Journal of Machine Learning Research EWRL (2015) Presentation-only Submitted ?; Published ?

Differentially private multi-agent multi-armed bandits

Aristide C. Y. Tossou ARISTIDE@CHALMERS.SE
Chalmers University of Technology

Christos Dimitrakakis CHRDIMI@CHALMERS.SE
Chalmers University of Technology

Editor: ?

Abstract

[1] We study the problem of privacy for distributed learning in Multi-Armed bandit (MAB) problem with multiple players. The players must co-ordinate, as choosing the same arm simultaneously results in a reduced reward. We wish to find a policy which maximises social welfare and individual utility, while being private. To achieve this goal, we derive two algorithms built upon decentralized Time Division Fair Sharing (TDFS) method and upper confidence bounds (UCB), where all decisions are taken based on private statistics. We provide regret guarantees that are almost as good as the non-private, multi-agent algorithm and demonstrate them empirically.

1. Introduction

The well-known stochastic N-armed bandit problem (Lai and Robbins, 1985) involves an agent sequentially choosing among a set of arms $\mathcal{A} = \{1, \dots, N\}$, and obtaining a sequence of scalar rewards $\{r_t\}$, such that if the agent's action at time t is $a_t = i$, then it obtains reward r_t drawn from some distribution P_{θ_i} with parameter θ_i, such that $\mathbb{E}(r_t \mid a_t = i) = \mu_i$. The goal of the decision maker is to draw arms so as to maximize total reward.

More generally, one could think of arms drawn at the same time, by a group of agents who are not aware of each other's actions. In particular, consider a co-ordination game where agents get reduced rewards (e.g. they must share the reward that the arm gives) whenever they pull the same arm. By co-ordinating, they can pull the M best arms, i.e. the ones with the highest expected reward μ_i. Communication constraints means that the performance achieved is lower than that the centralized model, however.

Liu and Zhao (2010) were the first to define the distributed multi-armed bandit problem. Consider an N-armed bandit problem, shared by M players. The i-th arm is associated with a local reward $x_{t,i}$ with distribution P_{θ_i}, and expectation $\mu_i = \mathbb{E}_{\theta_i} x_{t,i}$, where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_N)$ are the unknown parameters of the problem. In this setting, whenever a joint action a_t is taken by the agents at time t, they all observe the local reward of the arm. The actual reward received by the i-th agent is the reward of the arm drawn, as long as it is the unique agent to have drawn this arm. If more than one agent draws the same arm, then nobody receives any reward. We can denote this via the following reward function

$$\boldsymbol{r}_t = x_{t,i} \mathbb{I}\left\{a_{t,i} \neq a_{t,j} \forall j \neq i\right\}, \tag{1}$$

1. A version of this paper has been submitted to NIPS 2015.

where $\mathbb{I}\{\cdot\}$ is an indicator function. This forces the agents to co-ordinate. If the problem parameters are known, then there is a simple optimal policy that the agents can choose. This is is done by choosing a permutation σ such that the arms are ordered in decreasing mean: $\mu_{\sigma(1)} \geq \cdots \geq \mu_{\sigma(i)} \geq \mu_{\sigma(i+1)} \geq \cdots \geq \mu_{\sigma(N)}$ and allocating each agent to one of the arms. Now consider an arbitrary policy π for choosing a set of M arms. The expected regret of that policy is defined as

$$R_T^\pi(\boldsymbol{\theta}) = T \sum_{j=1}^{M} \mu_{\sigma(j)} - \mathbb{E}_{\boldsymbol{\theta}}^\pi \sum_{t=1}^{T} \boldsymbol{r}_t. \tag{2}$$

In our paper, we wish to study differentially private algorithms for multi-agent multi-armed bandits. We provide two algorithms based on a continual private sum mechanism that only suffer a poly-log regret.

This paper is organised as follows. Section 2 gives an overview of related work. Section 3 describes the setting and our algorithms. The time division algorithm by Liu and Zhao (2010) for multi-agent bandits, which we extend to the private setting, is explained in Section 3.1. Section 3.2 discusses our first algorithm, which employs a private version of UCB, while Section 3.3 describes a mechanism that does not add the private bound yet still achieves a logarithmic number of suboptimal arm pulls. Experiments that empirically show that indeed the algorithms suffer logarithmic loss are given in Section 4, and we conclude with Section 5. Finally, while we provide sketch proofs for all the main results; detailed proofs can be found in the appendix.

2. Related work

The distributed multi-armed bandit problem was first reported by Liu and Zhao (2010). Co-operative bandit problems were also considered by Stone and Kraus (2010), who formalised them as a *teacher-learner problem*, whereby the learner has a fixed greedy behaviour and the teacher has an adaptive strategy. Later Chen et al. (2011) examined the problem of how to provide incentives in order to induce another agent to perform certain actions.

Differential privacy (c.f. Dwork and Roth, 2013) characterises the amount of information that can be leaked about the input of the algorithm from its output.

Definition 1 (Differentially private algorithm) *An algorithm $\pi : X^* \to \mathcal{A}$ is (ϵ, δ)-differential private if, for any two neighbouring inputs $x, x' \in X^*$, it holds that $\forall A \subset \mathcal{A}$, $\pi(A \mid x) \leq \pi(A \mid x')e^\epsilon + \delta$.*

Differential privacy in bandits was first considered by Mishra and Thakurta (2014), who used a binary tree construction (Chan et al., 2010; Dwork et al., 2010a) to achieve privacy with low utility losses. In particular, the first paper built upon the following measure concentration result, which we also employ in this paper.

Fact 1 (Corollary 2.9 in Chan et al. (2010)) *If $\lambda_i \sim \mathcal{L}ap(b_i)$ and $Y = \sum_i \lambda_i$ then*

$$\mathbb{P}(|Y| \geq \|b\|_2 \ln \frac{1}{\delta}) \leq \delta. \tag{3}$$

Using a tree construction, they construct noisy partial sums of data in binary intervals, which ensures that we release statistics based on $\ln T$ partial sums. Setting the Laplace parameter to $\frac{\ln T}{\epsilon}$, results in an ϵ-DP release, and poly-log utility with probability $1-\delta$.

In our paper, we are interested in the distributed setting, where privacy is inherently more important. We show that in that case, algorithms can be made differentially private for all players, and that the regret is $O(\ln^2 T)$. The setting is explained below.

3. Private distributed multi-armed bandits

At each time step t, all players $p \in [M]$ select an action $a_{t,p} \in [N]$ and obtain a reward $r_{t,p} \in \{0,1\}$. Two collisions models are considered. Under *collision model 1*, when multiple players select the same arm k, then only one of them receives a reward. The other gets 0. Under *collision model 2*, when multiple players select the same arm, all of them obtain a reward of 0. In both models, the players observe the collision.

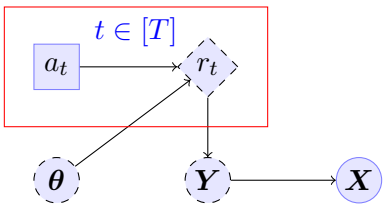


Figure 1: Graphical model for the empirical and private means, from the point of view of a single decision making agent. a_t is the action of the agent, while r_t is the reward obtained, which depends on the bandit parameters $\boldsymbol{\theta}$. The vector of empirical means $\boldsymbol{Y}$ is then made into a private vector $\boldsymbol{X}$ which each agent uses to select actions. The rewards are essentially hidden from each individual agent by the differentially private mechanism.

We wish each player to behave in a way that an external observer who can observe their actions, as well including other players, who can observe collisions, cannot infer the reward sequence they obtain. More specifically, we want the players actions to be differentially private with respect to the reward sequence. Then there should exist $\epsilon, \delta > 0$ such that

$$\mathbb{P}(a_t \mid a_{1:t-1}, r_{1:t}) \leq \mathbb{P}(a_t \mid a_{1:t-1}, r'_{1:t})e^{\epsilon} + \delta, \tag{4}$$

for all $r_{1:t}, r'_{1:t}$ that differ in one time step. To achieve this, we make the players base their actions on *differentially private* statistics. This ensures their actions are also differentially private. In particular, the differentially private mechanism keeps track of the vector of non-private statistics $\boldsymbol{Y}$ for each player, and outputs a vector of private statistics $\boldsymbol{X}$. This is then used by the player to select an action, as shown in Figure 1. This effectively also makes the rewards hidden to the player. While this architecture ensures privacy, we also need a method for the agents to co-ordinate. This can be done with a time-division algorithm.

3.1 Private Time Division Fair Sharing Algorithm

To solve the distributed multi-armed bandits problem, we use the time division sharing technique of Liu and Zhao (2010). In the first time step, the first player will target the best

arm, and the k-th player target the k-th best arm. In the second time step, the first player will target the second best arm, etc. This idea can be modeled by associating an offset o to each player p. At each time step t, player p will target the $(t-1+o-1) \mod M+1$ best action. In this setting, we assume a pre-agreement for the offset used by each player, but randomized offsets can be used to resolve conflicts otherwise as explained in Liu and Zhao (2010).

A *subsequence* consists of all time-steps $t \in [T]$ where all players have a given offset. These are further partitioned into *mini-sequences*, where arms are ranked differently. Of main interest is a *dominant* mini-sequence for player i. This is the set of time-steps where player i has correctly identified the best $i-1$ arms.

Our algorithms differ from the one in Liu and Zhao (2010) in the following ways. We use the single player UCB algorithm presented in Auer et al. (2002) instead of the Lai-Robbins Policy. This choice is motivated by the simplicity of UCB. At each time step, UCB based its action on an estimate of the expected reward of each arm. This estimate is the sum of the empirical mean and an upper bound confidence equal to $\sqrt{\frac{2\log t}{n_a}}$ where t is the time step and n_a the number of times arm a is played. The second difference is that we don't remove the previously played $j-1$ arms when targeting the jth best arm. Finally, we also employ an algorithm that calculates private means over intervals.

In order to get a differential private version of the time division sharing algorithm, we observe that it is enough to have a differential private version of the UCB. And to get a differential private version of UCB, it is enough to privately compute the empirical mean of each arm. This is so, because once the mean of each arms are computed the action which will be played is a deterministic function of the means.

We provide two different algorithms that use different techniques to privately compute the mean. All our algorithms apply naturally to the single player case by setting the number of player M to 1.

3.2 Private Hybrid MAB Algorithm with upper confidence bound (`UCB-Private-Bound`)

The idea of this algorithm is to compute the sum of the rewards of each arm using the Hybrid mechanism (Chan et al., 2010). The hybrid mechanism is similar to the binary tree construction but is time unbounded as it does not need an upper bound on T. However, using the hybrid mechanism requires us to change the UCB confidence bound as additional noise is added to the true empirical mean. To fix it, we add the upper bound confidence $\frac{1}{\epsilon}\log(\frac{1}{\delta})\frac{\log^{1.5} n_a}{n_a}$ of the Hybrid Mechanism to the UCB bound. This approach is summarized in Algorithm 1. We will now show that Algorithm 1 is both private and leads to poly-log regret. The first statement follows from the fact that each agent is using a differentially private statistic to select actions. The second follows from concentration inequalities that bound the error in each round of the algorithm and the fact that the confidence bound we used results in a poly-log number of erroneous arm pulls.

Let us first discuss the privacy of the algorithm.

Theorem 2 *Algorithm 1 is ϵ-DP for each player after any number of T plays.*

Algorithm 1 Private Hybrid Mechanism DMAB with bound (`UCB-Private-Bound`)

Without any loss of generality we can consider player p_i with pre-agreed offset i.
n_a be the total number of times arm a is played.
ϵ the differential privacy parameter.
Instantiate one private Hybrid Mechanism (Chan et al., 2010, Alg. 4) for each arm a.
for $t \leftarrow 1$ to T **do**
  if $t \leq N$ **then**
    play arm $a = (i + t) \mod N$ and observe the reward r_a
    Insert r_a to the hybrid mechanism corresponding to arm a
  else
    The player will target the jth best arm in the set $\mathcal{A}$
    Where $j = (t - 1 + i - 1) \mod M + 1$
    $s_a(t) \leftarrow$ total reward sum computed using the hybrid mechanism a
    $\delta \leftarrow t^{-4}$
    Pull the jth best arm using the ranking $\frac{s_a(t)}{n_a} + \sqrt{\frac{2 \log t}{n_a}} + \frac{1}{\epsilon}(\log \frac{1}{\delta}) \frac{\log^{1.5} n_a}{n_a}$
    Let's $\sigma(j)$ the jth best arm played.
    Observe the reward $r_\sigma(j)$
    Insert $r_\sigma(j)$ to the hybrid mechanism corresponding to arm $\sigma(j)$
  end if
end for

Proof This follows directly from the fact that the hybrid mechanism is ϵ-DP after any number of T plays and a single flip of one reward in the sequences of reward only affect one mechanism. Furthermore, the whole algorithm is a random mapping from the output of the hybrid mechanism to the action taken and using Proposition 2.1 of Dwork and Roth (2013) complete the proof. ■

Corollary 3 *Algorithm 1 is $M\epsilon$-DP with respect to an external observer that can see the actions of all players.*

Proof This follows directly from the composition property of differential privacy. ■

To perform the regret analysis, we need to count the number of times each player plays an incorrect arm (i.e. the i-th player in the subsequence does not play the i-th best arm). We do this in two parts. First, we count the number of times that an incorrect arm is pulled when the i-th player has correctly identified the top $i - 1$ arms. Secondly, we count the number of times that the i-th player has incorrectly identified the top arms.

Lemma 4 *Let $\tau^D_{a,\sigma(p)}$ denote the expected number of times player p plays arm a in the dominant mini-sequence of the first subsequence when Algorithm 1 is followed. Then, for any arm a with $\mu_a < \mu_{\sigma(p)}$, we have:*

$$\tau^D_{a,\sigma(p)} \leq 1 + \frac{1}{M}\left[\max\left(\left(\frac{8}{\Delta_{a,\sigma(p)}\lambda_0}\frac{1}{\epsilon}\log t\right)^{2.25}, \frac{8}{\lambda_0^2 \Delta^2_{a,\sigma(p)}}\log t\right)\right] + \frac{2\pi^2}{3M^2} \tag{5}$$

for any λ_0 *such that* $0 < \lambda_0 < 1$

Proof [Sketch] The differentially private algorithm we use has a high-probability bound on the error of order poly-log T. We bound the total error between the private mean and expectation of the arm, by the Laplace concentration inequality (3) for the privacy noise and by Chernoff-Hoeffding bounds for the observation noise. We select the confidence bounds appropriately so that the error probability at each step is t^{-4}. This also results in a poly-log number of steps where player p selects the wrong arm. ■

Lemma 5 *Let's denote by* $\hat{\tau}_{\overline{\sigma(p)}}$ *the expected total number of times player* p *does not play the* p*th best arm in the first subsequence up to time* t*. We have,*

$$\hat{\tau}_{\overline{\sigma(p)}} \leq \sum_{j=1}^{p} \sum_{a:\mu_a<\mu_{\sigma(j)}} \tau^D_{a,\sigma(j)} \tag{6}$$

Proof [Sketch] The proof is done by induction on p, using Lemma 4. Intuitively, the number of times that the i-th player does not identify the $i-1$ best arms correctly is bounded by the number of times any of the previous players j did not identify the j-best arm. ■

Theorem 6 *If Algorithm 1 is run on* N *machines and with* M *agents, having arbitrary reward distributions* $P_1, \ldots, P_K$ *with support in* $\{0, 1\}$*, then with probability at least* $1 - \delta$*, its expected regret* $\mathcal{R}$ *after any number* t *of plays is:*

$$\mathcal{R} \leq M \left(\sum_{p=1}^{M} \sum_{j=1}^{p} \sum_{a:\mu_a<\mu_{\sigma(j)}} \tau^D_{a,\sigma(j)} \mu_{\sigma(p)} \right) \tag{7}$$

under both collision models where $\mu_1, \ldots, \mu_K$ *are the expected values of* $P_1, \ldots, P_K$*;* $\tau^D_{a,\sigma(j)}$ *is the corresponding value of lemma 4.*

Plugging in the bounds for each arm, and setting $\delta = t^{-4}$, we obtain $O(M^2 \log^{2.25} T)$ regret.

3.3 Private Hybrid Algorithm with no bound (UCB-Private)

Instead of adding the hybrid mechanism bound to the UCB bound, we can make sure that the variance of the noise added for each arm is the same. To do that, when we play arm a and insert the return reward to its private sum, we also insert a 0 to the private sums of all other arms. This ensures that all the sums have the same variance due to the private mechanism. Consequently, there is no reason to add a bound related to the DP term when selecting an arm, as it is the same for all of them. This idea is sketched in Algorithm 2. Algorithm 2 also enjoys poly-log number of steps where player p selects the wrong arm.

Algorithm 2 Private Hybrib UCB with no bound(`UCB-Private`)

Run Algorithm 1 using the ranking $\frac{s_a(t)}{n_a} + \sqrt{\frac{2\log t}{n_a}}$
When a new reward is observed for arm a, increase n_a and insert the observed reward to the a-th arm's private sum and insert 0 to all arms $a' \neq a$.

Lemma 7 *Let $\tau^D_{a,\sigma(p)}$ denote the expected number of times player p plays arm a in the dominant mini-sequence of the first subsequence when Algorithm 2 is followed. Then, for any arm a with $\mu_a < \mu_{\sigma(p)}$, we have:*

$$\tau^D_{a,\sigma(p)} \leq 1 + \frac{1}{M} \max\left[\left(\frac{18(7+4\sqrt{3})}{\epsilon^2}\log^4 t\right), \frac{8}{\Delta^2_{a,\sigma(p)}}\log t\right] + \frac{4(1+\log t)}{M^2} \tag{8}$$

Proof [Sketch] The proof is similar to the one for Lemma 4. But this time we have to choose the error probability to be t^{-3}. ■

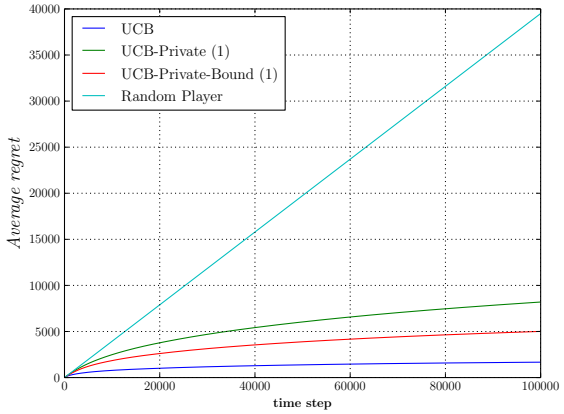
Let's note that theorem 6 also applies to Algorithm 2.

4. Experiments

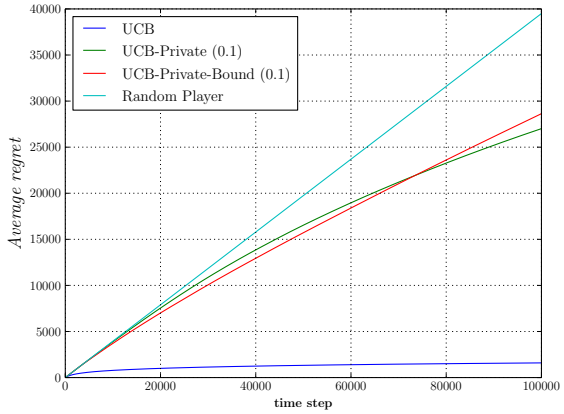
We perform experiments using arms with rewards drawn from independent Bernoulli distributions. We plot the cumulative regret for each algorithm averaged over 100 runs. We tested against private UCB and a random player. We targeted two levels of ϵ-differential privacy: 0.1 and 1. In Figure 2, we show the result for a single and a three player MAB. Figure 2(a) shows that the regret of all three algorithms is indeed logarithmic. Figure 2(b) shows that when the privacy is lower both the algorithm UCB-Private and UCB-Private-Bound takes a long time before starting to learn a good policy. We repeat the experiments for 3 players, shown in Figure 2(c,d), and there we see that the relationships between the algorithms remain the same.

5. Conclusion and future work

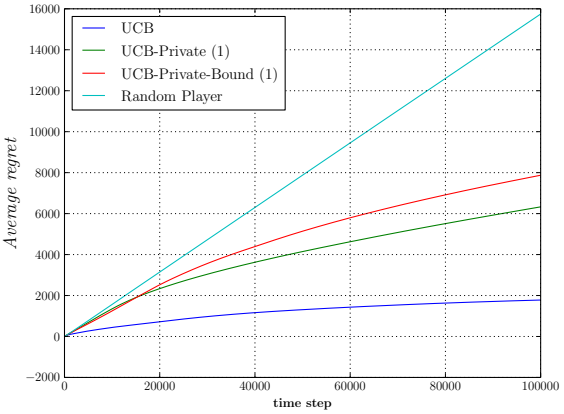
At the moment, we have analysed the achievable performance in the presence of differential privacy constraints from the set of agents to the external environment. Crucially, we have shown that by careful design of the algorithm, we can obtain nearly the same regret as the non-private, single agent version, i.e. poly-logarithmic regret with respect to the horizon, with a polynomial dependency on the privacy parameter and the number of agents. There are a number of different directions one could follow. For example, it would be interesting to analyse the performance of the system by relaxing the lack of communication between the agents so that there is a common signal, which is differentially private. We could also examine algorithms based on other mechanisms, such as posterior sampling, which is known to be differentially private (Dimitrakakis et al., 2014). Finally, we could analyse if by careful design of the algorithm, we can obtain the same order of regret as the non-private algorithm.



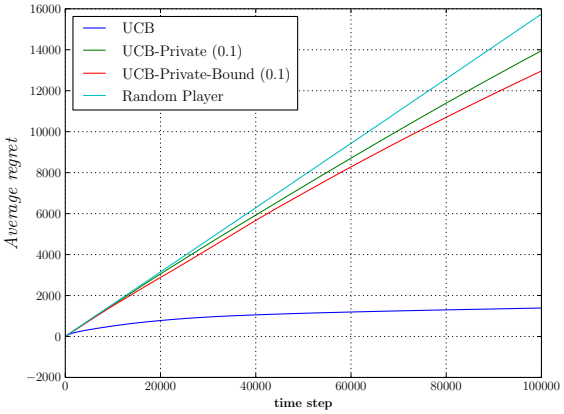
(a) Regret for $\epsilon = 1$, 1 player



(b) Regret for $\epsilon = 0.1$, 1 player



(c) Regret for $\epsilon = 1$, 3 players



(d) Regret for $\epsilon = 0.1$, 3 players

Figure 2: Experimental results with 100 runs, 1 or 3 players, 10 arms with rewards: $0.1 \ldots 0.2, 0.55, 0.1 \ldots .$

References

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2/3):235–256, 2002.

Wolfgang Borsch-Supan. On the evaluation of the function ... for real values of γ. *Journal of Research of the NBS.*, 65, 1961. URL `http://nvlpubs.nist.gov/nistpubs/jres/65B/jresv65Bn4p245_A1b.pdf`.

TH Hubert Chan, Elaine Shi, and Dawn Song. Private and continual release of statistics. In *Automata, Languages and Programming*, pages 405–417. Springer, 2010.

Yiling Chen, Jerry Kung, David C Parkes, Ariel D Procaccia, and Haoqi Zhang. Incentive design for adaptive agents. In *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pages 627–634. International Foundation for Autonomous Agents and Multiagent Systems, 2011.

Christos Dimitrakakis, Blaine Nelson, Aikaterini Mitrokotsa, and Benjamin Rubinstein. Robust and private Bayesian inference. In *Algorithmic Learning Theory*, 2014.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(34):211–407, 2013. ISSN 1551-305X. doi: 10.1561/0400000042. URL `http://dx.doi.org/10.1561/0400000042`.

Cynthia Dwork, Moni Naor, Toniann Pitassi, and Guy N Rothblum. Differential privacy under continual observation. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 715–724. ACM, 2010a.

Cynthia Dwork, Guy N. Rothblum, and Salil Vadhan. Boosting and differential privacy. In *Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, FOCS '10, pages 51–60, Washington, DC, USA, 2010b. IEEE Computer Society. ISBN 978-0-7695-4244-7. doi: 10.1109/FOCS.2010.12. URL `http://dx.doi.org/10.1109/FOCS.2010.12`.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

Keqin Liu and Qing Zhao. Distributed learning in multi-armed bandit with multiple players. *Signal Processing, IEEE Transactions on*, 58(11):5667–5681, Nov 2010. ISSN 1053-587X. doi: 10.1109/TSP.2010.2062509.

Nikita Mishra and Abhradeep Thakurta. Private stochastic multi-armed bandits: From theory to practice. In *ICML Workshop on Learning, Security, and Privacy*, 2014.

Peter Stone and Sarit Kraus. To teach or not to teach?: decision making under uncertainty in ad hoc teams. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1*, pages 117–124. International Foundation for Autonomous Agents and Multiagent Systems, 2010.

Appendix A. Collected proofs

A.1 Proof of lemma 4

Before proving lemma 4, let's define some notations. The symbol p will be used to indicate the index of a player. a will be used to indicate index of an arm. s is the index of a subsequence. ϵ is the differential privacy parameter. a_s^* is used to denote the sth-best arm. t is used to denote the time step. $n_{a,p,t}$ is used to denote the number of times an arm a is played by a given player p at time t, omitting subscripts for brevity whenever they are clear from context.

To prove lemma 4 let's observe that in each dominant mini-sequence D_{a_p}, the algorithm is doing a single player multi armed bandit where the best arm is a_p^* with expectation μ_p^*.

From corollary 4.8 in Chan et al. (2010) we know that the error between the empirical and private mean is bounded as $|Y - X| \leq h_n$, w.p. $1-\delta$ where "w.p. $1-\delta$" is used to mean that the bound is true with probability at least $1-\delta$, X is the empirical mean returned by the private mechanism, Y the true empirical mean, h_n the error due to the differentially private mechanism. It is defined as: $h_n = \mathcal{O}(\frac{1}{\epsilon}) \cdot (\ln n)^{1.5} \cdot \log \frac{1}{\delta} \cdot \frac{1}{n}$. We can rewrite this bound into equations 9 and 10.

$$\mathbb{P}(X \geq Y + h_n) \leq \delta \tag{9}$$
$$\mathbb{P}(X \leq Y - h_n) \leq \delta. \tag{10}$$

Let $T_{a,p}^1$ be the number of times arm a is played by player p in subsequence 1. Let's $c_{t,n} \triangleq \sqrt{(2 \ln t)/n}$ denote the confidence interval. Let's call $\mathcal{I}_i^n$ the set containing numbers from i to n with step size M: $\mathcal{I}_i^n = i, i+M, i+2M, ...n$. Similarly to the proof of the UCB algorithm in (Auer et al., 2002), we have:

$$\begin{aligned} T_{a,p}^1 &= 1 + \sum_{t \in \mathcal{I}_{M+1}^n} \{a_{t,p} = a\} \\ &\leq \ell + \sum_{t' \in \mathcal{I}_1^t} \sum_{n \in \mathcal{I}_1^{t'-1}} \sum_{n_a \in \mathcal{I}_\ell^{t'-1}} \left\{ X_n^* + c_{t',n} + h_n \leq X_{a,n_a} + c_{t',n_a} + h_{n_a} \right\} \end{aligned} \tag{11}$$

In equation 11, X_n^* is the mean returned by the private mechanism for the 1st best arm when it has been played n times. Now we can observe that $X_n^* + c_{t,n} + h_n \leq X_{a,n_a} + c_{t,n_a} + h_{n_a}$ implies that at least one of the following must hold

$$X_n^* \leq \mu^* - c_{t,n} - h_n \tag{12}$$
$$X_{a,n_a} \geq \mu_a + c_{t,n_a} + h_{n_a} \tag{13}$$
$$\mu^* < \mu_a + 2c_{t,n} + 2h_n \tag{14}$$

Now we can bound the probability of events (12) using equation (10), the union bound and the Chernoff-Hoeffding bound.

$$\begin{aligned} \mathbb{P}(X_n^* \leq \mu^* - c_n - h_n) &= \mathbb{P}(X_n^* \leq Y_n^* - h_n \vee Y_n^* \leq \mu^* - c_{t,n}) && (15) \\ &\leq \mathbb{P}(X_n^* \leq Y_n^* - h_n) + \mathbb{P}(Y_n^* \leq \mu^* - c_{t,n}) && (16) \\ &\leq \delta + \exp(-4 \log t) = \delta + t^{-4}. && (17) \end{aligned}$$

Similarly, to prove a bound on the probability (13) we use (9):

$$\begin{aligned} \mathbb{P}(X_{a,n_a} \geq \mu_a + c_{t,n_a} + h_{n_a}) &= \mathbb{P}(X_{a,n_a} \geq Y_{a,n_a} + h_{n_a} \vee Y_{a,n_a} \geq \mu_a + c_{t,n_a}) && (18) \\ &\leq \mathbb{P}(X_{a,n_a} \geq Y_{a,n_a} + h_{n_a}) + \mathbb{P}(Y_{a,n_a} \geq \mu_a + c_{t,n_a}) && (19) \\ &\leq \delta + \exp(-4 \log t) = \delta + t^{-4}. && (20) \end{aligned}$$

Now we choose $\delta = t^{-4}$ which leads to

$$\mathbb{P}(X_{a,n_a} \geq \mu_a + c_{t,n_a} + h_{n_a}) \leq 2t^{-4} \tag{21}$$
$$\mathbb{P}(X_n^* \leq \mu^* - c_{t,n} - h_n) \leq 2t^{-4} \tag{22}$$

Now consider the last condition. For this, we want to find the minimum number n for which event (14) is always false. Event (14) is false, implies that $\Delta_a > 2c_{t,n} + 2h_n$ where $\Delta_a = \mu_* - \mu_a$. We observe that for $\Delta_a > 2c_{t,n} + 2h_n$ to hold, it is enough that the following two conditions hold for any λ_0 such that $0 < \lambda_0 < 1$.

$$\lambda_0 \Delta_a > 2c_{t,n} \tag{23}$$
$$(1 - \lambda_0)\Delta_a > 2h_n \tag{24}$$

From equation (23), we have $n \geq \frac{8}{\lambda_0^2 \Delta_a^2} \log(t)$. Equation (24) leads to

$$n \geq \lambda_1(\Delta_a) \log^{1.5}(n), \qquad \lambda_1(\Delta_a) = \frac{2}{\Delta_a \lambda_0} \frac{1}{\epsilon} \log(\frac{1}{\delta}) = \frac{8}{\Delta_a \lambda_0} \frac{1}{\epsilon} \log(t).$$

Using two successive variable replacements, one using $n' = \log(n)$ and the other using $x = \frac{-n'}{1.5}$, it is easy to see that, the inequality leads to $\exp(-x) >= -1.5x\lambda_1(\Delta_a)^{1/1.5}$ which is a standard transcendental algebraic equations whose solutions are given by the Lambert W function. So,

$$n \geq \begin{cases} \exp(-1.5(W(-1, \frac{-1}{1.5\lambda_1(\Delta_a)^{1/1.5}}))), & \text{if } \lambda_1(\Delta_a) > (\frac{e}{1.5})^{1.5} \\ \emptyset(\text{Non real}), & \text{otherwise} \end{cases} = \lambda_2(\Delta_a) \tag{25}$$

where $W(k, x)$ is the Lambert function of x on branch k. Note that $\lambda_1(\Delta_a)$ can always be taken such that it is greater than $(\frac{e}{1.5})^{1.5}$.

Now using the approximation of the Lambert function provided in section 4 of Borsch-Supan (1961), we can conclude that $\lambda_2(\Delta_a) \approx (\lambda_1(\Delta_a))^{2.25} = \left(\frac{8}{\Delta_a \lambda_0} \frac{1}{\epsilon} \log t\right)^{2.25}$

Putting equations (23) and (24) yields, $n \geq \max\left(\lambda_2(\Delta_a), \frac{8}{\lambda_0^2 \Delta_a^2} \log(t)\right)$

This means that if the number of times we take a suboptimal arm is lower than $\lambda_2(\Delta_a)$, we will be picking it over the optimal one with high probability until n becomes larger than $\lambda_2(\Delta_a)$. In summary,

$$T_{a,p}^1 \leq \ell + \sum_{t' \in \mathcal{I}_1^t} \sum_{n \in \mathcal{I}_1^{t'-1}} \sum_{n_a \in \mathcal{I}_\ell^{t'-1}} \left\{ X_n^* + c_{t',n} + h_n \leq X_{a,n_a} + c_{t',n_a} + h_{n_a} \right\}$$

$$
\begin{aligned}
T^1_{a,p} &\leq \lceil \frac{1}{M} \left[\max \left(\lambda_2(\Delta_a), \frac{8}{\lambda_0^2 \Delta_a^2} \log(t) \right) \right] \rceil + \sum_{t' \in \mathcal{I}_1^\infty} \sum_{n \in \mathcal{I}_1^{t'-1}} \sum_{n_a \in \mathcal{I}_\ell^{t'-1}} 4t'^{-4} \\
&\leq \lceil \frac{1}{M} \left[\max \left(\lambda_2(\Delta_a), \frac{8}{\lambda_0^2 \Delta_a^2} \log(t) \right) \right] \rceil + 4 \lim_{t \to \infty} \sum_{t'=1}^{t} \frac{1}{M} \sum_{n=1}^{t'} \frac{1}{M} \sum_{n_a=1}^{t'} t'^{-4} \qquad (26) \\
&\leq \lceil \frac{1}{M} \left[\max \left(\lambda_2(\Delta_a), \frac{8}{\lambda_0^2 \Delta_a^2} \log(t) \right) \right] \rceil + \frac{4}{M^2} \lim_{t \to \infty} \sum_{t'=1}^{t} t'^{-2} \qquad (27) \\
&\leq 1 + \frac{1}{M} \left[\max \left(\lambda_2(\Delta_a), \frac{8}{\lambda_0^2 \Delta_a^2} \log(t) \right) \right] + \frac{2\pi^2}{3M^2} \qquad (28) \\
&\leq 1 + \frac{1}{M} \left[\max \left(\left(\frac{8}{\Delta_a \lambda_0} \frac{1}{\epsilon} \log t \right)^{2.25}, \frac{8}{\lambda_0^2 \Delta_a^2} \log t \right) \right] + \frac{2\pi^2}{3M^2} \qquad (29)
\end{aligned}
$$

In equation (29), note that the values of n found for event (14) are the number of consecutive steps we are going to play arm a. Now taking into account that for subsequence s, only time steps in $\mathcal{I}_1^n$ are present; it means that the bound for n should be divided by the step size M.

A.2 Proof for the Unbounded mechanism

Proof As much of the proof is quite similar to that Lemma 4, we shall omit some steps. Similarly to Lemma 4, we will take a bad arm when one of the following 3 events happens:

$$
\begin{aligned}
X^*_n &\leq \mu^* - c_{t,n} \qquad (30) \\
X_{a,n_a} &\geq \mu_a + c_{t,n_a} \qquad (31) \\
\mu^* &< \mu_a + 2c_{t,n} \qquad (32)
\end{aligned}
$$

We also know that

$$
\mathbb{P}\{|Y - X| \geq h_{t,n}\} \leq \delta \qquad (33)
$$

with $h_{t,n} = \mathcal{O}(\frac{1}{\epsilon}) \cdot \log^{1.5} t \cdot \log \frac{1}{\delta} \cdot \frac{1}{n}$. Now we can bound the probability of events (31) using equation (33), the union bound and the Chernoff-Hoeffding bound.

$$
\begin{aligned}
\mathbb{P}(X_{a,n_a} \geq \mu_a + c_{t,n_a}) &\leq \mathbb{P}(X_{a,n_a} \geq Y_{a,n_a} + h_{t,n_a} \vee Y_{a,n_a} \geq \mu_a + c_{t,n_a} - h_{t,n_a}) \qquad (34) \\
&\leq \mathbb{P}(X_{a,n_a} \geq Y_{a,n_a} + h_{t,n_a}) + \mathbb{P}(Y_{a,n_a} \geq \mu_a + c_{t,n_a} - h_{t,n_a}) \qquad (35) \\
&\leq \delta + \mathbb{P}(Y_{a,n_a} \geq \mu_a + c_{t,n_a} - h_{t,n_a}). \qquad (36) \\
&\leq \delta + \exp(-2n_a(c_{t,n_a} - h_{t,n_a})^2) \qquad (37) \\
&\leq \delta + t^{-3} \qquad (38) \\
&\leq 2t^{-3} \qquad (39)
\end{aligned}
$$

For the above inequalities to hold, we require that $h_{t,n_a} \leq \lambda_4 c_{t,n_a}$ with $\lambda_4 = 1 - \frac{\sqrt{3}}{2}$ and $\delta = t^{-3}$. Similarly, we prove a bound on the probability (30):

$$\begin{aligned}
\mathbb{P}(X_n^* \leq \mu^* - c_{t,n}) &= \mathbb{P}(X_n^* \leq Y_n^* - h_{n_a} \vee Y_n^* \leq \mu^* - c_{t,n} + h_{n_a}) && (40)\\
&\leq \mathbb{P}(X_n^* \leq Y_n^* - h_{n_a}) + \mathbb{P}(Y_n^* \leq \mu^* - c_{t,n} + h_{n_a}) && (41)\\
&\leq \delta + t^{-3} = 2t^{-3}. && (42)
\end{aligned}$$

Now, let's compute the minimum value for n_a which is consistent with our choice of λ_4 and δ. It is easy to show that $n_a \geq \left(\frac{18(7+4\sqrt{3})}{\epsilon^2}\log^4 t\right)$. The final event 32 is exactly the standard UCB event and 32 will be false for all $n_a \geq \frac{8}{\Delta_a^2}\log t$. Putting together those requirements for n_a, and adding up the probabilities for error when they are satisfied, we have (and similarly to lemma 4)

$$\begin{aligned}
T_{a,p}^1 &\leq \lceil \frac{1}{M}\left[\max\left(\frac{18(7+4\sqrt{3})}{\epsilon^2}\log^4 t, \frac{8}{\Delta_a^2}\log t\right)\right]\rceil + \sum_{t'\in\mathcal{I}_1^t}\sum_{n\in\mathcal{I}_1^{t'-1}}\sum_{n_a\in\mathcal{I}_\ell^{t'-1}} 4t'^{-3}\\
&\leq \lceil \frac{1}{M}\left[\max\left(\frac{18(7+4\sqrt{3})}{\epsilon^2}\log^4 t, \frac{8}{\Delta_a^2}\log t\right)\right]\rceil + \frac{4}{M^2}\sum_{t'=1}^{t} t'^{-1} && (43)\\
&\leq 1 + \frac{1}{M}\left[\max\left(\frac{18(7+4\sqrt{3})}{\epsilon^2}\log^4 t, \frac{8}{\Delta_a^2}\log t\right)\right] + \frac{4(1+\log t)}{M^2} && (44)
\end{aligned}$$

■

A.3 Proof of lemma 5

We will prove this by induction on p. Consider the case $p = 1$. Let $\hat{\tau}_{\overline{\sigma(p)}}$ denote the expected total number of times that player p does not play the p-th best arm up to time t in the first sequence.

For player 1 which targets the first best arm, this number is the sum of the number of times he plays all other arms having expectation lower than the best arm. In other words, by using lemma 4,

$$\hat{\tau}_{\overline{\sigma(1)}} = \sum_{a:\mu_a<\mu_{\sigma(1)}} \tau_{a,\sigma(1)}^D \leq \sum_{j=1}^{1}\sum_{a:\mu_a<\mu_{\sigma(j)}} \tau_{a,\sigma(j)}^D \tag{45}$$

which shows that lemma 5 is true for $p = 1$.

Let's denote by $\tau^{\overline{D_p}}$ the number of time steps not in the dominant subsequence of player p. Let's observe that $\tau^{\overline{D_2}}$ is exactly the number of times player 1 does not play the best arm. So,

$$\tau^{\overline{D_{p+1}}} \leq \sum_{j=1}^{p}\sum_{a:\mu_a<\mu_{\sigma(j)}} \tau_{a,\sigma(j)}^D \tag{46}$$

is true for $p = 1$. Let's assume that equations (46) and lemma 5 are true from 1 up to p. Let's show that there are also true for $p+1$. Let $\tau_{\sigma(p)}^{D_p}$ be the expected number of time step

where the pth best arm is played by player p in its dominant mini-sequence up to time t and let $\tau_{\sigma(p)}^{D_p}$ denote the expected number of time step where the pth best arm is not played by player p in its dominant mini-sequence up to time t. From lemma 4, we have that

$$\tau_{\sigma(p+1)}^{D_{p+1}} \leq \sum_{a:\mu_a<\mu_{\sigma(p+1)}} \tau_{a,\sigma(p+1)}^{D_{p+1}}. \tag{47}$$

Since

$$\tau^{D_{p+1}} = \tau_{\sigma(p+1)}^{D_{p+1}} + \tau_{\sigma(p+1)}^{D_{p+1}} \leq \tau_{\sigma(p+1)}^{D_{p+1}} + \hat{\tau}_{\sigma(p+1)} \tag{48}$$

we have,

$$\begin{aligned} t/M - \hat{\tau}_{\sigma(p+1)} &\leq t/M - \tau_{\sigma(p+1)}^{D_{p+1}} \\ &\leq t/M - \tau^{D_{p+1}} + \tau_{\sigma(p+1)}^{D_{p+1}} \\ &\leq \tau^{\overline{D_{p+1}}} + \tau_{\sigma(p+1)}^{D_{p+1}} \\ &\leq \sum_{j=1}^{p} \sum_{a:\mu_a<\mu_{\sigma(j)}} \tau_{a,\sigma(j)}^{D} + \sum_{a:\mu_a<\mu_{\sigma(p+1)}} \tau_{a,\sigma(p+1)}^{D} \end{aligned}$$

$$\hat{\tau}_{\overline{\sigma(p+1)}} \leq \sum_{j=1}^{p+1} \sum_{a:\mu_a<\mu_{\sigma(j)}} \tau_{a,\sigma(j)}^{D} \tag{49}$$

which completes the proof of lemma 5.

A.4 Proof of theorem 6

To found the regret bound, we can just focus in the first subsequence. Indeed, all subsequences will incur the same loss due to the symmetry (a simple permutation for the players index in any subsequence will lead to the first subsequence).

In the first subsequence, the regret loss is the sum of the regret incurs by each player. A player will only incur a loss if he or any other player does not play its corresponding pth best arm. And when that happens, the worst case is that there is a collision. Lemma 5 already gave us the number of time the pth best arm is not taken by any player. We just have to know the reward loss incurred each time the pth best arm is not taken.

Under collision model 1, in case of a collision, a player will not loose more than $\mu_{\sigma(p)}$. And this is also true under collision model 2.