The Price of Differential Privacy for Online Learning

Naman Agarwal [1] **Karan Singh** [1]

Abstract

We design differentially private algorithms for the problem of online linear optimization in the full information and bandit settings with optimal $\tilde{O}(\sqrt{T})$[1] regret bounds. In the full-information setting, our results demonstrate that ε-differential privacy may be ensured for free – in particular, the regret bounds scale as $O(\sqrt{T}) + \tilde{O}\left(\frac{1}{\varepsilon}\right)$. For bandit linear optimization, and as a special case, for non-stochastic multi-armed bandits, the proposed algorithm achieves a regret of $\tilde{O}\left(\frac{1}{\varepsilon}\sqrt{T}\right)$, while the previously known best regret bound was $\tilde{O}\left(\frac{1}{\varepsilon}T^{\frac{2}{3}}\right)$.

1. Introduction

In the paradigm of online learning, a learning algorithm makes a sequence of predictions given the (possibly incomplete) knowledge of the correct answers for the past queries. In contrast to statistical learning, online learning algorithms typically offer distribution-free guarantees. Consequently, online learning algorithms are well suited to dynamic and adversarial environments, where real-time learning from changing data is essential making them ubiquitous in practical applications such as servicing search advertisements. In these settings often these algorithms interact with sensitive user data, making privacy a natural concern for these algorithms. A natural notion of privacy in such settings is differential privacy (Dwork et al., 2006) which ensures that the outputs of an algorithm are indistinguishable in the case when a user's data is present as opposed to when it is absent in a dataset.

In this paper, we design differentially private algorithms for online linear optimization with near-optimal regret, both in the full information and partial information (bandit) settings. This result improves the known best regret bounds for a number of important online learning problems – including *prediction from expert advice* and *non-stochastic multi-armed bandits*.

1.1. Full-Information Setting: Privacy for Free

For the full-information setting where the algorithm gets to see the complete loss vector every round, we design ε-differentially private algorithms with regret bounds that scale as $O\left(\sqrt{T}\right) + \tilde{O}\left(\frac{1}{\varepsilon}\right)$ (Theorem 3.1), partially resolving an open question to improve the previously best known bound of $O\left(\frac{1}{\varepsilon}\sqrt{T}\right)$ posed in (Smith & Thakurta, 2013). A decomposition of the bound on the regret bound of this form implies that when $\varepsilon \geq \frac{1}{\sqrt{T}}$, the regret incurred by the differentially private algorithm matches the optimal regret in the non-private setting, i.e. differential privacy is *free*. Moreover even when $\varepsilon \leq \frac{1}{\sqrt{T}}$, our results guarantee a sub-constant regret per round in contrast to the vacuous constant regret per round guaranteed by existing results.

Concretely, consider the case of online linear optimization over the cube, with unit l_∞-norm-bounded loss vectors. In this setting, (Smith & Thakurta, 2013) achieves a regret bound of $O(\frac{1}{\varepsilon}\sqrt{NT})$, which is meaningful only if $T \geq \frac{N}{\varepsilon^2}$. Our theorems imply a regret bound of $\tilde{O}(\sqrt{NT} + \frac{N}{\varepsilon})$. This is an improvement on the previous bound regardless of the value of ε. Furthermore, when T is between $\frac{N}{\varepsilon}$ and $\frac{N}{\varepsilon^2}$, the previous bounds are vacuous whereas our results are still meaningful. Note that the above arguments show an improvement over existing results even for moderate value of ε. Indeed, when ε is very small, the magnitude of improvements are more pronounced.

Beyond the separation between T and ε, the key point of our analysis is that we obtain bounds for general regularization based algorithms which adapt to the geometry of the underlying problem optimally, unlike the previous algorithms (Smith & Thakurta, 2013) which utilizes euclidean regularization. This allows our results to get rid of a polynomial dependence on N (in the $\sqrt{T}$ term) in some cases. Online linear optimization over the sphere and prediction with expert advice are notable examples.

[*]Equal contribution [1]Computer Science, Princeton University, Princeton, NJ, USA. Correspondence to: Naman Agarwal <namana@cs.princeton.edu>, Karan Singh <karans@cs.princeton.edu>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

[1]Here the $\tilde{O}(\cdot)$ notation hides $polylog(T)$ factors.

We summarize our results in Table 1.1.

1.2. Bandits: Reduction to the Non-private Setting

In the partial-information (bandit) setting, the online learning algorithm only gets to observe the loss of the prediction it prescribed. We outline a reduction technique that translates a non-private bandit algorithm to a differentially private bandit algorithm, while retaining the $\tilde{O}(\sqrt{T})$ dependency of the regret bound on the number of rounds of play (Theorem 4.5). This allows us to derive the first ε-differentially private algorithm for bandit linear optimization achieving $\tilde{O}(\sqrt{T})$ regret, using the algorithm for the non-private setting from (Abernethy et al., 2012). This answers a question from (Smith & Thakurta, 2013) asking if $\tilde{O}(\sqrt{T})$ regret is attainable for differentially private linear bandits .

An important case of the general bandit linear optimization framework is the *non-stochastic multi-armed bandits* problem(Bubeck et al., 2012b), with applications for website optimization, personalized medicine, advertisement placement and recommendation systems. Here, we propose an ε-differentially private algorithm which enjoys a regret of $\tilde{O}(\frac{1}{\varepsilon}\sqrt{NT\log N})$ (Theorem 4.1), improving on the previously best attainable regret of $\tilde{O}(\frac{1}{\varepsilon}NT^{\frac{2}{3}})$(Smith & Thakurta, 2013).

We summarize our results in Table 1.2.

1.3. Related Work

The problem of differentially private online learning was first considered in (Dwork et al., 2010), albeit guaranteeing privacy in a weaker setting – ensuring the privacy of the individual entries of the loss vectors. (Dwork et al., 2010) also introduced the tree-based aggregation scheme for releasing the cumulative sums of vectors in a differentially private manner, while ensuring that the total amount of noise added for each cumulative sum is only poly-logarithmically dependent on the number of vectors. The stronger notion of privacy protecting entire loss vectors was first studied in (Jain et al., 2012), where gradient-based algorithms were proposed that achieve (ε, δ)-differntial privacy and regret bounds of $\tilde{O}\left(\frac{1}{\varepsilon}\sqrt{T}\log\frac{1}{\delta}\right)$. (Smith & Thakurta, 2013) proposed a modification of Follow-the-Approximate-Leader template to achieve $\tilde{O}\left(\frac{1}{\varepsilon}\log^{2.5}T\right)$ regret for strongly convex loss functions, implying a regret bound of $\tilde{O}\left(\frac{1}{\varepsilon}\sqrt{T}\right)$ for general convex functions. In addition, they also demonstrated that under bandit feedback, it is possible to obtain regret bounds that scale as $\tilde{O}\left(\frac{1}{\varepsilon}T^{\frac{2}{3}}\right)$. (Dwork et al., 2014a; Jain & Thakurta, 2014) proved that in the special case of *prediction with expert advice* setting, it is possible to achieve a regret of $O\left(\frac{1}{\varepsilon}\sqrt{T\log N}\right)$. While most algorithms for differentially private online learning are based on the regularization template, (Dwork et al., 2014b) used a perturbation-based algorithm to guarantee (ε, δ)-differential privacy for the problem of online PCA. (Tossou & Dimitrakakis, 2016) showed that it is possible to design ε-differentially private algorithms for the stochastic multi-armed bandit problem with a separation of ε, T for the regret bound. Recently, an independent work due to (Tossou & Dimitrakakis, 2017), which we were made aware of after the first manuscript, also demonstrated a $\tilde{O}\left(\frac{1}{\varepsilon}\sqrt{T}\right)$ regret bound in the *non-stochastic multi-armed bandits* setting. We match their results (Theorem 4.1), as well as provide a generalization to arbitrary convex sets (Theorem 4.5).

1.4. Overview of Our Techniques

Full Information Setting: We consider the two well known paradigms for online learning, *Folllow-the-Regularized-Leader (FTRL)* and *Folllow-the-Perturbed-Leader (FTPL)*. In both cases, we ensure differential privacy by restricting the mode of access to the inputs (the loss vectors). In particular, the algorithm can only retrieve estimates of the loss vectors released by a tree based aggregation protocol (Algorithm 2) which is a slight modification of the protocol used in (Jain et al., 2012; Smith & Thakurta, 2013). We outline a tighter analysis of the regret minimization framework by crucially observing that in case of linear losses, the expected regret of an algorithm that injects identically (though not necessarily independently) distributed noise per step is the same as one that injects a single copy of the noise at the very start of the algorithm.

The regret analysis of Follow-the-Leader based algorithm involves two components, a *bias* term due to the regularization and a *stability* term which bounds the change in the output of the algorithm per step. In the analysis due to (Smith & Thakurta, 2013), the stability term is affected by the variance of the noise as it changes from step to step. However in our analysis, since we treat the noise to have been sampled just once, the stability analysis does not factor in the variance and the magnitude of the noise essentially appears as an additive term in the bias.

Bandit Feedback: In the bandit feedback setting, we show a general reduction that takes a non-private algorithm and outputs a private algorithm (Algorithm 4). Our key observation here (presented as Lemma 4.3) is that on linear functions, in expectation the regret of an algorithm on a noisy sequence of loss vectors is the same as its regret on the original loss sequence as long as noise is zero mean. We now bound the regret on the noisy sequence by conditioning out the case when the noise can be large and using exploration techniques from (Bubeck et al., 2012a) and (Abernethy et al., 2008).

Function Class (N Dimensions)	Previous Best Known Regret	Our Regret Bound	Best Non-private Regret
Prediction with Expert Advice	$\tilde{O}\left(\frac{\sqrt{T\log N}}{\varepsilon}\right)$	$O\left(\sqrt{T\log N}+\frac{N\log N\log^2 T}{\varepsilon}\right)$	$O(\sqrt{T\log N})$
Online linear optimization over the Sphere	$\tilde{O}\left(\frac{\sqrt{NT}}{\varepsilon}\right)$	$O\left(\sqrt{T}+\frac{N\log^2 T}{\varepsilon}\right)$	$O(\sqrt{T})$
Online linear optimization over the Cube	$\tilde{O}\left(\frac{\sqrt{NT}}{\varepsilon}\right)$	$O\left(\sqrt{NT}+\frac{N\log^2 T}{\varepsilon}\right)$	$O(\sqrt{NT})$
Online Linear Optimization	$\tilde{O}\left(\frac{\sqrt{T}}{\varepsilon}\right)$	$O\left(\sqrt{T}+\frac{\log^2 T}{\varepsilon}\right)$	$O(\sqrt{T})$

Table 1. Summary of our results in the full-information setting. In the last row we suppress the constants depending upon $\mathcal{X}, \mathcal{Y}$.

Function Class (N Dimensions)	Previous Best Known Regret	Our Regret Bound	Best Non-private Regret
Bandit Linear Optimization	$\tilde{O}\left(\frac{T^{\frac{2}{3}}}{\varepsilon}\right)$	$\tilde{O}\left(\frac{\sqrt{T}}{\varepsilon}\right)$	$O(\sqrt{T})$
Non-stochastic Mult-armed Bandits	$\tilde{O}\left(\frac{NT^{\frac{2}{3}}}{\varepsilon}\right)$	$\tilde{O}\left(\frac{\sqrt{TN\log N}}{\varepsilon}\right)$	$O(\sqrt{NT})$

Table 2. Summary of our results in the bandit setting. In the first row we suppress the specific constants depending upon $\mathcal{X}, \mathcal{Y}$.

2. Model and Preliminaries

This section introduces the model of online (linear) learning, the distinction between full and partial feedback scenarios, and the notion of differential privacy in this model.

Full-Information Setting: *Online linear optimization* (Hazan et al., 2016; Shalev-Shwartz, 2011) involves repeated decision making over T rounds of play. At the beginning of every round (say round t), the algorithm chooses a point in $x_t \in \mathcal{X}$, where $\mathcal{X} \subseteq \mathbb{R}^N$ is a (compact) convex set. Subsequently, it observes the loss $l_t \in \mathcal{Y} \subseteq \mathbb{R}^N$ and suffers a loss of $\langle l_t, x_t\rangle$. The success of such an algorithm, across T rounds of play, is measured though **regret**, which is defined as

$$\text{Regret} = \mathbb{E}\left[\sum_{t=1}^{T}\langle l_t, x_t\rangle - \min_{x\in\mathcal{K}}\sum_{t=1}^{T}\langle l_t, x\rangle\right]$$

where the expectation is over the randomness of the algorithm. In particular, achieving a sub-linear regret ($o(T)$) corresponds to doing almost as good (averaging across T rounds) as the fixed decision with the least loss in hindsight. In the non-private setting, a number of algorithms have been devised to achieve $O(\sqrt{T})$ regret, with additional dependencies on other parameters dependent on the properties of the specific decision set $\mathcal{X}$ and loss set $\mathcal{Y}$. (See (Hazan et al., 2016) for a survey of results.)

Following are three important instantiations of the above framework.

- *Prediction with Expert Advice*: Here the underlying decision set is the simplex $\mathcal{X} = \Delta_N = \{x \in \mathbb{R}^n : x_i \geq 0, \sum_{i=1}^n x_i = 1\}$ and the loss vectors are constrained to the unit cube $\mathcal{Y} = \{l_t \in \mathbb{R}^N : \|l_t\|_\infty \leq 1\}$.
- *OLO over the Sphere*: Here the underlying decision is the euclidean ball $\mathcal{X} = \{x \in \mathbb{R}^n : \|x\|_2 \leq 1\}$ and the loss vectors are constrained to the unit euclidean ball $\mathcal{Y} = \{l_t \in \mathbb{R}^N : \|l_t\|_2 \leq 1\}$.
- *OLO over the Cube*: The decision is the unit cube $\mathcal{X} = \{x \in \mathbb{R}^n : \|x\|_\infty \leq 1\}$, while the loss vectors are constrained to the set $\mathcal{Y} = \{l_t \in \mathbb{R}^N : \|l_t\|_1 \leq 1\}$.

Partial-Information Setting: In the setting of bandit feedback, the critical difference is that the algorithm only gets to observe the value $\langle l_t, x_t\rangle$, in contrast to the complete loss vector $l_t \in \mathbb{R}^N$ as in the full information scenario. Therefore, the only feedback the algorithm receives is the value of the loss it incurs for the decision it takes. This makes designing algorithms for this feedback model challenging. Nevertheless for the general problem of bandit linear optimization, (Abernethy et al., 2008) introduced a computationally efficient algorithm that achieves an optimal dependence of the incurred regret of $O(\sqrt{T})$ on the number of rounds of play. The *non-stochastic multi-armed*

bandit (Auer et al., 2002) problem is the bandit version of the *prediction with expert advice* framework.

Differential Privacy: Differential Privacy (Dwork et al., 2006) is a rigorous framework for establishing guarantees on privacy loss, that admits a number of desirable properties such as graceful degradation of guarantees under composition and robustness to linkage acts (Dwork et al., 2014a).

Definition 2.1 ((ε, δ)-Differential Privacy)**.** *A randomized online learning algorithm $\mathcal{A}$ on the action set $\mathcal{X}$ and the loss set $\mathcal{Y}$ is (ε, δ)-differentially private if for any two sequence of loss vectors $L = (l_1, \dots l_T) \subseteq \mathcal{Y}^T$ and $L' = (l'_1, \dots l'_T) \subseteq \mathcal{Y}^T$ differing in at most one vector – that is to say $\exists t_0 \in [T], \forall t \in [T] - \{t_0\}, l_t = l'_t$ – for all $S \subseteq \mathcal{X}^T$, it holds that*

$$\mathbb{P}(\mathcal{A}(L) \in S) \leq e^{\varepsilon}\mathbb{P}(\mathcal{A}(L') \in S) + \delta$$

Remark 2.2. *The above definition of Differential Privacy is specific to the online learning scenario in the sense that it assumes the change of a complete loss vector. This has been the standard notion considered earlier in (Jain et al., 2012; Smith & Thakurta, 2013). Note that the definition entails that the entire sequence of predictions produced by the algorithm is differentially private.*

Notation: We define $\|\mathcal{Y}\|_p = \max\{\|l_t\|_p : l_t \in \mathcal{Y}\}$, $\|\mathcal{X}\|_p = \max\{\|x\|_p : x \in \mathcal{X}\}$, and $M = \max_{l \in \mathcal{Y}, x \in \mathcal{X}} |\langle l, x \rangle|$, where $\|\cdot\|_p$ is the l_p norm. By Holder's inequality, it is easy to see that $M \leq \|\mathcal{Y}\|_p \|\mathcal{X}\|_q$ for all $p, q \geq 1$ with $\frac{1}{p} + \frac{1}{q} = 1$. We define the distribution $Lap^N(\lambda)$ to be the distribution over $\mathbb{R}^N$ such that each coordinate is drawn independently from the Laplace distribution with parameter λ.

3. Full-Information Setting: Privacy for Free

In this section, we describe an algorithmic template (Algorithm 1) for differentially private online linear optimization, based on *Follow-the-Regularized-Leader* scheme. Subsequently, we outline the noise injection scheme (Algorithm 2), based on the Tree-based Aggregation Protocol (Dwork et al., 2010), used as a subroutine by Algorithm 1 to ensure input differential privacy. The following is our main theorem in this setting.

Theorem 3.1. *Algorithm 1 when run with $\mathcal{D} = Lap^N(\lambda)$ where $\lambda = \frac{\|\mathcal{Y}\|_1 \log T}{\varepsilon}$, regularization $R(x)$, decision set $\mathcal{X}$ and loss vectors $l_1, \dots l_t$, the regret of Algorithm 1 is bounded by*

$$Regret \leq \sqrt{D_R \sum_{t=1}^{T} \max_{x \in \mathcal{X}} (\|l_t\|^*_{\nabla^2 R(x)})^2 + D_{Lap}}$$

where

$$D_{Lap} = \mathbb{E}_{Z \sim \mathcal{D}'} \left[\max_{x \in X} \langle Z, x \rangle - \min_{x \in X} \langle Z, x \rangle \right]$$

$$D_R = \max_{x \in X} R(x) - \min_{x \in X} R(x)$$

and $\mathcal{D}'$ is the distribution induced by the sum of $\lceil \log T \rceil$ independent samples from $\mathcal{D}$, $\|\cdot\|^*_{\nabla^2 R(x)}$ represents the dual of the norm with respect to the hessian of R. Moreover, the algorithm is ε-differentially private, i.e. the sequence of predictions produced $(x_t : t \in [T])$ is ε-differentially private.

Algorithm 1 FTRL Template for OLO

input Noise distribution $\mathcal{D}$, Regularization $R(x)$

Initialize an empty binary tree B to compute differentially private estimates of $\sum_{s=1}^{t} l_s$.
Sample $n_0^1, \dots n_0^{\lceil \log T \rceil}$ independently from $\mathcal{D}$.
$\tilde{L}_0 \leftarrow \sum_{i=1}^{\lceil \log T \rceil} n_0^i$.
for $t = 1$ to T **do**
  Choose $x_t = argmin_{x \in \mathcal{X}} \left(\eta \langle x, \tilde{L}_{t-1} \rangle + R(x) \right)$.
  Observe $l_t \in \mathcal{Y}$, and suffer a loss of $\langle l_t, x_t \rangle$.
  $(\tilde{L}_t, B) \leftarrow \texttt{TreeBasedAgg}(l_t, B, t, \mathcal{D}, T)$.
end for

The above theorem leads to following corollary where we show the bounds obtained in specific instantiations of online linear optimization.

Corollary 3.2. *Substituting the choices of $\lambda, R(x)$ listed below, we specify the regret bounds in each case.*

1. ***Prediction with Expert Advice:*** *Choosing $\lambda = \frac{N \log T}{\varepsilon}$ and $R(x) = \sum_{i=1}^{N} x_i \log(x_i)$,*

$$Regret \leq O\left(\sqrt{T \log N} + \frac{N \log^2 T \log N}{\varepsilon} \right)$$

2. ***OLO over the Sphere*** *Choosing $\lambda = \frac{\sqrt{N} \log T}{\varepsilon}$ and $R(x) = \|x\|_2^2$*

$$Regret \leq O\left(\sqrt{T} + \frac{N \log^2 T}{\varepsilon} \right)$$

3. ***OLO over the Cube*** *With $\lambda = \frac{\log T}{\varepsilon}$ and $R(x) = \|x\|_2^2$*

$$Regret \leq O\left(\sqrt{NT} + \frac{N \log^2 T}{\varepsilon} \right)$$

Algorithm 2 `TreeBasedAgg`$(l_t, B, t, \mathcal{D}, T)$

input Loss vector l_t, Binary tree B, Round t, Noise distribution $\mathcal{D}$, Time horizon T

1: $(\tilde{L'}_t, B) \leftarrow$ `PrivateSum`$(l_t, B, t, \mathcal{D}, T)$ – Algorithm 5 ((Jain et al., 2012)) with the noise added at each node – be it internal or leaf – sampled independently from the distribution $\mathcal{D}$.
2: $s_t \leftarrow$ the binary representation of t as a string.
3: Find the minimum set $\mathcal{S}$ of *already* populated nodes in B that can compute $\sum_{s=1}^{t} l_s$.
4: Define $Q = |\mathcal{S}| \leq \lceil \log T \rceil$. Define $r_t = \lceil \log T \rceil - Q$.
5: Sample $n_t^1, \ldots n_t^{r_t}$ independently from $\mathcal{D}$.
6: $\tilde{L}_t \leftarrow \tilde{L'}_t + \sum_{i=1}^{r_t} n_t^i$.

output $(\tilde{L}_t, B)$.

3.1. Proof of Theorem 3.1

We first prove the privacy guarantee, and then prove the claimed bound on the regret. For the analysis, we define the random variable Z_t to be the net amount of noise injected by the TreeBasedAggregation (Algorithm 2) on the true partial sums. Formally, Z_t is the difference between cumulative sum of loss vectors and its differentially private estimate used as input to the arg-min oracle.

$$Z_t = \tilde{L}_t - \sum_{i=1}^{t} l_i$$

Further, let $\mathcal{D}'$ be the distribution induced by summing of $\lceil \log T \rceil$ independent samples from $\mathcal{D}$.

Privacy : To make formal claims about the quality of privacy, we ensure *input differential privacy* for the algorithm – that is, we ensure that the **entire sequence** of partial sums of the loss vectors $(\sum_{s=1}^{t} l_s : t \in [T])$ is ε-differentially private. Since the outputs of Algorithm 1 are strictly determined by the prefix sum estimates produced by `TreeBasedAgg`, by the post-processing theorem, this certifies that the entire sequence of choices made by the algorithm (across all T rounds of play) $(x_t : t \in [T])$ is ε-differentially private. We modify the standard Tree-based Aggregation protocol to make sure that the noise on each output (partial sum) is distributed identically (though not necessarily independently) across time. While this modification is not essential for ensuring privacy, it simplifies the regret analysis.

Lemma 3.3 (Privacy Guarantees with Laplacian Noise)**.** *Choose any $\lambda \geq \frac{\|\mathcal{Y}\|_1 \log T}{\varepsilon}$. When Algorithm 2 $\mathcal{A}(\mathcal{D}, T)$ is run with $\mathcal{D} = Lap^N(\lambda)$, the following claims hold true:*

- **Privacy:** *The sequence $(\tilde{L}_t : t \in [T])$ is ε-differentially private.*
- **Distribution:** *$\forall t \in [T], Z_t \sim \sum_{i=1}^{\lceil \log T \rceil} n_i$, where each n_i is independently sampled from $Lap^N(\lambda)$.*

Proof. By Theorem 9 ((Jain et al., 2012)), we have that the sequence $(\tilde{L'}_t : t \in [T])$ is ε-differentially private. Now the sequence $(\tilde{L}_t : t \in [T])$ is ε-differentially private because differential privacy is immune to post-processing(Dwork et al., 2014a).

Note that the `PrivateSum` algorithm adds exactly $|\mathcal{S}|$ independent draws from the distribution $\mathcal{D}$ to $\sum_{s=1}^{t} l_s$, where $\mathcal{S}$ is the minimum set of already populated nodes in the tree that can compute the required prefix sum. Due to Line 6 in Algorithm 2, it is made certain that every prefix sum released is a sum of the true prefix sum and $\lceil \log T \rceil$ independent draws from $\mathcal{D}$. □

Regret Analysis: In this section, we show that for linear loss functions any instantiation of the *Follow-the-Regularized-Leader* algorithm can be made differentially private with an additive loss in regret.

Theorem 3.4. *For any noise distribution $\mathcal{D}$, regularization $R(x)$, decision set $\mathcal{X}$ and loss vectors $l_1, \ldots l_t$, the regret of Algorithm 1 is bounded by*

$$\textit{Regret} \leq \sqrt{D_R \sum_{t=1}^{T} \max_{x \in \mathcal{X}} (\|l_t\|^*_{\nabla^2 R(x)})^2} + D_{\mathcal{D}'}$$

where $D_{\mathcal{D}'} = \mathbb{E}_{Z \sim \mathcal{D}'} [\max_{x \in \mathcal{X}} \langle Z, x \rangle - \min_{x \in \mathcal{X}} \langle Z, x \rangle]$, $D_R = \max_{x \in \mathcal{X}} R(x) - \min_{x \in \mathcal{X}} R(x)$, and $\| \cdot \|^*_{\nabla^2 R(x)}$ represents the dual of the norm with respect to the hessian of R.

Proof. To analyze the regret suffered by Algorithm 1, we consider an alternative algorithm that performs a one-shot noise injection – this alternate algorithm may not be differentially private. The observation here is that the alternate algorithm and Algorithm 1 suffer the same loss in expectation and therefore the same expected regret which we bound in the analysis below.

Consider the following alternate algorithm which instead of sampling noise Z_t at each step instead samples noise at the beginning of the algorithm and plays with respect to that. Formally consider the sequence of iterates $\hat{x}_t$ defined as follows. Let $Z \sim D$.

$$\hat{x}_1 \triangleq x_1, \quad \hat{x}_t \triangleq argmin_{x \in \mathcal{X}} \eta \langle x, Z + \sum_i l_i \rangle + R(x)$$

We have that

$$\mathbb{E}_{Z_1 \ldots Z_T \sim D} \left[\sum_{t=1}^{T} \langle l_t, x_t \rangle \right] = \mathbb{E}_{Z \sim D} \left[\sum_{t=1}^{T} \langle l_t, \hat{x}_t \rangle \right] \quad (1)$$

To see the above equation note that $\mathbb{E}_{Z_t \sim D} [\langle l_t, \hat{x}_t \rangle] = \mathbb{E}_{Z \sim D} [\langle l_t, x_t \rangle]$ since $x, \hat{x}_t$ have the same distribution.

Therefore it is sufficient to bound the regret of the sequence $\hat{x}_1 \dots \hat{x}_t$. The key idea now is to notice that the addition of one shot noise does not affect the stability term of the FTRL analysis and therefore the effect of the noise need not be paid at every time step. Our proof will follow the standard template of using the FTL-BTL (Kalai & Vempala, 2005) lemma and then bounding the stability term in the standard way. Formally define the augmented series of loss functions by defining

$$l_0(x) = \frac{1}{\eta}R(x) + \langle Z, x\rangle$$

where $Z \sim D$ is a sample. Now invoking the Follow the Leader, Be the Leader Lemma (Lemma 5.3, (Hazan et al., 2016)) we get that for any fixed $u \in \mathcal{X}$

$$\sum_{t=0}^{T} l_t(u) \geq \sum_{t=0}^{T} l_t(\hat{x}_{t+1})$$

Therefore we can conclude that

$$\sum_{t=1}^{T}[l_t(\hat{x}_t) - l_t(u)] \tag{2}$$

$$\leq \sum_{t=1}^{T}[l_t(\hat{x}_t) - l_t(\hat{x}_{t+1})] + l_0(u) - l_0(\hat{x}_1)$$

$$\leq \sum_{t=1}^{T}[l_t(\hat{x}_t) - l_t(\hat{x}_{t+1})] + \frac{1}{\eta}D_R + D_Z \tag{3}$$

where $D_Z \triangleq \max_{x\in X}(\langle Z, x\rangle) - \min_{x\in X}(\langle Z, x\rangle)$ Therefore we now need to bound the stability term $l_t(\hat{x}_t) - l_t(\hat{x}_{t+1})$. Now, the regret bound follows from the standard analysis for the stability term in the FTRL scheme (see for instance (Hazan et al., 2016)). Notice that the bound only depends on the change in the cumulative loss per step i.e. $\eta\left(\sum_t l_t + Z\right)$, for which the change is the loss vector ηl_{t+1} across time steps. Therefore we get that

$$l_t(\hat{x}_t) - l_t(\hat{x}_{t+1}) \leq \max_{x\in X} \|\eta l_t\|^2_{\eta\nabla^{-2}R(x)} \tag{4}$$

Combining Equations (1), (3), (4) we get the regret bound in Theorem 3.4.

□

3.2. Regret Bounds for FTPL

In this section, we outline algorithms based on the *Follow-the-Perturbed-Leader* template(Kalai & Vempala, 2005). FTPL-based algorithms ensure low-regret by perturbing the cumulative sum of loss vectors with noise from a suitably chosen distribution. We show that the noise added in the process of FTPL is sufficient to ensure differential privacy. More concretely, using the regret guarantees due to (Abernethy et al., 2014), for the full-information setting, we establish that the regret guarantees obtained scale as $O(\sqrt{T})+\tilde{O}(\frac{1}{\varepsilon}\log\frac{1}{\delta})$. While Theorem 3.5 is valid for all instances of online linear optimization and achieves $O(\sqrt{T})$ regret, it yields sub-optimal dependence on the dimension of the problem. The advantage of FTPL-based approaches over FTRL is that FTPL performs linear optimization over the decision set every round, which is possibly computationally less expensive than solving a convex program every round, as FTRL requires.

Algorithm 3 FTPL Template for OLO – $\mathcal{A}(\mathcal{D}, T)$ on the action set $\mathcal{X}$, the loss set $\mathcal{Y}$.

1: Initialize an empty binary tree B to compute differentially private estimates of $\sum_{s=1}^{t} l_s$.
2: Sample $n_0^1, \dots n_0^{\lceil \log T\rceil}$ independently from $\mathcal{D}$.
3: $\tilde{L}_0 \leftarrow \sum_{i=1}^{\lceil \log T\rceil} n_0^i$.
4: **for** $t = 1$ to T **do**
5: Choose $x_t = argmin_{x\in\mathcal{X}}\langle x, \tilde{L}_{t-1}\rangle$.
6: Observe the loss vector $l_t \in \mathcal{Y}$, and suffer $\langle l_t, x_t\rangle$.
7: $(\tilde{L}_t, B) \leftarrow \texttt{TreeBasedAgg}(l_t, B, t, \mathcal{D}, T)$.
8: **end for**

Theorem 3.5 (FTPL: Online Linear Optimization)**.** *Let* $\|\mathcal{X}\|_2 = \sup_{x\in\mathcal{X}}\|x\|_2$ *and* $\|\mathcal{Y}\|_2 = \sup_{l_t\in\mathcal{Y}}\|l_t\|_2$. *Choosing* $\sigma = \max\{\|\mathcal{Y}\|_2\sqrt{\frac{T}{\sqrt{N}\log T}}, \frac{\sqrt{N}}{\varepsilon}\log T\log\frac{\log T}{\delta}\}$ *and* $\mathcal{D} = \mathcal{N}(0, \sigma^2\mathbb{I}_N)$, *we have that* $Regret_{\mathcal{A}(\mathcal{D},T)}(T)$ *is*

$$O\left(N^{\frac{1}{4}}\|\mathcal{X}\|_2\|\mathcal{Y}\|_2\sqrt{T} + \frac{N\|\mathcal{X}\|_2}{\varepsilon}\log^{1.5}T\log\frac{\log T}{\delta}\right)$$

Moreover the algorithm is ε*-differentially private.*

The proof of the theorem is deferred to the appendix.

4. Differentially Private Multi-Armed Bandits

In this section, we state our main results regarding bandit linear optimization, the algorithms that achieve it and prove the associated regret bounds. The following is our main theorem concerning *non-stochastic multi-armed bandits*.

Theorem 4.1 (Differentially Private Multi-Armed Bandits)**.** *Fix loss vectors* $(l_1 \dots l_T)$ *such that* $\|l_t\|_\infty \leq 1$. *When Algorithm 4 is run with parameters* $\mathcal{D} = Lap^N(\lambda)$ *where* $\lambda = \frac{1}{\varepsilon}$ *and algorithm* $\mathcal{A}$ = *Algorithm 5 with the following parameters:* $\eta = \sqrt{\frac{\log N}{2NT(1+2\lambda^2\log NT)}}$, $\gamma = \eta N\sqrt{1+2\lambda^2\log NT}$ *and the exploration distribution* $\mu(i) = \frac{1}{N}$. *The regret of the Algorithm 4 is*

$$O\left(\frac{\sqrt{NT\log T\log N}}{\varepsilon}\right)$$

Moreover, Algorithm 4 is ε*-differentially private*

Bandit Feedback: Reduction to the Non-private Setting

We begin by describing an algorithmic reduction that takes as input a non-private bandit algorithm and translates it into an ε-differentially private bandit algorithm. The reduction works in a straight-forward manner by adding the requisite magnitude of Laplace noise to ensure differential privacy. For the rest of this section, for ease of exposition we will assume that both T and N are sufficiently large.

Algorithm 4 $\mathcal{A}'(\mathcal{A}, \mathcal{D})$ – Reduction to the Non-private Setting for Bandit Feedback

Input: Online Algorithm $\mathcal{A}$, Noise Distribution $\mathcal{D}$.

for $t = 0$ **to** T **do**
  Receive $\tilde{x_t} \in \mathcal{X}$ from $\mathcal{A}$ and output $\tilde{x}_t$.
  Receive a loss value $\langle l_t, \tilde{x_t} \rangle$ from the adversary.
  Sample $Z_t \sim \mathcal{D}$.
  Forward $\langle l_t\, \tilde{x_t} \rangle + \langle Z_t, \tilde{x_t} \rangle$ as input to $\mathcal{A}$.
end for

Algorithm 5 EXP2 with exploration μ

Input: learning rate η; mixing coefficient γ; distribution μ

$q_1 = \left(\frac{1}{N} \dots \frac{1}{N}\right) \in \mathbb{R}^N$.
for t = 1,2 ... T **do**
  Let $p_t = (1-\gamma)q_t + \gamma\mu$ and play $i_t \sim p_t$
  Estimate loss vector l_t by $\tilde{l_t} = P_t^+ e_{i_t} e_{i_t}^T l_t$, with $P_t = \mathbb{E}_{i \sim p_t}\left[e_i e_i^T\right]$
  Update the exponential weights,

$$q_{t+1}(i) = \frac{e^{-\eta \langle e_i, \tilde{l_t} \rangle} q_t(i)}{\sum_{i'} e^{-\eta \langle e_{i'}, \tilde{l_t} \rangle} q_t(i')}$$

end for

The following Lemma characterizes the conditions under which Algorithm 4 is ε differentially private

Lemma 4.2 (Privacy Guarantees)**.** *Assume that each loss vector l_t is in the set $\mathcal{Y} \subseteq \mathbb{R}^N$, such that $\max_{t, l \in \mathcal{Y}} |\frac{\langle l, \tilde{x}_t \rangle}{\|\tilde{x}_t\|_\infty}| \leq B$. For $\mathcal{D} = Lap^N(\lambda)$ where $\lambda = \frac{B}{\varepsilon}$, the sequence of outputs $(\tilde{x}_t : t \in [T])$ produced by the Algorithm $\mathcal{A}'(\mathcal{A}, \mathcal{D})$ is ε-differentially private.*

The following lemma charaterizes the regret of Algorithm 4. In particular we show that the regret of Algorithm 4 is, in expectation, same as that of the regret of the input algorithm $\mathcal{A}$ on a perturbed version of loss vectors.

Lemma 4.3 (Noisy Online Optimization)**.** *Consider a loss sequence $(l_1 \dots l_T)$ and a convex set $\mathcal{X}$. Define a perturbed version of the sequence as random vectors $(\tilde{l_t} : t \in [T])$ as $\tilde{l_t} = l_t + Z_t$ where Z_t is a random vector such that $\{Z_1, \dots Z_t\}$ are independent and $\mathbb{E}[Z_t] = 0$ for all $t \in [T]$.*

Let $\mathcal{A}$ be a full information (or bandit) online algorithm which outputs a sequence $(\tilde{x}_t \in \mathcal{X} : t \in [T])$ and takes as input $\tilde{l}_t$ (respectively $\langle \tilde{l}_t, \tilde{x}_t \rangle$) at time t. Let $x^* \in K$ be a fixed point in the convex set. Then we have that

$$\mathbb{E}_{\{Z_t\}}\left[\mathbb{E}_{\mathcal{A}}\left[\sum_{t=1}^{T}\left(\langle l_t, \tilde{x}_t \rangle - \langle l_t, x^* \rangle\right)\right]\right]$$

$$= \mathbb{E}_{\{Z_t\}}\left[\mathbb{E}_{\mathcal{A}}\left[\sum_{t=1}^{T}\left(\langle \tilde{l}_t . \tilde{x}_t \rangle - \langle \tilde{l}_t, x^* \rangle\right)\right]\right]$$

We provide the proof of Lemma 4.2 and defer the proof of Lemma 4.3 to the Appendix Section B.

Proof of Lemma 4.2. Consider a pair of sequence of loss vectors that differ at exactly one time step – say $L = (l_1, \dots l_{t_0} \dots, l_T)$ and $L' = (l_1, \dots, l'_{t_0}, \dots l_T)$. Since the prediction of produced by the algorithm at time step any time t can only depend on the loss vectors in the past $(l_1, \dots l_{t-1})$, it is clear that the distribution of the output of the algorithm for the first t_0 rounds $(\tilde{x}_1, \dots \tilde{x}_{t_0})$ is unaltered. We claim that $\forall \mathcal{I} \subseteq \mathbb{R}$, it holds that

$$\mathbb{P}(\langle l_{t_0} + Z_{t_0}, \tilde{x}_{t_0} \rangle \in \mathcal{I}) \leq e^{\varepsilon} \mathbb{P}(\langle l'_{t_0} + Z_{t_0}, \tilde{x}_{t_0} \rangle \in \mathcal{I})$$

Before we justify the claim, let us see how this implies that desired statement. To see this, note that conditioned on the value fed to the inner algorithm $\mathcal{A}$ at time t_0, the distribution of all outputs produced by the algorithm are completely determined since the feedback to the algorithm at other time steps (discounting t_0) stays the same (in distribution). By the above discussion, it is sufficient to demonstrate ε-differential privacy for each input fed (as feedback) to the algorithm $\mathcal{A}$.

For the sake of analysis, define l_t^{Fict} as follows. If $\tilde{x}_t = 0$, define $l_t^{Fict} = 0 \in \mathbb{R}^N$. Else, define $l_t^{Fict} \in \mathbb{R}^N$ to be such that $(l_t^{Fict})_i = \frac{\langle l_t, \tilde{x}_t \rangle}{\tilde{x}_i}$ if and only if $i = argmax_{i \in [d]} |\tilde{x}_i|$ and 0 otherwise, where $argmax$ breaks ties arbitrarily. Define $\tilde{l}_t^{Fict} = l_t^{Fict} + Z_t$. Now note that $\langle \tilde{l}_t^{Fict}, \tilde{x}_t \rangle = \langle l_t\, \tilde{x_t} \rangle + \langle Z_t, \tilde{x_t} \rangle$.

It suffices to establish that each $\tilde{l}_t^{Fict}$ is ε-differentially private. To argue for this, note that Laplace mechanism (Dwork et al., 2014a) ensures the same, since the l_1 norm of $\tilde{l}_t^{Fict}$ is bounded by B. □

4.1. Proof of Theorem 4.1

Privacy: Note that since $\max_{t, l \in \mathcal{Y}} |\frac{\langle l, \tilde{x}_t \rangle}{\|\tilde{x}_t\|_\infty}| \leq \|\mathcal{Y}\|_\infty \leq 1$ as $\tilde{x}_t \in \{e_i : i \in [N]\}$. Therefore by Lemma 4.2, setting $\lambda = \frac{1}{\varepsilon}$ is sufficient to ensure ε-differential privacy.

Regret Analysis: For the purpose of analysis we define the following pseudo loss vectors.

$$\tilde{l}_t = l_t + Z_t$$

where by definition $Z_t \sim Lap^N(\lambda)$. The following follows from Fact C.1 proved in the appendix.

$$\mathbb{P}(\|Z_t\|_\infty^2 \geq 10\lambda^2 \log^2 NT) \leq \frac{1}{T^2}$$

Taking a union bound, we have

$$\mathbb{P}(\exists t \;\; \|Z_t\|_\infty^2 \geq 10\lambda^2 \log^2 NT) \leq \frac{1}{T} \tag{5}$$

To bound the norm of the loss we define the event $F \triangleq \{\exists t : \|Z_t\|_\infty^2 \geq 10\lambda^2 \log^2 NT\}$. We have from (5) that $\mathbb{P}(F) \leq \frac{1}{T}$. We now have that

$$\mathbb{E}[Regret] \leq \mathbb{E}[Regret|\bar{F}] + \mathbb{P}(F)\mathbb{E}[Regret|F]$$

Since the regret is always bounded by T we get that the second term above is at most 1. Therefore we will concern ourselves with bounding the first term above. Note that Z_t remains independent and symmetric even when conditioned on the event $\bar{F}$. Moreover the following statements also hold.

$$\forall t \;\; \mathbb{E}[Z_t|\bar{F}] = 0 \tag{6}$$

$$\forall t \;\; \mathbb{E}[\|Z_t\|_\infty^2|\bar{F}] \leq 10\lambda^2 \log^2 NT \tag{7}$$

Equation (6) follows by noting that Z_t remains symmetric around the origin even after conditioning. It can now be seen that Lemma 4.3 still applies even when the noise is sampled from $Lap^N(\lambda)$ conditioned under the event $\bar{F}$ (due to Equation 6). Therefore we have that

$$\mathbb{E}[Regret|\bar{F}] = \mathbb{E}_{\{Z_t\}}\left[\mathbb{E}_{\mathcal{A}}\left[\sum_{t=1}^{T}\left(\langle \tilde{l}_t, \tilde{x}_t\rangle - \langle \tilde{l}_t, x^*\rangle\right)\right]\middle|\bar{F}\right] \tag{8}$$

To bound the above quantity we make use of the following lemma which is a specialization of Theorem 1 in (Bubeck et al., 2012a) to the case of multi-armed bandits.

Lemma 4.4 (Regret Guarantee for Algorithm 5)**.** *If η is such that $\eta|\langle e_i, \tilde{l}_t\rangle| \leq 1$, we have that the regret of Algorithm 5 is bounded by*

$$Regret \leq 2\gamma T + \frac{\log N}{\eta} + \eta\mathbb{E}\left[\sum_t \sum_i p_t(i)\langle e_i, \tilde{l}_t\rangle^2\right]$$

Now note that due to the conditioning $\|Z_t\|_\infty^2 \leq 10\lambda^2 \log^2 NT$ and therefore we have that

$$max_{t,x\in\Delta_N}|\langle Z_t, x\rangle| \leq 4\lambda \log NT.$$

It can be seen that the condition $\eta|\langle e_i, \tilde{l}_t\rangle| \leq 1$ in Theorem 4.4 is satisfied for exploration $\mu(i) = \frac{1}{N}$ and under the condition $\bar{F}$ as long as

$$\eta N(1 + 4\lambda \log NT) \leq \gamma$$

which holds by the choice of these parameters. Finally

$$\begin{aligned}
&\mathbb{E}[Regret|\bar{F}] \\
&= \mathbb{E}_{\{Z_t\}}\left[\mathbb{E}_{\mathcal{A}}\left[\sum_{t=1}^{T}\left(\langle \tilde{l}_t, \tilde{x}_t\rangle - \langle \tilde{l}_t, x^*\rangle\right)\right]\middle|\bar{F}\right] \\
&\leq \mathbb{E}_{\{Z_t\}}\left[\frac{\log N}{\eta} + \eta\sum_{t=1}^{T} N\|\tilde{l}_t\|_\infty^2 + 2T\gamma \;\middle|\bar{F}\right] \\
&\leq \mathbb{E}_{\{Z_t\}}\left[\frac{\log N}{\eta} + 2\eta\sum_{t=1}^{T} N(\|l_t\|_\infty^2 + \|Z_t\|_\infty^2) + 2T\gamma \;\middle|\bar{F}\right] \\
&\leq \frac{\log N}{\eta} + 2\eta TN(1 + \lambda^2 \log^2 NT) + 2T\gamma \\
&\leq O\left(\sqrt{TN \log N(1 + \lambda^2 \log^2 NT)}\right) \\
&\leq O\left(\frac{\sqrt{NT \log T \log N}}{\varepsilon}\right)
\end{aligned}$$

4.2. Differentially Private Bandit Linear Optimization

In this section we prove a general result about bandit linear optimization over general convex sets, the proof of which is deferred to the appendix.

Theorem 4.5 (Bandit Linear Optimization)**.** *Let $\mathcal{X} \subseteq \mathbb{R}^N$ be a convex set. Fix loss vectors $(l_1, \ldots l_T)$ such that $\max_{t,x\in\mathcal{X}} |\langle l_t, x\rangle| \leq M$. We have that Algorithm 4 when run with parameters $\mathcal{D} = Lap^N(\lambda)$ (with $\lambda = \frac{\|\mathcal{Y}\|_1}{\varepsilon}$) and algorithm $\mathcal{A} = SCRiBLe$(Abernethy et al., 2012) with step parameter $\eta = \sqrt{\frac{\nu \log T}{2N^2T(M^2+\lambda^2 N\|\mathcal{X}\|_2^2)}}$ we have the following guarantees that the regret of the algorithm is bounded by*

$$O\left(\sqrt{T \log T}\sqrt{N^2\nu\left(M^2 + \frac{N\|\mathcal{X}\|_2^2\|\mathcal{Y}\|_1^2}{\varepsilon^2}\right)}\right)$$

where ν is the self-concordance parameter of the convex body $\mathcal{X}$. Moreover the algorithm is ε-differentially private.

5. Conclusion

In this work, we demonstrate that ensuring differential privacy leads to only a constant additive increase in the incurred regret for online linear optimization in the full feedback setting. We also show nearly optimal bounds (in terms of T) in the bandit feedback setting. Multiple avenues for future research arise, including extending our bandit results to other challenging partial-information models such as semi-bandit, combinatorial bandit and contextual bandits. Another important unresolved question is whether it is possible to achieve an additive separation in ε, T in the adversarial bandit setting.

References

Abernethy, Jacob, Hazan, Elad, and Rakhlin, Alexander. Competing in the dark: An efficient algorithm for bandit linear optimization. In *COLT*, pp. 263–274, 2008.

Abernethy, Jacob, Lee, Chansoo, Sinha, Abhinav, and Tewari, Ambuj. Online linear optimization via smoothing. In *COLT*, pp. 807–823, 2014.

Abernethy, Jacob D, Hazan, Elad, and Rakhlin, Alexander. Interior-point methods for full-information and bandit online learning. *IEEE Transactions on Information Theory*, 58(7):4164–4175, 2012.

Auer, Peter, Cesa-Bianchi, Nicolo, Freund, Yoav, and Schapire, Robert E. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

Bubeck, Sébastien, Cesa-Bianchi, Nicolo, Kakade, Sham M, Mannor, Shie, Srebro, Nathan, and Williamson, Robert C. Towards minimax policies for online linear optimization with bandit feedback. In *COLT*, volume 23, 2012a.

Bubeck, Sébastien, Cesa-Bianchi, Nicolo, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012b.

Dwork, Cynthia, McSherry, Frank, Nissim, Kobbi, and Smith, Adam. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pp. 265–284. Springer, 2006.

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

Dwork, Cynthia, Roth, Aaron, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014a.

Dwork, Cynthia, Talwar, Kunal, Thakurta, Abhradeep, and Zhang, Li. Analyze gauss: optimal bounds for privacy-preserving principal component analysis. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, pp. 11–20. ACM, 2014b.

Hazan, Elad et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4): 157–325, 2016.

Jain, Prateek and Thakurta, Abhradeep G. (near) dimension independent risk bounds for differentially private learning. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 476–484, 2014.

Jain, Prateek, Kothari, Pravesh, and Thakurta, Abhradeep. Differentially private online learning. In *COLT*, volume 23, pp. 24–1, 2012.

Kalai, Adam and Vempala, Santosh. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, 2005.

Shalev-Shwartz, Shai. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2011.

Smith, Adam and Thakurta, Abhradeep Guha. (nearly) optimal algorithms for private online learning in full-information and bandit settings. In *Advances in Neural Information Processing Systems*, pp. 2733–2741, 2013.

Tossou, Aristide and Dimitrakakis, Christos. Algorithms for differentially private multi-armed bandits. In *AAAI 2016*, 2016.

Tossou, Aristide C. Y. and Dimitrakakis, Christos. Achieving privacy in the adversarial multi-armed bandit. In *14th International Conference on Artificial Intelligence (AAAI 2017)*, 2017.