Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Achieving Privacy in the Adversarial Multi-Armed Bandit

Aristide C. Y. Tossou
Chalmers University of Technology
Gothenburg, Sweden
aristide@chalmers.se

Christos Dimitrakakis
University of Lille, France
Chalmers University of Technology, Sweden
Harvard University, USA
christos.dimitrakakis@gmail.com

Abstract

In this paper, we improve the previously best known regret bound to achieve ϵ-differential privacy in oblivious adversarial bandits from $\mathcal{O}(T^{2/3}/\epsilon)$ to $\mathcal{O}(\sqrt{T}\ln T/\epsilon)$. This is achieved by combining a Laplace Mechanism with EXP3. We show that though EXP3 is already differentially private, it leaks a linear amount of information in T. However, we can improve this privacy by relying on its intrinsic exponential mechanism for selecting actions. This allows us to reach $\mathcal{O}(\sqrt{\ln T})$-DP, with a regret of $\mathcal{O}(T^{2/3})$ that holds against an adaptive adversary, an improvement from the best known of $\mathcal{O}(T^{3/4})$. This is done by using an algorithm that run EXP3 in a mini-batch loop. Finally, we run experiments that clearly demonstrate the validity of our theoretical analysis.

1 Introduction

We consider multi-armed bandit problems in the adversarial setting whereby an agent selects one from a number of alternatives (called arms) at each round and receives a gain that depends on its choice. The agent's goal is to maximize its total gain over time. There are two main settings for the bandit problem. In the stochastic one, the gains of each arm are generated i.i.d by some unknown probability law. In the adversarial setting, which is the focus of this paper, the gains are generated adversarially. We are interested in finding algorithms with a total gain over T rounds not much smaller than that of an oracle with additional knowledge about the problem. In both settings, algorithms that achieve the optimal (problem-independent) regret bound of $\mathcal{O}(\sqrt{T})$ are known (Auer, Cesa-Bianchi, and Fischer 2002; Burnetas and Katehakis 1996; Pandey and Olston 2006; Thompson 1933; Auer et al. 2003; Auer 2002; Agrawal and Goyal 2012).

This problem is a model for many applications where there is a need for trading-off exploration and exploitation. This is so because, whenever we make a choice, we only observe the gain generated by that choice, and not the gains that we could have obtained otherwise. An example is clinical trials, where arms correspond to different treatments or tests, and the goal is to maximize 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 formalized similarly (Pandey and Olston 2006).

Privacy can be a serious issue in the bandit setting (c.f. (Jain, Kothari, and Thakurta 2012; Thakurta and Smith 2013; Mishra and Thakurta 2015; Zhao et al. 2014)). For example, in clinical trials, we may want to detect and publish results about the best drug without leaking sensitive information, such as the patient's health condition and genome. *Differential privacy* (Dwork 2006) formally bounds the amount of information that a third party can learn no matter their power or side information.

Differential privacy has been used before in the stochastic setting (Tossou and Dimitrakakis 2016; Mishra and Thakurta 2015; Jain, Kothari, and Thakurta 2012) where the authors obtain optimal algorithms up to logarithmic factors. In the adversarial setting, (Thakurta and Smith 2013) adapts an algorithm called *Follow The Approximate Leader* to make it private and obtain a regret bound of $\mathcal{O}(T^{2/3})$. In this work, we show that a number of simple algorithms can satisfy privacy guarantees, while achieving nearly optimal regret (up to logarithmic factors) that scales naturally with the level of privacy desired.

Our work is also of independent interest for non-private multi-armed bandit algorithms, as there are competitive with the current state of the art against switching-cost adversaries (where we recover the optimal bound). Finally, we provide rigorous empirical results against a variety of adversaries.

The following section gives the main background and notations. Section 3.1 describes meta-algorithms that perturb the gain sequence to achieve privacy, while Section 3.2 explains how to leverage the privacy inherent in the EXP3 algorithm by modifying the way gains are used. Section 4 compares our algorithms with EXP3 in a variety of settings. The full proofs of all our main results are in the full version.

2 Preliminaries

2.1 The Multi-Armed Bandit problem

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 gain $g_{I_t,t} \in [0,1]$. The agent only observes the gain of arm I_t, and not that of any other arms. The goal of this agent is to maximize its total gain

after T rounds, $\sum_{t=1}^{T} g_{I_t,t}$. A randomized bandit algorithm $\Lambda : (\mathcal{A} \times [0,1])^* \to \mathscr{D}(\mathcal{A})$ maps every arm-gain history to a distribution over the next arm to take.

The nature of the adversary, and specifically, how the gains are generated, determines the nature of the game. For the *stochastic* adversary (Thompson 1933; Auer, Cesa-Bianchi, and Fischer 2002), the gain obtained at round t is generated i.i.d from a distribution P_{I_t}. The more general *fully oblivious* adversary (Audibert and Bubeck 2010) generates the gains independently at round t but not necessarily identically from a distribution $P_{I_t,t}$. Finally, we have the *oblivious* adversary (Auer et al. 2003) whose only constraint is to generate the gain $g_{I_t,t}$ as a function of the current action I_t only, i.e. ignoring previous actions and gains.

While focusing on oblivious adversaries, we discovered that by targeting differential privacy we can also compete against 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 gains. 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 $m = 1$).

Regret. Relying on the cumulative gain of an agent to evaluate its performance can be misleading. Indeed, consider the case where an adversary gives a zero gain for all arms at every round. The cumulative gain 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 gain and the one obtained by some hypothetical agent, called *oracle*, with additional information or computational power. This gap is called the *regret*.

There are also 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 :

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

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

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

where the expectation is taken with respect to the random choices of both the agent and adversary. There are other oracles like the *shifting oracle* but those are out of scope of this paper.

EXP3. The Exponential-weight for Exploration and Exploitation (EXP3 (Auer et al. 2003)) algorithm achieves the optimal bound (up to logarithmic factors) of $\mathcal{O}(\sqrt{TK \ln K})$ for the weak regret (i.e. the expected regret compared to the *fixed oracle*) against an *oblivious adversary*. EXP3 simply maintains an estimate $\tilde{G}_{i,t}$ for the cumulative gain of arm i up to round t with $\tilde{G}_{i,t} = \sum_{s=1}^{t} \frac{g_{i,t}}{p_{i,t}} \mathbb{1}_{I_t=i}$ where

$$p_{i,t} = (1-\gamma)\frac{\exp\left(\gamma/K\tilde{G}_{i,t}\right)}{\sum_{i=1}^{K} \exp\left(\gamma/K\tilde{G}_{i,t}\right)} + \frac{\gamma}{K} \tag{2.1}$$

with γ a well defined constant.

Finally, EXP3 plays one action randomly according to the probability distribution $p_t = \{p_{1,t}, \dots p_{K,t}\}$ with $p_{i,t}$ as defined above.

2.2 Differential Privacy

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

Definition 2.1 (Single round (ϵ, δ)-differentially private bandit algorithm)**.** A randomized bandit algorithm Λ is (ϵ, δ)-differentially private at round t, if for all sequence $g_{1:t-1}$ and $g'_{1:t-1}$ that differs in at most one round, we have for any action subset $S \subseteq \mathcal{A}$:

$$\mathbb{P}_\Lambda(I_t \in S \mid g_{1:t-1}) \leq \delta + \mathbb{P}_\Lambda(I_t \in S \mid g'_{1:t-1})e^\epsilon, \tag{2.2}$$

where $\mathbb{P}_\Lambda$ denotes the probability distribution specified by the algorithm and $g_{1:t-1} = \{g_1, \dots g_{t-1}\}$ with g_s the gains of all arms at round s. When $\delta = 0$, the algorithm is said to be ϵ-*differential 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 2.1 means that the output of the bandit algorithm at round t is almost insensible to any single change in the gains sequence. This implies that whether or not we remove a single round, replace the gains, the bandit algorithm will still play almost the same action. Assuming the gains at round t are 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) specifies how much the output is affected by a single user.

We would like Definition 2.1 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. 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 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 ((ϵ, δ)-differentially private bandit algorithm)**.** A randomized bandit algorithm Λ is (ϵ, δ)-differentially private up to round t, if for all $g_{1:t-1}$ and $g'_{1:t-1}$ that differs in at most one round, we have for any action subset $S \subseteq \mathcal{A}^t$:

$$\mathbb{P}_\Lambda(I_{1:t} \in S \mid g_{1:t-1}) \leq \delta + \mathbb{P}_\Lambda(I_{1:t} \in S \mid g'_{1:t-1})e^\epsilon, \tag{2.3}$$

where $\mathbb{P}_\Lambda$ and g are as defined in Definition 2.1.

Most of the time, we will refer to Definition 2.2 and whenever we need to use Definition 2.1, this will be made explicit.

The simplest mechanism to achieve differential privacy for a function is to add Laplace noise of scale proportional to its sensitivity. The sensitivity is the maximum amount by which the value of the function can change if we change a single element in the inputs sequence. For example, if the input is a stream of numbers in $[0, 1]$ and the function their sum, we can add Laplace noise of scale $\frac{1}{\epsilon}$ to each number and achieve ϵ-differential privacy with an error of $\mathcal{O}(\sqrt{T}/\epsilon)$ in the sum. However, (Chan, Shi, and Song 2010) introduced *Hybrid Mechanism*, which achieves ϵ-differential privacy with only poly-logarithmic error (with respect to the true sum). The idea is to group the stream of numbers in a binary tree and only add a Laplace noise at the nodes of the tree.

As demonstrated above, the main challenge with differential privacy is thus to trade-off optimally privacy and utility.

Notation. In this paper, i will be used as an index for an arbitrary arm in $[1, K]$, while k will be used to indicate an optimal arm and I_t is the arm played by an agent at round t. We use $g_{i,t}$ to indicate the gain of the i-th arm at round t. $\mathcal{R}_\Lambda(T)$ is the regret of the algorithm Λ after T rounds. The index and T are dropped when it is clear from the context. Unless otherwise specified, the regret is defined for oblivious adversaries against the fixed oracle. We use "$x \sim P$" to denote that x is generated from distribution P. $\mathcal{Lap}(\lambda)$ is used to denote the Laplace distribution with scale λ while $\mathcal{Bern}(p)$ denotes the Bernoulli distribution with parameter p.

3 Algorithms and Analysis

3.1 *DP-*Λ*-Lap*: Differential privacy through additional noise

We start by showing that the obvious technique to achieve a given ϵ-differential privacy in adversarial bandits already beat the state-of-the art. The main idea is to use any base bandit algorithm Λ as input and add a Laplace noise of scale $\frac{1}{\epsilon}$ to each gain before Λ observes it. This technique gives ϵ-DP differential privacy as the gains are bounded in $[0, 1]$ and the noises are added i.i.d at each round.

However, bandits algorithms require bounded gains while the noisy gains are not. The trick is to ignore rounds where the noisy gains fall outside an interval of the form $[-b, b+1]$. We pick the threshold b such that, with high probability, the noisy gains will be inside the interval $[-b, b + 1]$. More precisely, b can be chosen such that with high probability, the number of rounds ignored is lower than the upper bound R_Λ on the regret of Λ. Given that in the standard bandit problem, the gains are bounded in $[0, 1]$, the gains at accepted rounds are rescaled back to $[0, 1]$.

Theorem 3.2 shows that all these operations still preserve ϵ-DP while Theorem 3.1 demonstrates that the upper bound on the expected regret of *DP-*Λ*-Lap* adds some small additional terms to R_Λ. To illustrate how small those additional terms are, we instantiate *DP-*Λ*-Lap* with the *EXP3* algorithm. This leads to a mechanism called *DP-EXP3-Lap* described in Algorithm 1. With a carefully chosen threshold b, corollary 3.1 implies that the additional terms are such that the expected regret of *DP-EXP3-Lap* is $\mathcal{O}(\sqrt{T}\ln T/\epsilon)$ which is optimal in T up to some logarithmic factors. This result is a significant improvement over the best known bound so far of $\mathcal{O}(T^{2/3}/\epsilon)$ from (Thakurta and Smith 2013) and solves simultaneously the challenge (whether or not one can get ϵ-DP mechanism with optimal regret) posed by the authors.

Algorithm 1 *DP-EXP3-Lap*

Let $\tilde{G}_i = 0$ for all arms and $b = \frac{\ln T}{\epsilon}$, $\gamma = \sqrt{\frac{K \ln K}{(e-1)T}}$
for each round $t = 1, \cdots, T$ **do**
  Compute the probability distribution p over the arms with $p = (p_{1,t}, \cdots p_{K,t})$ and $p_{i,t}$ as in eq (2.1).
  Draw an arm I_t from the probability distribution p.
  Receive the reward $g_{I_t,t}$
  Let the noisy gain be $g'_{I_t,t} = g_{I_t,t} + \mathcal{N}_{I_t,t}$ with $\mathcal{N}_{I_t,t} \sim \mathcal{Lap}(\frac{1}{\epsilon})$
  if $g'_{I_t,t} \in [-b, b + 1]$ **then**
    Scale $g'_{I_t,t}$ to $[0, 1]$
    Update the estimated cumulative gain of arm I_t: $\tilde{G}_{I_t} = \tilde{G}_{I_t} + \frac{g'_{I_t,t}}{p_{I_t,t}}$
  end if
end for

Theorem 3.1. *If* DP-Λ-Lap *is run with input a base bandit algorithm* Λ*, the noisy reward* $g'_{I_t,t}$ *of the true reward* $g_{I_t,t}$ *set to* $g'_{I_t,t} = g_{I_t,t} + \mathcal{N}_{I_t,t}$ *with* $\mathcal{N}_{I_t,t} \sim \mathcal{Lap}(\frac{1}{\epsilon})$*, the acceptance interval set to* $[-b, b + 1]$ *with the scaling of the rewards* g'_{I_t} *outside* $[0, 1]$ *done using* $g'_{I_t,t} = \frac{g'_{I_t,t}+b}{2b+1}$*; then the regret* $R_{\text{DP-}\Lambda\text{-Lap}}$ *of* DP-Λ-Lap *satisfies:*

$$\mathbb{E}\, R_{\text{DP-}\Lambda\text{-Lap}} \leq \mathbb{E}\, R_\Lambda^{scaled} + 2TK\exp(-\epsilon b) + \frac{\sqrt{32T}}{\epsilon} \tag{3.1}$$

where R_Λ^{scaled} *is the upper bound on the regret of* Λ *when the rewards are scaled from* $[-b, b + 1]$ *to* $[0, 1]$

Proof Sketch. We observed that *DP-*Λ*-Lap* is an instance of Λ run with the noisy rewards g' instead of g. This means R_Λ^{scaled} is an upper bound of the regret L on g'. Then, we derived a lower bound on L showing how close it is to $R_{\textit{DP-}\Lambda\textit{-Lap}}$. This allows us to conclude. □

Corollary 3.1. *If* DP-Λ-Lap *is run with* EXP3 *as its base algorithm and* $b = \frac{\ln T}{\epsilon}$*, then its expected regret* $\mathbb{E}\, R_{\text{DP-EXP3-Lap}}$ *satisfies*

$$\mathbb{E}\, R_{\text{DP-EXP3-Lap}} \leq \frac{4\ln T}{\epsilon}\sqrt{(e-1)TK\ln K} + 2K + \frac{\sqrt{32T}}{\epsilon}$$

Proof. The proof comes by combining the regret of *EXP3* (Auer et al. 2003) with Theorem 3.1 □

Theorem 3.2. DP-Λ-Lap *is ϵ-differentially private up to round T.*

Proof Sketch. Combining the privacy of Laplace Mechanism with the parallel composition (McSherry 2009) and post-processing theorems (Dwork and Roth 2013) concludes the proof. □

3.2 Leveraging the inherent privacy of EXP3

On the differential privacy of *EXP3* (Dwork and Roth 2013) shows that a variation of EXP3 for the full-information setting (where the agent observes the gain of all arms at any round regardless of what he played) is already differentially private. Their results imply that one can achieve the optimal regret with only a sub-logarithmic privacy loss ($\mathcal{O}(\sqrt{128 \log T})$) after T rounds.

We start this section by showing a similar result for EXP3 in Theorem 3.3. Indeed, we show that EXP3 is already differentially private but with a per-round privacy loss of 2. [1] Our results imply that EXP3 can achieve the optimal regret albeit with a linear privacy loss of $\mathcal{O}(2T)$-DP after T rounds. This is a huge gap compared with the full-information setting and underlines the significance of our result in section 3.1 where we describe a concrete algorithm demonstrating that the optimal regret can be achieved with only a logarithmic privacy loss after T rounds.

Theorem 3.3. *The* EXP3 *algorithm is:*

$$\min\left\{2T, T \cdot \ln\frac{K(1-\gamma)+\gamma}{\gamma}, 2(1-\gamma)T + 2\sqrt{\frac{2\ln T}{T}}\right\}$$

differentially private up to round T.

In practice, we also want EXP3 *to have a sub-linear regret. This implies that $\gamma << 1$ and* EXP3 *is simply $2T$-DP over T rounds.*

Proof Sketch. The first two terms in the theorem come from the observation that EXP3 is a combination of two mechanisms: the Exponential Mechanism (McSherry and Talwar 2007) and a randomized response. The last term comes from the observation that with probability γ we enjoy a perfect 0-DP. Then, we use Chernoff to bound with high probability the number of times we suffer a non-zero privacy loss. □

We will now show that the privacy of EXP3 itself may be improved without any additional noise, and with only a moderate impact on the regret.

On the privacy of a *EXP3* wrapper algorithm The previous paragraph leads to the conclusion that it is impossible to obtain a sub-linear privacy loss with a sub-linear regret while using the original *EXP3*. Here, we will prove that an existing technique is already achieving this goal. The algorithm which we called *EXP3*$_\tau$ is from (Dekel, Tewari, and Arora 2012). It groups the rounds into disjoint intervals of fixed size τ where the j'th interval starts on round $(j-1)\tau+1$ and ends on round $j\tau$. At the beginning of interval j, *EXP3*$_\tau$ receives an action from *EXP3* and plays it for τ rounds. During that time, *EXP3* does not observe any feedback. At the end of the interval, *EXP3*$_\tau$ feeds *EXP3* with a single gain, the average gain received during the interval.

Theorem 3.4 borrowed from (Dekel, Tewari, and Arora 2012) specifies the upper bound on the regret *EXP3*$_\tau$. It is remarkable that this bound holds against the *m-memory bounded adaptive adversary*. While in theorem 3.5, we show the privacy loss enjoyed by this algorithm, one gets a better intuition of how good those results are from corollary 3.2 and 3.3. Indeed, we can observe that *EXP3*$_\tau$ achieves a sub-logarithmic privacy loss of $\mathcal{O}(\sqrt{\ln T})$ with a regret of $\mathcal{O}(T^{2/3})$ against a special case of the *m-memory bounded adaptive adversary* called the *switching costs adversary* for which $m = 1$. This is the optimal regret bound (in the sense that there is a matching lower bound (Dekel et al. 2014)). This means that in some sense we are getting privacy for free against this adversary.

Theorem 3.4 (Regret of *EXP3*$_\tau$ (Dekel, Tewari, and Arora 2012))**.** *The expected regret of* EXP3$_\tau$ *is upper bounded by:*

$$\sqrt{7T\tau K \ln K} + \frac{Tm}{\tau} + \tau$$

against the m-memory bounded adaptive adversary *for any $m < \tau$.*

Theorem 3.5 (Privacy loss of *EXP3*$_\tau$)**.** EXP3$_\tau$ *is* $\left(\frac{4T}{\tau^3} + \sqrt{8\ln(1/\delta')\frac{T}{\tau^3}}, \delta'\right)$*-DP up to round T.*

Proof. The sensitivity of each gain is now $\frac{1}{\tau}$ as we are using the average. Combined with theorem (3.3), it means the per-round privacy loss is $2\frac{T}{\tau}$. Given that *EXP3* only observes $\frac{T}{\tau}$ rounds, using the advanced composition theorem (Dwork, Rothblum, and Vadhan 2010) (Theorem III.3) concludes the final privacy loss over T rounds. □

Corollary 3.2. EXP3$_\tau$ *run with $\tau = (7K\log K)^{-1/3}T^{1/3}$ is (ϵ, δ') differentially private up to round T with $\delta' = T^{-2}$, $\epsilon = 28K\ln K + \sqrt{112K\ln K \ln T}$. Its expected regret against the* switching costs adversary *is upper bounded by* $2(7K\ln K)^{1/3}T^{2/3} + (7K\log K)^{-1/3}T^{1/3}$.

Proof. The proof is immediate by replacing τ and δ' in Theorem 3.4 and 3.5 and the fact that for the *switching costs adversary*, $m = 1$. □

Corollary 3.3. EXP3$_\tau$ *run with* $\tau = \left(\frac{4T\epsilon + 2T\ln\frac{1}{\delta}}{\epsilon^2}\right)^{1/3}$ *is (ϵ, δ) differentially private and its expected regret against the* switching costs adversary *is upper bounded by:*

$$\mathcal{O}\left(T^{2/3}\sqrt{K\ln K}\left(\frac{\sqrt{\ln\frac{1}{\delta}}}{\epsilon}\right)^{1/3}\right)$$

4 Experiments

We tested *DP-EXP3-Lap*, EXP3$_\tau$ together with the non-private *EXP3* against a few different adversaries. The privacy parameter ϵ of *DP-EXP3-Lap* is set as defined in corollary 3.2. This is done so that the regret of *DP-EXP3-Lap* and

[1] Assuming we want a sub-linear regret. See Theorem 3.3

EXP3$_\tau$ are compared with the same privacy level. All the other parameters of *DP-EXP3-Lap* are taken as defined in corollary 3.1 while the parameters of EXP3$_\tau$ are taken as defined in corollary 3.2.

For all experiments, the horizon is $T = 2^{18}$ and the number of arms is $K = 4$. We performed 720 independent trials and reported the *median-of-means* estimator[2] of the cumulative regret. It partitions the trials into a_0 equal groups and return the median of the sample means of each group. Proposition 4.1 is a well known result (also in (Hsu and Sabato 2013; Lerasle and Oliveira 2011)) giving the accuracy of this estimator. Its convergence is $\mathcal{O}(\sigma/\sqrt{N})$, with exponential probability tails, even though the random variable x may have heavy-tails. In comparison, the empirical mean can not provide such guarantee for any $\sigma > 0$ and confidence in $[0, 1/(2e)]$ (Catoni 2012).

Proposition 4.1. Let x be a random variable with mean μ and variance $\sigma^2 < \infty$. Assume that we have N independent sample of x and let $\hat{\mu}$ be the *median-of-means* computed using a_0 groups. With probability at least $1 - e^{-a_0/4.5}$, $\hat{\mu}$ satisfies $|\hat{\mu} - \mu| \leq \sigma\sqrt{6a_0/N}$.

We set the number of groups to $a_0 = 24$, so that the confidence interval holds w.p. at least 0.995.

We also reported the deviation of each algorithm using the Gini's Mean Difference (GMD hereafter) (Gini and Pearson 1912). GMD computes the deviation as $\sum_{j=1}^{N}(2j - N - 1)x_{(j)}$ with $x_{(j)}$ the j-th order statistics of the sample (that is $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(N)}$). As shown in (Yitzhaki and others 2003; David 1968), the GMD provides a superior approximation of the true deviation than the standard one. To account for the fact that the cumulative regret of our algorithms might not follow a symmetric distribution, we computed the GMD separately for the values above and below the *median-of-means*.

At round t, we computed the cumulative regret against the fixed oracle who plays the best arm assuming that the end of the game is at t. The oracle uses the actual sequence of gains to decide his best arm. For a given trial, we make sure that all algorithms are playing the same game by generating the gains for all possible pair of round-arm before the game starts.

Deterministic adversary. As shown by (Audibert and Bubeck 2010), the expected regret of any agent against an oblivious adversary can not be worse than that against the worst case deterministic adversary. In this experiment, arm 2 is the best and gives 1 for every even round. To trick the players into picking the wrong arms, the first arm always gives 0.38 whereas the third gives 1 for every round multiple of 3. The remaining arms always give 0. As shown by the figure, this simple adversary is already powerful enough to make the algorithms attain their upper bound.

[2]Used heavily in the streaming literature (Alon, Matias, and Szegedy 1996)

Stochastic adversary This adversary draws the gains of the first arm i.i.d from $\mathcal{Bern}(0.55)$ whereas all other gains are drawn i.i.d from $\mathcal{Bern}(0.5)$.

Fully oblivious adversary. For the best arm k, it first draws a number p uniformly in $[0.5, 0.5 + 2 \cdot \varepsilon]$ and generates the gain $g_{k,t} \sim \mathcal{Bern}(p)$. For all other arms, p is drawn from $[0.5-\varepsilon, 0.5+\varepsilon]$. This process is repeated at every round. In our experiments, $\varepsilon = 0.05$

An oblivious adversary. This adversary is identical to the fully oblivious one for every round multiple of 200. Between two multiples of 200 the last gain of the arm is given.

The Switching costs adversary This adversary (defined at Figure 1 in (Dekel et al. 2014)) defines a stochastic processes (including simple Gaussian random walk as special case) for generating the gains. It was used to prove that any algorithm against this adversary must incur a regret of $\mathcal{O}(T^{2/3})$.

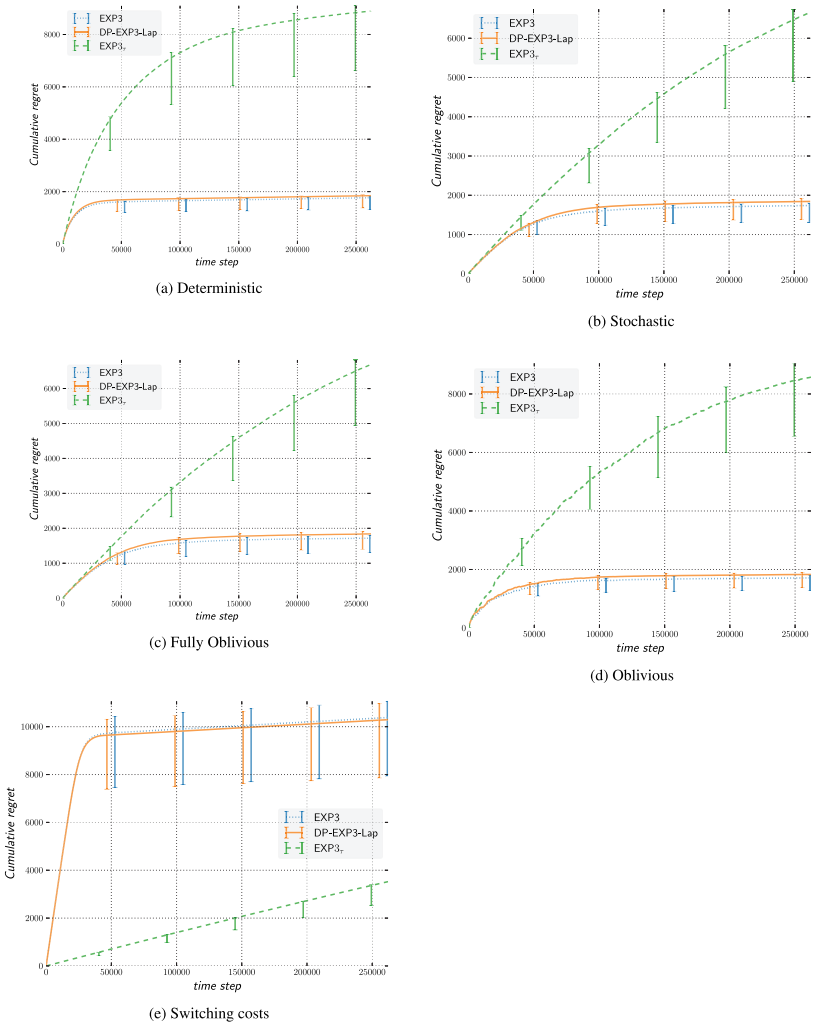
Discussion Figure 1 shows our results against a variety of adversaries, with respect to a *fixed oracle*. Overall, the performance (in term of regret) of *DP-EXP3-Lap* is very competitive against that of EXP3 while providing a significant better privacy. This means that *DP-EXP3-Lap* allows us to get privacy for free in the bandit setting against an adversary not more powerful than the oblivious one.

The performance of EXP3$_\tau$ is worse than that of *DP-EXP3-Lap* against an oblivious adversary or one less powerful. However, the situation is completely reversed against the more powerful switching cost adversary. In that setting, EXP3$_\tau$ outperforms both EXP3 and *DP-EXP3-Lap* confirming the theoretical analysis. We can see EXP3$_\tau$ as the algorithm providing us privacy for free against switching cost adversary and adaptive m-bounded memory one in general.

5 Conclusion

We have provided the first results on differentially private adversarial multi-armed bandits, which are optimal up to logarithmic factors. One open question is how differential privacy affects regret in the full reinforcement learning problem. At this point in time, the only known results in the MDP setting obtain differentially private algorithms for Monte Carlo policy evaluation (Balle, Gomrokchi, and Precup 2016). While this implies that it is possible to obtain policy iteration algorithms, it is unclear how to extend this to the full online reinforcement learning problem.

Acknowledgements. This research was supported by the SNSF grants "Adaptive control with approximate Bayesian computation and differential privacy" and "Swiss Sense Synergy", by the Marie Curie Actions (REA 608743), the Future of Life Institute "Mechanism Design for AI Architectures" and the CNRS Specific Action on Security.



(a) Deterministic

(b) Stochastic

(c) Fully Oblivious

(d) Oblivious

(e) Switching costs

Figure 1: Regret and Error bar against five different adversaries, with respect to the fixed oracle

References

Agrawal, S., and Goyal, N. 2012. Analysis of thompson sampling for the multi-armed bandit problem. In *COLT 2012*.

Alon, N.; Matias, Y.; and Szegedy, M. 1996. The space complexity of approximating the frequency moments. In *28th STOC*, 20–29. ACM.

Audibert, J.-Y., and Bubeck, S. 2010. Regret bounds and minimax policies under partial monitoring. *J. Mach. Learn. Res.* 11:2785–2836.

Auer, P.; Cesa-Bianchi, N.; Freund, Y.; and Schapire, R. E. 2003. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.* 32(1):48–77.

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

Auer, P. 2002. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research* 3:397–422.

Balle, B.; Gomrokchi, M.; and Precup, D. 2016. Differentially private policy evaluation. In *ICML 2016*.

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

Catoni, O. 2012. Challenging the empirical mean and empirical variance: A deviation study. *Annales de l'I.H.P. Probabilités et statistiques* 48(4):1148–1185.

Cesa-Bianchi, N.; Dekel, O.; and Shamir, O. 2013. Online learning with switching costs and other adaptive adversaries. In *NIPS*, 1160–1168.

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

David, H. 1968. Miscellanea: Gini's mean difference rediscovered. *Biometrika* 55(3):573–575.

Dekel, O.; Ding, J.; Koren, T.; and Peres, Y. 2014. Bandits with switching costs: T2/3 regret. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, STOC '14, 459–467. New York, NY, USA: ACM.

Dekel, O.; Tewari, A.; and Arora, R. 2012. Online bandit learning against an adaptive adversary: from regret to policy regret. In *ICML*. icml.cc / Omnipress.

Dwork, C., and Roth, A. 2013. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science* 9(3–4):211–407.

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.

Gini, C., and Pearson, K. 1912. *Variabilità e mutabilità: contributo allo studio delle distribuzioni e delle relazioni statistiche. Fascicolo 1*. tipografia di Paolo Cuppini.

Hsu, D., and Sabato, S. 2013. Loss minimization and parameter estimation with heavy tails. *arXiv preprint arXiv:1307.1827*.

Jain, P.; Kothari, P.; and Thakurta, A. 2012. Differentially private online learning. In Mannor, S.; Srebro, N.; and Williamson, R. C., eds., *COLT 2012*, volume 23, 24.1–24.34.

Lerasle, M., and Oliveira, R. I. 2011. Robust empirical mean estimators. *arXiv preprint arXiv:1112.3914*.

McSherry, F., and Talwar, K. 2007. Mechanism design via differential privacy. In *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '07, 94–103. Washington, DC, USA: IEEE Computer Society.

McSherry, F. D. 2009. Privacy integrated queries: An extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data*, SIGMOD '09, 19–30. New York, NY, USA: ACM.

Merhav, N.; Ordentlich, E.; Seroussi, G.; and Weinberger, M. J. 2002. On sequential strategies for loss functions with memory. *IEEE Trans. Information Theory* 48(7):1947–1958.

Mishra, N., and Thakurta, A. 2015. (nearly) optimal differentially private stochastic multi-arm bandits. *Proceedings of the 31th UAI*.

Pandey, S., and Olston, C. 2006. Handling advertisements of unknown quality in search advertising. In Schölkopf, B.; Platt, J. C.; and Hoffman, T., eds., *Twentieth NIPS*, 1065–1072.

Thakurta, A. G., and Smith, A. D. 2013. (nearly) optimal algorithms for private online learning in full-information and bandit settings. In *NIPS*, 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. 2016. Algorithms for differentially private multi-armed bandits. In *AAAI*, 2087–2093. AAAI Press.

Yitzhaki, S., et al. 2003. Gini's mean difference: A superior measure of variability for non-normal distributions. *Metron* 61(2):285–316.

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.