Decentralized Exploration in Multi-Armed Bandits

Raphaël Féraud [1] **Réda Alami** [1] **Romain Laroche** [2]

Abstract

We consider the *decentralized exploration problem*: a set of players collaborate to identify the best arm by asynchronously interacting with the same stochastic environment. The objective is to ensure privacy in the best arm identification problem between asynchronous, collaborative, and thrifty players. In the context of a digital service, we advocate that this decentralized approach allows a good balance between conflicting interests: the providers optimize their services, while protecting privacy of users and saving resources. We define the privacy level with respect to the amount of information an adversary could infer by intercepting all the messages concerning a single user. We provide a generic algorithm DECENTRALIZED ELIMINATION, which uses any best arm identification algorithm as a subroutine. We prove that this algorithm ensures privacy, with a low communication cost, and that in comparison to the lower bound of the best arm identification problem, its sample complexity suffers from a penalty depending on the inverse of the probability of the most frequent players. Then, thanks to the generality of the approach, we extend the proposed algorithm to the non-stationary bandits. Finally, experiments illustrate and complete the analysis.

1. Introduction

1.1. Motivations

We consider a collaborative exploration problem, the *decentralized exploration problem*. The main motivation of this new problem setting comes from sequential A/B and multivariate testing applications. For instance, most digital applications perform sequential A/B and multivariate testing in order to optimize the value of their audience. When a device is connecting to the application, the application presents an option to the user of the device. The aim is to maximize the clicks of users on the proposed options. Using the standard centralized exploration approach, the click stream of users is gathered and processed to choose the option which generates the most clicks. In this paper, we formulate this standard exploration problem in a decentralized way.

When the event *"player n is active"* occurs, player n reads the messages received from other players and then chooses an arm to play. The reward of the played arm is revealed to player n. Finally, she may send a message to the other players for sharing information about the arms.

The decentralized approach presents significant advantages. First, the clicks of users contain information that may be embarrassing when revealed, or that can be used by a third party in an undesirable way. The decentralization of exploration favors privacy since the click stream is not transmitted. However, it is not sufficient. The messages sent by a user may still contain private information such as her favorite topics, and therefore her political views, sexual orientation... As the players broadcast messages to other players, a malicious adversary can pretend to be a player, and then listening the exchanged messages. To ensure privacy one must guarantee that no useful information can be inferred from the messages sent by a single user. Second, the decentralization of exploration reduces the communication cost. This is a significant requirement for the Internet of Thing applications, since the smart devices often run on batteries. Third and finally, for all digital applications and in particular for the mobile phone applications, the decentralization with a low communication cost increases the responsiveness of applications by minimizing the number of interactions between the application server and the devices.

Finally, the objective of the *decentralized exploration problem* is threefold:

1. *sample efficiency*: finding a near-optimal arm with high-probability using a minimal number of interactions with the environment.
2. *user privacy*: protecting information contained in the interaction history of a single player.
3. *low communication cost*: minimizing the number of exchanged messages.

*Equal contribution [1]Orange Labs [2]Microsoft Research. Correspondence to: Raphaël Féraud <raphael.feraud@orange.com>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

1.2. Related works

The problem of the best arm identification has been studied in two distinct settings in the literature:

- the fixed budget setting: the duration of the exploration phase is fixed and is known by the forecaster, and the objective is to maximize the probability of returning the best arm (Bubeck et al, 2009; Audibert et al, 2010; Gabillon et al, 2013);
- the fixed confidence setting: the objective is to minimize the number of rounds needed to achieve a fixed confidence to return the best arm (Even-Dar et al, 2006; Kalyanakrishnan et al, 2012; Gabillon et al, 2013; Kaufmann and Kalyanakrishnan, 2013).

In this paper, we focus on the fixed confidence setting. Its theoretical analysis is based on the *Probably Approximately Correct* framework (Valiant, 1984), and focuses on the sample complexity to identify a near-optimal arm with high probability. This theoretical framework has been used to analyze the best arm identification problem in (Even-Dar et al, 2006), the dueling bandit problem in (Urvoy et al, 2013), the batched bandit problem in (Perchet et al, 2015), the linear bandit problem in (Soare et al, 2014), the contextual bandit problem in (Féraud et al, 2016), and the non-stationary bandit problem in (Allesiardo et al, 2017).

The *decentralized multi-player multi-armed bandits* have been studied for opportunistic spectrum access in (Liu and Zhao, 2010; Avner and Mannor, 2014; Nayyar et al, 2015) or for optimizing communications in Internet of Things, even when no sensing information is available (Besson and Kaufmann, 2018). The objective is to avoid collisions between concurrent players that share the same channels, while choosing the best channels and minimizing the communication cost between players.

Recent years have seen an increasing interest for the study of the distributed collaborative scheme, where there is no collision when players choose the same arm at the same time. The distributed collaborative multi-armed bandits have been studied when the agents communicate through a neighborhood graph in (Szörényi et al, 2013; Landgren et al, 2016). Here, we allow each player to broadcast messages to all players. In (Chakraborty et al, 2017), a team of agents collaborate to handle the same multi-armed bandit problem. At each step the agent can broadcast her last obtained reward for the chosen arm to the team or pull an arm. The communication cost corresponds to the lost of the potential reward. As the pull of the arm of the agent is broadcasted, this approach does not ensure privacy of users. The tradeoff between the communication cost and the regret has been studied in the case of distributed collaborative non-stochastic experts (Kanade et al, 2012). In (Hernandez-Lobato et al, 2017), the best arm identification task with fixed budget is distributed using Thompson Sampling in order to accelerate the exploration of the chemical space. In (Hillel et al, 2013), the best arm identification task with fixed confidence is distributed on a parallel processing architecture. The analysis focuses on the trade-off between the number of communication rounds and the number of pulls per player. Here, we consider here that the players activation is under the control of the environment. As a consequence, synchronized communication rounds can no longer be used to control the communication cost. In our paper, the cost of communications is assessed by the number of exchanged messages.

Moreover, our purpose is also to protect privacy of players. In the current context of massive storage of personal data and massive usage of models inferred from personal data, privacy is an issue. Even if individual data are anonymized, the pattern of data associated with an individual is itself uniquely identifying. The k-anonymity approach (Sweeney, 2002) provides a guarantee to resist to direct linkage between stored data and the individuals. However, this approach can be vulnerable to composition attacks: an adversary could use side information that combined with the k-anonymized data allows to retrieve a unique identifier (Ganta et al, 2008). The *differential privacy* (Dwork et al, 2006) provides an alternative approach. The sensitive data are hidden. The guarantee is provided by algorithms that allow to extract information from data. An algorithm is differentially private if the participation of any record in the database does not alter the probability of any outcome by very much. The *differential privacy* has been extended to *local differential privacy* in which the data remains private even from the learner (Duchi et al, 2014). In (Gajane et al, 2018), the authors propose an approach which handles the stochastic multi-armed bandit problem, while ensuring *local differential privacy*. The ϵ-differential privacy is ensured to the players by using a stochastic corruption of rewards. As all the rewards are transmitted to a centralized bandit algorithm, this approach has the maximum communication cost. Here, we define the privacy level with respect to the information about the preferred arms of a player, that an adversary could infer by intercepting the messages of this player. The messages could be corrupted feedbacks as in (Gajane et al, 2018), or as we choose a more compact representation of the same information.

1.3. Our contribution

In Section 2, we propose a new problem setting for ensuring privacy in the best arm identification problem between asynchronous, collaborative, and thrifty players. In Section 3, we propose a generic algorithm, DECENTRALIZED ELIMINATION, which handles the *decentralized exploration problem* using any best arm identification algorithm as a subroutine.

Theorem 1 states that DECENTRALIZED ELIMINATION ensures privacy, finds an approximation of the best arm with high probability, and requires a low communication cost. Furthermore, Theorem 2 states a generic upper bound of the sample complexity of DECENTRALIZED ELIMINATION. More specifically, Corollary 1 and 2 state the sample complexity bound when respectively MEDIAN ELIMINATION and SUCCESSIVE ELIMINATION (Even-Dar et al, 2006) are used as subroutine. Then, in Section 4, we extend the algorithmic approach to the *decentralized exploration in non-stationary bandit problem*. In Section 5, to illustrate and complete the analysis, we empirically compare the performances of DECENTRALIZED ELIMINATION with two natural baselines (Kanade et al, 2012): an algorithm that does not share any information between the players, and hence that ensures privacy with a zero communication cost, and a centralized algorithm that shares all the information between players, and hence that does not ensure privacy and that has the maximum communication cost.

2. The decentralized exploration problem

Let $\mathcal{N} = \{1, ..., N\}$ be a set of N players. Let $x \in \mathcal{N}$ be a discrete random variable which realization denotes the index n of the *active* player (the player for which an event occurs). Let P_x be the probability distribution of x which is assumed to be stationary and unknown to the players. Let $\mathcal{K} = \{1, ..., K\}$ be a set of K arms. Let $y_k^n \in [0, 1]$ be the bounded random variable which realization denotes the reward of arm k for player n, and μ_k^n be its mean reward. Let $\mathbf{y}_{x=n} = \{y_k^n\}_{k \in \mathcal{K}}$ be the vector of independent random variables y_k^n. Let $P_{\mathbf{y}}$ and $P_{x,\mathbf{y}}$ be respectively the probability distribution of $\mathbf{y}$ and the joint probability distribution of x and $\mathbf{y}$, which are assumed to be unknown to the players.

Assumption 1 (stationary rewards). The mean reward of arms does not depend on time: $\forall t$, $\forall n \in \mathcal{N}$,and $\forall k \in \mathcal{K}$, $\mu_k^n(t) = \mu_k^n$.

Assumption 2 (multi-armed bandits). The mean reward of arms does not depend on the player: $\forall n \in \mathcal{N}$ and $\forall k \in \mathcal{K}$, $\mu_k^n = \mu_k$.

Assumption 1 and 2 are used to focus on the stochastic multi-armed bandits. This section lays the theoretical foundations of the *decentralized exploration problem* in its elementary form. The next section proposes an extension to the *decentralized exploration in non-stationary bandits*. The extension to the *decentralized exploration in contextual bandits* is discussed in future works.

Definition 1 (ϵ-optimal arm). An arm $k \in \mathcal{K}$ is said to be ϵ-optimal, if $\mu_k \geq \mu_{k^*} - \epsilon$, where $k^* = \arg\max_{k \in \mathcal{K}} \mu_k$ and $\epsilon \in (0, 1]$. $\mathcal{K}_\epsilon$ denotes the set of ϵ-optimal arms.

Definition 2 (message). A message $\lambda_k^n \in \{0, 1\}$ is a binary random variable, that is sent by player n to other players, and where $\lambda_k^n = 1$ means that player n estimates that k is not an ϵ-optimal arm.[1]

Let $\mathcal{M}_n$ be the set of sent messages by player n at stopping time. Let $\mathcal{K}^n(l^n) \subseteq \mathcal{K}$ be the set of remaining arms at epoch $l^n \in \{1, ..., L\}$ for player n, where L is the maximal number of epochs.

Definition 3 ((ϵ, η)-private). The decentralized algorithm $\mathcal{A}$ is (ϵ, η)-private for finding an ϵ-optimal arm, if for any player n, an adversary, that knows $\mathcal{M}_n$, the set of messages of player n, and the algorithm $\mathcal{A}$, cannot infer what arm is ϵ-optimal for player n with a probability higher than $1 - \eta$:

$$\forall n \in \mathcal{N}, \forall l^n \in \{1, ..., L\}, \nexists \eta_1, 0 \leq \eta_1 < \eta \leq 1,$$
$$\mathbb{P}\left(\mathcal{K}^n(l^n) \subseteq \mathcal{K}_\epsilon | \mathcal{M}_n, \mathcal{A}\right) \geq 1 - \eta_1.$$

$1 - \eta$ is the confidence level associated to the decision of the adversary. If η is small, then the adversary can use the set of messages $\mathcal{M}_n$ to infer with high probability which arm is an ϵ-optimal arm for player n. If η is high, the only information, that can be inferred by the adversary, is that the probability that an arm is an ϵ-optimal of arm for player n is a little bit higher than 0, which can be much lesser than the random choice $1/K$. η is a parameter which allows to tune the level of privacy: the higher η, the higher the privacy protection.

The goal of the *decentralized exploration problem* (see Algorithm 1) is to design an algorithm, that, when run on each player, samples effectively to find an ϵ-optimal arm for each player, while ensuring (ϵ, η)-privacy to players, and minimizing the number of exchanged messages.

Algorithm 1 DECENTRALIZED EXPLORATION PROBLEM

Inputs: $\mathcal{K}$, $\epsilon \in [0, 1]$, $\eta \in [0, 1]$
Output: an arm in each set $\mathcal{K}^n(l^n)$
Initialization: $l^n := 1$, $\mathcal{K}^n(l^n) := \mathcal{K}$

repeat
  a player is sampled: $n \sim P_x$
  player n gets the messages of other players
  arm $k \in \mathcal{K}^n(l^n)$ is played by player n
  player n receives reward $y_k^n \sim P_{x=n,y}$
  if player n updates $\mathcal{K}^n(l^n)$ **then** $l^n := l^n + 1$
  player n sends a message to other players
until $(\forall n \in \mathcal{N}, |\mathcal{K}^n(l^n)| = 1)$

The lower bound of the number of samples in $P_{x,\mathbf{y}}$ needed to find with high probability an ϵ-optimal arm, which is

[1] We choose a Bernoulli random variable for the sake of clarity. Notice that any random variable could be used as message.

$\Omega\left(\frac{K}{\epsilon^2}\log\frac{1}{\delta}\right)$ (Mannor and Tsitsiklis, 2004), holds for the *decentralized exploration problem*, since a message can be sent at each time an arm is sampled by a player. The number of messages, that has to be exchanged in order to find with high probability an ϵ-optimal arm, could be zero if each player independently handles the best arm identification problem.

Assumption 3 (all players are active). $\forall n \in \mathcal{N}, P_x(x = n) \neq 0$.

Assumption 3 is a sanity check assumption for the *decentralized exploration problem*. Indeed, if it exists a player n such that $P_x(x = n) = 0$, then Algorithm 1 never stops (the stopping condition line 8 never happens).

3. Decentralized Elimination

3.1. ArmSelection subroutine

Before describing a generic algorithm for the *decentralized exploration problem*, we need to define an ArmSelection subroutine that handles all best arm identification algorithms. Let $\overline{\mathcal{K}^n}(l^n)$ and $\mathcal{K}^n(l^n)$ be respectively the set of eliminated arms and the set of remaining arms of player n at elimination epoch l^n, such that $\overline{\mathcal{K}^n}(l^n) \cup \mathcal{K}^n(l^n) = \mathcal{K}^n(l^n - 1)$.

Definition 4 (ArmSelection subroutine). an ArmSelection subroutine takes as parameters an approximation factor ϵ, a confidence level $1 - \eta$, and a set of remaining arm $\mathcal{K}^n(l^n)$. It samples a remaining arm in $\mathcal{K}^n(l^n)$ and returns the set of eliminated arms $\overline{\mathcal{K}^n}(l^n)$. An ArmSelection subroutine satisfies Properties 1 and 2.

Let t^n be the number of calls of the ArmSelection subroutine. Let $\mathcal{H}_{t^n}$ be the sequence of rewards of chosen arms $\{(k_1, y^n_{k_1}), (k_2, y^n_{k_2}), ..., (k_{t^n}, y^n_{k_{t^n}})\}$. Let $f : \{1, ..., L\} \to [0, 1]$ be a function such that $\sum_{l^n=1}^{L} f(l^n) = 1$.

Property 1. (remaining ϵ-optimal arm)

$$\forall l^n \in \{1, ..., L\}, \mathcal{K}^n\left(l^n\right) \subset \mathcal{K}^n\left(l^n - 1\right),$$
$$\mathbb{P}\left(\{\mathcal{K}^n(l^n) \cap \mathcal{K}_\epsilon = \emptyset\} | \mathcal{H}_{t^n}, \mathcal{K}^n\left(l^n - 1\right) \cap \mathcal{K}_\epsilon \neq \emptyset\right) \leq \eta \times f(l^n).$$

Property 2. (finite sample complexity)

$$\exists t^n \geq 1, \forall \eta \in (0, 1), \forall \epsilon \in (0, 1],$$
$$\mathbb{P}\left(\{\mathcal{K}^n(L) \subset \mathcal{K}_\epsilon\} | \mathcal{H}_{t^n}\right) \geq 1 - \eta.$$

Property 1 ensures that with high probability at least an ϵ-optimal arm remains in the set of arms $\mathcal{K}^n(l^n)$, while Property 2 ensures that the ArmSelection subroutine finds in a finite time an ϵ-optimal arm whatever the confidence level $1 - \eta$ and the approximation factor ϵ. To the best of our knowledge, all best arm identification algorithms can be used as ArmSelection subroutine with straightforward transformations. We consider three classes of best arm identification algorithms.

The fixed-design algorithms use *uniform sampling* during a predetermined number of samples. NAIVE ELIMINATION ($L = 1$ and $f(l^n) = 1$) and MEDIAN ELIMINATION ($L = \log_2 K$ and $f(l^n) = 1/2^{l^n}$) (Even-Dar et al, 2006) are *fixed-design* algorithms which can be used as ArmSelection subroutines.

The successive elimination algorithms are based on *uniform sampling* and *arm eliminations*. At each time step a remaining arm is uniformly sampled. The empirical mean of the played arm is updated. The arms, which cannot be an ϵ-optimal arm with high probability, are discarded. If suboptimal arms are discarded the epoch l is increased by one. SUCCESSIVE ELIMINATION ($L = K$ and $f(l^n) = 1/K$) (Even-Dar et al, 2006), KL-RACING ($L = K$ and $f(l^n) = 1/K$) (Kaufmann and Kalyanakrishnan, 2013) are *successive elimination* algorithms which can be used as ArmSelection subroutines.

The explore-then-commit algorithms are based on *adaptive sampling* and *a stopping rule*. Rather than choosing arms uniformly, the *explore-then-commit* algorithms play one of the two critical arms: the empirical best arm, and the empirical suboptimal arm associated with the maximum upper confidence bound. The stopping rule simply tests if the difference, between the maximum of upper confidence bound of suboptimal arms and the lower confidence bound of the empirical best arm, is higher than the approximation factor ϵ. When the algorithm stops it returns the best arm. LUCB (Kalyanakrishnan et al, 2012), KL-LUCB (Kaufmann and Kalyanakrishnan, 2013), UGAPEC (Gabillon et al, 2013) can also be used as ArmSelection subroutines by returning the set of eliminated arms when the stopping event occurs ($L = 1$ and $f(l^n) = 1$).

3.2. Algorithm description

The basic idea of DECENTRALIZED ELIMINATION is to use the vote of independent players, which communicate the arm they would like to eliminate with a high probability of failure for ensuring privacy. As the players are independent, the probability of failure of the vote is the multiplication of the individual probability of failures. The number of players needed for eliminating an arm is provided by the analysis.

DECENTRALIZED ELIMINATION (see Algorithm 2) takes as parameters the privacy level η, the failure probability δ, the approximation factor ϵ, and an ArmSelection subroutine. It outputs an ϵ-optimal arm for each player with high

probability. The algorithm sketch is described below.

When player n is active (i.e. when player n is sampled):

- player n gets messages from other players (line 3).
- When enough players have eliminated an arm, it is eliminated from the shared set of arms $\mathcal{K}(l)$ and from the set of arms $\mathcal{K}^n(l^n)$ of player n with a low probability of failure (lines 5-10).
- When there is only one arm in $\mathcal{K}(l)$, it is an ϵ-optimal arm with high probability $1-\delta$, and the set of arms of player n is $\mathcal{K}(l)$ (line 11).
- An ArmSelection subroutine, run with a low confidence level $1-\eta$ (i.e. high privacy level) on the set $\mathcal{K}^n(l^n)$, samples an arm and returns $\overline{\mathcal{K}}^n(l^n)$ the set of arms that player n has eliminated at step t^n (line 13).
- When player n has eliminated an arm, she communicates to other players the index of the arm (lines 14-20).

Algorithm 2 DECENTRALIZED ELIMINATION

Inputs: $\epsilon \in (0,1]$, $\eta \in (0,1)$, $\delta \in [\eta^N, \eta^2]$, $\mathcal{K}$, an ArmSelection subroutine
Output: an arm in each set $\mathcal{K}^n(l^n)$
Initialization: $l := 1$, $\mathcal{K}(l) := \mathcal{K}$, $\forall n\ t^n := 1$, $l^n := 1$, $\mathcal{K}^n(l^n) := \mathcal{K}$, $\forall\,(k,n)\ \lambda_k^n := 0$

1: **repeat**
2: player n is sampled: $n \sim P_x$
3: player n gets the messages λ_k^j from other players
4: **if** $|\mathcal{K}(l)| > 1$ **then**
5: **for** all $k \in \mathcal{K}(l)$ **do**
6: **if** $\sum_{j=1}^N \lambda_k^j \geq \lfloor \frac{\log \delta}{\log \eta} \rfloor$ **then**
7: $\mathcal{K}(l) := \mathcal{K}(l) \setminus \{k\}$, $l := l+1$
8: $\mathcal{K}^n(l^n) := \mathcal{K}^n(l^n) \setminus \{k\}$
9: **end if**
10: **end for**
11: **else** $\mathcal{K}^n(l^n) := \mathcal{K}(l)$
12: **end if**
13: $\overline{\mathcal{K}}^n(l^n) := \mathrm{ArmSelection}(\epsilon, \eta, \mathcal{K}^n(l^n))$
14: **if** $|\overline{\mathcal{K}}^n(l^n)| > 1$ **then**
15: $l^n := l^n + 1$
16: **for** all $k \in \overline{\mathcal{K}}^n(l^n)$ **do**
17: $\mathcal{K}^n(l^n) := \mathcal{K}^n(l^n) \setminus \{k\}$
18: $\lambda_k^n := 1$, λ_k^n is sent to other players
19: **end for**
20: **end if**
21: $t^n := t^n + 1$
22: **until** $\forall n\ |\mathcal{K}^n(l^n)| = 1$

3.3. Analysis of the algorithm

Theorem 1 states the upper bound of the communication cost for obtaining with high probability an ϵ-optimal arm while ensuring (ϵ, η)-privacy to the players. The communication cost depends only on the problem parameters: the privacy constraint η, the probability of failure δ, the number of actions, and notably not on the number of samples. Notice that the probability of failure is low since the failure probability is lower than the level of privacy guarantee: $\delta < \eta$.

Theorem 1. *Using any* ArmSelection *subroutine,* DECENTRALIZED ELIMINATION *is an (ϵ, η)-private algorithm, that finds an ϵ-optimal arm with a failure probability $\delta \leq \eta^{\lfloor \frac{\log \delta}{\log \eta} \rfloor}$ and that exchanges at most $\lfloor \frac{\log \delta}{\log \eta} \rfloor K - 1$ messages.*

To finely analyze the sample complexity of DECENTRALIZED ELIMINATION algorithm, one needs to handle the randomness of the voting process. Let $T_{P_{x,\mathbf{y}}}$ be the number of samples in $P_{x,\mathbf{y}}$ at stopping time. Let $T_{P_\mathbf{y}}$ be the number of samples in $P_\mathbf{y}$ needed by the ArmSelection subroutine to find an ϵ-optimal arm with high probability. Let $\mathcal{N}_M$ be the set of the $M = \lfloor \frac{\log \delta}{\log \eta} \rfloor$ most likely players, let $p^* = \min_{n \in \mathcal{N}_M} P_x(x=n)$, and let $p^\dagger = \min_{n \in \mathcal{N}} P_x(x=n)$.

Theorem 2. *Using any* ArmSelection *subroutine, with a probability higher than* $(1-\delta)\left(1 - I_{1-p^*}\left(T_{P_{x,\mathbf{y}}} - T_{P_\mathbf{y}}, 1 + T_{P_\mathbf{y}}\right)\right)^{\lfloor \frac{\log \delta}{\log \eta} \rfloor}$ DECENTRALIZED ELIMINATION *stops after:*

$$\mathcal{O}\left(\frac{1}{p_*}\left(T_{P_\mathbf{y}} + \sqrt{\frac{1}{2}\log\frac{1}{\delta}}\right)\right) \text{ samples in } P_{x,\mathbf{y}},$$

where $I_a(b,c)$ denotes the incomplete beta function evaluated at a with parameters b, c.

As the number of players involved in the vote is set as small as possible $\lfloor \frac{\log \delta}{\log \eta} \rfloor$, Theorem 2 provides with high probability [2] the sample complexity of DECENTRALIZED ELIMINATION. Notice, that when the number of players is high, and when the distribution of players is far from the uniform distribution, we have $p^* \gg p^\dagger$.

Corollary 1. *With a probability higher than* $(1-\delta)\left(1 - I_{1-p^*}\left(T_{P_{x,\mathbf{y}}} - T_{P_\mathbf{y}}, 1 + T_{P_\mathbf{y}}\right)\right)^{\lfloor \frac{\log \delta}{\log \eta} \rfloor}$ DECENTRALIZED MEDIAN ELIMINATION *stops after:*

$$\mathcal{O}\left(\frac{1}{p^*}\left(\frac{K}{\lfloor \frac{\log \delta}{\log \eta} \rfloor \epsilon^2}\log\frac{1}{\delta} + \sqrt{\frac{1}{2}\log\frac{1}{\delta}}\right)\right) \text{ samples in } P_{x,\mathbf{y}}.$$

[2] for instance, $I_{0.99}(500, 500) = 1.47 \times 10^{-302}$

Corollary 2. *With a probability higher than* $(1 - \delta)\left(1 - I_{1-p^*}\left(T_{P_{x,\mathbf{y}}} - T_{P_{\mathbf{y}}}, 1 + T_{P_{\mathbf{y}}}\right)\right)^{\lfloor \frac{\log \delta}{\log \eta} \rfloor}$ DECENTRALIZED SUCCESSIVE ELIMINATION *stop after:*

$$\mathcal{O}\left(\frac{1}{p^*}\left(\frac{K}{\epsilon^2}\left(\log K + \frac{1}{\lfloor \frac{\log \delta}{\log \eta} \rfloor}\log\frac{1}{\delta}\right) + \sqrt{\frac{1}{2}\log\frac{1}{\delta}}\right)\right)$$

samples in $P_{x,\mathbf{y}}$.

Corollary 1 and 2 state the number of samples in $P_{x,\mathbf{y}}$ needed to find an ϵ-optimal arm by DECENTRALIZED ELIMINATION using respectively MEDIAN ELIMINATION and SUCCESSIVE ELIMINATION as ArmSelection subroutines.

To illustrate these results, we consider the case of the uniform distribution of players. With a failure probability at most $\delta = \eta^N$ the number of sample in $P_{x,\mathbf{y}}$ needed by DECENTRALIZED MEDIAN ELIMINATION to find an ϵ-optimal arm is:

$$\mathcal{O}\left(\frac{K}{\epsilon^2}\log\frac{1}{\delta} + N\sqrt{\frac{1}{2}\log\frac{1}{\delta}}\right) \text{ samples in } P_{x,\mathbf{y}}.$$

In comparison to an optimal best arm identification algorithm, which communicates all the messages and does not provide privacy protection guarantee, which has a sample complexity in $\mathcal{O}\left(\frac{K}{\epsilon^2}\log\frac{1}{\delta}\right)$, the sample complexity of DECENTRALIZED ELIMINATION mostly suffers from a penalty depending on the inverse of the probability of the most frequent players, that in the case of uniform distribution of players is linear with respect to the number of players. The proofs of Theorem 2, Corollary 1 and 2 are provided in the appendix.

4. Decentralized exploration in non-stationary bandits

Recently, the best arm identification problem has been studied in the case of non-stationary bandits, where Assumption 1 does not hold (Allesiardo et al, 2017; Abbasi-Yadkori et al, 2018). In the first reference, the authors analyze the non-stationary stochastic best-arm identification in the fixed confidence setting by splitting the game into independent sub-games where the best arm does not change. In the second reference, the authors propose a simple and anytime algorithm, which is analyzed for stochastic and adversarial rewards in the case of fixed budget setting. For the consistency of the paper, which focuses on fixed confidence setting, we choose to extend DECENTRALIZED ELIMINATION to SUCCESSIVE ELIMINATION with RANDOMIZED ROUND-ROBIN (SER3 (Allesiardo et al, 2017)). Basically, SER3 consists in shuffling the set of arms at each step of SUCCESSIVE ELIMINATION. SER3 works for the sequences where Assumption 4 holds.

Assumption 4 (Positive mean-gap) For any $k \in \mathcal{K} \setminus \{k^*\}$ and any $[\tau] \in \mathbb{T}(\tau)$ with $\tau \geq \log\frac{K}{\eta}$, we have:

$$\Delta_k^*([\tau]) = \frac{1}{\tau}\sum_{i=1}^{\tau}\sum_{j=i}^{i+K_i-1}\frac{\Delta_{k^*,k}(j)}{K_i} > 0,$$

where $\mathbb{T}(\tau)$ is the set containing all possible realizations of τ round-robin steps, $\Delta_{k^*,k}(t)$ is the difference between the mean reward of the best arm and the mean reward of arm k at time t, and K_t is the number of remaining arms at time t.

We provide below the sample complexity bound of DECENTRALIZED SUCCESSIVE ELIMINATION with RANDOMIZED ROUND-ROBIN (DSER3), which is simply DECENTRALIZED ELIMINATION using SER3 as the ArmSelection subroutine.

Theorem 3. *For* $K \geq 2$, $\delta \in (0, 0.5]$, *for the sequences of rewards where Assumption 4 holds, DSER3 is an* (ϵ, η)*-private algorithm, that exchanges at most* $\lfloor \frac{\log \delta}{\log \eta} \rfloor K - 1$ *messages, that finds an* ϵ*-optimal arm with a probability at least* $(1 - \delta)\left(1 - I_{1-p^*}\left(T_{P_{x,\mathbf{y}}} - T_{P_{\mathbf{y}}}, 1 + T_{P_{\mathbf{y}}}\right)\right)^{\lfloor \frac{\log \delta}{\log \eta} \rfloor}$, *and that stops after:*

$$\mathcal{O}\left(\frac{1}{p^*}\left(\frac{K}{\epsilon^2}\left(\log K + \frac{1}{\lfloor \frac{\log \delta}{\log \eta} \rfloor}\log\frac{1}{\delta}\right) + \sqrt{\frac{1}{2}\log\frac{1}{\delta}}\right)\right)$$

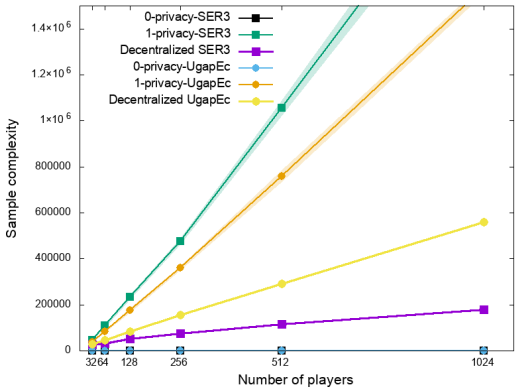
samples in $P_{x,\mathbf{y}}$.

Finally DECENTRALIZED SUCCESSIVE ELIMINATION with RANDOMIZED ROUND-ROBIN AND RESET (DSER4) handles any sequence of rewards: when Assumption 4 does not hold a switch occurs. DSER4 consists in using SER4 (Allesiardo et al, 2017) as the ArmElimination subroutine in DECENTRALIZED ELIMINATION. In addition, when a reset occurs in SER4, DECENTRALIZED ELIMINATION is reset.

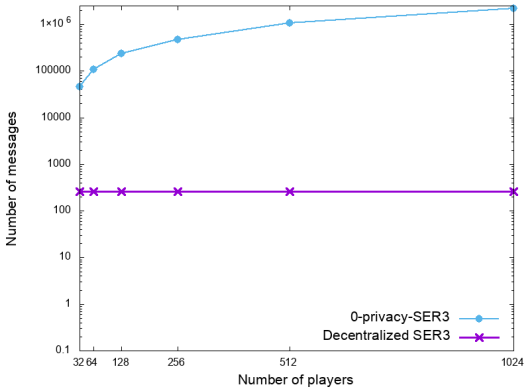
Theorem 4. *For* $K \geq 2$, $\epsilon \geq \frac{\eta}{K}$, $\varphi \in (0, 1]$, *for any sequences of rewards,* DSER4 *is an* (ϵ, η)*-private algorithm, that exchanges on average at most* $\varphi T(\lfloor \frac{\log \delta}{\log \eta} \rfloor K - 1)$ *messages, and that plays, with an expected probability at most* $\delta + \varphi T I_{1-p^*}\left(T_{P_{x,\mathbf{y}}} - T_{P_{\mathbf{y}}}, 1 + T_{P_{\mathbf{y}}}\right)^{\lfloor \frac{\log \delta}{\log \eta} \rfloor}$, *a suboptimal arm on average no more than:*

$$\mathcal{O}\left(\frac{1}{p^*}\left(\frac{1}{\epsilon^2}\sqrt{\frac{SK\log K + \frac{1}{\lfloor \frac{\log \delta}{\log \eta} \rfloor}\log\frac{1}{\delta}}{\delta^{\frac{1}{\lfloor \frac{\log \delta}{\log \eta} \rfloor}}}} + \sqrt{\frac{1}{2}\log\frac{1}{\delta}}\right)\right)$$

times, where S *is the number of switches of best arms,* φ *is the probability of reset in* SER4, T *is the time horizon, and the expected values are taken with respect to the randomization of resets.*

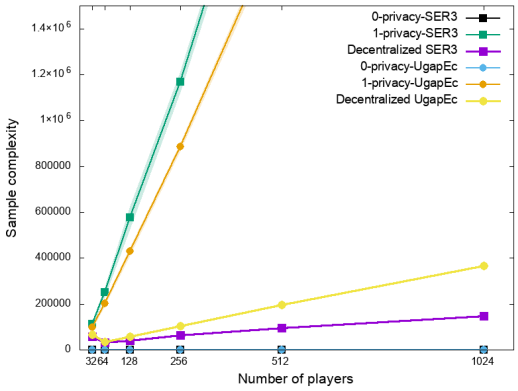


(a) Problem 1 - Sample Complexity

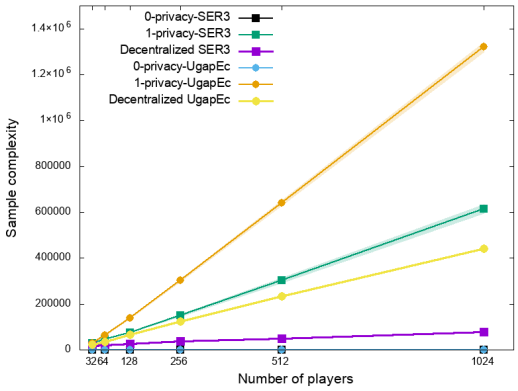


(b) Problem 1 - Number of messages

Figure 1: Uniform distribution of players. The sample complexities of 0-PRIVACY baselines are the same: 800.



(a) Problem 2 - Sample Complexity



(b) Problem 3 - Sample Complexity

Figure 2: 50% of players generates 80% of events (a), and the mean rewards of suboptimal arms linearly decrease (b).

5. Experiments

5.1. Experimental setting

To illustrate and complete the analysis of DECENTRALIZED ELIMINATION, we run three synthetic experiments:

- **Problem 1: Uniform distribution of players.** There are 10 arms. The optimal arm has a mean reward $\mu_1 = 0.7$, the second one $\mu_2 = 0.5$, the third one $\mu_3 = 0.3$, and the others have a mean reward of 0.1. Each player has a probability equal to $1/N$.
- **Problem 2: 50% of players generates 80% of events.** The same 10 arms are reused with an unbalanced distribution of players. The players are split in two groups of sizes $N/2$. When a player is sampled, a uniform random variable $z \in [0, 1]$ is drawn. If $z < 0.8$ the player is uniformly sampled from the first group, otherwise it is uniformly sampled from the second group.
- **Problem 3: non-stationary rewards.** The distribution of players is uniform. The same 10 arms are reused. The mean reward of the optimal arm does not change during time. The mean reward of suboptimal arms linearly decrease: $\mu(t) = \mu(0) - 10^{-5}t$.

As comparison points, we include two natural baselines:

- 1-PRIVACY: an $(\epsilon, 1)$-private algorithm that does not share any information between the players, and hence that runs at a zero communication cost. The ArmSelection subroutine is run with parameters $(\epsilon, \delta/N)$ to ensure that all the players find with a probability $1 - \delta$ an ϵ-optimal arm.
- 0-PRIVACY: an $(\epsilon, 0)$-private algorithm that shares all the information between players, and hence that runs at a minimal privacy and a maximal communication cost. This algorithm does not meet the original goal but is interesting as a reference to assess the sample efficiency loss stemming from the privacy constraint.

As ARMSELECTION subroutines, We choose two frequentist algorithms [3] based on Hoeffding inequality: a *explore-then-commit* algorithm UGAPEC (Gabillon et al, 2013) and a *successive elimination* algorithm SER3 (Allesiardo et al, 2017), which handles non-stationary rewards. Combining DECENTRALIZED ELIMINATION and the two baselines with the two ARMSELECTION subroutines, we compare 6 algorithms (DECENTRALIZED SER3, DECENTRALIZED UGAPEC, 1-PRIVACY-SER3, 1-PRIVACY-UGAPEC, 0-PRIVACY-SER3, 0-PRIVACY-UGAPEC) on the three problems. The algorithms are compared with respect to two key performance indicators: the sample complexity and the communication cost. For all the experiments, ϵ is set to 0.25, and δ is set to 0.05. The privacy level η is set to 0.9. All the curves and the measures are averaged over 20 trials.

5.2. Results

The results reveal that the sample efficiency of 1-PRIVACY baselines is horrendous on both problems: it increases superlinearly as the number of players increases. Worse, when the distribution of players moves away from the uniformity, which is the case in most of digital applications, the performances of 1-PRIVACY baselines decreases (see Figure 1a, 2a). Contrary to 1-PRIVACY baselines, the performances of DECENTRALIZED UGAPEC and DECENTRALIZED SER3 increases in Problem 2 (see Figure 2a). More precisely, the sample complexity curves of DECENTRALIZED UGAPEC and DECENTRALIZED SER3 exhibit two regimes: first the sample complexity decreases (between 32 to 64 players), and then the sample complexity linearly increases with the number of players. The values of hyper-parameters: $\delta = 0.05$ and $\eta = 0.9$, imply that the number $M = \lfloor \frac{\log \delta}{\log \eta} \rfloor$ of player votes required to eliminate an arm is 28. In Problem 2 with 32 players, it means that the algorithm has to wait for infrequent players votes to terminate. When the number of players is 64, this issue disappears. This is the reason why the sample complexity for 64 players is lower than for 32 players. The linear dependency of the sample complexity with respect to the number of players of the second regime is due to the fact that in the considered problems, the probability of the most likely player p^*decreases in $1/N$.

Concerning the ARMSELECTION subroutines, we observe that 1-PRIVACY-UGAPEC clearly outperforms 1-PRIVACY-SER3 on stationary problems (see Figures 1a and 2a). Moreover, the performance gain of 1-PRIVACY-UGAPEC increases with the number of players. This is due to the adaptive sampling strategy of UGAPEC: by sampling alternatively the empirical best arm and the most loosely estimated suboptimal arm, 1-PRIVACY-UGAPEC reduces the variance of the sample complexity, and thus reduces the maximum of sample complexities of players. However, when used as a subroutine in DECENTRALIZED ELIMINATION, the *successive elimination* algorithms such as SER3 are more efficient: thanks to the different suboptimal arms which are progressively eliminated by different groups of voting players, DECENTRALIZED SER3 clearly outperforms DECENTRALIZED UGAPEC (see Figure 1a and 2a).

When the mean rewards of suboptimal arms are decreasing (Figure 2b), in comparison to SER3 the performances of UGAPEC, which is not designed for non-stationary rewards, collapse: 1-PRIVACY-UGAPEC and DECENTRALIZED UGAPEC are respectively outperformed by 1-PRIVACY-SER3 and DECENTRALIZED SER3. The optimistic approach used in the sampling rule of UGAPEC is too optimistic when the mean reward are decreasing.

The communication cost is the number of exchanged messages: 1-PRIVACY baselines send zero messages, while 0-PRIVACY baselines send $N - 1$ messages per time step until the ϵ-optimal arm is found. DECENTRALIZED SER3 needs three to four orders of magnitude less messages than 0-PRIVACY-SER3 (see Figure 1b).

6. Conclusion an future works

We have provided a new definition of privacy for the decentralized algorithms. We have proposed a new problem, the *decentralized exploration problem*, where players sampled from a distribution collaborate to identify a near-optimal arm with a fixed confidence, while ensuring privacy to players and minimizing the communication cost. We have designed and analyzed a generic algorithm for this problem: DECENTRALIZED ELIMINATION uses any best arm identification algorithm as an ArmSelection subroutine. Thanks to the generality of the approach, we have extended the analysis of the algorithm to the case where the distributions of rewards are not stationary. Finally, our experiments suggest that *successive elimination* algorithms are better suited for the *decentralized exploration problem* than *explore-then-commit* algorithms.

Future work may focus on user-dependent best arms. When Assumption 2 does not hold, DECENTRALIZED ELIMINATION finds with high probability the best arm of the most frequent players. However, in lot of applications the players can observe a context before choosing an arm. The extension of the proposed approach to *contextual bandits* is not straightforward because to collaborate for building a model, the players have to exchange messages about their favorite arms and their contextual variables, that also contain private information.

[3] due to high values of sampling complexity obtained by MEDIAN ELIMINATION which flatten the differences between algorithms, we report its performance in appendix.

References

Abbasi-Yadkori, Y., Bartlett, P., Gabillon, V., Malek, A., Valko, M.:Best of both worlds: Stochastic adversarial best-arm identification, COLT, 2018.

Allesiardo, R., Féraud, R., Maillard, O. A.: The Non-Stationary Stochastic Multi-Armed Bandit Problem, International Journal of Data Science and Analytics, 2017.

Audibert, J.Y., Bubeck, S., Munos, R.: Best Arm Identification in Multi-Armed Bandits, COLT, 2010.

Avner, O., Mannor, S.: Concurrent bandits and cognitive radio networks, ECML PKDD, 2014.

Besson, L., Kaufmann, E.: Multi-Player Bandits Revisited, ALT, 2018.

Bubeck, S., Wang, T., Stoltz, G.: Pure exploration in multi-armed bandits problems, COLT, 2009.

Chakraborty, M., Chua, K. Y. P., Das, S., Juba, B.: Coordinated versus decentralized exploration in multi-armed bandits, IJCAI, 2017.

Even-Dar, E., Mannor, S., Mansour Y.: Action Elimination and Stopping Conditions for the Multi-Armed Bandit and Reinforcement Learning Problems, JMLR, 1079-1105, 2006.

Duchi, J. C., Jordan, M. I., Wainwright, M. J.: Privacy aware learning, Journal of the ACM, 2014.

Dwork, C., Mcsherry, F., Nissim, K., and Smith, A.: Calibrating noise to sensitivity in private data analysis. In Proceedings of the 3rd Theory of Cryptography Conference, 2006.

Féraud, R., Allesiardo, R., Urvoy, T., Clérot, F.: Random Forest for the Contextual Bandit Problem, AISTATS, 2016.

Gabillon, V., Ghavamzadeh M., Lazaric, A.: Best Arm Identification: A Unified Approach to Fixed Budget and Fixed Confidence, NIPS, 2013.

Gajane, P., Urvoy, T., Kaufmann, E.: Corrupt Bandits for Preserving Local Privacy, ALT, 2018.

Ganta, S., R., Kasiviswanathan, S., P. and Smith, A.: Composition attacks and auxiliary information in data privacy, KDD, 2008.

Hernandez-Lobato, J. M., Requeima, J., Pyzer-Knapp, E. O., Aspuru-Guzik, A.: Parallel and Distributed Thompson Sampling for Large-scale Accelerated Exploration of Chemical Space, ICML, 2017.

Hillel, E., Karnin, Z., Koren, T. Lempel, R., Somekh, O.: Distributed Exploration in Multi-Armed Bandits, NIPS, 2013.

Kauffman, E., Kalyanakrishnan, S.: Information Complexity in Bandit Subset Selection, COLT, 2013.

Kalyanakrishnan, S., Tewari, A., Auer, P., Stone, P.: PAC subset selection in stochastic multi-armed bandits, ICML, 2012.

Kanade, V., Liu Z., Radunovič, B.: Distributed Non-Stochastic Experts, NIPS, 2012.

Landgren, P., Srivastava, V., Leonard, N. E.: Distributed Cooperative Decision Making in Multiarmed Bandits: Frequentist and Bayesian Algorithms, CDC, 2016.

Mannor, S., and Tsitsiklis, J. N.: The sample complexity of Exploration in the Multi-Armed Bandit Problem, JMLR, 2004.

Nayyar, N., Katathil, D., Jain, R.: On Regret-Optimal Learning in Decentralized Multi-player Multi-armed Bandits, IEEE Transactions on Control of Network Systems, 2015.

Perchet, V., Rigollet, P., Chassang, S., Snowberg, E.: Batched Bandit Problems, The Annals of Statistics,Vol. 44, No. 2, 2016.

Soare, M., Lazaric, A., Munos, R.: Best-Arm Identification in Linear Bandits, NIPS, 2014.

Sweeney, L.: k-anonymity: A model for protecting privacy, International Journal on Uncertainty, Fuzziness and Knowledge based Systems, 10(5):557-570, 2002.

Szörényi, B., Busa-Fekete, R., Hegedũs, I., Ormàndi, R., Jelasity, M., Kègl, B.: Gossip-based distributed stochastic bandit algorithms, ICML, 2013.

Urvoy, T., Clérot, F., Féraud, R., Naamane, S.: Generic Exploration and K-armed Voting Bandits, ICML, 2013.

Valiant, L.: A theory of the learnable, Communications of the ACM, 27, 1984.

Liu, K. and Zhao, Q.: Distributed Learning in Multi-Armed Bandit With Multiple Players, IEEE Transactions on Signal Processing, vol. 58, N. 11, 2010.