Differentially Private Empirical Risk Minimization with Smooth Non-Convex Loss Functions: A Non-Stationary View*

Di Wang, Jinhui Xu
Deparment of Computer Science and Engineering
State University of New York at Buffalo
Buffalo, New York 14260

Abstract

In this paper, we study the Differentially Private Empirical Risk Minimization (DP-ERM) problem with non-convex loss functions and give several upper bounds for the utility in different settings. We first consider the problem in low-dimensional space. For DP-ERM with non-smooth regularizer, we generalize an existing work by measuring the utility using ℓ_2 norm of the projected gradient. Also, we extend the error bound measurement, for the first time, from empirical risk to population risk by using the expected ℓ_2 norm of the gradient. We then investigate the problem in high dimensional space, and show that by measuring the utility with Frank-Wolfe gap, it is possible to bound the utility by the Gaussian Width of the constraint set, instead of the dimensionality p of the underlying space. We further demonstrate that the advantages of this result can be achieved by the measure of ℓ_2 norm of the projected gradient. A somewhat surprising discovery is that although the two kinds of measurements are quite different, their induced utility upper bounds are asymptotically the same under some assumptions. We also show that the utility of some special non-convex loss functions can be reduced to a level (*i.e.,* depending only on $\log p$) similar to that of convex loss functions. Finally, we test our proposed algorithms on both synthetic and real world datasets and the experimental results confirm our theoretical analysis.

Introduction

Learning from sensitive data is a frequently encountered challenging task in many data analytic applications. It requires the learning algorithm to not only learn effectively from the data but also provide a certain level of guarantee on privacy preserving. As a rigorous notion for statistical data privacy, differential privacy has received a great deal of attentions in recent years (Dwork et al. 2006; Dwork 2008). It works by injecting random noise into the statistical results obtained from the sensitive data so that the distribution of the perturbed results are incentive to any single record change in the original dataset.

As one of the most commonly used supervised learning models, Empirical Risk Minimization (ERM) has been extensively studied in recent years. Its Differentially Private (DP) version (DP-ERM) can be formally defined as follows.

Definition 1 (DP-ERM (Wang, Gaboardi, and Xu 2018; Wang, Smith, and Xu 2019))**.** Given a dataset $D = \{z_1, \cdots, z_n\}$ from a data universe $\mathcal{X}$ and a closed convex set $\mathcal{C} \subseteq \mathbb{R}^p$, DP-ERM is to find $x^{\text{priv}} \in \mathcal{C}$ so as to minimize the empirical risk, *i.e.* $F^r(x, D) = \frac{1}{n}\sum_{i=1}^{n} f(x, z_i) + r(x)$, with the guarantee of being differentially private, where f is the loss function and r is some simple (non-)smooth convex function called **regularizer**. When the inputs are drawn i.i.d from an unknown underlying distribution $\mathcal{P}$ on $\mathcal{X}$, we also consider the population risk $\mathbb{E}_{z\sim\mathcal{P}}[f(x, z)]$. If the loss function is convex, the utility of the algorithm is measured by the expected excess empirical risk, *i.e.* $\mathbb{E}_{\mathcal{A}}[F^r(x^{\text{priv}}, D)] - \min_{x\in\mathcal{C}} F^r(x, D)$, or the expected excess population risk (generalization error), *i.e.* $\mathbb{E}_{z\sim\mathcal{P},\mathcal{A}}[f(x^{\text{priv}}, z)] - \min_{x\in\mathcal{C}} \mathbb{E}_{z\sim\mathcal{P}}[f(x, z)]$, where the expectation of $\mathcal{A}$ is taking over all the randomness of the algorithm.

Previous research on DP-ERM has mainly focused on convex loss functions, starting from the work in (Chaudhuri and Monteleoni 2009). However, several empirical studies have revealed that non-convex loss functions can achieve better classification accuracy than convex loss functions (Nguyen and Sanner 2013), and recent developments in Deep Neural Networks (Goodfellow et al. 2016) have further suggested that the loss functions are more likely to be non-convex in real world applications. Thus, there is an urgent need for the research community to shift its focus from convex to non-convex loss functions. However, due to the fact that finding the global minimum for non-convex functions is NP-hard, which implies that measuring the utility by the expected excess empirical risk may not always be a good choice. So far, only a few papers (Zhang et al. 2017; Wang, Ye, and Xu 2017) have considered the utility of DP-ERM with non-convex loss functions, but all of them measure the utility by ℓ_2 norm of the gradient, instead of the expected excess empirical risk.

Despite the aforementioned progresses on this problem, there are still quite a few remaining issues. 1) Previous work has obtained the error bounds for the smooth loss functions with smooth regularizer; it is not clear whether they can be extended to non-smooth regularizer, such as ℓ_1 norm. 2) Even though existing work has considered the error bound

*This research was supported in part by the National Science Foundation (NSF) through grants CCF-1422324 and CCF-1716400.

Method	Assumption	Utility Upper Bd	Non-smooth Regularizer	Measurement
(Zhang et al. 2017)	Smooth, ℓ_2-norm Lipschitz	$O(\frac{\sqrt[4]{p\ln(\frac{1}{\delta})\ln(\frac{n}{\delta})}}{\sqrt{n\epsilon}})$	No	ℓ_2 norm of gradient
(Wang, Ye, and Xu 2017)	Smooth, ℓ_2-norm Lipschitz	$O(\frac{\sqrt[4]{p\ln(\frac{1}{\delta})}}{\sqrt{n\epsilon}})$	No	ℓ_2 norm of gradient
Algorithm 1	Smooth, ℓ_2-norm Lipschitz	$O(\frac{\sqrt[4]{p\ln(\frac{1}{\delta})}}{\sqrt{n\epsilon}})$	Yes	ℓ_2 norm of projected gradient
Algorithm 2	Smooth, ℓ_2-norm Lipschitz, $\mathcal{C}$ bounded	$O(\frac{\sqrt[4]{(\|\mathcal{C}\|_2^2+G_{\mathcal{C}}^2)\ln\frac{1}{\delta}}}{\sqrt{n\epsilon}})$	No	Frank-Wolfe gap
Algorithm 3	Smooth, ℓ_2-norm Lipschitz, $\mathcal{C}$ bounded	$O(\frac{\sqrt[4]{(\|\mathcal{C}\|_2^2+G_{\mathcal{C}}^2)\ln\frac{1}{\delta}}}{\sqrt{n\epsilon}})$	Yes	ℓ_2 norm of projected gradient
Algorithm 4	Smooth, ℓ_1-norm Lipschitz, $\mathcal{C}$ is ℓ_1 norm ball (or polytope)	$O(\frac{\sqrt[4]{\ln(\frac{1}{\delta})}\sqrt{\ln(np)}}{\sqrt{n\epsilon}})$	No	Frank-Wolfe gap

Table 1: Comparisons with previous (ϵ, δ)-DP algorithms for DP-ERM with non-convex loss function. We assume that the Lipschitz and smooth parameters are 1, and $\|\mathcal{C}\|_2 \leq 1$.

measured by empirical risk, it is not clear what is the generalization property of the problem. Particularly, it is unknown what is the error bound measured by population risk for non-convex loss functions and its difference with the convex ones (Bassily, Smith, and Thakurta 2014). 3) Existing work mainly focuses on the low dimensional case, where $n \gg p$. It is still unknown what can be done for the high dimensional case. In this paper, we will address the above issues. Our main results are listed in **Table 1**. Below is a more detailed description of our contributions.

1. For low dimensional space, we consider the general case for DP-ERM with non-convex loss function and non-smooth regularizer. For this case (see **Algorithm 1**), we generalize the approaches in (Zhang et al. 2017; Wang, Ye, and Xu 2017), which consider only smooth regularizer and unconstrained domain, *i.e.* $\mathcal{C} = \mathbb{R}^p$). Particularly, we use as the utility the ℓ_2 norm of the projected gradient, while (Zhang et al. 2017; Wang, Ye, and Xu 2017) use the ℓ_2 norm of the gradient. Then, we resolve some practical issues in (Zhang et al. 2017; Wang, Ye, and Xu 2017) by using zero Concentrated Differential Privacy. Finally, we study the generalization property of the private estimator. By using ℓ_2 norm of the gradient in the empirical risk, we show an upper bound of the population risk with non-convex loss functions at the point θ^{priv} based on **the expected ℓ_2-norm of the gradient**, *i.e.* $\mathbb{E}_{\mathcal{A}}\|\mathbb{E}_{z\sim\mathcal{P}}[\nabla f(x^{\text{priv}}, z)]\|_2$.
2. For high dimensional space (*i.e.* $p \gg n$), we first show that by using the differentially private version of Frank-Wolfe method, it is possible to measure the utility by Frank-Wolfe gap (see **Algorithm 2**), and the utility upper bound depends only on the Gaussian Width of the constraint set $\mathcal{C}$ (see Definition 8), instead of the dimensionality p of the underlying space. Then, we improve the robustness of the above approach for non-smooth regularizer, while still maintain the same utility upper bound (see **Algorithm 3**) for the case of $\|\mathcal{C}\|_2 \leq 1$ by using the ℓ_2 norm of the projected gradient. Finally, we consider a special case where $\mathcal{C}$ is a polytope and the loss function is ℓ_1-Lipschitz, which has been studied in (Talwar, Thakurta, and Zhang 2015) for the convex case. For this case (see **Algorithm 4**), we present a method which uses Frank-Wolfe gap to measure the utility and achieves an upper bound depending only on $\log p$, instead of the Gaussian Width or the dimensionality of the underlying space.

Due to the space limit, all the proofs and additional experiments are left in the supplemental materials.

Related Work

There is a long list of works on differentially private ERM in the last decade which attack the problem from different perspectives. Below we mainly discuss those which are related to our problem and have theoretical guarantees on the utility.

Quite a number of approaches exist for DP-ERM with convex loss function, which can be roughly classified into three categories. The first type of approaches is to perturb the output of a non-DP algorithm. (Chaudhuri and Monteleoni 2009) first proposed the output-perturbation approach, which is later extended by (Zhang et al. 2017) and achieved the optimal bound for smooth and strongly convex functions. The second type of approaches is to perturb the objective function(Chaudhuri and Monteleoni 2009), which is later extended by (Kifer, Smith, and Thakurta 2012) to more general cases. Note that the above two approaches depend on the convexity of the loss function which in general cannot be used for the non-convex case.

The third type of approaches is to perturb the gradients in gradient descent algorithms. (Bassily, Smith, and Thakurta 2014) proposed the gradient perturbation approach and gave the lower bound of the utility for both general convex and strongly convex loss functions. Recently (Wang, Ye, and Xu 2017) combined this idea with variance reduction method and obtained faster algorithms. They also extended their approach for strongly convexity to PL-condition. In this paper, we mainly follow this type of approaches.

For DP-ERM with convex loss functions in high dimensional space, (Talwar, Thakurta, and Zhang 2014) showed that the lower bound given in (Bassily, Smith, and Thakurta 2014) can actually be broken by adding more restrictions on the convex domain $\mathcal{C}$ of the problem. Their lower bound depends on the the Gaussian Width and is independent of the dimensionality p. Later, (Talwar, Thakurta, and Zhang 2015) considered a special case where the loss function is Lipschitz under ℓ_1 norm and the constraint set $\mathcal{C}$ is a polytope. They demonstrated that the utility bound in this case still depends only on $\log p$.

Preliminaries

Definition 2 (Lipschitz Function over x)**.** A loss function $f : \mathcal{C} \times \mathcal{X} \to \mathbb{R}$ is G-Lipschitz under ℓ_2-norm over x, if for any $z \in \mathcal{X}$ and $x_1, x_2 \in \mathcal{C}$, $|f(x_1, z) - f(x_2, z)| \leq G||x_1 - x_2||_2$ holds. Similarly, f is G-Lipschitz under ℓ_1-norm over θ, if for any $z \in \mathcal{X}$ and $x_1, x_2 \in \mathcal{C}$, $|f(x_1, z) - f(x_2, z)| \leq G||x_1 - x_2||_1$ holds.

We will consider the Lipschitz property under ℓ_2-norm for the first three results, and under ℓ_1-norm for the fourth result. A good example for the Lipschitz property under ℓ_1-norm is linear regression $F(\theta, (X, y)) = \frac{1}{n}\|X\theta - y\|^2$; for $|x_{ij}|, |y_j| = O(1)$, it is $O(1)$-Lipschitz under ℓ_1-norm.

Definition 3 (L-smooth Function over x)**.** A loss function $f : \mathcal{C} \times \mathcal{X} \to \mathbb{R}$ is L-smooth over x with respect to the norm $||\cdot||$ if for any $z \in \mathcal{X}$ and $x_1, x_2 \in \mathcal{C}$, the following holds

$$||\nabla f(x_1, z) - \nabla f(x_2, z)||_* \leq L||x_1 - x_2||,$$

where $||\cdot||_*$ is the dual norm of $||\cdot||$. If f is differentiable, this yields $f(x_1, z) \leq f(x_2, z) + \langle \nabla f(x_2, z), x_1 - x_2\rangle + \frac{L}{2}||x_1 - x_2||^2$.

We say that two datasets D and D' are neighbors to each other if they differ only by one entry, denoted as $D \sim D'$.

Definition 4 (Differentially Private)**.** A randomized algorithm $\mathcal{A}$ is (ϵ, δ)-differentially private if for all neighboring datasets $D, D' \in \mathcal{X}^n$ and for all events S in the output space of $\mathcal{A}$, the following holds $\Pr(\mathcal{A}(D) \in S) \leq e^\epsilon Pr(\mathcal{A}(D') \in S) + \delta$. when $\delta = 0$, $\mathcal{A}$ is ϵ-differentially private.

Lemma 1. [Advanced Composition Theorem (Dwork, Rothblum, and Vadhan 2010)] Given target privacy parameters $0 < \epsilon' < 1$ and $\delta > 0$, to ensure $(\epsilon', k\delta + \delta')$ cumulative privacy loss over k mechanisms, it suffices that each mechanism is (ϵ, δ)- differentially private, where $\epsilon = \frac{\epsilon'}{\sqrt{8k\ln(\frac{1}{\delta'})}}$.

Definition 5 (Gaussian Mechanism)**.** Given a function $q : \mathcal{X}^n \to \mathbb{R}^p$, the Gaussian Mechanism is defined as: $\mathcal{M}_G(D, q, \epsilon) = q(D) + Y$, where Y is drawn from a Gaussian Distribution $\mathcal{N}(0, \sigma^2 I_p)$ with $\sigma \geq \frac{\sqrt{2\ln(1.25/\delta)}\Delta_2(q)}{\epsilon}$. $\Delta_2(q)$ is the ℓ_2-sensitivity of the function q, *i.e.,* $\Delta_2(q) = \sup_{D\sim D'} ||q(D) - q(D')||_2$. Gaussian Mechanism preserves (ϵ, δ)-differentially private.

Moments Accountant (MA) is a method for accumulating the privacy cost to achieve a tighter bound on ϵ and δ. Roughly speaking, when we use the Gaussian Mechanism on the (stochastic) gradient descent, Moments Accountant allows us to save a factor of $\sqrt{\ln(T/\delta)}$ in the asymptotic bound on the standard deviation of noise compared with those achieved by using the Advanced Composition Theorem.

Lemma 2. (Abadi et al. 2016) For G-Lipschitz loss function, there exist constants c_1 and c_2 so that given the sampling probability $q = l/n$ and the number of steps T, for any $\epsilon < c_1 q^2 T$, a DP stochastic gradient algorithm with batch size l that injects Gaussian Noise with standard deviation $\frac{G}{n}\sigma$ to the gradients (Algorithm 1 in (Abadi et al. 2016)), is (ϵ, δ)-differentially private for any $\delta > 0$ if $\sigma \geq c_2 \frac{q\sqrt{T\ln(1/\delta)}}{\epsilon}$.

Definition 6 (Exponential Mechanism)**.** The Exponential Mechanism allows differentially private computation over arbitrary domains and range $\mathcal{R}$, parametrized by a score function $u(D, r)$ which maps a pair of input data set D and candidate result $r \in \mathcal{R}$ to a real valued score. With the score function u and privacy budget ϵ, the mechanism yields an output with exponential bias in favor of high scoring outputs. Let $\mathcal{M}(D, x, R)$ denote the exponential mechanism, and Δ be the sensitivity of u in the range R, $\Delta = \max_{r\in\mathcal{R}} \max_{D\sim D'} |u(D, r) - u(D', r)|$. Then if $\mathcal{M}(D, x, R)$ selects and outputs an element $r \in \mathcal{R}$ with probability proportional to $\exp(\frac{\epsilon u(D,r)}{2\Delta u})$, it preserves ϵ-differential privacy.

Lemma 3. (Dwork, Roth, and others 2014) For the exponential mechanism $\mathcal{M}(D, u, \mathcal{R})$, we have

$$\Pr\{u(\mathcal{M}(D, u, \mathcal{R})) \leq OPT_u(x) - \frac{2\Delta u}{\epsilon}(\ln|\mathcal{R}| + t)\} \leq e^{-t}.$$

where $OPT_u(x)$ is the highest score in the range $\mathcal{R}$, *i.e.* $\max_{r\in\mathcal{R}} u(D, r)$.

Low Dimension Case

Extending to Non-Smooth Regularizer

In this section, we consider DP-ERM with non-convex loss function and non-smooth convex regularizer, *i.e.*,

$$\min_{x\in\mathcal{C}} F^r(x, D) = \frac{1}{n}\sum_{i=1}^{n} f(x, z_i) + r(x). \tag{1}$$

For convenience, we let $F(x) = \frac{1}{n}\sum_{i=1}^n f(x, z_i)$ and $F^r(x) = F^r(x, D)$.

Assumption 1. $F(x)$ is assumed to be differentiable and L-smooth over x with respect to ℓ_2 norm. Also, the loss function $f(\cdot, z)$ is assumed to be G-Lipschitz over x with respect to ℓ_2-norm for all $z \in \mathcal{X}$.

In order to measure the utility for (1), we define the **generalized projected gradient** as $\mathcal{P}_\mathcal{C}(x, g, \gamma) = \frac{1}{\gamma}(x - x^+)$, where

$$x^+ = \arg\min_{u\in\mathcal{C}}\{\langle g, u\rangle + \frac{1}{2\gamma}\|x - u\|_2^2 + r(u)\}. \tag{2}$$

Note that this measurement has been widely used in the optimization community for studying the convergence and non-stationarity, such as (Ghadimi and Lan 2016; Ghadimi, Lan, and Zhang 2016). Actually, if $\mathcal{C} = \mathbb{R}^p$ and $r(x) \equiv 0$, we have $\mathcal{P}_\mathcal{C}(x, \nabla F(x), \gamma) = \nabla F(x) = \nabla F^r(x)$.

Based on the Projected Gradient Descent, we have the following algorithm for DP-ERM with non-convex loss function and non-smooth convex regularizer.

Theorem 1. There exist constants c, c_1, such that for any $0 < \epsilon < c_1 T, 0 < \delta < 1$, **DP-PGD** (Algorithm 1) is (ϵ, δ)-differentially private if

$$\sigma^2 = c\frac{G^2 T\ln(\frac{1}{\delta})}{n^2\epsilon^2}. \tag{3}$$

Algorithm 1 DP-PGD($F, x_1, T, \sigma, \{\gamma_k\}_{k=1}^T$)

Input: T is the maximum number of iterations, x_1 is the initial point, and $\{\gamma_k\}_{k=1}^T$ is the step size. ϵ and δ are privacy parameters.

1: **for** $k = 1, \cdots, T$ **do**
2: Compute $x_{k+1} = \arg\min_{u\in\mathcal{C}}\{\langle \nabla F(x_k) + \epsilon_k, u\rangle + \frac{1}{2\gamma_k}\|u - x_k\|_2^2 + r(u)\}$, where $\epsilon_k \sim N(0, \sigma^2 I_p)$, here σ can be chosen by Theorem 1 or as the following:
3: Compute ρ which satisfies $\rho + 2\sqrt{\rho\log(\frac{1}{\delta})} = \epsilon$. Then set $\sigma^2 = \frac{2L^2T}{n^2\rho^2}$.
4: **end for**
5: **return** $x_R \in \{x_1, \cdots, x_T\}$ such that R is uniformly sampled from $\{1, 2, \cdots, T\}$.

Theorem 2. Under **Assumption** 1, if we take σ^2 as in (3), $\{\gamma\}_{k=1}^T = \frac{1}{2L}$ and $T = O(\frac{n\epsilon}{\sqrt{p\ln(\frac{1}{\delta})}})$ in Algorithm 1, the following inequality holds,

$$\mathbb{E}\|g_{\mathcal{C},R}\|_2 \le O\Big(\frac{\sqrt[4]{p\ln(\frac{1}{\delta})}}{\sqrt{n\epsilon}}\Big), \tag{4}$$

where $g_{\mathcal{C},R} = \frac{1}{\gamma_k}(x_R - x_{R+1})$.

Remark 1. Note that if we remove the non-smoothness restriction on the regularizer and assume that $\mathcal{C} = \mathbb{R}^p$, the upper bound in Theorem 2 becomes the same as in (Wang, Ye, and Xu 2017). Thus Theorem 2 can be viewed as a generalization of theirs.

Also it is worth noting that if we use the output in classical non-convex optimization algorithm directly, such as the one on Page 26 in (Nesterov 2013), *i.e.* $\|g_{\mathcal{C},R}\|_2 = \min_{1\le k\le T}\|g_{\mathcal{C},k}\|_2$, the algorithm will not be differentially private. Thus, here we use another randomizer on R. This is a main difference between our algorithm and those optimization algorithms.

It is notable that the variance of noise (3) in Theorem 1, which is based on Moment Accountant (Lemma 2), just states the existence of such constant c without specifying it. The constant also has not been mentioned in the previous work in (Wang, Ye, and Xu 2017), while in (Zhang et al. 2017) the noise added in each iteration is $\sigma_1^2 = \frac{32\ln(1/\delta)\ln(T/\delta)}{n^2\epsilon^2}$. Since differential privacy is a rigorous mathematical definition, it is important to select the appropriate constant c (Fredrikson et al. 2014) in practice. Actually we can follow the way in (Abadi et al. 2016) which is based on grid search for finding this hidden constant. However, this procedure is costy and complex, here we propose a more practical approach by transforming zero Concentrated Different Privacy (zCDP) (Bun and Steinke 2016) to (ϵ, δ)-DP, which corresponds to the step 3 in Algorithm 1[1].

[1]Recently, (Lee and Kifer 2018) also proposed a similar way of reducing the noise in DP-GD based on zCDP. However, here we do not compare with it since there is no theoretical guarantee in their paper.

The idea is that we first make the algorithm to be ρ-zCDP and then transfer to (ϵ, δ)-DP, *i.e.* we first compute the number ρ which satisfies $\rho + 2\sqrt{\rho\log(\frac{1}{\delta})} = \epsilon$. Then we perform Algorithm 1 for T iterations. We can easily get in this case the variance satisfies $\sigma_2^2 = \frac{2L^2T}{n^2(\sqrt{\log(1/\delta)+\epsilon}-\sqrt{\log(1/\delta)})^2}$. When $\frac{\epsilon}{\log(1/\delta)} \ll 1$ (this case will always holds since in practice we select $\epsilon = 0.1 - 0.5$ and $\delta = \frac{1}{n}$), by expanding Taylor series of $\sqrt{1+x}$, we have $(\sqrt{\log(1/\delta)+\epsilon} - \sqrt{\log(1/\delta)})^2 \simeq \frac{\epsilon^2}{4\log(1/\delta)}$, so $\sigma_1^2 \simeq \frac{8L^2T\log(1/\delta)}{n^2\epsilon^2}$. We can see that compared with moment accountant method, our method is much more practical and simpler, compared with advanced composition theorem, it adds less noise in each iteration (see Experiment section for details).

Extension to Population Risk

An important problem in machine learning is to use population risk to measure the performance of an estimator. It indicates how well the estimator performs on unseen examples from the same distribution. Based on the idea of measuring the utility of θ^{priv} by the ℓ_2 norm of the gradient of the empirical risk, in this section, we show an upper bound of θ^{priv} on the population risk based on the ℓ_2 norm of the gradient for non-convex loss functions, *i.e.* $\mathbb{E}_{\mathcal{A}}\|\mathbb{E}_{z\sim\mathcal{P}}[\nabla f(x^{\text{priv}}, z)]\|_2$, where $\mathcal{A}$ is the randomized algorithm which outputs the private estimator x^{priv}.

Due to the hardness of the problem even in non-private settings, we need to make some assumptions. Below, we only consider the non-regularizer case.

Assumption 2. The gradient of the loss function is τ-sub-Gaussian. That is, for any $\lambda \in \mathbb{R}^p$ and $x \in \mathbb{R}^p$, we have $\mathbb{E}\{\exp(\langle\lambda, \nabla f(x,z) - \mathbb{E}[\nabla f(x,z)]\rangle)\} \le \exp(\frac{\tau^2\|\lambda\|^2}{2})$.

Assumption 3. The Hessian of the population risk is bounded. That is, there exists an H such that $\|\nabla^2\mathbb{E}_{z\sim\mathcal{P}}[f(x_0, z)]\|_2 \le H$ for all $x_0 \in \mathbb{R}^p$. Also, the Hessian of the loss function is L-Lipschitz. That is, for every z and $x_1, x_2 \in \mathbb{R}^p$, we have $\frac{\|\nabla^2 f(x_1,z) - \nabla^2 f(x_2,z)\|_2}{\|x_1 - x_2\|_2} \le L$, where the ℓ_2 norm of the Hessian is the operator norm. Furthermore, we assume that the constant H, L cannot be too large with respect to τ and p. This means that there exists a constant c such that $H \le \tau^2 p^c$ and $L \le \tau^3 p^c$.

Note that the first assumption is quite standard for analyzing the population risk (Chen, Su, and Xu 2017). The second assumption is very common in many non-convex loss functions, such as robust regression and binary classification. The examples can be found in (Mei, Bai, and Montanari 2016). Based on recent results on non-convex learning, we now have the following theorem.

Theorem 3. Under Assumption 1, 2 and 3, if $n \ge \Omega(p\log(p))$, then for any $0 < \epsilon, \delta, \beta \le 1$, there is an (ϵ, δ)-differentially private algorithm $\mathcal{A}$ which outputs x_R satisfy-

ing the following with probability at least $1-\beta$,

$$\mathbb{E}_{\mathcal{A}}\|\mathbb{E}_{z\sim\mathcal{P}}[\nabla f(x_R,z)]\|_2 \le O(\sqrt{\frac{\tau^2 p\log(\frac{\tau}{\beta})\log n}{n}}+ \frac{\sqrt[4]{p\log(\frac{1}{\delta})}}{\sqrt{n\epsilon}}) = O(\tau\sqrt{\frac{p\log(\frac{\tau}{\beta})\log n\sqrt{\log\frac{1}{\delta}}}{n\epsilon}}). \quad (5)$$

Remark 2. As we can see from above theorem, compared with the uniform convergence error, *i.e.* the first term in the right side of (5), the error due to differential privacy, *i.e.* the second term in the right side of (5), is less when we consider ϵ, δ as constants. Thus, the effect of differential privacy on the convergence error is just making the efficient sample complexity n become $n\epsilon$. This is quite different from the population risk in convex loss functions under differential privacy, where the error caused by privacy plays a much more important role, *i.e.* there is additional factor of $\sqrt{p}$ in the population risk under differential privacy compared with non-private case. For details, please refer to Appendix F in (Bassily, Smith, and Thakurta 2014). An open problem is that whether this bound is tight, or whether we can deal with the high dimensional case, we left these for future works.

High Dimension Case

Error Bounded by Frank-Wolfe Gap

The utility bound in (4) depends on the dimensionality p. In high dimensional (*i.e.,* $p \gg n$) space, such a dependence may no longer be desirable. For convex loss functions, (Talwar, Thakurta, and Zhang 2014) showed that it is possible to make the utility bound (using the expected excess empirical risk as the measurement) depend only on the Gaussian Width of the constrained set $\mathcal{C}$, which could be considerably smaller than $O(\sqrt{p})$ when $\mathcal{C}$ is a bounded closed centrally symmetric convex set $\mathcal{C} \subseteq \mathbb{R}^p$ (such as l_1-norm ball). Thus, a natural question is whether such an improvement can also be achievable for non-convex loss functions. Below we give an affirmative answer by showing that this is indeed possible for non-convex loss function (without considering the non-smoothness constraint on the regularizer, *i.e.*, $r(x) \equiv 0$).

We start our discussion with some definitions and lemmas which will be used in this and next section.

Definition 7 (Minkowski Norm)**.** The Minkowski norm (denoted by $||\cdot||_{\mathcal{C}}$) with respect to a centrally symmetric convex set $\mathcal{C} \subseteq \mathbb{R}^p$ is defined as follows. For any vector $v \in \mathbb{R}^p$, $||\cdot||_{\mathcal{C}} = \min\{r \in \mathbb{R}^+ : v \in r\mathcal{C}\}$. The dual norm of $||\cdot||_{\mathcal{C}}$ is denoted as $||\cdot||_{\mathcal{C}^*}$; for any vector $v \in \mathbb{R}^p$, $||v||_{\mathcal{C}^*} = \max_{w\in\mathcal{C}} |\langle w, v\rangle|$.

Definition 8 (Gaussian Width)**.** Let $b \sim \mathcal{N}(0, I_p)$ be a Gaussian random vector in $\mathbb{R}^p$. The Gaussian width for a set $\mathcal{C}$ is defined as $G_{\mathcal{C}} = \mathbb{E}_b[\sup_{w\in\mathcal{C}}\langle b, w\rangle]$.

Compared with the dimensionality p, Gaussian Width of a convex set $\mathcal{C} \subseteq \mathbb{R}^p$ could be much smaller. For example, when $\mathcal{C}$ is l_1-norm unit ball, $G_{\mathcal{C}} = O(\sqrt{\log p})$; when $\mathcal{C}$ is the set of all unit s-sparse vectors on $\mathbb{R}^p$, $G_{\mathcal{C}} = O(\sqrt{s\log(p/s)})$.

Lemma 4. (Talwar, Thakurta, and Zhang 2014) For $W = (\max_{w\in\mathcal{C}}\langle w, v\rangle)^2$, where $v \sim \mathcal{N}(0, I_p)$, we have $\mathbb{E}_v[W] = O(G_{\mathcal{C}}^2 + ||\mathcal{C}||_2^2)$.

For simplicity, we let $\|\cdot\|$ denote $\|\cdot\|_{\mathcal{C}}$ and $\|\cdot\|_*$ denote $\|\cdot\|_{\mathcal{C}^*}$ in this section.

Our algorithm is based on the Frank-Wolfe method, where a differentially private version of Frank-Wolfe has been studied in (Talwar, Thakurta, and Zhang 2015) for LASSO. Frank-Wolfe method can be viewed as a greedy algorithm which moves towards the optimum solution in the first order approximation. It reduces the problem to solving a minimization problem of linear function, which exploits the geometric property of the constrained set $\mathcal{C}$. It also provides a new measurement of the non-stationarity, called Frank-Wolfe gap, for the utility, which has already been used in (Lacoste-Julien 2016; Reddi et al. 2016). Formally, the Frank-Wolfe gap at a point x of the function F is defined as: $\mathcal{G}(x) = \max_{v\in\mathcal{C}}\langle v - x, -\nabla F(x)\rangle, x \in \mathcal{C}$. Since the gap $\mathcal{G}(x) = 0$ if and only if x is a stationary point, it could provide of stationarity for a point. Our following algorithm uses the Frank-Wolfe gap as a measurement for DP-ERM with non-convex smooth loss functions.

Algorithm 2 DP-FW-L2($F, x_1, T, \sigma, \{\gamma_t\}_{t=1}^T$)

Input: T is the maximum of iterations, x_1 is the initial point, and $\{\gamma_t\}_{t=1}^T$ is the step size.

for $k = 1, \cdots, T$ **do**

Compute $v_t = \arg\max_{v\in\mathcal{C}}\langle v, -(\nabla F(x) + \epsilon_t)\rangle$, where $\epsilon_k \sim N(0, \sigma^2 I_p)$.

$x_{t+1} = x_t + \gamma_t(v_t - x_t)$.

end for

return $x_R \in \{x_1, \cdots, x_T\}$ such that R is uniformly sampled from $\{1, \cdots, T\}$.

Theorem 4. Let $\mathcal{C}$ be a bounded, closed, centrally symmetric convex set. Then, there exist constants c, c_1, under **Assumption** 1 and for any $0 < \epsilon < c_1 T, 0 < \delta < 1$, **DP-FW-L2** (Algorithm 2) is (ϵ, δ)-differentially private if σ^2 is chosen as in (3). Moreover, if taking $\{\gamma_t\}_{t=1}^T = O(\frac{\sqrt[4]{(\|\mathcal{C}\|_2^2+G_{\mathcal{C}}^2)\ln\frac{1}{\delta}}}{\|\mathcal{C}\|_2\sqrt{n\epsilon}})$ and $T = O(\frac{n\epsilon}{\sqrt{(\|\mathcal{C}\|_2^2+G_{\mathcal{C}}^2)\ln\frac{1}{\delta}}})$, the following holds,

$$\mathbb{E}[\mathcal{G}_R] \le O\Big(\frac{\|\mathcal{C}\|_2\sqrt[4]{(\|\mathcal{C}\|_2^2 + G_{\mathcal{C}}^2)\ln\frac{1}{\delta}}}{\sqrt{n\epsilon}}\Big), \quad (6)$$

where $\mathcal{G}_t = \max_{v\in\mathcal{C}}\langle -\nabla F(x_t), v - x_t\rangle$.

Error Bounded by Norm of Gradient

So far we have presented two methods for the general non-convex case and the high dimension case, respectively. Theorem 4 enables us to bound the utility using Gaussian Width, but has some robustness issue with non-smooth regularizer. Contrarily, Theorem 2 can handle non-smooth regularizer, but its utility depends on the dimensionality of the space. Below we show in **Algorithm 3** that it is actually possible to combine the advantages of both methods by using Mirror Descent.

Definition 9. A function $w : \mathcal{C} \to \mathbb{R}$ is said to be a distance generating function with modulus $\alpha > 0$ (w.r.t. $\|\cdot\|$ norm), if w is continuously differentiable and strongly convex satisfying the following inequality for any $x, z \in \mathcal{C}$, $\langle x - z, \nabla w(x) - \nabla w(z)\rangle \geq \alpha\|x - z\|^2$. The Bregman Divergence associated with w is defined as $V(x, z) = w(x) - w(z) - \langle \nabla w(z), x - z\rangle$.

Similar to (2), we define the generalized projected gradient as $\mathcal{P}_{\mathcal{C}}(x, g, \gamma) = \frac{1}{\gamma}(x - x^+)$, where $x^+ = \arg\min_{u\in\mathcal{C}}\{\langle g, u\rangle + \frac{1}{\gamma}V(u, x) + r(u)\}$. Note that (2) is a special case in which $w(x) = \frac{1}{2}\|x\|_2^2$.

Algorithm 3 DP-PMD($F, x_1, T, \sigma, \{\gamma_k\}_{k=1}^T, w(\cdot)$)

Input: T is the maximum number of iterations, x_1 is the initial point, $w : \mathcal{C} \to \mathbb{R}$ is a distance generating function with modulus 1 (w.r.t. $\|\cdot\|$ norm) and $V(\cdot, \cdot)$ is its Bregman Divergence, $\{\gamma_k\}_{k=1}^T$ is the step size.

1: **for** $k = 1, \cdots, T$ **do**
2: Compute $x_{k+1} = \arg\min_{u\in\mathcal{C}}\{\langle \nabla F(x_k) + \epsilon_k, u\rangle + \frac{1}{\gamma_k}V(u, x) + r(u)\}$, where $\epsilon_k \sim N(0, \sigma^2 I_p)$.
3: **end for**
4: **return** $x_R \in \{x_1, \cdots, x_T\}$ where R is uniformly sampled from $\{1, \cdots, T\}$.

Theorem 5. Let $\mathcal{C}$ be a bounded closed centrally symmetric convex set. Then, under **Assumption** 1 and for any $0 < \epsilon < c_2 T, \delta > 0$, **DP-PMD** (Algorithm 3) is (ϵ, δ)-differentially private if σ^2 is chosen as in (3). Moreover, if taking $\{\gamma\}_{k=1}^T = \frac{1}{2L\|\mathcal{C}\|_2^2}$ and $T = O(\frac{n\epsilon\|\mathcal{C}\|_2}{\sqrt{(\|\mathcal{C}\|_2^2 + G_{\mathcal{C}}^2)\ln(\frac{1}{\delta})}})$, the following holds

$$\mathbb{E}\|g_{\mathcal{C},R}\|_2 \leq O(\frac{\|\mathcal{C}\|_2^{\frac{3}{2}}\sqrt[4]{(\|\mathcal{C}\|_2^2 + G_{\mathcal{C}}^2)\ln(\frac{1}{\delta})}}{\sqrt{n\epsilon}}), \quad (7)$$

where $g_{\mathcal{C},k} = \frac{1}{\gamma_k}(x_k - x_{k+1})$.

Remark 3. If $\|\mathcal{C}\|_2 \leq 1$, $\mathcal{G}_{\mathcal{C}} = o(\sqrt{p})$, from Theorems 5 and 2, we can see that the utility bound of (7) is always less than (4). One of the main reasons for us to have Theorem 5 is the fact that we can exploit the geometric structure of the problem (by Remark 2 and the Mirror Descent). Moreover, when we ignore the terms related to $\mathcal{C}$, the upper bounds in Theorem 5 and 4 actually achieve the same upper bound, although the utilities are measured quite differently.

Further Reducing the Utility

Theorem 5 allows us to bound the utility quite well for the general non-convex case. However, as shown in (Talwar, Thakurta, and Zhang 2015; Kifer, Smith, and Thakurta 2012), the utility can be further reduced for some convex loss functions to a level depending only on $\log(p)$, rather than $G_{\mathcal{C}}$ or p. This inspires us to ask whether there is any special case for non-convex loss functions to achieve the same. In this section, we give an affirmative answer to this by showing (in **Algorithm 4**) that there is indeed a case where the Frank-Wolf gap depends only on $\log(p)$. We consider problem (1) without the regularizer term.

Assumption 4. $F(x)$ is assumed to be differentiable and L-smooth over x w.r.t ℓ_2-norm, and $f(\cdot, z)$ is assumed to be G-Lipschitz over x with respect to ℓ_1-norm for all $z \in \mathcal{X}$. $\mathcal{C} \subseteq \mathbb{R}^p$ is assumed to be a closed convex set. Furthermore, $\mathcal{C}$ is assumed to be the convex hull of some finite set A, i.e., $\mathcal{C} = \text{Conv(A)}$ and bounded. (For example, $\mathcal{C}$ could be a polytope.)

Algorithm 4 DP-FW-L1($F, x_1, T, \sigma, \{\gamma_t\}_{t=1}^T$)

Input: T is the iteration number and x_1 is the initial point. $\{\gamma_t\}_{t=1}^T$ is the step size. $\mathcal{C} \subseteq \mathbb{R}^p$ be the convex hull of a compact set $A \subseteq \mathbb{R}^p$.

1: **for** $t = 1, \cdots, T$ **do**
2: Use exponential mechanism $\mathcal{M}(D, u, \mathcal{R})$, where $\mathcal{R} = A$, $u(D, s) = -\langle s, \nabla F(x_t, D)\rangle$, to ensure $(\frac{\epsilon}{\sqrt{8T\ln(\frac{1}{\delta})}}, 0)$-differentially private. Denote the output as $\tilde{x_t}$.
3: Compute $x_{t+1} = (1 - \gamma_t)x_t + \gamma_t\tilde{x_t}$.
4: **end for**
5: **return** $x_R \in \{x_1, \cdots, x_T\}$ where R is uniformly sampled from $\{1, 2\cdots, T\}$.

Theorem 6. Assume A is a finite set. Then, for any $\epsilon, \delta > 0$, **DP-FW-L1** (Algorithm 4) ensures (ϵ, δ)-differentially private. Furthermore, if we set $T = O(\frac{n\epsilon}{\sqrt{\ln(\frac{1}{\delta})}\ln(|A|n/\eta)})$ and $\{\gamma_t\}_{t=1}^T = \sqrt{\frac{2}{T\|\mathcal{C}\|_2^2}}$. Then with probability at least $1 - \eta$, the following holds

$$\mathbb{E}[\mathcal{G}_R] \leq O(\frac{\|\mathcal{C}\|_1\sqrt[4]{\ln(\frac{1}{\delta})}\sqrt{\ln\frac{n|A|}{\eta}}}{\sqrt{n\epsilon}}), \quad (8)$$

where $\mathcal{G}_t = \max_{v\in\mathcal{C}}\langle -\nabla F(x_t), v - x_t\rangle$.

Corollary 1. If $\mathcal{C}$ is an ℓ_1-norm ball or a simplex in $\mathbb{R}^p$, then we can see that A is the set of the vertices of $\mathcal{C}$, in this case, the Frank-Wolf gap in (8) is $\mathbb{E}\mathcal{G}_R = O(\frac{\sqrt[4]{\ln(\frac{1}{\delta})}\sqrt{\ln(np)}}{\sqrt{n\epsilon}})$.

Note that since A in step 2 of Algorithm 4 is finite and u is a linear function, it could run in $O(|A|p)$ time; also we can use Report-Noisy-Max in (Dwork, Roth, and others 2014) instead of the exponential mechanism, see (Lou and Cheung 2018) for details. The above bound could be the smallest among all the results presented so far. For example, when $\mathcal{C}$ contains the unit Euclidean ball, $G_{\mathcal{C}} = \Omega(\sqrt{p})$. Thus, all the previous results depend on p while (8) depends only on $\log(p)$.

Experiments

In this section, we experimentally study the behavior of differentially private gradient descent method with non-convex loss functions. Particularly, we will consider the sigmoid function as the loss and with an ℓ_1-regularizer, *i.e.*

$$\min_{\theta\in\mathbb{R}^p} F^r(\theta, D) = \frac{1}{n}\sum_{i=1}^n \frac{1}{1 + \exp(-y_i\theta^T x_i)} + \frac{\lambda}{2}\|\theta\|_1.$$

Experiment Settings

Due to the hardness of computing the Frank-Wolfe gap, we will measure the ℓ_2-norm of the generalized projected gradient. W set $\lambda = 0.01$, and test our algorithms on both synthetic and real world datasets. The synthetic dataset is generated by $\Pr(y_i|x_i) = \frac{1}{1+\exp(-y_i\theta^* x_i)}$. That is, we first randomly choose θ^*, and then for each random vector x_i, we set $y_i = 1$ if $\frac{1}{1+\exp(-y_i\theta^* x_i)} > \frac{1}{2}$. The size of the synthetic dataset is 10000×100. We use Covertype dataset (Dheeru and Karra Taniskidou 2017) as the real world dataset, which is used for binary classification. We choose 2×10^5 samples for optimization, and thus the size is $(2 \times 10^5, 54)$. We normalize all the above datasets so that the loss function is 1-Lipschitz.

For the parameter of differential privacy, we choose $\epsilon = 0.1, 0.5, 2, 5$, respectively, with fixed $\delta = 0.001$. For the optimization algorithm, the initial vector is selected randomly. Also since the step size does not affect differential privacy, we use the way of choosing step size in http://cvxr.com/tfocs/. All the experiments are performed on MATLAB.

For our methods, we use the practical way proposed in the previous section. The methods to be compared are the ones in (Zhang et al. 2017) and in (Wang, Ye, and Xu 2017), where the constant behind the noise in (Wang, Ye, and Xu 2017) is determined by the approach in (Abadi et al. 2016).

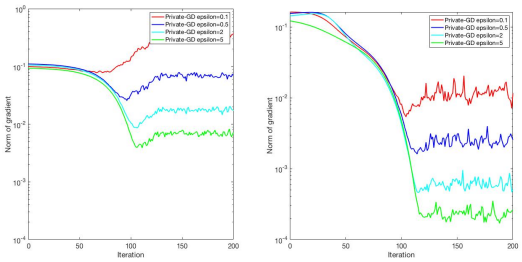

Figure 1: Results of DP-GD with different level of ϵ. The left one is for the synthetic dataset, and the right one is for Covertype dataset; all iteration numbers $T = 200$.

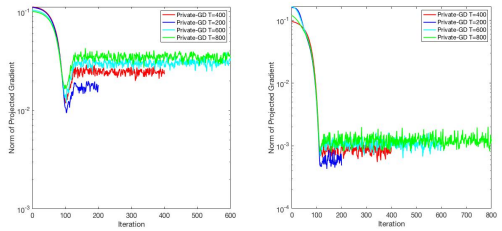

Figure 2: Result of DP-GD with different iteration number T. The left one is for the synthetic dataset, and the right one is for Covertype dataset; all the privacy parameter $\epsilon = 2$.

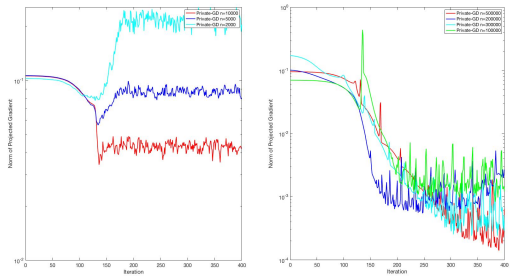

Figure 3: Result of DP-GD with different sample size n. The left one is for the synthetic dataset, and the right one is for Covertype dataset; all the privacy parameter $\epsilon = 2$ and iteration $T = 400$.

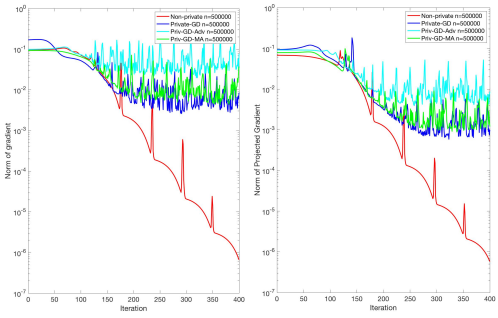

Figure 4: Results on the Covertype dataset of different methods of DP-GD with different privacy levels. The left one is for the case of $\epsilon = 0.5$, and the right one is for the case of $\epsilon = 2$.

Experiments Results

We can see the results from Figure 1 that with lower privacy, which means that ϵ is larger, there is less error, i.e. the norm of gradient will be smaller. Also, compared with the size of the two datasets, we can see that with larger n, the error is smaller.

Figure 2 studies the norm of gradient with different iteration number T, which affects the magnitude of noise added in each iteration; here we fix $\epsilon = 2$. We can see that unlike the effect of ϵ in Figure 1, the effect of iteration number is less, since all the upper bounds are independent of T, although it indeed increases the magnitude of noise.

Figure 3 studies the error with respect to the sample size n with fixed T and ϵ. As shown in our analysis, the norm of gradient is smaller with larger sample size, and the results in Figure 3 for both synthetic and real world dataset confirm it.

Figure 4 studies the norm of gradient with respect to different methods of DP-GD. We can see that whenever ϵ is large or small, our method is better than the previous ones. Also, as mentioned before, our method is more practical and simpler than the previous ones.

References

Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H. B.; Mironov, I.; Talwar, K.; and Zhang, L. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 308–318. ACM.

Bassily, R.; Smith, A.; and Thakurta, A. 2014. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*, 464–473. IEEE.

Bun, M., and Steinke, T. 2016. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, 635–658. Springer.

Chaudhuri, K., and Monteleoni, C. 2009. Privacy-preserving logistic regression. In *Advances in Neural Information Processing Systems*, 289–296.

Chen, Y.; Su, L.; and Xu, J. 2017. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 1(2):44.

Dheeru, D., and Karra Taniskidou, E. 2017. UCI machine learning repository.

Dwork, C.; McSherry, F.; Nissim, K.; and Smith, A. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, 265–284. Springer.

Dwork, C.; Roth, A.; et al. 2014. 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 *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, 51–60. IEEE.

Dwork, C. 2008. Differential privacy: A survey of results. In *International Conference on Theory and Applications of Models of Computation*, 1–19. Springer.

Fredrikson, M.; Lantz, E.; Jha, S.; Lin, S.; Page, D.; and Ristenpart, T. 2014. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In *USENIX Security Symposium*, 17–32.

Ghadimi, S., and Lan, G. 2016. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming* 156(1-2):59–99.

Ghadimi, S.; Lan, G.; and Zhang, H. 2016. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming* 155(1-2):267–305.

Goodfellow, I.; Bengio, Y.; Courville, A.; and Bengio, Y. 2016. *Deep learning*, volume 1. MIT press Cambridge.

Kifer, D.; Smith, A.; and Thakurta, A. 2012. Private convex empirical risk minimization and high-dimensional regression. In *COLT*, 25–1.

Lacoste-Julien, S. 2016. Convergence rate of frank-wolfe for non-convex objectives. *arXiv preprint arXiv:1607.00345*.

Lee, J., and Kifer, D. 2018. Concentrated differentially private gradient descent with adaptive per-iteration privacy budget. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1656–1665. ACM.

Lou, J., and Cheung, Y. 2018. Uplink communication efficient differentially private sparse optimization with feature-wise distributed data. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018*.

Mei, S.; Bai, Y.; and Montanari, A. 2016. The landscape of empirical risk for non-convex losses. *arXiv preprint arXiv:1607.06534*.

Nesterov, Y. 2013. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media.

Nguyen, T., and Sanner, S. 2013. Algorithms for direct 0–1 loss optimization in binary classification. In *International Conference on Machine Learning*, 1085–1093.

Reddi, S. J.; Sra, S.; Póczos, B.; and Smola, A. 2016. Stochastic frank-wolfe methods for nonconvex optimization. In *Communication, Control, and Computing (Allerton), 2016 54th Annual Allerton Conference on*, 1244–1251. IEEE.

Talwar, K.; Thakurta, A.; and Zhang, L. 2014. Private empirical risk minimization beyond the worst case: The effect of the constraint set geometry. *arXiv preprint arXiv:1411.5417*.

Talwar, K.; Thakurta, A.; and Zhang, L. 2015. Nearly optimal private lasso. In *Advances in Neural Information Processing Systems*, 3025–3033.

Wang, D.; Gaboardi, M.; and Xu, J. 2018. Empirical risk minimization in non-interactive local differential privacy revisited. *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, 3-8 December 2018, Montreal, QC, Canada*.

Wang, D.; Smith, A.; and Xu, J. 2019. Differentially private empirical risk minimization in non-interactive local model via polynomial of inner product approximation. In *Algorithmic Learning Theory, ALT 2019, 22-24 March 2019, Chicago, IL, USA*.

Wang, D.; Ye, M.; and Xu, J. 2017. Differentially private empirical risk minimization revisited: Faster and more general. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, 2719–2728.

Zhang, J.; Zheng, K.; Mou, W.; and Wang, L. 2017. Efficient private ERM for smooth objectives. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, 3922–3928.