THESIS FOR THE DEGREE OF LICENTIATE OF ENGINEERING

Privacy In The Age Of Artificial Intelligence

ARISTIDE C. Y. TOSSOU

Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY

Gothenburg, Sweden 2017

Privacy In The Age Of Artificial Intelligence
ARISTIDE C. Y. TOSSOU

Thesis for the degree of Licentiate of Engineering
ISSN 1652-876X
Technical Report No. 169L
Department of Computer Science and Engineering

Department of Computer Science and Engineering
Chalmers University of Technology
SE-412 96 Gothenburg
Sweden
Telephone: +46 (0)31-772 1000

Cover:
This shows how to reply with a fake answer by adding Laplace Noise in order to protect privacy while still giving useful answer.

Chalmers Reproservice
Gothenburg, Sweden 2017

To Bernadette, Rose and Armand.

Abstract

An increasing number of people are using the Internet in their daily life. Indeed, more than 40% of the world population have access to the Internet, while Facebook (one of the top social network on the web) is actively used by more than 1.3 billion users each day. This huge amount of customers creates an abundance of user data containing personal information. These data are becoming valuable to companies and used in various ways to enrich user experience or increase revenue.

This has led many citizens and politicians to be concerned about their privacy on the Internet to such an extent that the European Union issued a "Right to be Forgotten" ruling, reflecting the desire of many individuals to restrict the use of their information. As a result, many online companies pledged to collect or share user data anonymously. However, anonymisation is not enough and makes no sense in many cases. For example, an MIT graduate was able to easily re-identify the private medical data of Governor William Weld of Massachusetts from supposedly anonymous records released by the Group Insurance Commission. All she did was to link the insurance data with the publicly available voter registration list and some background knowledge.

Those shortcomings have led to the development of a more rigorous mathematical framework for privacy: Differential privacy. Its main characteristic is to bound the information one can gain from released data, no matter what side information they have available.

In this thesis, we present differentially private algorithms for the multi-armed bandit problem. This is a well known multi-round game, that originally stemmed from clinical trials applications and is now one promising solution to enrich user experience in the booming online advertising and recommendation systems. However, as recommendation systems are inherently based on user data, there is always some private information leakage. In our work, we show how to minimise this privacy loss, while maintaining the effectiveness of such algorithms. In addition, we show how one can take advantage of the correlation structure inherent in a user graph such as the one arising from a social network.

Acknowledgements

I would not have been able to write this licentiate thesis without the support of many people surrounding me. This is why I want to thank a few of them here for the time they gave me.

First, I would like to thank my supervisor Christos Dimitratakis for the huge support he has given me so far. You were always available and guiding me into the correct direction. The pieces of advice about my professional career you have been relentlessly giving me were really helpful. Next, is my co-supervisor Katerina Mitrokotsa for her huge support and for the awesome research visit in Tokyo you found for me. Furthermore, I would like to express my sincere gratitude to Kobbi Nissim for taking his time and energy traveling from the US to Sweden and leading the discussions on my licentiate thesis.

I cannot forget to thank Devdatt Dubashi for suggesting many useful and interesting research directions during my time so far; Graham Kemp for making me aware of an interesting conference that allowed me to disseminate my work back in country, Benin.

I am grateful to Pablo Picazo-Sanchez, Petre Mihail Anton and Ashkan Panahi for their bits of advice about life in Sweden and the various benefits at Chalmers; Hannes Eriksson for the many fruitful discussions about differential privacy and bandit algorithms; my lab mates Mikael, Olof, Prasanth for all the discussions and help given for writing this thesis. Furthermore, I show my gratitude to the many people whose name are not mentioned here but worked behind the scenes to help me.

Finally, I am indebted to my parents and friends who increased my motivation to continue this thesis.

List of publications

This thesis is based on the following manuscripts.

Paper I A. C. Y. Tossou and C. Dimitrakakis (2016a). "Algorithms for Differentially Private Multi-Armed Bandits". *AAAI*. AAAI Press, pp. 2087–2093

Paper II A. C. Y. Tossou and C. Dimitrakakis (2017). "Achieving Privacy in the Adversarial Multi-Armed Bandit". *AAAI*. AAAI Press, pp. 2653–2659

Paper III A. C. Y. Tossou, C. Dimitrakakis, and D. P. Dubhashi (2017). "Thompson Sampling for Stochastic Bandits with Graph Feedback". *AAAI*. AAAI Press, pp. 2660–2666

The following manuscripts have been published, but are not included in this work.

Paper IV A. C. Y. Tossou and C. Dimitrakakis (2015). "Optimal Advertisement Strategies for Small and Big Companies". *AFRICOMM*. vol. 171. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, pp. 94–98

Paper V P. Ekman et al. (2017). Learning to Match. *CoRR* **abs/1707.09678**

Paper VI A. Hossmann-Picu et al. (2016). "Synergistic user $\longleftrightarrow$ context analytics". *ICT Innovations 2015*. Vol. 399. Springer, pp. 163–172

Contents

List of Figures

List of Tables

Part I
Extended summary

Chapter 1

Introduction

The Digital Revolution spurred by computers and the Internet is now giving place to a 4th industrial revolution mainly driven by Artificial Intelligence and Machine Learning (Schwab 2016).

The main characteristic of this revolution is emphasized by its scale. For example, in the second quarter of 2017, Facebook hit more than 1.3 billion daily active users; Amazon net sales reached 136 billion USD and Netflix has more than 103 million subscribers (Statista 2017). This has led to a huge amount of data, creating new opportunities for companies to both improve their revenue and increase the satisfaction of their customers. One of the most widely used opportunity is to heavily customize the browsing experience for returning customers by recommending items they will like. For example, Amazon deploys a recommendation system to suggest interesting products to buy; Netflix lists movies that users would like to watch and Facebook displays advertisements that would be useful to read.

For recommendations to be effective, one needs to have a great understanding of user's preferences. A video games enthusiast might be presented with products and devices that will improve his gaming experience, while someone learning a new programming language might see book suggestions. But it is particularly important to also take into account the current needs of a user. So, it is useless to propose the new FIFA 18 [1] to a gamer while he/she is just looking for tools to fix his broken PlayStation 4 [2].

Apart from having accurate recommendations, one of the main challenges is how long it takes to display relevant suggestions. Indeed, we want to decrease the number of pages the user visits and the time he/she spends browsing before getting useful recommendations for fear of losing him/her. If we knew the user's immediate needs, we would just recommend the best products. But we are uncertain about their needs and as a result, have to explore. And we cannot explore too much either as the time is limited.

One of the well studied mathematical frameworks to solve this dilemma is the multi-armed bandit. According to Wikipedia 2017, *"The multi-armed bandit problem is the*

[1] FIFA 18 is a football simulation video game developed and published by Electronic Arts.

[2] PlayStation 4 (PS4) is a home video console that is used as a device to play games on. FIFA 18 is one of the games that can be played on the PS4.

problem a gambler faces at a row of slot machines when deciding which machines to play, how many times to play each machine and in which order to play them. When played, each machine provides a random reward from a distribution specific to that machine. The objective of the gambler is to maximize the sum of rewards earned through a sequence of lever pulls." For the case of products recommendation, it means that each time a user requests a web page view, we display one set of products out of K possible choices. Then, the user sends his/her feedback: by clicking a link; by buying or rating the product recommended; or by making a new search query. Next, we proceed either to another user or to the next page view of this user. Our goal here could be to improve the overall click-through rate, the amount of sales generated or the time spent on the website.

However, the feedback provided by users can carry private information. For example, in experimental clinical trials, it could be used to know if a user has a particular disease. Someone clicking on an advertisement to solve alcoholic addictions is indicating his medical condition that might be used by an insurance company against him. A woman suddenly purchasing a set of products following a specific pattern could indicate that she is pregnant. Target, a US discount store, used clues such as vitamin supplements, large quantities of lotion, and hand sanitizers, typical to many pregnant women to predict a girl pregnancy even before her father knew about it (Forbes 2012).

The previous examples show that users face serious privacy breaches by the companies collecting their feedbacks. These breaches are made possible through two main routes: (1) the direct revelation of sensitive personal information (2) their misuse for other purposes. More worryingly, there is another type of privacy breach by malicious third parties (people neither the company nor the users intended to reveal the private data to). This is done by inferring the private information of the users with the help of record linkage. This is possible because many companies using recommendation systems, earn money by releasing some information about their users. Although this information alone is not enough to breach privacy, armed with some moderate side information a malicious person can learn the private data of a specific user. For example, if you make an advertisement campaign on Facebook, it will disclose the number of people who have clicked on your advertisement as well as their demographic and other statistics. Based on that, (Korolova 2010) was able to detect the age of a specific user who had explicitly configured it to be hidden. Amazon lists the reviews of users publicly and products' recommendations are also publicly available. Armed with that information, (Calandrino et al. 2011) was able to infer what a specific user bought on a specific day even when this user did not make any review (or public comment) for his buy. All those attacks were passive and could be performed by any user on the Internet.

Many solutions have been proposed to solve this privacy issue. The first and most used one is data anonymisation. Given a dataset consisting of a list of fields for each individual, anonymisation encrypts or removes fields containing personally identifiable information (PII) before releasing the dataset to the public [3]. However, it is easy to cross-reference this anonymised data with other sources and de-anonymise it. An obvious example is a famous case where the private medical records of Governor William Weld of Massachusetts were identified in supposedly anonymous records released by the Group

[3]Anonymisation can also be applied when collecting the data for internal use or in all cases where someone needs to access it.

Insurance Commission (Ohm 2009). Also, by themselves, some fields may not be PII, but combined with other non-PII fields they can identify a user. K-anonymity (Sweeney 2002) tries to solve this problem by changing fields such that the information for each person contained in the release cannot be distinguished from at least $K-1$ individuals with respect to the PII-fields. But this still does not prevent the de-anonymization issue (See Machanavajjhala et al. 2007 for example attacks). Aggregation which released a single statistic, averaged over many users, is another technique believed to protect privacy. But as demonstrated by the Facebook attack mentioned earlier, aggregation alone is not enough to prevent third parties from inferring users' data. The main issue in all those techniques is that they are implicitly or explicitly restricting the side information that a third party can get. As soon as someone has more side information, they can breach those techniques.

This is what motivated the development of another framework for formalizing privacy: Differential Privacy (DP) (Dwork 2006). DP is a property that provides privacy guarantees for data release no matter what side information is available to a third party. It provides an upper bound on the information a third party can gain about the data after the release. We will now give an example algorithm that satisfies the DP property. To answer the question: "*Did you vote for Trump?*", a coin is flipped in secret. If it is heads, the true answer is given otherwise one answers randomly. This means that no matter how powerful a third party is, they will not know if a user actually voted for Trump.

In order for any algorithm to be differentially private, it has to be randomized. So, noise is inherent to DP; otherwise, there is always some side information that could lead to total privacy loss. The more noise is added, the more private we are. However, the amount of noise could affect the usefulness of the data released. For example, we still want a recommendation to be useful to users and if the algorithm is based on a version of the feedback too noisy it might affect the accuracy. This is the main goal of this thesis. How can we simultaneously achieve strong privacy guarantees while still recommending very accurate choices for sequential decision problems?

Main contributions. Our main contribution is the development of multi-armed bandit algorithms whose choices are more optimal while achieving a given privacy loss. We derived formal bounds that improved on previous results, for both the optimality and privacy of our algorithms.

Thesis outline. In Chapter 2, we formally define multi-armed bandits and differential privacy, then we introduce the necessary background that will help understand the remaining of this thesis.

In Chapter 3, we summarize the contributions of this thesis. Section 3.1 discusses the algorithms we introduced in the stochastic multi-armed bandit that improves on the regret for a given privacy level. Section 3.2 demonstrates that one can reuse the inherent noise in existing adversarial multi-armed bandits algorithms to achieve difference privacy. In section 3.3, we present algorithms demonstrating that knowing some correlation between users can improve the recommendations made by a multi-armed bandit algorithms. This raises an important question about how privacy can be achieved with respect to such correlations.

Chapter 4 concludes the thesis and discusses some interesting future work. The remainder of the thesis is a reprint of the full version of the three papers (Tossou, Dimitrakakis, and Dubhashi 2017; Tossou and Dimitrakakis 2017; Tossou and Dimitrakakis 2016a) included in this work.

Chapter 2

Background

In this chapter, we will formally define what is a multi-armed bandit problem. Then, we will talk about the different types of bandit problems we dealt with in our work. Finally, we will define differential privacy and explain what it means in the context of bandit algorithms.

2.1 Multi-Armed Bandits

Formally, a bandit game is defined between an adversary and an agent as follows: there is a set of K arms $\mathcal{A}$, and at each round t, the agent plays an arm $I_t \in \mathcal{A}$. Given the choice I_t, the adversary grants the agent a reward $r_{I_t,t} \in [0,1]$. Whereas the adversary selects a reward for each arm, the agent only observes the reward of arm I_t, and not that of any other arms. The goal of this agent is to maximize its total reward after T rounds, $\sum_{t=1}^{T} r_{I_t,t}$. A randomized bandit algorithm $\Lambda : (\mathcal{A} \times [0,1])^* \to \mathscr{D}(\mathcal{A})$ maps every arm-reward history to a distribution over the next arms to take.

Relying on the total (cumulative) reward of an agent to evaluate its performance can be misleading. Indeed, consider the case where an adversary gives zero as reward for all arms at every round. The cumulative reward of the agent would look bad but no other agents could have done better. This is why one compares the gap between the agent's cumulative reward and the one obtained by some hypothetical agent, called *oracle*, with additional information or computational power. This gap is called the *regret*.

There are many variants of the oracle that are considered in the literature. The most common variant is the *fixed oracle*, which always plays the best fixed arm in hindsight. The regret $\mathcal{R}$ against this *oracle* is :

$$\mathcal{R} = \max_{i=1,\dots K} \sum_{t=1}^{T} r_{i,t} - \sum_{t=1}^{T} r_{I_t,t}$$

In practice, we either prove a high probability bound on $\mathcal{R}$ or an expected value $\mathbb{E}\,\mathcal{R}$ with:

$$\mathbb{E}\,\mathcal{R} = \mathbb{E}\left[\max_{i=1,\dots K}\sum_{t=1}^{T} r_{i,t} - \sum_{t=1}^{T} r_{I_t,t}\right]$$

where the expectation is taken with respect to the random choices of both the agent and adversary.

The nature of the adversary, and specifically, how the rewards are generated, determines the nature of the game. We have two main adversaries: the stochastic and the non-stochastic discussed below.

2.1.1 Stochastic Adversary

In the stochastic multi-armed bandit problem (Thompson 1933; Auer, Cesa-Bianchi, and Fischer 2002), the reward obtained at round t is generated i.i.d from a distribution P_{I_t}. More precisely, at each round t, the agent plays an arm $I_t \in \mathcal{V}$ and receives a reward $r_t = R(Y_{t,I_t})$, where $Y_{t,I_t} : \Omega \to \mathcal{Y}$ is a random variable defined on some probability space (P, Ω, Σ) and $R : \mathcal{Y} \to \mathbb{R}$ is a reward function.

Each arm i has mean reward $\mu_i(P) = \mathbb{E}_P\, R(Y_{t,i})$. Our goal is to maximize its expected cumulative reward after T rounds. An equivalent notion is to minimize the expected regret against an oracle which knows P. More formally, the expected regret $\mathbb{E}_P^\pi\,\mathcal{L}$ of an agent policy π for a bandit problem P is defined as:

$$\mathbb{E}_P^\pi\,\mathcal{L} = T\mu_*(P) - \mathbb{E}_P^\pi \sum_{t=1}^{T} r_{I_t}, \tag{2.1.1}$$

where $\mu_*(P) = \max_{i\in\mathcal{V}} \mu_i(P)$ is the mean of the optimal arm and $\pi(I_t|h_t)$ is the policy of the agent, defining a probability distribution on the next arm I_t given the history $h_t = \langle I_{1:t-1}, r_{1:t-1}\rangle$ of previous arms and rewards.

The main challenge in this model is that the agent does not know P, and it only observes the reward of the arm it played. As a consequence, the agent must trade-off exploitation (taking the apparently best arm) with exploration (trying out other arms to assess their quality).

The Bayesian setting offers a natural way to model this uncertainty, by assuming that the underlying probability law P is in some set $\mathcal{P} = \{\, P_\theta \mid \theta \in \Theta \,\}$ parametrised by θ, over which we define a prior probability distribution $\mathbb{P}$. In that case, we can define the Bayesian regret:

$$\mathbb{E}^\pi\,\mathcal{L} = \int_\Theta \mathbb{E}_{P_\theta}^\pi(\mathcal{L})\,\mathrm{d}\,\mathbb{P}(\theta). \tag{2.1.2}$$

It is the regret the agent expects to obtain given its uncertainty about the true parameter if it uses the policy π to select actions.

2.1.2 Non-Stochastic Adversary

If the rewards are not generated independently and identically at each round, then we are in the non-stochastic case. It comprises of several adversaries. The *fully oblivious*

adversary (Audibert and Bubeck 2010) generates the rewards independently at round t but not necessarily identically from a distribution $P_{I_t,t}$. There is also the more general *oblivious* adversary (Auer, Cesa-Bianchi, Freund, et al. 2003) whose only constraint is to generate the reward $r_{I_t,t}$ as a function of the current action I_t only, i.e. ignoring previous actions and rewards. Furthermore we have the stronger *m-bounded memory adaptive adversary* (Cesa-Bianchi, Dekel, and Shamir 2013; Merhav et al. 2002; Dekel, Tewari, and Arora 2012) who can use up to the last m rewards. The oblivious adversary is a special case with $m = 0$. Another special case of this adversary is the one with *switching costs*, who penalises the agent whenever he switches arms, by giving the lowest possible gain of 0 (here $m = 1$).

2.2 Differential privacy

The following definition (Dwork and Roth 2013) formally introduces what it means for an algorithm to be differentially private.

Definition 2.2.1 ((ϵ, δ)-differentially private algorithm)**.** A randomized algorithm $\mathcal{M}$ with domain a dataset D is (ϵ, δ)-differentially private if for all $S \subseteq \text{Range}(\mathcal{M})$ and for all $x, y \in D$ such that x and y differs in a single element:

$$\mathbb{P}\left[\mathcal{M}(x) \in S\right] \leq e^{\epsilon}\, \mathbb{P}\left[\mathcal{M}(y) \in S\right] + \delta \tag{2.2.1}$$

When $\delta = 0$, the algorithm is said to be ϵ-*differentially private.*

The ϵ and δ parameters quantify the amount of privacy loss. Lower (ϵ,δ) indicate higher privacy and consequently we will also refer to (ϵ,δ) as the privacy loss. Definition 2.2.1 means that the output of the algorithm is almost insensible to any single change in its input sequence. This implies that whether or not we remove a single element, or replace it, the algorithm will still produce almost the same output. Assuming that a single element is linked to a user private data (for example his cancer status or the advertisement he clicked), the definition preserves the privacy of that user against any third parties looking at the output. This is the case because the choices or the participation of that user would not almost affect the output. Equation (2.2.1) specifies how much the output is affected by a single user.

2.2.1 Single round differentially private bandit algorithm

The following definition (from Tossou and Dimitrakakis 2016b) specifies what is meant when we call a bandit algorithm differentially private at a single round t:

Definition 2.2.2 (Single round (ϵ, δ)-differentially private bandit algorithm)**.** A randomized bandit algorithm Λ is (ϵ, δ)-differentially private at round t, if for all sequences $r_{1:t-1}$ and $r'_{1:t-1}$ that differs in at most one round, we have for any action subset $S \subseteq \mathcal{A}$:

$$\mathbb{P}_{\Lambda}(I_t \in S \mid r_{1:t-1}) \leq \delta + \mathbb{P}_{\Lambda}(I_t \in S \mid r'_{1:t-1})e^{\epsilon}, \tag{2.2.2}$$

where $\mathbb{P}_{\Lambda}$ denotes the probability distribution specified by the algorithm and $r_{1:t-1} = \{r_1, \ldots r_{t-1}\}$ with r_s the rewards of all arms at round s.

This is by no means the only definition possible. One could also define a bandit algorithm that would be differentially private not with respect to the reward *sequence*, but with respect to the reward *function*, when the arms generate outcomes over which the player has some preferences. The latter definition would be natural if what we wanted to hide would be the player's preferences. However, in our setting, we want to hide individual rewards, as they may be connected to side information – such as patient data in the clinical trial example.

2.2.2 Differentially private bandit algorithm

We would like Definition 2.2.2 to hold for all rounds, so as to protect the privacy of all users. If it does for some (ϵ, δ), then we say the algorithm has *per-round* or *instantaneous* privacy loss (ϵ, δ). Such an algorithm also has a *cumulative* privacy loss of at most (ϵ', δ') with $\epsilon' = \epsilon T$ and $\delta' = \delta T$ after T steps [1]. Our goal is to design bandit algorithm such that their cumulative privacy loss (ϵ', δ') are as low as possible while achieving simultaneously a very low regret. In practice, we would like ϵ' and the regret to be sub-linear while δ' should be a very small quantity. Definition 2.2.3 formalizes clearly the meaning of this cumulative privacy loss and for ease of presentation, we will ignore the term "cumulative" when referring to it.

Definition 2.2.3 ((ϵ, δ)-differentially private bandit algorithm)**.** A randomized bandit algorithm Λ is (ϵ, δ)-differentially private up to round t, if for all $r_{1:t-1}$ and $r'_{1:t-1}$ that differs in at most one round, we have for any action subset $S \subseteq \mathcal{A}^t$:

$$\mathbb{P}_\Lambda(I_{1:t} \in S \mid r_{1:t-1}) \leq \delta + \mathbb{P}_\Lambda(I_{1:t} \in S \mid r'_{1:t-1})e^\epsilon, \tag{2.2.3}$$

where $\mathbb{P}_\Lambda$ and r are as defined in Definition 2.2.2.

[1] This is due to the basic composition theorem of differential privacy in Theorem 3.1 from Dwork, Rothblum, and Vadhan 2010.

Chapter 3

Differential privacy in multi-armed bandits

The following sections will outline the contributions of this thesis.

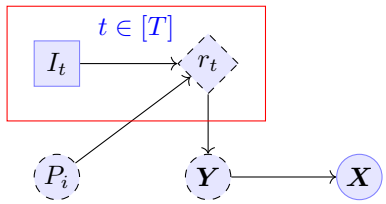

Figure 3.1.1: *Graphical model for the empirical and private means. I_t is the action of the agent, while r_t is the reward obtained, which is drawn from the bandit distribution P_i. The vector of empirical means $\boldsymbol{Y}$ is then made into a private vector $\boldsymbol{X}$ which the agent uses to select actions. The rewards are essentially hidden from the agent by the differentially private mechanism.*

3.1 How to achieve privacy in the stochastic bandit problem

We developed three algorithms that are based on the non-private UCB (Auer, Cesa-Bianchi, and Fischer 2002) algorithm. At each round t, UCB takes the best action according to an optimistic estimate of the expected reward of each arm. This is the sum of the empirical mean of the arm and an upper bound confidence equal to $\sqrt{\frac{2\log t}{n_{a,t}}}$ where $n_{a,t}$ is the number of times arm a has been played till round t. As this estimate is the only quantity depending on the reward, it is enough to make it private to achieve differential privacy (See Figure 3.1.1 for an illustration).

Our first algorithm (Tossou and Dimitrakakis 2016b) called DP-UCB-Bound shows that by using the hybrid mechanism [1] (Chan, Shi, and Song 2010) to compute a private version of the sum of rewards we can achieve ϵ-differential privacy with only a regret of $O(\epsilon^{-1}\log t \cdot \log^2 \log t)$ compared to a regret of $O(\log t)$ for the non private UCB.

Our second algorithm called DP-UCB builds on the first one and makes sure the variance of the noise on the sum of rewards is the same for all arms. This is done by adding a fake reward of 0 to all arms not played at round t; without increasing the number of times they have been played. This simple modification is enough to get a better regret of $O(\epsilon^{-1}\log t \cdot \log\log t)$.

Our next algorithm called DP-UCB-Int simply adds Laplace noise of appropriate scale to the true empirical mean of each arm. However, it only updates this mean every f_t rounds where f_t is a decreasing sequence. This algorithm achieves an optimal regret $(O(\epsilon^{-1} + \log t))$ with only additive constant. A key idea in the proof is to see that when we are not updating the reward we don't suffer any privacy loss. Table 3.1.1 summarizes our results in the stochastic case and also shows the improvement over the differentially private UCB algorithm (called Private-UCB) presented by (Mishra and Thakurta 2015).

[1] an online algorithm that can compute the sum of a stream of number while preserving differential privacy.

Algorithms	**Privacy**	**Regret**
DP-UCB-BOUND	ϵ	$O(\frac{\log t \cdot \log^2 \log t}{\epsilon})$
DP-UCB	ϵ	$O(\frac{\log t \cdot \log \log t}{\epsilon})$
DP-UCB-INT	$(\epsilon, \delta \leq T^{-4})$	$O(\epsilon^{-1} + \log t)$
UCB	None	$O(\log t)$
Private-UCB	ϵ	$O(\frac{\log^3 t}{\epsilon})$

Table 3.1.1: Summary of our bound for the stochastic DP bandit

3.2 How to achieve privacy in the adversarial bandit problem

We developed two algorithms for the oblivious adversarial case [2]. While focusing on oblivious adversaries, we discovered that by targeting differential privacy we can also compete against the stronger *m-bounded memory adaptive adversary*.

Our algorithms are a variant of the EXP3 algorithm (Auer, Cesa-Bianchi, Freund, et al. 2003). EXP3 takes an action according to a discrete probability distribution proportional to the exponential of the weight of each arm. This weight is the discounted sum of the non biased rewards received by each arm.

Our first algorithm (*DP-EXP3-Lap*) shows that by simply adding a Laplace noise to each reward (See Figure 3.2.1 for more details), we incur a regret of $O(\sqrt{T}\log T/\epsilon)$ in the oblivious case to achieve ϵ-differential privacy. This significantly improves on the previous state of the art regret of $O(T^{2/3}/\epsilon)$ for the same ϵ-differential privacy.

Next, we observed that the EXP3 algorithm is an instance of the exponential mechanism (McSherry and Talwar 2007) and as a result, is by itself differentially private. We then improved its intrinsic privacy/regret trade-off by dividing the rounds into disjoint mini-batches of size τ. The policy of EXP3 is updated only at the beginning of each mini-batch by using the mean of the rewards received so far. So, for the remaining of the mini-batch we play actions using the same policy obtained from the previous mini-batch (See Figure 3.2.2 for an illustration). This algorithm [3] called *EXP3*$_\tau$ achieves an impressive privacy loss of $\mathcal{O}(\sqrt{\log T})$ for a regret of $\mathcal{O}(T^{2/3})$ in the oblivious case. This regret and privacy also hold against the stronger 1-memory bounded adversarial case.

3.3 Thompson Sampling for Stochastic Bandits with Graph Feedback

Thompson Sampling maintains a distribution over the problem parameters (for example the rewards of each arm in the multi-armed bandit problem). At each round, it selects an

[2] See section 2.1.2 for a definition of oblivious adversary.

[3] The algorithm is inspired from Dekel, Tewari, and Arora 2012.

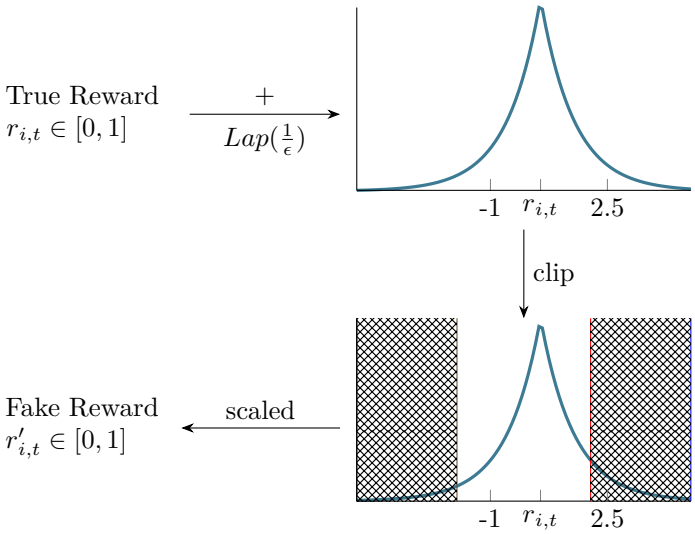

Figure 3.2.1: *Illustration of how* DP-EXP3-Lap *algorithm works. A Laplace noise of scale* $\frac{1}{\epsilon}$ *is added to the true reward* $r_{i,t}$*. Then the result is clipped into a bounded interval before being scaled to* $[0,1]$*. This gives* $r'_{i,t}$ *which is used instead of* $r_{i,t}$*.*

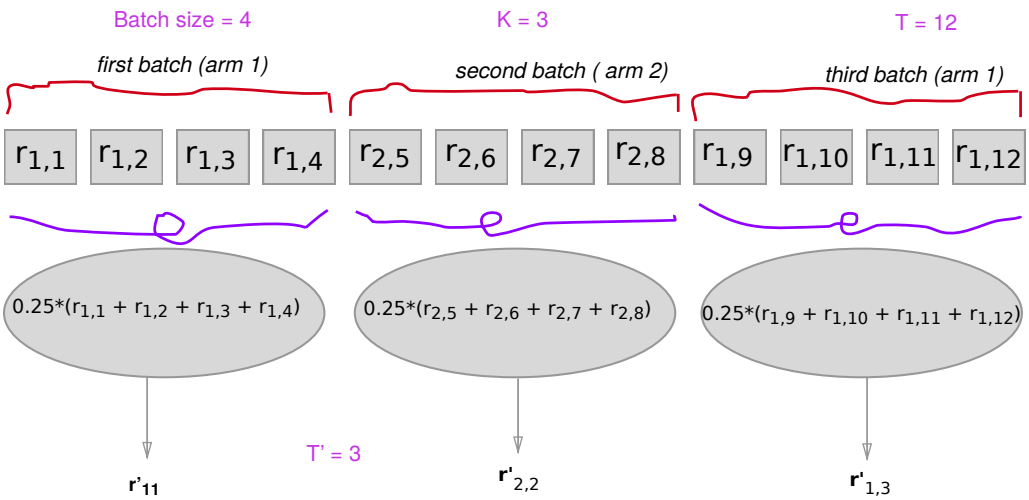

Figure 3.2.2: *Illustration of how the* EXP3$_\tau$ *algorithm works. We have 12 rounds (T is 12) and there are grouped into 3 batches of size 4. The first 4 rewards in the first batch are averaged and gives a reward* $r'_{1,1}$*. Similarly for the remaining two batches giving the rewards* $r'_{2,2}$ *and* $r'_{1,3}$*.* EXP3$_\tau$ *only observe those 3 rewards (*$r'_{1,1}$ $r'_{2,2}$ *and* $r'_{1,3}$*) and for him, the number of rounds is 3.*

Algorithms	**Privacy**	**Regret**
DP-EXP3-Lap	ϵ	$\mathcal{O}(\frac{\sqrt{T}\log T}{\epsilon})$
$EXP3_\tau$	$\left(\mathcal{O}(\sqrt{\log T}), \delta \leq T^{-2}\right)$	$\mathcal{O}(T^{2/3})$
EXP3	$\mathcal{O}(T)$	$\mathcal{O}(\sqrt{T})$
Previous Best	ϵ	$\mathcal{O}(\frac{T^{2/3}}{\epsilon})$

Table 3.2.1: Summary of our bound for the oblivious DP bandit

arm according to the probability of its mean being the largest under this distribution. It then observes a set of rewards which it uses to update its probability distribution over the parameters.

We develop a simple extension to Thompson Sampling for multi-armed bandits with side information, where choosing an action provides additional information for a subset of the remaining actions. Formally, we assume the existence of an undirected graph $G = (\mathcal{V}, \mathcal{E})$ [4] with vertices corresponding to arms. By taking an arm $I_t \in \mathcal{V}$, not only we receive the reward of the arm we played, but we also observe the rewards of all neighbouring arms $\mathcal{N}_{I_t} = \{\, a' \in \mathcal{V} \mid (I_t, a') \in E \,\}$. More precisely, at each round t we observe $Y_{t,a'}$ for all $a' \in \mathcal{N}_{I_t}$, while our reward is still $r_t = R(Y_{t,I_t})$.

Our fist policy, called TS-N, simply updates the distribution of all arms observed separately at each round. The second policy, called TS-MaxN, fully exploits the graphical structure. For example, as noted by Caron et al. 2012, instead of doing exploration on arm i we could explore an apparently better neighbour, which would give us the same information. More precisely, instead of picking arm i, we pick the arm $j \in \mathcal{N}_i$ with the best empirical mean. The intuition behind it is that, if we take any arm in $\mathcal{N}_i$, we are going to observe anyway the reward of i. So, it is always better to exploit the best arm in $\mathcal{N}_i$.

We then performed extensive experiments both with simulated and real-world graphs and compared against similar algorithms built on top of UCB. Our main observation is that TS-N even while not explicitly exploiting the full structure of the graph managed to outperform UCB-MaxN that fully exploit the graph but uses UCB instead of Thompson Sampling.

Finally, our theoretical results extend Russo and Roy 2016 to graph-structured feedback. We thus obtain a problem-independent bound of $O(\sqrt{\frac{1}{2}\chi(\overline{G})T})$ for the Bayesian regret defined in equation (2.1.2).

[4] This graph is not assumed to be fixed. It can change at every round. All our algorithms and analysis apply to this case.

Chapter 4

Concluding remarks

In this thesis, we introduced a definition of privacy for the multi-armed bandit problem. Then, we presented various algorithms that can simultaneously minimise the regret and privacy loss incurred.

More precisely, for the stochastic multi-armed bandit we developed three algorithms on top of UCB. The first two (DP-UCB, DP-UCB-Bound) compute the sum of rewards for each arm in a differentially private way using the Hybrid Mechanism. They only differ in how they treat the arms that are not observed at each round. We show that their regret already improves on the state-of-art for a given privacy loss ϵ. The third (DP-UCB-Int) makes its output less dependent on individual rewards by updating the mean of each arm less frequently. The frequency of updates is chosen such that the privacy loss and regret are efficiently traded off.

In the oblivious adversarial case, our algorithms are built on top of EXP3. The first (*DP-EXP3-Lap*) simply adds Laplace noise to the rewards before it is observed. We show that this simple technique is already enough to achieve near-optimal regret (up to logarithmic factors) for a given privacy loss. Finally, after observing that EXP3 is by itself differential private albeit with a linear privacy loss, we presented a technique ($EXP3_\tau$) that is sub-linear both in regret and privacy loss. The key idea is very similar to that of DP-UCB-Int. We depend less on individual rewards by updating the parameters of each arm less frequently. $EXP3_\tau$ achieved this by dividing the rounds into disjoints sets. Furthermore, each set is treated as a single big round in the sense that we only observe one reward [1].

Finally, we built algorithms on top of Thompson Sampling that take advantage of the side information available when each arm is played.

So far, we have built algorithms whose bounds improved on previous results in the multi-armed bandit problem. However, a natural question is whether one can do better. The best way to address it is to work on deriving a lower bound for differentially private bandits algorithms. Another interesting future work is to check if one can re-use inherent noise in Thompson Sampling to prove a differentially private bound. This would be similar to how we re-use the noise in EXP3 to prove a privacy bound. Next, we would

[1] the mean of all rewards for the rounds in the set.

like to extend the results obtained in this thesis to take into account correlation between users (for example in the form of a social graph). This is a more realistic setting as many users information are correlated, either through their friends or because they use multiple devices. We could, for example, take these correlations into account by satisfying pufferfish privacy (Kifer and Machanavajjhala 2014), a generalisation of differential privacy. Finally, we would like to explore the feasibility of distributed and local algorithms that can achieve privacy while taking into account those correlations. This looks like a very tough problem and we can start by analysing what can actually be learned privately (Kasiviswanathan et al. 2011).

References

Audibert, J.-Y. and S. Bubeck (2010). Regret Bounds and Minimax Policies Under Partial Monitoring. *J. Mach. Learn. Res.* **11**, 2785–2836. ISSN: 1532-4435. URL: http://dl.acm.org/citation.cfm?id=1756006.1953023.

Auer, P., N. Cesa-Bianchi, and P. Fischer (2002). Finite Time Analysis of the Multiarmed Bandit Problem. *Machine Learning* **47**.2/3, 235–256.

Auer, P., N. Cesa-Bianchi, Y. Freund, et al. (2003). The Nonstochastic Multiarmed Bandit Problem. *SIAM J. Comput.* **32**.1, 48–77. ISSN: 0097-5397. DOI: 10.1137/S0097539701398375. URL: http://dx.doi.org/10.1137/S0097539701398375.

Calandrino, J. A. et al. (2011). ""You Might Also Like: " Privacy Risks of Collaborative Filtering". *32nd IEEE Symposium on Security and Privacy*, pp. 231–246.

Caron, S. et al. (2012). Leveraging side observations in stochastic bandits. *UAI*.

Cesa-Bianchi, N., O. Dekel, and O. Shamir (2013). "Online Learning with Switching Costs and Other Adaptive Adversaries". *NIPS*, pp. 1160–1168.

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

Dekel, O., A. Tewari, and R. Arora (2012). "Online Bandit Learning against an Adaptive Adversary: from Regret to Policy Regret". *ICML*. icml.cc / Omnipress.

Dwork, C. (2006). "Differential privacy". *ICALP*. Springer, pp. 1–12.

Dwork, C. and A. Roth (2013). The Algorithmic Foundations of Differential Privacy. *Foundations and Trends® in Theoretical Computer Science* **9**.3–4, 211–407.

Dwork, C., G. N. Rothblum, and S. Vadhan (2010). "Boosting and Differential Privacy". *Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science*. FOCS '10, pp. 51–60.

Ekman, P. et al. (2017). Learning to Match. *CoRR* **abs/1707.09678**.

Forbes (2012). *How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did*. https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/. Accessed: 2017-09-28.

Hossmann-Picu, A. et al. (2016). "Synergistic user $\longleftrightarrow$ context analytics". *ICT Innovations 2015*. Vol. 399. Springer, pp. 163–172.

Kasiviswanathan, S. P. et al. (2011). What can we learn privately? *SIAM Journal on Computing* **40**.3, 793–826.

Kifer, D. and A. Machanavajjhala (2014). Pufferfish: A framework for mathematical privacy definitions. *ACM Transactions on Database Systems (TODS)* **39**.1, 3.

Korolova, A. (2010). "Privacy Violations Using Microtargeted Ads: A Case Study". *ICDMW 2010, The 10th IEEE International Conference on Data Mining Workshops*, pp. 474–482.

Machanavajjhala, A. et al. (2007). *L*-diversity: Privacy beyond *k*-anonymity. *TKDD* **1**.1, 3.

McSherry, F. and K. Talwar (2007). "Mechanism Design via Differential Privacy". *48th IEEE Symposium on Foundations of Computer Science*. FOCS '07. Washington, DC, USA: IEEE Computer Society, pp. 94–103. ISBN: 0-7695-3010-9. DOI: 10.1109/FOCS.2007.41. URL: http://dx.doi.org/10.1109/FOCS.2007.41.

Merhav, N. et al. (2002). On sequential strategies for loss functions with memory. *IEEE Trans. Information Theory* **48**.7, 1947–1958.

Mishra, N. and A. Thakurta (2015). (Nearly) Optimal Differentially Private Stochastic Multi-Arm Bandits. *Proceedings of the 31th UAI*.

Ohm, P. (2009). Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. English. *UCLA Law Review, Vol. 57, p. 1701, 2010*.

Russo, D. and B. V. Roy (2016). An Information-Theoretic Analysis of Thompson Sampling. *Journal of Machine Learning Research* **17**.68, 1–30. URL: http://jmlr.org/papers/v17/14-087.html.

Schwab, K. (2016). *The Fourth Industrial Revolution*. World Economic Forum. ISBN: 9781944835002. URL: https://books.google.se/books?id=mQQwjwEACAAJ.

Statista (2017). *The Statistics Portal - Statistics and Studies from more than 18,000 Sources*. https://www.statista.com. Accessed: 2017-09-28.

Sweeney, L. (2002). k-Anonymity: A Model for Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* **10**.5, 557–570.

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 C. Dimitrakakis (2015). "Optimal Advertisement Strategies for Small and Big Companies". *AFRICOMM*. Vol. 171. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, pp. 94–98.

— (2016a). "Algorithms for Differentially Private Multi-Armed Bandits". *AAAI*. AAAI Press, pp. 2087–2093.

— (2016b). "Algorithms for Differentially Private Multi-Armed Bandits". *AAAI*. AAAI Press, pp. 2087–2093.

Tossou, A. C. Y., C. Dimitrakakis, and D. P. Dubhashi (2017). "Thompson Sampling for Stochastic Bandits with Graph Feedback". *AAAI*. AAAI Press, pp. 2660–2666.

Tossou, A. C. Y. and C. Dimitrakakis (2017). "Achieving Privacy in the Adversarial Multi-Armed Bandit". *AAAI*. AAAI Press, pp. 2653–2659.

Wikipedia (2017). *Multi-armed bandit — Wikipedia, The Free Encyclopedia*. [Online; accessed 2-October-2017]. URL: %5Curl%7Bhttps://en.wikipedia.org/w/index.php?title=Multi-armed_bandit&oldid=801364866%7D.