Differentially Private Empirical Risk Minimization with Non-convex Loss Functions

Di Wang [1] **Changyou Chen** [1] **Jinhui Xu** [1]

Abstract

We study the problem of Empirical Risk Minimization (ERM) with (smooth) non-convex loss functions under the differential-privacy (DP) model. We first study the expected excess empirical (or population) risk, which was primarily used as the utility to measure the quality for convex loss functions. Specifically, we show that the excess empirical (or population) risk can be upper bounded by $\tilde{O}(\frac{d\log(1/\delta)}{\log n\epsilon^2})$ in the (ϵ,δ)-DP settings, where n is the data size and d is the dimensionality of the space. The $\frac{1}{\log n}$ term in the empirical risk bound can be further improved to $\frac{1}{n^{\Omega(1)}}$ (when d is a constant) by a highly non-trivial analysis on the time-average error. To obtain more efficient solutions, we also consider the connection between achieving differential privacy and finding approximate local minimum. Particularly, we show that when the size n is large enough, there are (ϵ,δ)-DP algorithms which can find an approximate local minimum of the empirical risk with high probability in both the constrained and non-constrained settings. These results indicate that one can escape saddle points privately.

1. 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 (DP) has received a great deal of attentions in the past decade (Dwork et al., 2006). DP works by injecting random noise into the statistical results obtained from sensitive data so that the distribution of the perturbed results is insensitive to any single-record change in the original dataset. A number of methods with DP guarantees have been discovered and recently adopted in industry (Near, 2018; Erlingsson et al., 2014).

As a fundamental supervised-learning model in machine learning, Empirical Risk Minimization (ERM) has been extensively studied in recent years. Previous research on DP-ERM (*i.e.,* DP version of ERM) mainly focuses on convex loss functions (Chaudhuri & Monteleoni, 2009) (see more details in the Related Work section). However, empirical studies have revealed that non-convex loss functions typically achieve better classification accuracy than the convex ones (Mei et al., 2018). Furthermore, recent developments in deep learning (Goodfellow et al., 2016) also suggest that 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 our focus from convex to non-convex loss functions. So far, very few papers (Zhang et al., 2017; Wang et al., 2017; Wang & Xu, 2019) have considered DP-ERM with non-convex loss functions. This is probably due to the fact that finding the global minimum of a non-convex loss function is NP-hard. Almost all of them (Zhang et al., 2017; Wang et al., 2017; Wang & Xu, 2019) used the ℓ_2 gradient-norm of a private estimator, *i.e.*, $\|\nabla\hat{L}(w^{\text{priv}})\|_2$, to measure the error bound. Despite some obvious advantages with such an approach, it also endows a few limitations: 1) although (Zhang et al., 2017; Wang et al., 2017; Wang & Xu, 2019) showed that the gradient norm tends to 0 as n goes to infinity, there is no guarantee that such an estimator will be close to any non-degenerate local minimum (Agarwal et al., 2017); 2) the gradient-norm estimator is not always consistent with the excess empirical (population) risk of the loss function, *i.e.*, $\hat{L}(w^{\text{priv}}) - \hat{L}(w^*)$, where w^* is the optimal solution (Bassily et al., 2014; Chaudhuri et al., 2011). Thus, it is difficult to compare the obtained solution with either the global or local minima. This propels us to the following interesting question.

Can the excess empirical (population) risk be used to measure the error of non-convex loss functions under differential privacy?

Due to the intrinsic challenge of finding global minima, re-

[1]Department of Computer Science and Engineering, State University of New York at Buffalo, Buffalo, USA. Emails: {*dwang45,changyou,jinhui*}*@buffalo.edu*. Correspondence to: Di Wang <dwang45@buffalo.edu>.

Proceedings of the 36^{th} *International Conference on Machine Learning*, Long Beach, California, PMLR 97, 2019.

cent research on deep neural network training (Ge et al., 2018; Kawaguchi, 2016) and many other machine learning problems (Ge et al., 2015; 2016; 2017; Bhojanapalli et al., 2016) has shifted their attentions to obtaining local minima. It has been shown that fast convergence to a local minimum is actually sufficient for such tasks, but convergence to critical points (*i.e.,* points with vanished gradients) is often not acceptable. This motivates us to investigate efficient techniques for finding local minima. However, as shown in (Anandkumar & Ge, 2016), computing a local minimum could be quite challenging as it is actually NP-hard for non-convex functions. Fortunately, many non-convex functions in machine learning are known to be strict saddle (Ge et al., 2015), meaning that a second-order stationary point (or approximate local minimum) is sufficient to obtain a close enough point to some local minimum.

To find (approximate) local minima, Ge et al. (2015) have recently proposed an elegant approach using a noisy version of gradient descent. Their method adds some scaled Gaussian noise in each iteration to the gradient before updating, rather than directly using SGD. Such a way of finding local minima resembles the idea used by the DP community for achieving differential privacy for SGD (Bassily et al., 2014; Wang et al., 2017; Wang & Xu, 2019). In DP-SGD, some Gaussian noise is also added to the gradient in each iteration to make it (ϵ, δ)-DP. Although these two algorithms focus on different perspectives (one for escaping saddle points while the other for making the algorithm DP), they both inject random Gaussian noise to the gradients in each iteration. This naturally leads us to another question:

Can we find some approximate local minimum which escapes saddle points, while keeping the algorithm (ϵ, δ)-differentially private?

In this paper, we study the above two questions and give positive answer to each of them. Below is a summary of our contributions.

- We first propose an (ϵ, δ)-DP algorithm, named DP-GLD (Algorithm 1), and prove that its excess empirical (or population) risk is upper bounded by $\tilde{O}(\frac{d \log(1/\delta)}{\log n \epsilon^2})$ when $\log n \geq O(d)$, where n is the data size and d is the dimensionality of the space. Our techinique is based on some recent developments in Bayesian learning and (stochastic) Gradient Langevin Dynamics (Raginsky et al., 2017; Chen et al., 2015; Xu et al., 2018a; Tzen et al., 2018). Interestingly, we show that the $\frac{1}{\log n}$ term in the empirical risk bound can be further improved to $\frac{1}{n^{\Omega(1)}}$ by a highly non-trivial analysis on the time-average error of a dynamic system.

- We also show that when the data size n is large enough, there exist polynomial-time [1] (ϵ, δ)-DP algorithms that can find an α-approximate local minimum of the empirical risk in both constrained and non-constrained settings. To the best of our knowledge, this is the first result that reveals a connection between differential privacy and saddle-point escaping.

Due to space limit, all proofs and some related background are left to the Supplementary Material.

2. Related Work

DP-ERM is a fundamental problem in machine learning and differential privacy. There are quite a number of results on differentially-private ERM with convex loss functions, which investigate the problem from different perspectives. For example, (Wang et al., 2018; 2019) considered ERM in the non-interactive local model. The problem has been well-studied under the central model both theoretically and practically (Chaudhuri & Monteleoni, 2009; Chaudhuri et al., 2011; Bassily et al., 2014; Wang et al., 2017; Zhang et al., 2017; Kifer et al., 2012), as well as in high dimensions (Talwar et al., 2015; 2014; Kasiviswanathan & Jin, 2016).

For general non-convex loss functions, existing results have mainly adopted gradient norm as a measurement to bound the error of a private estimator (Zhang et al., 2017; Wang et al., 2017; Wang & Xu, 2019). Since gradient norm does not guarantee quality of solutions in general, it is thus not very meaningful to compare these bounds with the ones in this paper. Excess empirical (or population) risk based utility has been applied only to some special non-convex loss functions. For example, Wang et al. (2017) showed a near optimal bound for some special non-convex loss functions satisfying the Polyak-Lojasiewicz condition. (Balcan et al., 2018) studied the problem of optimizing privately piecewise Lipschitz functions in online settings. However, the loss functions should satisfy the dispersion condition, which is quite different from ours; thus their result is also not comparable with ours.

Previous works on the DP version of SGLD have focused on Bayesian learning, such as (Wang et al., 2015; Li et al., 2019), which differ from our work considerably. Firstly, our work mainly focuses on achieving (ϵ, δ)-DP for ERM with non-convex loss functions, and on measuring the error of a private estimator with respect to the global or local minima. Secondly, existing works assume that the temperature-parameter β in the gradient Langevin dynamics is one or some constant, while β in our problem is not even a constant, making the analysis significantly more challenging in our work than in previous ones (see Remark 2 for more details).

[1] For the constrained case, polynomial-time solutions are only for some specified sets, see Remark 3 for details.

3. Preliminaries

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

Definition 1 (Differential Privacy (Dwork et al., 2006))**.** A randomized algorithm $\mathcal{A}$ is (ϵ, δ)-differentially private if for all neighboring datasets D, D', and for all events S in the output space of $\mathcal{A}$, we have $\Pr(\mathcal{A}(D) \in S) \leq e^{\epsilon}\Pr\mathcal{A}(D') \in S) + \delta$. When $\delta = 0$, $\mathcal{A}$ is called ϵ-differentially private.

Problem Setting (Bassily et al., 2014) Given a dataset $D = \{z_1 = (x_1, y_1), z_2 = (x_2, y_2) \cdots, z_n = (x_n, y_n)\}$ from a data universe $\mathcal{Z}$ and a closed convex set $\mathcal{C} \subseteq \mathbb{R}^d$, where $\{x_i\}_{i=1}^n$ are feature vertors and $\{y_i\}_{i=1}^n$ are labels or responses. DP-ERM is to find $w^{\text{priv}} \in \mathcal{C}$ by minimizing the empirical risk defined as $\hat{L}^r(w, D) \triangleq \hat{L}(w, D) + r(w) \triangleq \frac{1}{n}\sum_{i=1}^n \ell(w, z_i) + r(w)$, with the guarantee of being differentially private (defined below). Here ℓ is the loss function; and $r(\cdot)$ is some simple (non)-smooth convex regularizer. The utility of an algorithm is measured by the **expected excess empirical risk** (which we call *empirical risk*), *i.e.*,

$$\text{Err}_D^r(w^{\text{priv}}) = \mathbb{E}[\hat{L}^r(w^{\text{priv}}, D)] - \min_{w\in\mathcal{C}} \hat{L}^r(w, D),$$

where the expectation is taking over the randomness of the algorithm. When the data are drawn i.i.d. from an unknown underlying distribution $\mathcal{P}$ on $\mathcal{Z}$, we also seek to minimize the population risk, defined as $L_{\mathcal{P}}^r(w) = \mathbb{E}_{z\in\mathcal{P}}[\ell(w, z)] + r(w)$. The **expected excess population risk** (which we call *population risk*) becomes [2]

$$\text{Err}_{\mathcal{P}}^r(w^{\text{priv}}) = \mathbb{E}[L_{\mathcal{P}}^r(w^{\text{priv}})] - \min_{w\in\mathcal{C}} L_{\mathcal{P}}^r(w).$$

Definition 2 (Lipschitz function over w)**.** A loss function $\ell : \mathcal{C} \times \mathcal{Z} \mapsto \mathbb{R}$ is called L-Lipschitz (under ℓ_2-norm) over w, if for any $z \in \mathcal{Z}$ and $w_1, w_2 \in \mathcal{C}$, we have $|\ell(w_1, z) - \ell(w_2, z)| \leq L||w_1 - w_2||_2$.

Definition 3 (Continuous smooth function over w)**.** A loss function $\ell : \mathcal{C} \times \mathcal{Z} \mapsto \mathbb{R}$ is called M-smooth over w with respect to the ℓ_2-norm if for any $z \in \mathcal{Z}$ and $w_1, w_2 \in \mathcal{C}$, we have $||\nabla\ell(w_1, z) - \nabla\ell(w_2, z)||_2 \leq M||w_1 - w_2||_2$ for some positive M. If ℓ is differentiable, this yields

$$\ell(w_1, z) \leq \ell(w_2, z) + \langle\nabla\ell(w_2, z), w_1 - w_2\rangle + \frac{M}{2}||w_1 - w_2||_2^2.$$

Definition 4. A twice-differentiable loss function $\ell : \mathcal{C}\times\mathcal{Z}$ is called ρ-Hessian Lipschitz if for any $z \in \mathcal{Z}$ and $w_1, w_2 \in \mathcal{C}$ we have

$$\|\nabla^2\ell(w_1, z) - \nabla^2\ell(w_2, z)\|_2 \leq \rho\|w_1 - w_2\|_2.$$

Definition 5. For two Borel measures μ, ν on $\mathbb{R}^d$ with finite second moments, the 2-Wasserstein distance, $\mathcal{W}_2(\mu, \nu)$, is defined as: $\mathcal{W}_2(\mu, \nu) := \inf\{(\mathbb{E}\|V - W\|_2^2)^{\frac{1}{2}} : \mu = \mathcal{L}(V), \nu = \mathcal{L}(W)\}$, where the infinitum is taken over all the random couples (V, W) whose values are taken in $\mathbb{R}^d \times \mathbb{R}^d$ with marginals $V \sim \mu$ and $W \sim \nu$. $\mathcal{L}(V)$ means the probability law of the random vector V.

Definition 6 (Gaussian Mechanism)**.** Given any function $q : \mathcal{Z}^n \to \mathbb{R}^d$, the Gaussian Mechanism is defined as: $\mathcal{M}_G(D, q, \epsilon) = q(D) + Y$, where Y is drawn from a Gaussian Distribution $\mathcal{N}(0, \sigma^2 I_d)$ with $\sigma \geq \frac{\sqrt{2\ln(1.25/\delta)}\Delta_2(q)}{\epsilon}$, and $\Delta_2(q)$ 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 preservers (ϵ, δ)-differential privacy.

The moments accountant proposed in (Abadi et al., 2016) is a method to accumulate the privacy cost, producing tighter bounds for ϵ and δ. For example, when the Gaussian Mechanism is used on (stochastic) gradient descent, one can save a factor of $\sqrt{\ln(T/\delta)}$ in the asymptotic bound of standard deviation of noise, compared with the advanced composition theorem in (Dwork et al., 2010).

Lemma 1 ((Abadi et al., 2016))**.** For an L-Lipschitz loss function, there exist constants c_1 and c_2 so that given the sampling probability $q = l/n$, the number of steps T and any $\epsilon < c_1 q^2 T$, a DP stochastic gradient algorithm with batch size l, which injects zero-mean Gaussian noise with standard deviation $L\sigma$ 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 7 (Exponential Mechanism (McSherry & Talwar, 2007))**.** 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, $\Delta = \max_{r\in\mathcal{R}} \max_{D\sim D'} |u(D, r) - u(D', r)|$. The exponential mechanism $\mathcal{M}(D, x, R)$ is defined as selecting and outputting an element $r \in \mathcal{R}$ with probability proportional to $\exp(\frac{\epsilon u(D,r)}{2\Delta u})$, which preserves ϵ-differential privacy.

4. Excess Risk of DP-ERM with Non-convex Loss Functions

We make the following assumptions in this section unless specified otherwise.

Assumption 1. 1. The hypothesis space $\mathcal{C} = \mathbb{R}^d$, regularizer is ℓ_2 norm, *e.g.*, $r(\cdot) = \frac{\lambda}{2}\|\cdot\|^2$ for some $\lambda > 0$.

2. For any $z \in \mathcal{Z}$, $\ell(\cdot, z)$ is L-Lipschitz, and $\ell(0, z) \leq A$.

[2] If there is no regularizer, we will simply denote as Err_p.

3. For each $z \in \mathcal{Z}$, $\ell(\cdot, z)$ is twice differentiable and is M-smooth.

These assumptions are quite standard in the DP-ERM literature with convex loss functions (Chaudhuri et al., 2011; Chaudhuri & Monteleoni, 2009). For some non-convex loss functions such as the sigmoid function, it is easy to see that these assumptions are satisfied. For convenience, we assume that A, λ, L, M are all constants, which will be omitted in the big O notation. Also the big $\tilde{O}$ terms omit the log terms.

We first review Gradient Langevin Dynamics (GLD), a popular generalization of the gradient descent algorithm. For ERM, the GLD algorithm executes the following recursion for w at iteration k:

$$w_k = w_{k-1} - \eta_{k-1}\nabla \hat{L}^r(w_{k-1}, D) + \sqrt{\frac{2\eta_{k-1}}{\beta}}\xi_{k-1}, \quad (1)$$

where ξ_{k-1} is a standard d-dimensional Gaussian random vector, η_{k-1} is the step size and $\beta > 0$ is the inverse temperature parameter. Actually, GLD can be viewed as a discrete-time approximation of a continuous-time Langevin diffusion, described by the following stochastic differential equation (SDE):

$$dW_t = -\nabla \hat{L}^r(W_t, D)dt + \sqrt{2\beta^{-1}}dB_t, \quad (2)$$

where $(B_t)_{t\geq 0}$ is a standard Brownian motion. It has been shown that the distribution of diffusion process in (2) converges to its stationary distribution, *i.e.* the Gibbs measure $\pi(dw) \propto \exp(-\beta \hat{L}^r(w, D))$ (Chiang et al., 1987). Moreover, when $\beta \to \infty$, the distribution concentrates around the minimizer of $\hat{L}^r(w, D)$. By choosing the step size η properly, GLD can maintain differential privacy, as described in Algorithm 1.

Algorithm 1 DP-GLD

Input: T is the iteration number. ϵ, δ are privacy parameters.

1: Choose an arbitrary point w_0 from distribution density $p_0(w)$ or fix the initial point w_0.
2: Denote $\eta = \frac{cn^2\epsilon^2}{L^2\beta T \log(1/\delta)}$, where $c = \frac{1}{c_2^2}$ is from Lemma 1.
3: **for** $k = 1, 2, \cdots, T$ **do**
4: $\quad w_k = w_{k-1} - \eta\nabla \hat{L}(w_{k-1}, D) + \sqrt{\frac{2\eta}{\beta}}\xi_{k-1}$, where $\xi_{k-1} \sim \mathcal{N}(0, I_d)$
5: **end for**
6: Return w_T or randomly sample $j \in [T]$ and return w_j.

It can be shown that Algorithm 1 ensures DP under certain conditions, as stated in Theorem 1.

Theorem 1. There exist constant numbers c_1 and c_2, such that for any $0 < \epsilon < c_1 T$ and $0 < \delta < 1$, Algorithm 1 is (ϵ, δ)-differentially private.

Our idea for proving an upper bound of the excess risk of ERM is based on the analysis of the convergence rate of GLD as in (Dalalyan, 2017; Dalalyan & Karagulyan, 2019; Raginsky et al., 2017). Let μ_k be the probability law of w_k in (1), and $\nu_{k\eta}$ the law of $W_{k\eta}$ in (2). Our main step is to analyze the 2-Wasserstein distance $\mathcal{W}_2(\mu_k, \pi)$, which can be decomposed into $\mathcal{W}_2(\mu_k, \nu_{k\eta})$ and $\mathcal{W}_2(\nu_{k\eta}, \pi)$. The key observation of our analysis is that in DP-GLD with $k = T$, ηT is a fixed number according to Algorithm 1, *i.e.*, $\eta T = \Theta(\frac{n^2\epsilon^2}{\beta \log(1/\delta)L^2})$. This means that the term $\mathcal{W}_2(\nu_{T\eta}, \pi)$ is always fixed, no matter how large T is. For the $\mathcal{W}_2(\mu_T, \nu_{T\eta})$ term, since w_T is a discretized version of $W_{T\eta}$, when η approaches 0, $\mathcal{W}_2(\mu_T, \nu_{T\eta})$ will also approach 0. Thus, it appears that the best of what we can do for bounding $\mathcal{W}_2(\mu_T, \pi)$ in DP-GLD is to bound it by $\mathcal{W}_2(\mu_T, \nu_{T\eta})$, *i.e.*,

$$\lim_{T\to\infty} \mathcal{W}_2(\mu_T, \pi) \leq \mathcal{W}_2(\nu_{T\eta}, \pi)\,.$$

The above distance can be bounded as shown in a recent work by using Logarithmic Sobolev inequality (Raginsky et al., 2017). For our problem, Theorem 2 extends the results by adapting recent non-asymptotic GLD theory (Raginsky et al., 2017; Xu et al., 2018a) and giving an upper bound of the excess risk for some initial points.

Theorem 2. Under the conditions of Theorem 1, if take $T \geq \Theta(\frac{(M+\lambda)^2 n^2\epsilon^2}{\lambda\beta \log(1/\delta)L^2})$ and $\beta \geq \max\{\frac{4}{\lambda}, d\}$ in Algorithm 1, and assume that the probability law, μ_0, of the initial hypothesis w_0 has a bounded and strictly positive density function w.r.t. Lebesgue measure on $\mathbb{R}^d$, and $k_0 = \log \int_{\mathbb{R}^d} e^{\|w\|^2} p_0(w)dw < \infty$, then the population risk at w_T is bounded by

$$\mathrm{Err}^r_{\mathcal{P}}(w_T) \leq O\Big(\frac{n^{\frac{5}{2}}\epsilon^{\frac{5}{2}}}{\beta^{\frac{3}{4}}T^{\frac{1}{4}}\log(1/\delta)} + \exp(O(\beta))\times \exp[-\frac{n^2\epsilon^2}{\beta\log(1/\delta)\exp(O(\beta))}] + \frac{\exp(O(\beta))}{n} + \frac{d\log(\beta)}{\beta}\Big). \quad (3)$$

The above bound implies that $\lim_{T\to\infty} \mathrm{Err}^r_{\mathcal{P}}(w_T) \leq O\Big(\frac{\exp(O(\beta))\log(1/\delta)}{n^2\epsilon^2} + \frac{\exp(O(\beta))}{n} + \frac{d\log(\beta)}{\beta}\Big)$ by the constraint on T.

For the empirical risk, we have

$$\mathrm{Err}^r_{D}(w_T) \leq O\Big(\exp(O(\beta))\exp[-\frac{n^2\epsilon^2}{\beta\log(1/\delta)\exp(O(\beta))}] + \frac{n^{\frac{5}{2}}\epsilon^{\frac{5}{2}}}{\beta^{\frac{3}{4}}T^{\frac{1}{4}}\log(1/\delta)} + \frac{d\log(\beta)}{\beta}\Big). \quad (4)$$

Remark 1. We can see from Theorem 2 that the excess risk is only meaningful when $\beta \geq O(d)$. If set $\beta = O(\log n)$, or equivalently $\log n \geq O(d)$, both the excess *population*

and empirical risks are bounded by $\tilde{O}(\frac{\log(1/\delta)}{n\epsilon^2} + \frac{d}{\log(n)}) = \tilde{O}(\frac{\log(1/\delta)d}{\log(n)\epsilon^2})$ when $T \to \infty$. These bounds are larger than the ones for convex loss functions, which are $O(\frac{\sqrt{d\log(1/\delta)}}{n\epsilon})$ and $O(\frac{d\log(1/\delta)}{n^2\epsilon^2})$ for population and empirical risks, respectively (Bassily et al., 2014).

Next, we improve the bounds in Theorem 2 by using a finer analysis of the time-average error for the SDE (2). We show that Algorithm 1 achieves a lower error bound in term of n, *i.e.*, $O(\frac{1}{n^{\Omega(1)}})$ instead of $O(\frac{1}{\log n})$ for the empirical risk when fixing the initial point for w.

Theorem 3. With the same assumption as in Theorem 2 and a fixed initial point for w, if we return w_j in Algorithm 1 instead of w_T, where j is uniformly sampled from $\{1, \cdots, T\}$, then the empirical risk is bounded, for sufficiently large T, by:

$$\mathrm{Err}_D^r(w_j) \le O\big(C[\frac{\beta^2\log(1/\delta)}{n^2\epsilon^2} + \frac{n^2\epsilon^2}{T\beta^2\log(1/\delta)}] + \frac{d}{\beta}\log(\beta)\big), \tag{5}$$

where $C = C(d,\beta)$ is a function of d, β. Moreover, the bound is polynomially depending on β (assuming that d is a constant) with degree independent of d. In other words, if β satisfies $\Theta(C\frac{\beta^2\log(1/\delta)}{n^2\epsilon^2}) = \Theta(\frac{d}{\beta}\log(\beta))$ in (5), then there exists a constant $0 < \tau < 1$, such that

$$\lim_{T\to\infty} \mathrm{Err}_D^r(w_T) \le \tilde{O}\big(\frac{C_0(d)\log(1/\delta)}{n^\tau\epsilon^\tau}\big), \tag{6}$$

where $C_0(d)$ is a function of d.

Remark 2. Theorem 3 is a significant improvement over Theorem 2, which is derived based on a novel and non-trivial analysis on the time-average error of SDEs. Specifically, three points are worth emphasizing: 1) Although the time-average-error analysis of an SDE has been studied, for example in (Vollmer et al., 2016; Chen et al., 2015), the non-asymptotic bounds in those results cannot be applied directly to our problem. This is because β in those results is assumed to be a constant. However, in our problem β is not even a constant, as it can be seen from (3) and (4). Furthermore, those results are based on the boundedness assumption on the solution of a Poisson equation (*e.g.*, Assumption 1 in (Chen et al., 2015) and Theorem 9 in (Vollmer et al., 2016)), which is too strong for our problem. Note that if β were a parameter, the hidden constant C in the bounds of (Vollmer et al., 2016; Chen et al., 2015) would depend on β. Fortunately, through a rather non-trivial analysis, we are able to show that the constant is at most polynomially depending on β. Even though the exact degree of the polynomial is unknown, it is independent of d. 2) Our result is significant in the sense that it provides new bounds for diffusion-based Bayesian sampling such as (Vollmer et al., 2016; Chen et al., 2015), where the dependency on d in their error bounds can be quantified, a key missing piece in previous results. 3) We reveal in Theorem 3 that if a random w_j, instead of the final w_T, is returned, one can improve the term related to n in the empirical risk bound from $1/\log n$ to $n^{-\tau}$. It can be seen from (5) that the relationships between β, d, and the constant C play an important role in proving the bound of the empirical risk. Since we are mainly targeting at the rate in terms of n, it suffices to consider only the relationship between β and C. We leave as an open problem to determine whether it is possible to obtain an even tighter or explicit bound for the empirical risk. The ideal scenario is that C is independent of β. In this case, a better and more accurate bound of $\tilde{O}(C_1(d)/(n\epsilon)^{\frac{2}{3}})$ can be obtained, where $C_1(d)$ is a function of d.

From Theorem 2 and 3, we can see that the error bound for the *excess population risk* in terms of n is $\frac{1}{\log n}$ (see Remark 1), while for the empirical risk it is $\frac{1}{n^\tau}$, where ideally $\tau \le \frac{2}{3}$ (see Remark 2). A natural question is thus to determine whether these bounds are tight. In the following, we first show that for loss functions satisfying Assumption 1, there is an ϵ-DP algorithm whose error bound of the empirical risk is $\tilde{O}(\frac{d}{n\epsilon})$ (and whose time complexity is exponential).

Theorem 4. For any $\beta < 1$, there is an ϵ-differentially private algorithm, whose output w^{priv} satisfies, with probability at least $1-\beta$, $\hat{L}^r(w^{\mathrm{priv}}, D) - \hat{L}^r(w^*, D) \le \tilde{O}(\frac{d}{n\epsilon})$. The time complexity is $O((1 + \frac{2Ln\epsilon}{\lambda d})^d n)$.

Note that since $\Theta(\frac{d}{n\epsilon})$ is the optimal bound for general convex functions (Bassily et al., 2014), our empirical-risk bound of $\tilde{O}(\frac{d}{n\epsilon})$ is thus near optimal.

In general, we can use an α-net and the exponential mechanism to obtain a private estimator, which has an upper bound of $\tilde{O}(\max\{\frac{d}{n\epsilon}, \alpha\})$ for the empirical risk with a time complexity of $O((1 + \frac{2L}{\lambda\alpha})^d n)$. Now consider the case that d is a constant. We can see that for the exponential mechanism, the bound in (6) can be obtained if we take $\frac{1}{\alpha} = O(n^\tau)$. However, in this case, the running time of exponential mechanism is $O(n^{\tau d+1})$ compared to $\tilde{O}(\mathrm{Poly}(n,d))$ with Algorithm 1. Alternatively, the running time for achieving error γ in Algorithm 1 is polynomial in $\frac{1}{\gamma}$, while it is $O((\frac{1}{\gamma}))^d$ with an exponential mechanism (for sufficient large n). This means that Algorithm 1 is much more efficient when d is large.

Next, we consider upper bounding the excess population risk. Instead of determining the optimal bound, we show how to improve the bounds for some specific problems. Particularly, we focus on the generalized linear model with non-convex loss functions and the robust regressions problem with additional assumptions, and present an (ϵ, δ)-DP algorithm for them with population risk $O(\frac{\sqrt[4]{d}}{\sqrt{n\epsilon}})$. Note that

these problems have been extensively studied in literature related to non-convex learning theory, such as (Mei et al., 2018; Foster et al., 2018; Loh & Wainwright, 2013; Lozano et al., 2016). Here, we adopt the same assumptions as in (Foster et al., 2018).

Generalized Linear Model We consider the problem of learning a generalized linear model (GLM) with squared loss. We assume that $\mathcal{X} = \{x \in \mathbb{R}^d | \|x\|_2 \leq 1\}$, $\mathcal{C} = \{w \in \mathbb{R}^d | \|w\|_2 \leq 1\}$ and $\mathcal{Y} = \{0, 1\}$, $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. With a link function σ, GLM endows a loss function: $\ell(w, (x, y)) = (\sigma(\langle w, x\rangle) - y)^2$. We further make the following assumptions on the link function, which includes the sigmoid and probit functions [3].

Assumption 2. Let $S = [-1, 1]$, we assume that

1. $\exists$ constant $C_\sigma \geq 1$ s.t. $\max\{\sigma'(s).\sigma''(s)\} \leq C_\sigma$, for $\forall s \in S$.
2. $\exists$ constant $c_\sigma > 0$ s.t. $\sigma'(s) \geq c_\sigma$, for $\forall s \in S$.
3. There exists some $\|w^*\|_2 \leq 1$ such that $\mathbb{E}[y|x] = \sigma(\langle w^*, x\rangle)$.
4. $|\sigma(s)| \leq B$ for some constant $B > 0$, for $\forall s \in S$.

Robust Regression Let $\mathcal{Z}$ and $\mathcal{C}$ be the same as in GLM, and $\mathcal{Y} = [-Y, Y]$ for some constant Y. For a non-convex positive loss function ψ, the loss of robust regression is defined as $\ell(w, (x, y)) = \psi(\langle x, w\rangle - y)$. We make the following assumptions on ψ, which includes the biweight loss function [4] (Loh & Wainwright, 2013).

Assumption 3. Let $S = [-(1 + Y), (1 + Y)]$.

1. $\exists C_\psi \geq 1$, s.t. $\max\{\psi'(s), \psi''(s)\} \leq C_\psi$, for $\forall s \in S$.
2. $\psi'(\cdot)$ is odd with $\psi'(s) > 0$, for $\forall s > 0$; and $h(s) := \mathbb{E}_\xi[\psi'(s + \xi)]$ satisfies $h'(0) > c_\psi$, where $c_\psi > 0$.
3. There is $w^* \in \mathcal{C}$ such that $y = \langle w^*, x\rangle + \xi$, where ξ is symmetric noise with a zero-mean given x.

Algorithm 2 solves both problems and is motivated by the fact that the population risk satisfies the inequality, $L_\mathcal{P}(w) - L_\mathcal{P}(w^*) \leq \mu\langle \nabla L_\mathcal{P}(w), w - w^*\rangle$, for some constant $\mu > 0$ and $\forall w \in \mathcal{C}$. Thus, it suffices to get an upper bound of $\langle \nabla L_\mathcal{P}(w), w - w^*\rangle$. It turns out that this can be obtained via a DP version of the Frank-wolfe method.

[3] The probit function is $\sigma(s) = \Phi(s)$, where Φ is the Gaussian cumulative distribution function.

[4] For a fixed parameter $c > 0$, the biweight loss is defined as $\psi(s) = \frac{c^2}{6} \cdot \begin{cases} 1 - (1 - (\frac{s}{c})^2)^3, |t| \leq c \\ 1, |t| \geq c. \end{cases}$

Algorithm 2 DP-FW-L2

Input: T is the number of iterations, w_1 is the initial point, and $\{\gamma_t\}_{t=1}^T$ is the step size. ϵ and δ are privacy parameters.

1: **for** $t = 1, \cdots, T$ **do**
2: Compute $v_t = \arg\max_{v \in \mathcal{C}} \langle v, -(\nabla \hat{L}(w_t, D) + \epsilon_t)\rangle$, where $\epsilon_t \sim N(0, \sigma^2 I_d)$ for some σ.
3: $w_{t+1} = w_t + \gamma_t(v_t - w_t)$.
4: **end for**
5: Return $w_R \in \{w_1, \cdots, w_T\}$ such that R is uniformly sampled from $\{1, \cdots, T\}$.

Theorem 5. For the general linear model with Assumption 2, there exist constants c_1 and $c_2 > 0$ such that for any $0 < \epsilon < c_1 T$ and $0 < \delta < 1$, Algorithm 2 is (ϵ, δ)-DP when $\sigma^2 = c_2 \frac{C_\sigma^2 (B+1)^2 T \log \frac{1}{\delta}}{n^2 \epsilon^2}$. Moreover, if taking $\gamma_t = O(\frac{\sqrt[4]{d \ln \frac{1}{\delta}}}{\sqrt{n\epsilon}})$ for all $t \in [1, \cdots, T]$ with $T = O(\frac{n\epsilon}{\sqrt{d \ln \frac{1}{\delta}}})$, we have $\text{Err}_\mathcal{P}(w_R) \leq O\big(\frac{\sqrt[4]{d \ln \frac{1}{\delta}}}{\sqrt{n\epsilon}}\big)$, where the big-$O$ notations omit other terms.

For the case of robust regression with Assumption 3, if we take $\sigma^2 = c_2 \frac{C_\psi^2 T \log \frac{1}{\delta}}{n^2 \epsilon^2}$, the algorithm is (ϵ, δ)-DP. Moreover, with the same conditions on $T, \{\gamma_t\}_{t=1}^T$ as above, it can be derived that $\text{Err}_\mathcal{P}(w_R) \leq O\big(\frac{\sqrt[4]{d \ln \frac{1}{\delta}}}{\sqrt{n\epsilon}}\big)$, where the big-$O$ notations omit other terms.

Motivated by Algorithm 2, under the conditions of $\mathcal{X} = \{x \in \mathbb{R}^d | \|x\|_\infty \leq 1\}$ and $\mathcal{C} = \{w \in \mathbb{R}^d | \|w\|_1 \leq 1\}$, we can actually derive an upper bound of the population risk that depends only logarithmically on d (*i.e.,* $\log d$), indicating that it is suitable for high dimensional applications. Note that the conditions on $\mathcal{X}$ and $\mathcal{C}$ have been considered in linear regression (Talwar et al., 2015). We adopt them to our problem and extend their DP-Frank-Wolfe algorithm to Algorithm 3.

Theorem 6. Let $\mathcal{X} = \{x \in \mathbb{R}^d | \|x\|_\infty \leq 1\}$ and $\mathcal{C} = \{w \in \mathbb{R}^d | \|w\|_1 \leq 1\}$. For the GLM and robust regression problems, Algorithm 3 is (ϵ, δ)-DP with sensitivities $\Delta = O(\frac{C_\sigma (B+1)}{n})$ and $\Delta = O(\frac{C_\psi}{n})$, respectively. Furthermore, if we set $T = O(\frac{n\epsilon}{\sqrt{\ln(\frac{1}{\delta}) \ln(dn/\eta)}})$ and $\{\gamma_t\}_{t=1}^T = O(\sqrt{\frac{2}{T}})$, then with probability at least $1 - \eta$, we have $\text{Err}_\mathcal{P}(w_R) \leq O(\frac{\sqrt[4]{\ln(\frac{1}{\delta})} \sqrt{\ln \frac{nd}{\eta}}}{\sqrt{n\epsilon}})$. Here the big-$O$ notations omit other terms.

Algorithm 3 DP-FW-L1

Input: T is the iteration number and w_1 is the initial point. $\{\gamma_t\}_{t=1}^T$ is the step size. A is the set of vertices of $\mathcal{C}$. ϵ and δ are privacy parameters.

1: **for** $t = 1, \cdots, T$ **do**
2: Use exponential mechanism $\mathcal{M}(D, u, \mathcal{R})$, where $\mathcal{R} = A$, $u(D, s) = -\langle s, \nabla \hat{L}(w_t, D)\rangle$, to ensure $(\frac{\epsilon}{\sqrt{8T\ln(\frac{1}{\delta})}}, 0)$-differential privacy. Denote the output as $\tilde{w}_t$.
3: Compute $w_{t+1} = (1-\gamma_t)w_t + \gamma_t \tilde{w}_t$.
4: **end for**
5: Return $w_R \in \{w_1, \cdots, w_T\}$, where R is uniformly sampled from $\{1, 2 \cdots, T\}$.

5. Finding Approximate Local Minimum Privately

In this section, instead of measuring the error w.r.t the global minimum, we show that when the size of the dataset n is large enough, there exist (ϵ, δ)-DP algorithms that can find some approximate local minimum (in terms of second-order stationary points). We first impose the following assumption on the loss function considered in this section.

Assumption 4. The loss function is L-Lipschitz, M-smooth and ρ-Hessian Lispchtiz. We further assume that the empirical risk $\hat{L}(w, D)$ is bounded by a constant B [5]. If $\mathcal{C}$ is closed, we denote the diameter of $\mathcal{C}$ as $D = \max_{x,x' \in \mathcal{C}} \|x - x'\|_2$.

5.1. Unconstrained Case

Definition 8. w is called a second-order stationary point (SOSP) of a twice differentiable function F if

$$\|\nabla F(w)\|_2 = 0 \text{ and } \lambda_{\min}(\nabla^2 F(w)) \geq 0\,,$$

where $\lambda_{\min}$ denotes its smallest eigenvalue.

Since it is extremely challenging to find an exact SOSP (Ge et al., 2015), we turn to its approximation. The following defintion of α-approximate SOSP relaxes the first- and second-order optimality conditions.

Definition 9 ((Agarwal et al., 2017))**.** w is an α-second-order stationary point (α-SOSP) or α-approximate local minimum of a twice differentiable function F, if [6]

$$\|\nabla F(w)\|_2 \leq \alpha \text{ and } \lambda_{\min}(\nabla^2 F(w)) \geq -\sqrt{\rho\alpha}. \quad (7)$$

[5]Note that if the empirical risk is not bounded, we can still use the same proof after replacing the constant by the term $\hat{L}(w_1, D) - \hat{L}(w^*, D)$. We make such an assumption for convenience.

[6]This is a special version of (ϵ, γ)-SOSP (Ge et al., 2015). Our results can be easily extended to the general definition. The same applies to the constrained case.

To find an α-SOSP privately, we present Algorithm 4. Comparing with the first-order noisy gradient descent methods, such as those in (Ge et al., 2015; Jin et al., 2017; Xu et al., 2018b; Jin et al., 2018), the main difference is that the noises added should be in the scale of $O(\frac{\sqrt{T}}{n\epsilon})$, which depends on the iteration number T. This dependency makes Algorithm 4 more complex than previous related algorithms.

Algorithm 4 DP-GD

Input: T is the iteration number and w_1 is the initial point. $\{\gamma_t\}_{t=1}^T$ is the step size. ϵ and δ are privacy parameters.

1: **for** $t = 1, \cdots, T$ **do**
2: Compute $w_{t+1} = w_t - \eta_t(\nabla \hat{L}(w_t, D) + \epsilon_t)$, where $\epsilon_t \sim N(0, \sigma^2 I_d)$ for some σ.
3: **end for**
4: Return $\{w_1, \cdots, w_{T+1}\}$.

To prove that Algorithm 4 has the ability of escaping saddle points, we first show that the iteration number satisfies $T = \tilde{O}(\frac{MB}{\alpha^2})$ when the magnitude of the noise is small enough (*i.e.*, when n is large enough). Based on this fact, we then prove that Algorithm 4 can find an α-SOSP with high probability. Our results are summarized in the following theorem.

Theorem 7. Under Assumption 4, there exist constants c_1, c_2, such that for any $0 < \epsilon < c_1 T$, Algorithm 4 is (ϵ, δ)-DP if $\sigma^2 = c_2 \frac{L^2 \log\frac{1}{\delta} T}{n^2\epsilon^2}$. Moreover, if the data size n is large enough such that

$$n \geq \tilde{\Omega}\Big(\frac{\sqrt{MB}\sqrt{\log\frac{1}{\delta} d}\log\frac{1}{\xi} L}{\epsilon\alpha^2}\Big), \quad (8)$$

and choose $T = \tilde{O}(\frac{MB}{\alpha^2})$, $\{\eta_t\}_{t=1}^T = \frac{1}{M}$, then with probability $1-\zeta$, one of the outputs is an α-SOSP of the empirical risk $\hat{L}(\cdot, D)$. Here the $\tilde{O}$ and $\tilde{\Omega}$ terms omit other log factors (see Supplemental Material for a complete version).

Recently, Wang & Xu (2019); Wang et al. (2017) show that there are (ϵ, δ)-DP algorithms satisfying $\|\nabla \hat{L}(w^{\text{priv}}, D)\|_2 \leq O(\frac{\sqrt[4]{d\log\frac{1}{\delta}}}{\sqrt{n\epsilon}})$. Thus, to achieve an ϵ-first-order stationary point, the size n should satisfy the condition of $n \geq \Omega(\frac{\sqrt{d\log\frac{1}{\delta}}}{\epsilon\alpha^2})$. Comparing to the sample complexity in (8) for ϵ-SOSP, we can see that they are actually asymptotically almost the same (up to some log factors).

Theorem 7 ensures the existence of an approximate SOSP among $\{w_1, \cdots, w_{T+1}\}$. To find such a SOSP with high probability, we propose Algorithm 5, which incurs an additional $O(\sqrt{d})$ factor in the sample size n in (8).

Theorem 8. There exist constants c_1, c_2 such that when $\sigma_1^2 = c_1 \frac{\log \frac{1}{\delta} T L^2}{n^2 \epsilon^2}$ and $\sigma_2^2 = c_2 \frac{\log \frac{1}{\delta} M^2 d T}{n^2 \epsilon^2}$, Algorithm 5 is (ϵ, δ)-DP. Furthermore, with probability at least $1 - \xi - \frac{T}{p^C}$ for some sufficiently large $C > 0$, the output is an α-SOSP when the sample size satisfies $n \geq \tilde{\Omega}(\frac{Md\sqrt{MB}\sqrt{\log \frac{1}{\delta}} \log \frac{1}{\xi} L}{\rho \epsilon \alpha^2})$.

Algorithm 5 Selecting SOSP

1: Run Algorithm 4 to ensure $(\frac{\epsilon}{2}, \frac{\delta}{2})$-differential privacy on finding an $\frac{\alpha}{2}$-SOSP with probability at least $1 - \frac{\xi}{2}$. Let the output be $\{w_1, \cdots, w_{T+1}\}$.
2: **for** $t = 1, \cdots, T+1$ **do**
3: Let $g_t = \nabla \hat{L}(w_t, D) + \epsilon_t$, where $\epsilon_t \sim \mathcal{N}(0, \sigma_1^2 I_d)$. $\tilde{H}_t = \nabla^2 \hat{L}(w_t, D) + H_t$, where H_t a symmetric matrix with its upper triangle (including the diagonal) being i.i.d samples from $\mathcal{N}(0, \sigma_2^2)$ and each lower triangle entry is copied from its upper triangle counterpart.
4: **if** $\|g_t\|_2 \leq \alpha$ and $\lambda_{\min}(\tilde{H}_t) \geq -\sqrt{\rho \alpha}$ **then**
5: Return w_t.
6: **end if**
7: **end for**

5.2. Constrained Case

In this section we consider a constrained-version of SOSP studied in last section (see Definition 10).

Definition 10 ((Mokhtari et al., 2018))**.** For a twice differentiable function F and a closed convex set $\mathcal{C}$, w^* is an α-second-order stationary point in the constraint set $\mathcal{C}$ if: 1) $\nabla F(w^*)^T (w - w^*) \geq -\alpha$, for $\forall w \in \mathcal{C}$, and 2) $(w - w^*)^T \nabla^2 F(w^*)(w - w^*) \geq -\sqrt{\rho \alpha}$, for $\forall w \in \mathcal{C}$, s.t. $\nabla F(w^*)^T (w - w^*) = 0$.

Recently, Mokhtari et al. (2018) proposed an algorithm for escaping the saddle points in the above constrained case. Motivated by their algorithm and the ideas in the proof of Theorem 7, we propose Algorithm 6 as a DP-version of the problem with a theoretical guarantee presented in Theorem 9.

Theorem 9. There exist constants c_1, c_2, c_3 and sufficiently large C such that for any $0 < \epsilon < c_1 T, 0 < \delta < 1$, if $\sigma_1^2 = c_2 \frac{\log \frac{1}{\delta} L^2 T}{n^2 \epsilon^2}$ and $\sigma_2^2 = c_3 \frac{\log \frac{1}{\delta} d M^2 T}{n^2 \epsilon^2}$, Algorithm 6 is (ϵ, δ)-DP. Moreover, taking $T = O(\max\{\frac{D^2 MB}{\alpha^2}, \frac{B\rho^{1/2} D^6}{\Phi^3 \alpha^{3/2}}\}) = O(\frac{BM\rho^{1/2} D^6}{\Phi^3 \alpha^2})$, $\theta = \frac{\Phi \alpha}{2D^3}$, $0 < \Phi \leq \frac{9}{5}$, $\{\eta_t\}_{t=1}^T = \frac{\alpha}{2D^2 M}$ and $r = \frac{\Phi^2 \Phi \alpha}{72 \rho D^3}$, we have that for any $0 < \xi < 1$, with probability at least $1 - \xi - \frac{T}{p^C}$, Algorithm 6 outputs w_t, which is an α-SOSP of the empirical risk $\hat{L}(\cdot, D)$, if the sample size

Algorithm 6 DP-GD-SO

Input: T is the iteration number and x_1 is the initial point. $\{\gamma_t\}_{t=1}^T$ is the step size. ϵ and δ are privacy parameters. $\theta, \sigma_1, \sigma_2$ are parameters to be specified later.

1: **for** $t = 1, \cdots, T$ **do**
2: Compute $g_t = \nabla \hat{L}(w_t, D) + \epsilon_t$,where $\epsilon_t \sim \mathcal{N}(0, \sigma_1^2 I_d)$ for some σ_1.
3: Compute $v_t = \arg\max_{v \in \mathcal{C}} \{-g_t^T v\}$.
4: **if** $g_t^T (v_t - w_t) < -\frac{\alpha}{2}$ **then**
5: Compute $w_{t+1} = (1 - \eta_t) w_t + \eta_t v_t$.
6: **else**
7: Let $\tilde{H}_t = \nabla^2 \hat{L}(w_t, D) + H_t$, where H_t is a symmetric matrix with its upper triangle (including the diagonal) being i.i.d samples from $\mathcal{N}(0, \sigma_2^2)$ and each lower triangle entry is copied from its upper triangle counterpart.
8: Find u_t, a Φ-approximate solution of

$$\min_u q(u) = (u - w_t)^T \tilde{H}_t (u - w_t)$$
$$\text{s.t. } u \in \mathcal{C}, g_t^T (u - w_t) \leq r$$

9: **if** $q(u_t) \leq \frac{-\Phi \sqrt{\rho \alpha}}{2}$ **then**
10: Compute $w_{t+1} = (1 - \theta) w_t + \theta u_t$.
11: **else**
12: Return w_t.
13: **end if**
14: **end if**
15: **end for**

n satisfies:

$$n \geq \tilde{\Omega}\Big(\max\{ \frac{LD^7 \sqrt{dMB \log \frac{1}{\delta}} \log \frac{1}{\xi} \rho^{1/4}}{\epsilon \alpha^2}, \frac{\sqrt{\log \frac{1}{\delta} dBM} LD^4 \log \frac{1}{\xi} \rho^{1/4}}{\epsilon \alpha^2}, \frac{d \sqrt{BM^3 \log \frac{1}{\delta}} D^5 \log \frac{1}{\xi}}{\rho^{1/4} \alpha^{3/2} \epsilon} \} \Big).$$

Here the $\tilde{\Omega}$-notation omits Φ and other log terms.

Remark 3. Firstly, we note that when omitting other terms in the bound in Theorem 9 such as L, B, Φ, D, G, ρ, the sample complexity for escaping saddle points in the constrained case is $\tilde{\Omega}(\frac{d}{\epsilon \alpha^2})$. Compared with the unconstrained case in Theorem 8, they are asymptotically the same. Secondly, a quadratic programming problem needs to be solved in step 8 of Algorithm 6. For a general constraint set $\mathcal{C}$, solving the quadratic problem is NP-hard. However, for some specified sets such as intersection of ellipsoids or balls, an approximate solution can be obtained in polynomial time. See (Mokhtari et al., 2018) for more details.

Acknowledgements

The research of the first and third authors was supported in part by NSF through grant CCF-1716400. Part of the work was done when Di Wang was visiting the Simons Institute for the Theory of Computing.

References

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

Agarwal, N., Allen-Zhu, Z., Bullins, B., Hazan, E., and Ma, T. Finding approximate local minima faster than gradient descent. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1195–1199. ACM, 2017.

Anandkumar, A. and Ge, R. Efficient approaches for escaping higher order saddle points in non-convex optimization. In *Conference on Learning Theory*, pp. 81–102, 2016.

Balcan, M.-F., Dick, T., and Vitercik, E. Dispersion for data-driven algorithm design, online learning, and private optimization. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 603–614. IEEE, 2018.

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

Bhojanapalli, S., Neyshabur, B., and Srebro, N. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, pp. 3873–3881, 2016.

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

Chaudhuri, K., Monteleoni, C., and Sarwate, A. D. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(Mar):1069–1109, 2011.

Chen, C., Ding, N., and Carin, L. On the convergence of stochastic gradient mcmc algorithms with high-order integrators. In *Advances in Neural Information Processing Systems*, 2015.

Chiang, T.-S., Hwang, C.-R., and Sheu, S. J. Diffusion for global optimization in r^n. *SIAM Journal on Control and Optimization*, 25(3):737–753, 1987.

Dalalyan, A. Further and stronger analogy between sampling and optimization: Langevin monte carlo and gradient descent. In *Conference on Learning Theory*, pp. 678–689, 2017.

Dalalyan, A. S. and Karagulyan, A. User-friendly guarantees for the langevin monte carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 2019.

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

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

Erlingsson, Ú., Pihur, V., and Korolova, A. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pp. 1054–1067. ACM, 2014.

Foster, D. J., Sekhari, A., and Sridharan, K. Uniform convergence of gradients for non-convex learning and optimization. In *Advances in Neural Information Processing Systems*, pp. 8759–8770, 2018.

Ge, R., Huang, F., Jin, C., and Yuan, Y. Escaping from saddle points-online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pp. 797–842, 2015.

Ge, R., Lee, J. D., and Ma, T. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pp. 2973–2981, 2016.

Ge, R., Jin, C., and Zheng, Y. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *International Conference on Machine Learning*, pp. 1233–1242, 2017.

Ge, R., Lee, J. D., and Ma, T. Learning one-hidden-layer neural networks with landscape design. In *International Conference on Learning Representations*, 2018.

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

Jin, C., Ge, R., Netrapalli, P., Kakade, S. M., and Jordan, M. I. How to escape saddle points efficiently. In *International Conference on Machine Learning*, pp. 1724–1732, 2017.

Jin, C., Liu, L. T., Ge, R., and Jordan, M. I. On the local minima of the empirical risk. In *Advances in Neural Information Processing Systems*, pp. 4901–4910, 2018.

Kasiviswanathan, S. P. and Jin, H. Efficient private empirical risk minimization for high-dimensional learning. In *International Conference on Machine Learning*, pp. 488–497, 2016.

Kawaguchi, K. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems*, pp. 586–594, 2016.

Kifer, D., Smith, A., and Thakurta, A. Private convex empirical risk minimization and high-dimensional regression. In *Conference on Learning Theory*, 2012.

Li, B., Chen, C., Liu, H., and Carin, L. On connecting stochastic gradient mcmc and differential privacy. In Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of Machine Learning Research*, volume 89, pp. 557–566, 2019.

Loh, P.-L. and Wainwright, M. J. Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. In *Advances in Neural Information Processing Systems*, pp. 476–484, 2013.

Lozano, A. C., Meinshausen, N., Yang, E., et al. Minimum distance lasso for robust high-dimensional regression. *Electronic Journal of Statistics*, 10(1):1296–1340, 2016.

McSherry, F. and Talwar, K. Mechanism design via differential privacy. In *Foundations of Computer Science, 2007. FOCS'07. 48th Annual IEEE Symposium on*, pp. 94–103. IEEE, 2007.

Mei, S., Bai, Y., Montanari, A., et al. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A):2747–2774, 2018.

Mokhtari, A., Ozdaglar, A., and Jadbabaie, A. Escaping saddle points in constrained optimization. In *Advances in Neural Information Processing Systems*, pp. 3633–3643, 2018.

Near, J. Differential privacy at scale: Uber and berkeley collaboration. In *Enigma 2018 (Enigma 2018)*, Santa Clara, CA, 2018. USENIX Association.

Raginsky, M., Rakhlin, A., and Telgarsky, M. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. In *Proceedings of the 2017 Conference on Learning Theory*, 2017.

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

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

Tzen, B., Liang, T., and Raginsky, M. Local optimality and generalization guarantees for the langevin algorithm via empirical metastability. In *Conference On Learning Theory*, pp. 857–875, 2018.

Vollmer, S. J., Zygalakis, K. C., and Teh, Y. W. Exploration of the (non-) asymptotic bias and variance of stochastic gradient langevin dynamics. *Journal of Machine Learning Research*, 17(159):1–48, 2016.

Wang, D. and Xu, J. Differentially private empirical risk minimization with smooth non-convex loss functions: A non-stationary view. *Thirty-Third AAAI Conference on Artificial Intelligence, (AAAI-19), Honolulu, Hawaii, USA, January 27-February 1, 2019*, 2019.

Wang, D., Ye, M., and Xu, J. Differentially private empirical risk minimization revisited: Faster and more general. In *Advances in Neural Information Processing Systems*, 2017.

Wang, D., Gaboardi, M., and Xu, J. 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*, 2018.

Wang, D., Smith, A., and Xu, J. Noninteractive locally private learning of linear models via polynomial approximations. In *Algorithmic Learning Theory*, pp. 897–902, 2019.

Wang, Y.-X., Fienberg, S., and Smola, A. Privacy for free: Posterior sampling and stochastic gradient monte carlo. In *International Conference on Machine Learning*, 2015.

Xu, P., Chen, J., Zou, D., and Gu, Q. Global convergence of langevin dynamics based algorithms for nonconvex optimization. In *Advances in Neural Information Processing Systems*, pp. 3126–3137, 2018a.

Xu, Y., Rong, J., and Yang, T. First-order stochastic algorithms for escaping from saddle points in almost linear time. In *Advances in Neural Information Processing Systems*, pp. 5535–5545, 2018b.

Zhang, J., Zheng, K., Mou, W., and Wang, L. Efficient private erm for smooth objectives. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pp. 3922–3928. AAAI Press, 2017.