Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

Algorithms for Differentially Private Multi-Armed Bandits

Aristide C. Y. Tossou and Christos Dimitrakakis

Chalmers University of Technology, Gothenburg, Sweden

{aristide, chrdimi}@chalmers.se

Abstract

We present differentially private algorithms for the stochastic Multi-Armed Bandit (MAB) problem. This is a problem for applications such as adaptive clinical trials, experiment design, and user-targeted advertising where private information is connected to individual rewards. Our major contribution is to show that there exist (ϵ, δ) differentially private variants of Upper Confidence Bound algorithms which have optimal regret, $\mathcal{O}(\epsilon^{-1} + \log T)$. This is a significant improvement over previous results, which only achieve poly-log regret $\mathcal{O}(\epsilon^{-1} \log^3 T)$, because of our use of a novel interval-based mechanism. We also substantially improve the bounds of previous family of algorithms which use a continual release mechanism. Experiments clearly validate our theoretical bounds.

1 Introduction

The well-known stochastic K-armed bandit problem (Thompson 1933; Robbins and others 1952) involves an agent sequentially choosing among a set of arms $\mathcal{A} = \{1, \dots, K\}$, 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_i with expectation $\mu_i \triangleq \mathbb{E}(r_t \mid a_t = i)$. The goal of the decision maker is to draw arms so as to maximize the total reward $\sum_{t=1}^{T} r_t$ obtained.

This problem is a model for many applications where there is a need for trading-off exploration and exploitation. This occurs because we only see the reward of the arm we pull. An example is clinical trials, where arms correspond to different treatments or tests, and the goal can be to maximise the number of cured patients over time while being uncertain about the effects of treatments. Other problems, such as search engine advertisement and movie recommendations can be formalised similarly (Pandey and Olston 2006).

It has been previously noted (Jain, Kothari, and Thakurta 2012; Thakurta and Smith 2013; Mishra and Thakurta 2015; Zhao et al. 2014) that privacy is an important consideration for many multi-armed bandit applications. Indeed, privacy can be easily violated by observing changes in the prediction of the bandit algorithm. This has been demonstrated for recommender systems such as Amazon by (Calandrino et al. 2011) and for user-targeted advertising such as Facebook by (Korolova 2010). In both cases, with a moderate amount of side information and by tracking changes in the output of the system, it was possible to learn private information of any targeted user.

Differential privacy (DP) (Dwork 2006) provides an answer to this privacy issue by making the output of an algorithm almost insensitive to any single user information. That is, no matter what side information is available to an outside observer, he can not have more information about a user than he already had by observing the outputs released by the algorithm. This goal is achieved by formally bounding the loss in privacy through the use of two parameters (ϵ, δ) as shown in Definition 2.1.

For bandit problems, differential privacy implies that the actions taken by the bandit algorithm do not reveal information about the sequence of rewards obtained. In the context of clinical trials and diagnostic tests, it guarantees that even an adversary with arbitrary side information, such as the identity of each patient, cannot learn anything from the output of the learning algorithm about patient history, condition, or test results.

1.1 Related Work

Differential privacy (DP) was introduced by (Dwork et al. 2006); a good overview is given in (Dwork and Roth 2013). While initially the focus in DP was static databases, interest in its relation to online learning problems has increased recently. In the full information setting, (Jain, Kothari, and Thakurta 2012) obtained differentially private algorithms with near-optimal bounds. In the bandit setting, (Thakurta and Smith 2013) were the first to present a differentially private algorithm, for the adversarial case, while (Zhao et al. 2014) present an application to smart grids in this setting. Then, (Mishra and Thakurta 2015) provided a differentially private algorithm for the stochastic bandit problem. Their algorithms are based on two non private stochastic bandit algorithms: Upper Confidence Bound (UCB, (Auer, Cesa-Bianchi, and Fischer 2002)) and Thompson sampling (Thompson 1933). Their results are sub-optimal: although simple index-based algorithms achieving $\mathcal{O}(\log T)$ regret exist (Burnetas and Katehakis 1996; Auer, Cesa-Bianchi, and Fischer 2002), these differentially private algo-

rithms additional poly-log terms in time T, as well further linear terms in the number of arms compared to the non-private optimal regret $\mathcal{O}(\log T)$.

We provide a significantly different and improved UCB-style algorithm whose regret only adds a constant, privacy-dependent term to the optimal. We also improve upon previous algorithms by relaxing the need to know the horizon T ahead of time, and as a result we obtain a uniform bound. Finally, we also obtain significantly improved bounds for a variant of the original algorithm of (Mishra and Thakurta 2015), by using a different proof technique and confidence intervals. Let's note that similarly to their result, we only make distributional assumptions on the data for the regret analysis. To ensure privacy, our algorithms do not make any assumption on the data. We summarize our contributions in the next section.

1.2 Our Contributions

- We present a novel differentially private algorithm (DP-UCB-INT) in the stochastic bandit setting that is almost optimal and only add an *additive constant* term (depending on the privacy parameter) to the optimal non private version. Previous algorithms had in large *multiplicative factors* to the optimal.
- We also provide an incremental but important improvement to the regret of existing differentially private algorithm in the stochastic bandit using the same family of algorithms as previously presented in the literature. This is done by using a simpler confidence bound and a more sophisticated proof technique. These bounds are achieved by DP-UCB-BOUND and DP-UCB algorithms.
- We present the first set of differentially private algorithm in the bandit setting which are unbounded and do not require the knowledge of the horizon T. Furthermore, all our regret analysis holds for any time step t.

2 Preliminaries

2.1 Multi-Armed Bandit

The well-known stochastic K-armed bandit problem (Thompson 1933; Lai and Robbins 1985; Auer, Cesa-Bianchi, and Fischer 2002) involves an agent sequentially choosing among a set of K arms $\mathcal{A} = \{1, \dots, K\}$. At each time step t, the player selects an action $a_t = i \in \mathcal{A}$ and obtains a reward $r_t \in [0, 1]$. The reward r_t is drawn from some fixed but unknown distribution P_i such that $\mathbb{E}(r_t \mid a_t) = \mu_i$. The goal of the decision maker is to draw arms so as to maximize the total reward obtained after T interactions. An equivalent notion is to minimize the total regret against an agent who knew the arm with the maximum expectation before the game starts and always plays it. This is defined by:

$$\mathcal{R} \triangleq T\mu_* - \mathbb{E}^{\pi} \sum_{t=1}^{T} r_t. \tag{2.1}$$

where $\mu_* \triangleq \max_{a \in \mathcal{A}} \mu_a$ is the mean reward of the optimal arm and $\pi(a_t \mid a_{1:t-1}, r_{1:t-1})$ is the policy of the decision maker, defining a probability distribution on the next actions a_t given the history of previous actions $a_{1:t-1} = a_1, \dots, a_{t-1}$ and rewards $r_{1:t-1} = r_1, \dots r_{t-1}$. Our goal is to bound the regret uniformly over T.

2.2 Differential Privacy

Differential privacy was originally proposed by (Dwork 2006), as a way to formalise the amount of information about the *input* of an algorithm, that is leaked to an adversary observing its *output*, no matter what the adversary's side information is. In the context of our setup, the algorithm's input is the sequence of rewards, and its output the actions. Consequently, we use the following definition of differentially private bandit algorithms.

Definition 2.1 ((ϵ, δ)-differentially private bandit algorithm)**.** A bandit algorithm π is (ϵ, δ)-differentially private if for all sequences $r_{1:t-1}$ and $r'_{1:t-1}$ that differs in at most one time step, we have for all $S \subseteq \mathcal{A}$:

$$\pi(a_t \in S \mid a_{1:t-1}, r_{1:t-1}) \leq \pi(a_t \in S \mid a_{1:t-1}, r'_{1:t-1}) e^{\epsilon} + \delta$$

where $\mathcal{A}$ is the set of actions. When $\delta = 0$, the algorithm is said to be ϵ-*differential private*.

Intuitively, this means that changing any reward r_t for a given arm, will not change too much the best arm released at time t or later on. If each r_t is a private information or a point associated to a single individual, then the definition aboves means that the presence or absence of that individual will not affect too much the output of the algorithm. Hence, the algorithm will not reveal any extra information about this individual leading to a privacy protection. The privacy parameters (ϵ, δ) determines the extent to which an individual entry affects the output; lower values of (ϵ, δ) imply higher levels of privacy.

A natural way to obtain privacy is to add a noise such as Laplace noise (Lap) to the output of the algorithm. The main challenge is how to get the maximum privacy while adding a minimum amount of noise as possible. This leads to a trade off between privacy and utility. In our paper, we demonstrated how to optimally trade-off this two notions.

2.3 Hybrid Mechanism

The hybrid mechanism is an online algorithm used to continually release the sum of some statistics while preserving differential privacy. More formally, there is a stream $\sigma_t = r_1, r_2 \dots r_t$ of statistics with r_i in $[0, 1]$. At each time step t a new statistic r_t is given. The goal is to output the partial sum ($y_t = \sum_{i=1}^{t} r_i$) of the statistics from time step 1 to t without compromising privacy of the statistics. In other words, we wish to find a randomised mechanism $M(y_t \mid \sigma_t, y_{1:t-1})$ that is (ϵ, δ)-differential private.

The hybrid mechanism solves this problem by combining the Logarithm and Binary Noisy Sum mechanisms. Whenever $t = 2^k$ for some integer k, it uses the Logarithm mechanism to release a noisy sum by adding Laplace noise of scale ϵ^{-1}. It then builds a binary tree B, which is used to release noisy sums until $t = 2^{k+1}$ via the Binary mechanism. This uses the leaf nodes of B to store the inputs r_i, while all other nodes store partial sums, with the root containing the

sum from 2^k to $2^{k+1}-1$. Since the tree depth is logarithmic, there is only a logarithmic amount of noise added for any given sum, more specifically Laplace noise of scale $\frac{\log t}{\epsilon}$ and mean 0 which is denoted by $\mathit{Lap}(\frac{\log t}{\epsilon})$.

(Chan, Shi, and Song 2010) proves that the hybrid mechanism is ϵ-differential private for any n where n is the number of statistics seen so far. They also show that with probability at least $1-\gamma$, the error in the released sum is upper bounded by $\frac{1}{\epsilon}\log(\frac{1}{\gamma})\log^{1.5} n$. In this paper, we derived and used a tighter bound for this same mechanism (see Appendix, Lemma (B.2) in (Tossou and Dimitrakakis 2015)) which is:

$$\frac{\sqrt{8}}{\epsilon}\log\frac{2}{\gamma}\cdot(\log n+1) \tag{2.2}$$

3 Private Stochastic Multi-Armed Bandits

We describe here the general technique used by our algorithms to obtain differential privacy. Our algorithms are based on the non-private UCB algorithm by (Auer, Cesa-Bianchi, and Fischer 2002). At each time step, UCB based its action according to an optimistic estimate of the expected reward of each arm. This estimate is the sum of the empirical mean and an upper bound confidence equal to $\sqrt{\frac{2\log t}{n_{a,t}}}$ where t is the time step and $n_{a,t}$ the number of times arm a has been played till time t. We can observe that the only quantity using the value of the reward is the empirical mean. To achieve differential privacy, it is enough to make the player based its action on *differentially private* empirical means for each arm. This is so, because, once the mean of each arm is computed, the action which will be played is a deterministic function of the means. In particular, we can see the differentially private mechanism as a black box, which keeps track of the vector of non-private empirical means Y for the player, and outputs a vector of private empirical means X. This is then used by the player to select an action, as shown in Figure 1.

We provide three different algorithms that use different techniques to privately compute the mean and calculate the index of each arm. The first, DP-UCB-BOUND, employs the Hybrid mechanism to compute a private mean and then adds a suitable term to the confidence bound to take into account the additional uncertainty due to privacy. The second, DP-UCB employs the same mechanism, but in such a way so as all arms have the same privacy-induced uncertainty; consequently the algorithm then uses the same index as standard UCB. The final one, employs a mechanism that only releases a new mean once at the beginning of each interval. This allows us to obtain the optimal regret rate.

3.1 The DP-UCB-BOUND Algorithm

In Algorithm 1, we compute the sum of the rewards of each arm using the hybrid mechanism (Chan, Shi, and Song 2010). However, the number and the variance of Laplace noise added by the hybrid mechanism increases as we keep pulling an arm. This means that the sum of each arm get added different amount of noise bigger than the original confidence bound used by UCB. This makes it difficult to identify the best arms. To solve this issue, we add a tight upper bound defined in equation (2.2) on the noise added by the hybrid mechanism.

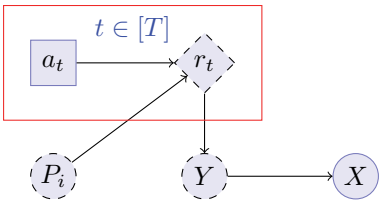


Figure 1: Graphical model for the empirical and private means. a_t is the action of the agent, while r_t is the reward obtained, which is drawn from the bandit distribution P_i. The vector of empirical means Y is then made into a private vector X which the agent uses to select actions. The rewards are essentially hidden from the agent by the DP mechanism.

Algorithm 1 DP-UCB-BOUND

Input ϵ, the differential privacy parameter.
Instantiate K Hybrid Mechanisms (Chan, Shi, and Song 2010); one for each arm a.
for $t \leftarrow 1$ to T **do**
  if $t \leq K$ **then**
    play arm $a = t$ and observe the reward r_t
    Insert r_t to the hybrid mechanism a
  else
    for all arm a **do**
      $s_a(t) \leftarrow$ total sum computed using the hybrid mechanism a
      $n'_{a,t} \leftarrow n_{a,t} - 2^{\lfloor \log_2 n_{a,t} \rfloor}$
        (Remove the closest power of 2 from $n_{a,t}$)
      $\nu_a \leftarrow \frac{4\sqrt{8}}{\epsilon}\log t \cdot (\log_2 n'_{a,t}+1)$
    end for
    Pull arm $a_t = \arg\max_a \frac{s_a(t)}{n_{a,t}} + \frac{\nu_a}{n_{a,t}}$
    Observe the reward r_t
    Insert r_t to the hybrid mechanism for arm a_t
  end if
end for

Theorem 3.2 validates this choice by showing that only $\mathcal{O}(\epsilon^{-1}\log\log t)$ factors are added to the optimal non private regret. In theorem 3.1, we demonstrate that Algorithm 1 is indeed ϵ-differential private.

Theorem 3.1. *Algorithm 1 is ϵ-differential private after any number of t of plays.*

Proof. This follows directly from the fact that the hybrid mechanism is ϵ-DP after any number t of plays and a single flip of one reward in the sequence of rewards 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) completes the proof. □

Theorem 3.2 gives the regret for algorithm 1. While here we give only a sketch proof, the complete derivation can be

found in (Tossou and Dimitrakakis 2015).

Theorem 3.2. *If Algorithm 1 is run with K arms having arbitrary reward distributions, then, its expected regret $\mathcal{R}$ after any number t of plays is bounded by:*

$$\mathcal{R} \leq \sum_{a:\mu_a<\mu_*} \max\left(B\left(\ln B + 7\right), \frac{8}{\lambda_0^2 \Delta_a}\log t\right) + \sum_{a:\mu_a<\mu_*} \left(\Delta_a + \pi^2 \Delta_a\right) \tag{3.1}$$

$$B = \frac{4\sqrt{8}}{\epsilon(1-\lambda_0)} \cdot \ln t$$

for any λ_0 such that $0 < \lambda_0 < 1$ where $\mu_1, \ldots, \mu_K$ are the expected values of $P_1, \ldots, P_K$ and $\Delta_a = \mu_* - \mu_a$.

Proof Sketch. We used the bound on the hybrid mechanism defined in equation 2.2 together with the union and Chernoff-Hoeffding bounds. We then select the error probability at each step to be $3\gamma = 6t^{-4}$. This leads to a transcendental inequality solved using the Lambert W function and approximated using section 3.1 of (Barry et al. 2000). □

3.2 The DP-UCB Algorithm

The key observation used in Algorithm 2 is that if at each time step we insert a reward to all hybrid mechanisms, then the scale of the noise will be the same. This means that there is no need anymore to compensate an additional bound. More precisely, every time we play an arm $a_t = a$ and receive the reward r_t, we not only add it to the hybrid mechanism corresponding to arm a but we also add a reward of 0 to the hybrid mechanism of all other arms. As these calculate a sum, it doesn't affect subsequent calculations.

Theorem 3.3 shows the validity of this approach by demonstrating a regret bound with only an additional factor of $\mathcal{O}(\epsilon^{-1}\log t)$ to the optimal non private regret.

Algorithm 2 DP-UCB

Run Algorithm 1 with $\nu_a = 0$
When arm $a_t = a$ is played, insert 0 to all hybrid mechanisms corresponding to arm $a' \neq a$ (Do not increase $n_{a',t}$)

Theorem 3.3. *If Algorithm 2 is run with K arms having arbitrary reward distributions, then, its expected regret $\mathcal{R}$ after any number t of plays is bounded by:*

$$\mathcal{R} \leq \sum_{a:\mu_a<\mu_*} \max\left(\frac{32\sqrt{2}\log^2 t}{\epsilon}, \frac{32\log t}{\Delta_a}\right) + \sum_{a:\mu_a<\mu_*} \left(\Delta_a + \pi^2 \Delta_a\right)$$

Proof Sketch. The proof is similar to the one for Theorem 3.2 with the error probability chosen to be $6t^{-4}$. □

3.3 The DP-UCB-INT Algorithm

Both Algorithms 1 and 2 enjoy a logarithmic regret with only a small additional factor in the time step t to the optimal non-private regret. However, this includes a multiplicative factor of ϵ^{-1}. Consequently, increasing privacy scales the total regret proportionally. A natural question is whether or not it is possible to get a differentially private algorithm with only an *additive* constant to the optimal regret. Algorithm 3 answers positively to this question by using novel tricks to achieve differential privacy. Looking at regret analysis of Algorithms 1 and 2, we observe that by adding noise proportional to ϵ, we will get a multiplicative factor to the optimal. In other words, to remove this factor, the noise should not depend on ϵ. But how can we get ϵ-DP in this case?

Note that if we compute and use the mean at each time step with an $\epsilon'_{n_{a,t}}$-DP algorithm, then after time step t, our overall privacy is roughly the sum $\mathcal{E}'$ of all $\epsilon'_{n_{a,t}}$. We then change the algorithm so that it only uses a released mean once every $\frac{1}{\epsilon}$ times, making privacy $\epsilon\mathcal{E}'$. In any case, $\epsilon'_{n_{a,t}}$ needs to decrease, at least as $n_{a,t}^{-1}$, for the sum to be bounded by $\log n_{a,t}$. However, $\epsilon'_{n_{a,t}}$ should also be big enough such that the noise added keeps the UCB confidence interval used at the same order, otherwise, the regret will be higher.

A natural choice for $\epsilon'_{n_{a,t}}$ is a p-series. Indeed, by making $\epsilon'_{n_{a,t}}$ to be of the form $\frac{1}{n_{a,t}^{v/2}}$, where $n_{a,t}$ is the number of times action a has been played until time t, its sum will converge to the Riemann zeta function when v is appropriately chosen. This choice of $\epsilon'_{n_{a,t}}$ leads to the addition of a Laplace noise of scale $\frac{1}{n_{a,t}^{1-v/2}}$ to the mean (See Lemma 3.1). Now our trade-off issue between high privacy and low regret is just reduced into choosing a correct value for v. Indeed, we can pick $v > 2$, for the privacy to converge; but the noise added at each time step will be increasing and greater than the UCB bound; which is not desirable. To overcome this issue, we used the more sophisticated k-fold adaptive composition theorem (III-3 in (Dwork, Rothblum, and Vadhan 2010)). Roughly speaking, this theorem shows that our overall privacy after releasing the mean a number of times depends on the sum of the *square* of each individual privacy parameter $\epsilon'_{n_{a,t}}$. So, $v > 1$ is enough for convergence and with $v \leq 1.5$, the noise added will be decreasing and will eventually become lower than the UCB bound.

In summary, we just need to *lazily update the mean* of each arm every $\frac{1}{\epsilon}$ times. However, we show that the interval of release is much better than $\frac{1}{\epsilon}$ and follows a series f as defined by Lemma (B.1) in the technical report (Tossou and Dimitrakakis 2015). Algorithm 3 summarizes the idea developed in this section.

The next lemma establishes the privacy ϵ' each time a new mean is released for a given arm a.

Lemma 3.1. *The mean $\hat{x}_a$ computed by Algorithm 3 for a given arm a at each interval is $n_{a,t}^{-v/2}$-differential private with respect to the reward sequence observed by that arm.*

Proof. Sketch This follows directly from the fact that we add Laplace noise of scale $n_{a,t}^{v/2-1}$. □

Algorithm 3 DP-UCB-INT (ϵ, v, K, $\mathcal{A}$)

Input $\epsilon \in (0,1]$ $v \in (1, 1.5]$; privacy rate.
K is the number of arms and $\mathcal{A}$ the set of all arms.
$f \leftarrow \lceil \frac{1}{\epsilon} \rceil$; $\hat{x} \leftarrow 0$
(For simplicity, we take the interval f to be $\lceil \frac{1}{\epsilon} \rceil$ here)
for $t \leftarrow 1$ to T **do**
  if $t \leq Kf$ **then**
    play arm $a = (t-1) \mod K + 1$ and observe r_t
  else
    for all $a \in \mathcal{A}$ **do**
      if $n_{a,t} \mod f = 0$ **then**
        $\hat{x}_a \leftarrow \frac{s_a}{n_{a,t}} + \mathcal{L}ap(0, \frac{1}{n_{a,t}^{1-v/2}}) + \sqrt{\frac{2 \log t}{n_{a,t}}}$
      end if
    end for
    Pull arm $a_t = \arg\max_a \hat{x}_a$ and observe r_t
    Update sum $s_a \leftarrow s_a + r_t$.
  end if
end for

The next theorem establishes the overall privacy after having played for t time steps.

Theorem 3.4. *After playing for any t time steps, Algorithm 3 is (ϵ', δ')-differential private with*

$$\epsilon' \leq \epsilon \cdot \min\left(\sum_{n=1}^{t} \frac{1}{\sqrt{n^v}}, \epsilon \sum_{n=1}^{t} \frac{e^{\frac{1}{\sqrt{n^v}}} - 1}{\sqrt{t^v}} + \sqrt{\sum_{n=1}^{t} \frac{2 \ln \frac{1}{\delta'}}{n^v}} \right)$$

for any $\delta' \in (0,1]$, $\epsilon \in (0,1]$

Proof Sketch. We begin by using similar observations as in Theorem 3.1. Then, we compute the privacy of the mean of an arm using the k-fold adaptive composition theorem in (Dwork, Rothblum, and Vadhan 2010) (see (Tossou and Dimitrakakis 2015)). □

The next corollary gives a nicer closed form for the privacy parameter which is needed in practice.

Corollary 3.1. *After playing for t time steps, Algorithm 3 is (ϵ', δ')-differential private with*

$$\epsilon' \leq \epsilon \cdot \min\left(\frac{t^{1-v/2} - v/2}{1 - v/2}, 2\epsilon\zeta(v) + \sqrt{2\zeta(v) \ln(1/\delta')} \right)$$

with ζ the Riemann Zeta Function for any $\delta' \in (0,1]$, $\epsilon \in (0,1]$, $v \in (1, 1.5]$.

Proof Sketch. We upper bounded the first term in theorem 3.4 by the integral test, then for the second term we used $e^x \leq 1 + 2x$ for all $x \in [0,1]$ to conclude the proof. □

The following corollary gives the parameter ϵ with which one should run Algorithm 3 to achieve a given ϵ' privacy.

Corollary 3.2. *If you run Algorithm 3 with parameter $\epsilon' = \sqrt{\frac{\log \frac{1}{\delta} + 4\epsilon}{8\zeta(v)}} - \sqrt{\frac{\log \frac{1}{\delta}}{8\zeta(v)}}$ for any $\delta \in (0,1]$, $\epsilon \in (0,1]$, $v \in (1, 1.5]$, you will be at least (ϵ, δ)-differential private.*

Proof. The proof is obtained by inverting the term using the Riemann zeta function in corollary 3.1. □

Finally, we present the regret of Algorithm 3 in theorem 3.5 . A simple observation shows us that it has the same regret as the non private UCB with just an additive constant.

Theorem 3.5. *If Algorithm 3 is run with K arms having arbitrary reward distributions, then, its expected regret $\mathcal{R}$ after any number t of plays is bounded by:*

$$\mathcal{R} \leq \sum_{a:\mu_a < \mu_*} \Delta_a \left[f_0 + \frac{8}{\Delta_a^2} \log t + 1 + 4\zeta(1.5) \right]$$

where $f_0 \leq \left(\sqrt{\frac{\log \frac{1}{\delta} + 4\epsilon}{8\zeta(v)}} - \sqrt{\frac{\log \frac{1}{\delta}}{8\zeta(v)}} \right)^{-1}$. More precisely, f_0 is the first value of the series f defined in Lemma B.1 in (Tossou and Dimitrakakis 2015).

Sketch. This is proven using a Laplace concentration inequality to bound the estimate of the mean then we selected the error probability to be $t^{-3.5}$. □

4 Experiments

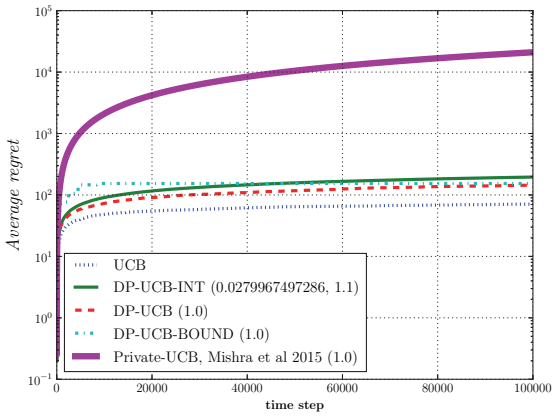
We perform experiments using arms with rewards drawn from independent Bernoulli distribution. The plot, in logarithmic scale, shows the regret of the algorithms over $100,000$ time steps averaged over 100 runs. We targeted 2 different ϵ privacy levels : 0.1 and 1. For DP-UCB-INT, we pick ϵ' according to corollary 3.2 such that the overall privacy is (ϵ, δ)-DP with $\delta = e^{-10}$, $\epsilon \in \{0.1, 1\}$, $v = 1.1$. We put in parenthesis the *input* privacy of each algorithm.

We compared against the non private UCB algorithm and the algorithm presented in (Mishra and Thakurta 2015) (*Private-UCB*) with a failure probability chosen to be t^{-4}.

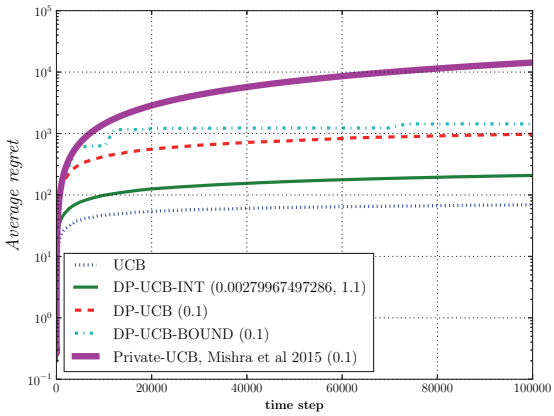
We perform two scenarios. Firstly we used two arms: one with expectation 0.9 and the other 0.6. The second scenario is a more challenging one with 10 arms having all an expectation of 0.1 except two with 0.55 and 0.2.

As expected, the performance of DP-UCB-INT is significantly better than all other private algorithms. More importantly, the gap between the regret of DP-UCB-INT and the non private UCB does not increase with time confirming the theoretical regret. We can notice that DP-UCB is better than DP-UCB-BOUND for small time steps. However, as the time step increases DP-UCB-BOUND outperforms DP-UCB and eventually catches its regret. The reason for that is: DP-UCB spends less time to distinguish between arms with close rewards due to the fact that the additional factor in its regret depends on $\Delta_a = \mu_* - \mu_a$ which is not the case for DP-UCB. Private-UCB performs worse than all other algorithms which is not surprising.

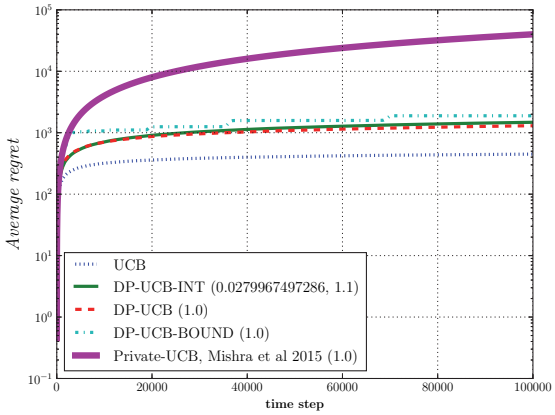
Moreover, we noticed that the difference between the best regret (after 100 runs) and worst regret is very consistent for all ours algorithms and the non private UCB (it is under 664.5 for the 2 arms scenario). However, this gap reaches $30,000$ for Private-UCB. This means that our algorithms are able to correctly trade-off between exploration and exploitation which is not the case for Private-UCB.



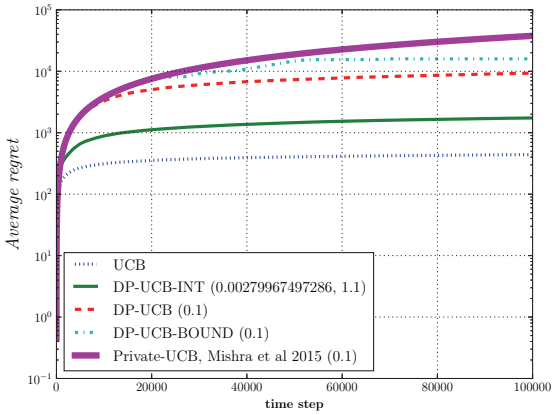
(a) Regret for $\epsilon = 1$, 2 arms



(b) Regret for $\epsilon = 0.1$, 2 arms



(c) Regret for $\epsilon = 1$, 10 arms



(d) Regret for $\epsilon = 0.1$, 10 arms

Figure 2: Experimental results with 100 runs, 2 or 10 arms with rewards: $\{0.9, 0.6\}$ or $\{0.1 \ldots 0.2, 0.55, 0.1 \ldots\}$.

5 Conclusion and Future Work

In this paper, we have proposed and analysed differentially private algorithms for the stochastic multi-armed bandit problem, significantly improving upon the state of the art. The first two, (DP-UCB and DP-UCB-BOUND) are variants of an existing private UCB algorithm (Mishra and Thakurta 2015), while the third one uses an interval-based mechanism.

Those first two algorithms are only within a factor of $\mathcal{O}(\epsilon^{-1} \log \log t)$ and $\mathcal{O}(\epsilon^{-1} \log t)$ to the non-private algorithm. The last algorithm, DP-UCB-INT, efficiently trades off the privacy level and the regret and is able to achieve the same regret as the non-private algorithm up to an additional *additive* constant. This has been achieved by using two key tricks: updating the mean of each arm lazily with a frequency proportional to the privacy ϵ^{-1} and adding a noise independent of ϵ. Intuitively, the algorithm achieves better privacy without increasing regret, because its output is less dependent on individual reward.

Perhaps it is possible to improve our bounds further if we are willing to settle for asymptotically low regret (Cowan and Katehakis 2015). A natural future work is to study if we can use similar methods for other mechanisms such as Thompson sampling (known to be differentially private (Dimitrakakis et al. 2014)) instead of UCB. Another question is whether a similar analysis can be performed for *adversarial* bandits.

We would also like to connect more to applications by two extensions of our algorithms. The first natural extension is to consider some side-information. In the drug testing example, this could include some information about the drug, the test performed and the user examined or treated. The second extension would relate to generalising the notion of neighbouring databases to take into account the fact that multiple observations in the sequence (say m) can be associated with a single individual. Our algorithms can be easily extended to deal with this setting (by re-scaling the privacy parameter to $\frac{\epsilon}{m}$). However, in practice, m could be quite large and it will

be an interesting future work to check if we could get sub linearity in the parameter m under certain conditions.

References

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

Barry, D.; Parlange, J.-Y.; Li, L.; Prommer, H.; Cunningham, C.; and Stagnitti, F. 2000. Analytical approximations for real values of the lambert w-function. *Mathematics and Computers in Simulation* 53(12):95 – 103.

Burnetas, A. N., and Katehakis, M. N. 1996. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics* 17(2):122–142.

Calandrino, J. A.; Kilzer, A.; Narayanan, A.; Felten, E. W.; and Shmatikov, V. 2011. "you might also like: " privacy risks of collaborative filtering. In *32nd IEEE Symposium on Security and Privacy*, 231–246.

Chan, T. H.; Shi, E.; and Song, D. 2010. Private and continual release of statistics. In *Automata, Languages and Programming*. Springer. 405–417.

Cowan, W., and Katehakis, M. N. 2015. Asymptotic behavior of minimal-exploration allocation policies: Almost sure, arbitrarily slow growing regret. *arXiv preprint arXiv:1505.02865*.

Dimitrakakis, C.; Nelson, B.; Mitrokotsa, A.; and Rubinstein, B. 2014. Robust and private Bayesian inference. In *Algorithmic Learning Theory*.

Dwork, C., and Roth, A. 2013. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science* 9(34):211–407.

Dwork, C.; McSherry, F.; Nissim, K.; and Smith, A. 2006. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Conference on Theory of Cryptography*, TCC'06, 265–284.

Dwork, C.; Rothblum, G. N.; and Vadhan, S. 2010. Boosting and differential privacy. In *Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, FOCS '10, 51–60.

Dwork, C. 2006. Differential privacy. In *ICALP*, 1–12. Springer.

Jain, P.; Kothari, P.; and Thakurta, A. 2012. Differentially private online learning. In Mannor, S.; Srebro, N.; and Williamson, R. C., eds., *COLT 2012 - The 25th Annual Conference on Learning Theory*, volume 23, 24.1–24.34.

Korolova, A. 2010. Privacy violations using microtargeted ads: A case study. In *ICDMW 2010, The 10th IEEE International Conference on Data Mining Workshops*, 474–482.

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

Mishra, N., and Thakurta, A. 2015. (nearly) optimal differentially private stochastic multi-arm bandits. *Proceedings of the 31th International Conference on Conference on Uncertainty in Artificial Intelligence (UAI-2015)*.

Pandey, S., and Olston, C. 2006. Handling advertisements of unknown quality in search advertising. In Schölkopf, B.; Platt, J. C.; and Hoffman, T., eds., *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing*, 1065–1072.

Robbins, H., et al. 1952. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society* 58(5):527–535.

Thakurta, A. G., and Smith, A. D. 2013. (nearly) optimal algorithms for private online learning in full-information and bandit settings. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013.*, 2733–2741.

Thompson, W. 1933. On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of two Samples. *Biometrika* 25(3-4):285–294.

Tossou, A. C. Y., and Dimitrakakis, C. 2015. Algorithms for Differentially Private Multi-Armed Bandits. Technical Report hal-01234427, Chalmers/INRIA.

Zhao, J.; Jung, T.; Wang, Y.; and Li, X. 2014. Achieving differential privacy of data disclosure in the smart grid. In *2014 IEEE Conference on Computer Communications, INFOCOM 2014*, 504–512.