi_{\hat{\boldsymbol{\theta}}}^*)\right) &= \min_{\hat{k}_1 = k_2^*, \hat{k}_2 = k_1^*} \left(\mu_{\boldsymbol{\theta}}(\pi_{\boldsymbol{\theta}}^*) - \mu_{\boldsymbol{\theta}}(\pi_{\hat{\boldsymbol{\theta}}}^*)\right), && (34)\\
&= \min_{k_1, k_2, k_1 < k_2} \left(p_{k_1}\theta^{k_1} + p_{k_2}\theta^{k_2} - p_{k_1}\theta^{k_2} - p_{k_2}\theta^{k_1}\right), && (35)\\
&= \epsilon \min_{k_1, k_2, k_1 < k_2} (p_{k_1} - p_{k_2}), && (36)\\
&= \epsilon\Delta_p, && (37)
\end{aligned}$$

where $\hat{k}$ and k^* respectively denote the arm assigned to player k by the optimal policy runs on estimated mean reward $\pi_{\hat{\theta}}^*$, and the optimal arm to be assigned to player k.

Then, the expected number of mistakes in assignment is:

$$E\left[\left| k \in [K], \hat{\theta}^{k+1} > \hat{\theta}^k b \right|\right] \geq (K - 1)P(\hat{\theta}^{k+1} > \hat{\theta}^k). \tag{38}$$

Now considering the opposite event, we have:

$$P(\hat{\theta}^k \geq \hat{\theta}^{k+1}) = \int_0^1 P(u \leq \hat{\theta}^k \leq \hat{\theta}^{k+1})du + \int_0^1 P(\hat{\theta}^k \leq u \leq \hat{\theta}^{k+1})du + \int_0^1 P(\hat{\theta}^k \leq \hat{\theta}^{k+1} \leq u)du, \quad (39)$$

$$\leq \int_0^1 \left(P(u \leq \hat{\theta}^k) + P(\hat{\theta}^k \leq u) P(u \leq \hat{\theta}^{k+1}) + P(\hat{\theta}^{k+1} \leq u) \right) du, \quad (40)$$

$$\leq \int_{-\theta^k}^{1-\theta^k} \left(P(\hat{\theta}^k \geq \theta^k + u) + P(\hat{\theta}^k \geq \theta^k + u) P(\hat{\theta}^{k+1} \leq \theta^k + u) + P(\hat{\theta}^{k+1} \leq \theta^k + u) \right) du, \quad (41)$$

$$\leq \int_{-\theta^k}^{1-\theta^k} \left(P(\hat{\theta}^k \geq \theta^k + u) + P(\hat{\theta}^k \geq \theta^k + u) P(\hat{\theta}^{k+1} \leq \theta^{k+1} + \epsilon + u) + P(\hat{\theta}^{k+1} \leq \theta^{k+1} + \epsilon + u) \right) du. \quad (42)$$

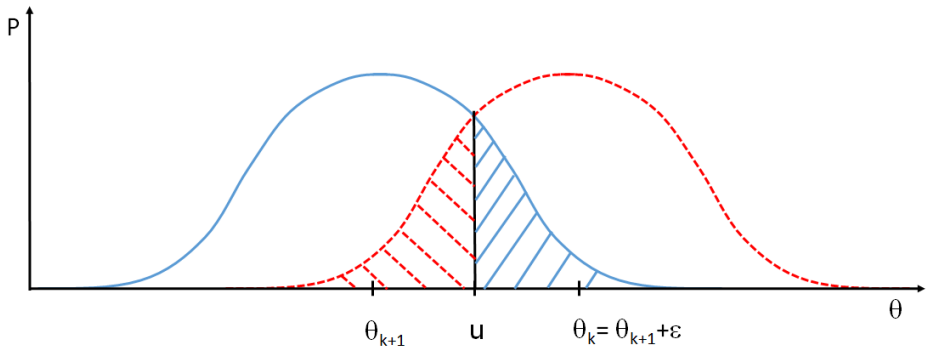

Figure 6: Satisfying equation (42) necessitates to cut the tail of the distribution of $\hat{\theta}_k$ in red dotted line, and/or the tail of the distribution of $\hat{\theta}_{k+1}$ in blue line.

Applying Hoeffding inequality to both terms (see figure 6), we obtain:

$$P(\hat{\theta}^k > \hat{\theta}^{k+1}) \leq \int_{-\theta^k}^{1-\theta^k} \left(\exp\left(-2t^* u^2\right) + \exp\left(-2t^* u^2\right) \exp\left(-2t^* \left(u+\epsilon\right)^2\right) + \exp\left(-2t^* \left(u+\epsilon\right)^2\right) \right) du, \quad (43)$$

$$\leq \int_{-\theta^k}^{1-\theta^k} \left(\exp\left(-2t^* u^2\right) + \exp\left(-4t^* u^2\right) + \exp\left(-2t^* u^2\right) \right) du, \quad (44)$$

$$\leq \int_{-\infty}^{+\infty} \left(2\exp\left(-2t^* \epsilon_1^2\right) + \exp\left(-4t^* \epsilon_1^2\right) \right) du, \quad (45)$$

$$\leq 2\sqrt{\frac{\pi}{2t^*}} + \sqrt{\frac{\pi}{4t^*}} = \sqrt{\frac{\pi}{2t^*}} \left(2 + \sqrt{\frac{1}{2}} \right). \quad (46)$$

Then, injecting equation (46) in equation (38), we have:

$$E\left[\left| k \in [K], \hat{k} \neq k^* \right| \right] \geq (K-1) \left(1 - \sqrt{\frac{\pi}{2t^*}} \left(2 + \sqrt{\frac{1}{2}} \right) \right). \quad (47)$$

Finally, the right term of equation 27 is lower bounded by:

$$\mu_{\boldsymbol{\theta}}(\pi^*_{\boldsymbol{\theta}}) - \mu_{\boldsymbol{\theta}}(\pi^*_{\hat{\boldsymbol{\theta}}}) \geq \Delta_p (K-1)\epsilon \left(1 - \sqrt{\frac{\pi}{2t^*}} \left(2 + \sqrt{\frac{1}{2}} \right) \right) = c_1 K \epsilon \left(1 - \sqrt{\frac{\pi}{t^*}} \right). \quad (48)$$

Lower bound of the left term. The left term of equation (27) is the instantaneous regret of the optimal exploration policy π_E^*. The optimal (in term of sample complexity) exploration policy cannot be an optimal target policy since estimating ϵ-approximations of arms necessitates to play exactly the same number of times the arms:

$$\mu_{\boldsymbol{\theta}}(\pi_{\boldsymbol{\theta}}^*) - \mu_{\boldsymbol{\theta}}(\pi_E^*) \geq c_2, \tag{49}$$

where $c_2 > 0$ is a constant depending on the problem parameters $\boldsymbol{\theta}$ and $p_1, ..., p_N$.

Lower bound of the regret. Now, injecting the lower bound of $\mu_{\boldsymbol{\theta}}(\pi_{\boldsymbol{\theta}}^*) - \mu_{\boldsymbol{\theta}}(\pi_E^*)$ (equation (49)) and the lower bound of $\mu_{\boldsymbol{\theta}}(\pi_{\boldsymbol{\theta}}^*) - \mu_{\boldsymbol{\theta}}(\pi_{\hat{\boldsymbol{\theta}}}^*)$ (equation (48)) in the regret decomposition (equation (27)), we obtain:

$$E[R(T)] \geq t^* c_2 + (T - t^*) c_1 K \epsilon \left(1 - \sqrt{\frac{\pi}{t^*}}\right), \tag{50}$$

$$\geq t^* c_2 + T c_1 K \epsilon \left(1 - \sqrt{\frac{\pi}{t^*}}\right) - c_1 t^* K \epsilon \left(1 - \sqrt{\frac{\pi}{t^*}}\right). \tag{51}$$

The lower bound of number of samples for finding a bias ϵ of a coin is $\Omega\left(1/\epsilon^2 \log 1/\delta\right)$ [Anthony and Bartlett, 1999]. At each time step, a maximum of $N = K$ players are sampled. Hence, the time t^* where π_E^* finds exactly an ϵ-approximation of each arm θ^k is at least:

$$\Omega\left(\frac{K}{N\epsilon^2} \log \frac{1}{\delta}\right) \Leftrightarrow \exists c_3 > 0, t^* = c_3 \frac{K}{N\epsilon^2} \log \frac{1}{\delta}. \tag{52}$$

We have:

$$E[R(T)] \geq c_3 c_2 \frac{K}{N\epsilon^2} \log \frac{1}{\delta} + T c_1 K \epsilon \left(1 - \epsilon \sqrt{\frac{N\pi}{c_3 K \log \frac{1}{\delta}}}\right) - c_1 c_3 \frac{K^2}{N\epsilon} \log \frac{1}{\delta} \left(1 - \epsilon \sqrt{\frac{N\pi}{c_3 K \log \frac{1}{\delta}}}\right). \tag{53}$$

Finally setting $\delta = 1/T$ and $\epsilon = \sqrt{K}/\sqrt[3]{T}$, obtain:

$$E[R(T)] \geq \Omega\left(T^{2/3} \frac{\log T}{N} + T^{2/3} K^{3/2} \left(1 - \sqrt{\frac{N\pi}{T^{2/3} \log T}}\right) - \frac{K^{3/2}}{N} T^{1/3} \log T \left(1 - \sqrt{\frac{N\pi}{T^{2/3} \log T}}\right)\right). \tag{54}$$

Hence, we have:

$$E[R(T)] \geq \Omega\left(T^{2/3} \left(\frac{\log T}{N} + K^{3/2}\right)\right). \tag{55}$$

C.8 Proof of Theorem 5

Lemma 4. *In Algorithm 3, so that player n sends successfully m messages, with a probability $1 - \delta$ player n needs to send a number of messages $C(m)$, which is at most:*

$$m \left\lceil \frac{\log \delta/m}{\log\left(1 - \sum_{k=1}^K \frac{\rho_n(\pi_u)}{K} \theta^k\right)} + 1 \right\rceil \text{ messages.}$$

Proof. Let $C(1)$ be the random variable corresponding to the number of trials of player n to send a message. $C(1)$ follows a geometric distribution with a probability of success $p = \mu_n(\pi_u) = \sum_{k=1}^K \frac{\rho_n(\pi_u)}{K} \theta^k$,

and probability of failure $q = 1 - p$. Let F be the number of failures before the success. We have:

$$\mathbb{P}(C(1) \leq F + 1) = 1 - q^F = 1 - \delta,$$
$$\implies F = \left\lceil \frac{\log \delta}{\log q} \right\rceil$$

Consequently, for sending m messages, with a probability $1 - \delta$ player n needs at most :

$$C(m) \leq m \left\lceil \frac{\log \delta/m}{\log(1 - \sum_{k=1}^{K} \frac{\rho_n(\pi_u)}{K} \theta^k)} + 1 \right\rceil$$

□

Theorem 5 *The total number of sent messages during Algorithm 3 is with probability $1 - \delta$ less than:*

$$2N \left\lceil \frac{\log \delta/2N}{\log(1 - \sum_{k=1}^{K} \frac{\rho_n(\pi_u)}{K} \theta^k)} + 1 \right\rceil \text{ messages.}$$

Proof. Using Lemma 4, the total number of messages sent by all players to send successfully their probabilities of sending messages and their estimations is with probability $1 - \delta$:

$$C(2N) \leq 2N \left\lceil \frac{\log \delta/2N}{\log(1 - \sum_{k=1}^{K} \frac{\rho_n(\pi_u)}{K} \theta^k)} + 1 \right\rceil \tag{56}$$

□

C.9 Proof of Theorem 6

With a probability at least $1 - \delta$, Algorithm 3 stops while finding the ϵ-approximations of θ at:

$$t^* \leq \frac{K \log 2K/\delta}{2\epsilon^2 (1 - p_1/K)^{2N-2} \sum_{i=1}^{N} p_i} + \frac{K}{p_N} \left(\sqrt{\frac{1}{2} \log \frac{NK}{\delta}} + C(2) \right),$$

where p_N is the lowest probability of sending a message among the players, and $C(2)$ is the needed number of sent messages to successfully send 2 messages.

Proof. A player n stops while finding its estimations when it plays each arm k at least t_n^* times (Lemma 2). Let t_n^k be the number of plays of arm k by player n before the algorithm stops at time t^*. t_n^k is a binomial random variable with parameters t^* and p_n/K. Then we have:

$$\mathbb{E}[t_n^k] = \frac{p_n}{K} . t^* \tag{57}$$

The estimation does not terminate if this event occurs: $E = \{\exists n \in [N], \exists k \in [K], t_n^k < t_n^* + C(2)\}$. Applying Hoeffding's inequality we get:

$$\mathcal{P}(t_n^k - \frac{p_n}{K} . t^* \leq -\epsilon) \leq \exp^{-2\epsilon^2} = \frac{\delta}{NK} \tag{58}$$

Hence, when E does not occur $\implies \forall n$ we have with probability at most δ:

$$
\begin{aligned}
& t_n^* + C(2) - \frac{p_n}{K}.t^* \leq -\sqrt{\frac{1}{2}\log\frac{NK}{\delta}} \\
& \implies \forall n \quad t^* \geq \left(\sqrt{\frac{1}{2}\log\frac{NK}{\delta}} + C(2) + p_n \frac{\log 2K/\delta}{2\epsilon^2(\prod_{n'\neq n}(1-p'_n/K))^2 \sum_{i=1}^N p_i}\right)\frac{K}{p_n} \\
& \implies \forall n \quad t^* \geq \frac{K}{p_n}\left(\sqrt{\frac{1}{2}\log\frac{NK}{\delta}} + C(2)\right) + K\frac{\log 2K/\delta}{2\epsilon^2(\prod_{n'\neq n}(1-p'_n/K))^2 \sum_{i=1}^N p_i} \\
& \implies t^* \geq \frac{K}{p_N}\left(\sqrt{\frac{1}{2}\log\frac{NK}{\delta}} + C(2)\right) + K\frac{\log 2K/\delta}{2\epsilon^2(1-p_1/K)^{2N-2} \sum_{i=1}^N p_i},
\end{aligned}
$$

Then, when E does not occur and hence the estimation terminates, we have $\forall n$ with probability at least $1-\delta$:

$$
t^* < \frac{K}{p_N}\left(\sqrt{\frac{1}{2}\log\frac{NK}{\delta}} + C(2)\right) + K\frac{\log 2K/\delta}{2\epsilon^2(1-p_1/K)^{2N-2} \sum_{i=1}^N p_i},
$$

where p_N and p_1 are respectively the lowest and the greatest probability of sending a message among the players.

□

C.10 Proof of Theorem 7

Lemma 5. *The expected instantaneous regret in the model $\boldsymbol{\theta}$ of the target policy $\pi^*_{\hat{\boldsymbol{\theta}}}$ using the estimated model $\hat{\boldsymbol{\theta}}$ with respect to the optimal policy $\pi^*_{\boldsymbol{\theta}}$ using the true model $\boldsymbol{\theta}$ is upper bounded by:*

$$
\mu_{\boldsymbol{\theta}}(\pi^*_{\boldsymbol{\theta}}) - \mu_{\boldsymbol{\theta}}(\pi^*_{\hat{\boldsymbol{\theta}}}) \leq 2K\epsilon, \tag{59}
$$

where $\mu_{\boldsymbol{\theta}}(\pi)$ denotes the mean reward of the policy π in the model $\boldsymbol{\theta}$.

Proof.

$$
\mu_{\boldsymbol{\theta}}(\pi^*_{\boldsymbol{\theta}}) - \mu_{\boldsymbol{\theta}}(\pi^*_{\hat{\boldsymbol{\theta}}}) = \mu_{\boldsymbol{\theta}}(\pi^*) - \mu_{\hat{\boldsymbol{\theta}}}(\pi^*) + \mu_{\hat{\boldsymbol{\theta}}}(\pi^*) - \mu_{\hat{\boldsymbol{\theta}}}(\pi^*_{\hat{\boldsymbol{\theta}}}) + \mu_{\hat{\boldsymbol{\theta}}}(\pi^*_{\hat{\boldsymbol{\theta}}}) - \mu_{\boldsymbol{\theta}}(\pi^*_{\hat{\boldsymbol{\theta}}}) \tag{60}
$$

Then, we have:

- $\mu_{\boldsymbol{\theta}}(\pi^*) - \mu_{\hat{\boldsymbol{\theta}}}(\pi^*) = \sum_{k=1}^K z^k l^k \theta^k - \sum_{k=1}^K z^k l^k \hat{\theta}^k \leq K\epsilon$,
- $\mu_{\hat{\boldsymbol{\theta}}}(\pi^*_{\boldsymbol{\theta}}) - \mu_{\hat{\boldsymbol{\theta}}}(\pi^*_{\hat{\boldsymbol{\theta}}}) \leq 0$, since $\pi^*_{\hat{\boldsymbol{\theta}}}$ is the best policy in the model $\hat{\boldsymbol{\theta}}$.
- $\mu_{\hat{\boldsymbol{\theta}}}(\pi^*_{\hat{\boldsymbol{\theta}}}) - \mu_{\boldsymbol{\theta}}(\pi^*_{\hat{\boldsymbol{\theta}}}) = \sum_{k=1}^K \hat{z}^k \hat{l}^k \hat{\theta}^k - \sum_{k=1}^K \hat{z}^k \hat{l}^k \theta^k \leq K\epsilon$.

□

Theorem 7: *For $\delta = 1/T, \epsilon = K/\sqrt[3]{T}$, when $\forall n \in [N], p_n = p$, the expected cumulative regret with respect to the target policy π^* of Algorithm 3 followed by a policy $\pi^*_{\hat{\theta}}$ is upper bounded by:*

$$
E[R(T)] \leq T^{2/3}\left(2K^2 + \frac{\log 2KT}{2(1-p/K)^{2N-2}Np}\right) + \frac{K^2}{p}\left(\sqrt{\frac{1}{2}\log NKT} + C(2)\right) + K
$$

Proof. Let T be the time horizon, π_u be the uniform policy used in Algorithm 3, which outputs an ϵ-approximation with high probability of $\boldsymbol{\theta}$, and $\pi^*_{\boldsymbol{\theta}}$ be the optimal policy. Let t^* be stopping time of the exploration phase. Then, the expected cumulative regret with respect to a target policy $\pi^*_{\boldsymbol{\theta}}$ of Algorithm 3 is expressed as:

$$
E[R(T)] = t^*((\mu_{\boldsymbol{\theta}}(\pi^*_{\boldsymbol{\theta}}) - (\mu_{\boldsymbol{\theta}}(\pi_u)) + (T-t^*)((\mu_{\boldsymbol{\theta}}(\pi^*_{\boldsymbol{\theta}}) - \mu_{\boldsymbol{\theta}}(\pi^*_{\hat{\boldsymbol{\theta}}})), \tag{61}
$$

where $\mu_{\boldsymbol{\theta}}(\pi^*_{\hat{\boldsymbol{\theta}}})$ denotes the mean reward in the model $\boldsymbol{\theta}$ of the optimal policy using the estimated model $\hat{\boldsymbol{\theta}}$. The left term of equation 61 is the instantaneous regret of the exploration policy $\pi_{\mathcal{U}}$, and the right term is the instantaneous regret of the estimated optimal policy $\pi^*_{\hat{\boldsymbol{\theta}}}$.

Theorem 6 allows us to upper-bound the stopping time of Algorithm 3 with t^* on an event of high probability $1-\delta$:

$$t^* \leq \frac{K}{p_N}\left(\sqrt{\frac{1}{2}\log\frac{NK}{\delta}} + C(2)\right) + K\frac{\log 2K/\delta}{2\epsilon^2(1-p_1/K)^{2N-2}\sum_{i=1}^N p_i} \tag{62}$$

When $\forall n \in [N], p_n = p$, we have:

$$t^* \leq \frac{K}{p}\left(\sqrt{\frac{1}{2}\log\frac{NK}{\delta}} + C(2)\right) + K\frac{\log 2K/\delta}{2\epsilon^2(1-p/K)^{2N-2}Np} \tag{63}$$

The regret of uniform policy with respect to the optimal policy $\pi^*_{\boldsymbol{\theta}}$ is upper bounded by:

$$\mu_\theta(\pi^*_{\boldsymbol{\theta}}) - \mu_{\boldsymbol{\theta}}(\pi_u) \leq K$$

and on the other hand we know by Lemma 5 that:

$$\mu_{\boldsymbol{\theta}}(\pi^*_{\boldsymbol{\theta}}) - \mu_{\boldsymbol{\theta}}(\pi^*_{\hat{\boldsymbol{\theta}}}) \leq 2K\epsilon \tag{64}$$

Then the expected cumulative regret is controlled by the trivial upper bound KT on the complementary event of probability less than δ:

$$E[R(T)] \leq t^*(\mu_\theta(\pi^*_{\boldsymbol{\theta}}) - \mu_{\boldsymbol{\theta}}(\pi_u)) + (T-t^*)(\mu_{\boldsymbol{\theta}}(\pi^*_{\boldsymbol{\theta}}) - \mu_{\boldsymbol{\theta}}(\pi^*_{\hat{\boldsymbol{\theta}}})) + \delta KT \tag{65}$$

$$\tag{66}$$

Then, by setting $\delta = 1/T$, the expected regret of Algorithm 3 followed by a policy $\pi^*_{\hat{\boldsymbol{\theta}}}$ is:

$$\begin{aligned} E[R(T)] &\leq Kt^* + (T-t^*)\times 2K\epsilon + K && (67)\\ &\leq Kt^* + 2K\epsilon T + K && (68)\\ &\leq \frac{K^2}{p}\left(\sqrt{\frac{1}{2}\log NKT} + C(2)\right) + \frac{K^2\log 2KT}{2\epsilon^2(1-p/K)^{2N-2}Np} + 2K\epsilon T + K && (69)\\ & && (70)\end{aligned}$$

Finally, by setting $\epsilon = K/\sqrt[3]{T}$, we conclude the proof:

$$E[R(T)] \leq T^{2/3}\left(2K^2 + \frac{\log 2KT}{2(1-p/K)^{2N-2}Np}\right) + \frac{K^2}{p}\left(\sqrt{\frac{1}{2}\log NKT} + C(2)\right) + K \tag{71}$$

□

C.11 Proof of Theorem 8

Applying Algorithm 2 on a model estimate $\hat{\boldsymbol{\theta}}$ returns an α-fair policy in the true model $\boldsymbol{\theta}$:

$$\alpha \geq 1 - \max_{n\in[N]} p_n - \frac{2\|\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\|_\infty}{\max_{n\in[N]}\frac{\hat{\theta}^{k_n}z^{k_n}}{1-p_n}} \tag{72}$$

Proof. Theorem 3 states that the policy returned by Algorithm 2, denoted as $\pi^\dagger$ has the following fairness guarantees:

$$\hat{\alpha} = \frac{\min_{n\in[N]} \mu_{n,\hat{\boldsymbol{\theta}}}(\pi^\dagger)}{\max_{n\in[N]} \mu_{n,\hat{\boldsymbol{\theta}}}(\pi^\dagger)} \geq 1 - \max_{n\in[N]} p_n, \tag{73}$$

with $\mu_{n,\hat{\boldsymbol{\theta}}}(\pi^\dagger)$ denoting the expectation of rewards received by player n in estimated model $\hat{\boldsymbol{\theta}}$ when following policy $\pi^\dagger$. We may write it as follows:

$$\mu_{n,\hat{\boldsymbol{\theta}}}(\pi^\dagger) = \hat{\theta}^{k_n} \prod_{n',\text{ s.t. } k_{n'}=k_n} (1-p_{n'}) = \frac{\hat{\theta}^{k_n} z^{k_n}}{1-p_n}. \tag{74}$$

We therefore get:

$$\begin{aligned}
\alpha &= \frac{\min_{n\in[N]} \mu_{n,\boldsymbol{\theta}}(\pi^\dagger)}{\max_{n\in[N]} \mu_{n,\boldsymbol{\theta}}(\pi^\dagger)} && (75)\\
&= \frac{\min_{n\in[N]} \frac{\theta^{k_n} z^{k_n}}{1-p_n}}{\max_{n\in[N]} \frac{\theta^{k_n} z^{k_n}}{1-p_n}} && (76)\\
&\geq \frac{\min_{n\in[N]} \frac{\hat{\theta}^{k_n} z^{k_n}}{1-p_n} - \|\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\|_\infty}{\max_{n\in[N]} \frac{\hat{\theta}^{k_n} z^{k_n}}{1-p_n} + \|\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\|_\infty} \qquad \text{since } \frac{z^{k_n}}{1-p_n} \leq 1, \forall n && (77)\\
&= \hat{\alpha} - \frac{2\|\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\|_\infty}{\max_{n\in[N]} \frac{\hat{\theta}^{k_n} z^{k_n}}{1-p_n} + \|\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\|_\infty} && (78)\\
&\geq 1 - \max_{n\in[N]} p_n - \frac{2\|\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\|_\infty}{\max_{n\in[N]} \frac{\hat{\theta}^{k_n} z^{k_n}}{1-p_n}} && (79)
\end{aligned}$$

□