Escaping Saddle Points of Empirical Risk Privately and Scalably via DP-Trust Region Method

Di Wang[1,2] and Jinhui Xu[1]

[1] Department of Computer Science and Engineering
State University of New York at Buffalo, Buffalo, NY 14260, USA
[2] King Abdullah University of Science and Technology, Thuwal, Saudi Arabia
{dwang45,jinhui}@buffalo.edu

Abstract. It has been shown recently that many non-convex objective/loss functions in machine learning are known to be strict saddle. This means that finding a second-order stationary point (*i.e.,* approximate local minimum) and thus escaping saddle points are sufficient for such functions to obtain a classifier with good generalization performance. Existing algorithms for escaping saddle points, however, all fail to take into consideration a critical issue in their designs, that is, the protection of sensitive information in the training set. Models learned by such algorithms can often implicitly memorize the details of sensitive information, and thus offer opportunities for malicious parties to infer it from the learned models. In this paper, we investigate the problem of privately escaping saddle points and finding a second-order stationary point of the empirical risk of non-convex loss function. Previous result on this problem is mainly of theoretical importance and has several issues (*e.g.,* high sample complexity and non-scalable) which hinder its applicability, especially, in big data. To deal with these issues, we propose in this paper a new method called Differentially Private Trust Region, and show that it outputs a second-order stationary point with high probability and less sample complexity, compared to the existing one. Moreover, we also provide a stochastic version of our method (along with some theoretical guarantees) to make it faster and more scalable. Experiments on benchmark datasets suggest that our methods are indeed more efficient and practical than the previous one.

Keywords: Differential Privacy · Empirical Risk Minimization · Private Machine Learning

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 [12, 11]. 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 [14].

As a fundamental supervised-learning problem in machine learning, Empirical Risk Minimization (ERM) has been extensively studied in recent years since it encompasses a large family of classical models such as linear regression, LASSO, SVM, logistic regression and ridge regression. Its Differentially Private (DP) version (DP-ERM) can be formally defined as follows.

Definition 1 (DP-ERM [34]). *Given a dataset $D = \{x_1, \cdots, x_n\}$ from a data universe $\mathcal{X}$, DP-ERM is to find an estimator $w^{priv} \in \mathbb{R}^p$ so as to minimize the empirical risk,* i.e.

$$L(w, D) = \frac{1}{n}\sum_{i=1}^{n} \ell(w, x_i), \tag{1}$$

with the guarantee of being differentially private, where $\ell(\cdot, \cdot)$ is the loss function.

If the loss function is convex, the utility of the private estimator is measured by the expected excess empirical risk, i.e. $\mathbb{E}_{\mathcal{A}}[L(w^{priv}, D)] - \min_{x\in\mathbb{R}^p} L(w, D)$*, 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 [7] (see the section of Related Work for more details). However, empirical studies have revealed that non-convex loss functions typically achieve better classification accuracy than the convex ones [25]. Furthermore, recent developments in deep learning [17] 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 [37, 34, 31, 28] have considered DP-ERM with non-convex loss functions. This is mainly due to the fact that finding the global minimum of a non-convex loss function is NP-hard.

Different from convex loss functions, non-convex functions have adopted a few different ways to measure the utility. The authors of [37] studied the problem with smooth loss function and proposed using the ℓ_2 gradient-norm of a private estimator, *i.e.*, $\|\nabla L(w^{\text{priv}}, D)\|_2$, to measure the utility, which was then extended in [34, 31] to the cases of non-smooth loss functions and high dimensional space. It is well known that ℓ_2 gradient-norm can estimate only the first-order stationary point (or critical point) [3], and thus may lead to inferior generalization performance [10]. The authors of [28] are the first to show that the utility of general non-convex loss functions can also be measured in the same way as convex loss functions by the expected excess empirical risk. However, their upper bound $O(\frac{p}{\log n\epsilon^2})$ is quite large compared with the convex case and needs to assume that $n \geq O(\exp(p))$, which may not be satisfied in some real-world datasets. They also showed that for some special loss functions such as sigmoid loss, the bound can be further improved. But such improvements are dependent on the special structures or some assumptions of the loss functions and thus cannot be extended to the general case.

Due to the intrinsic challenge of approximating global minimum and issues related to saddle points, recent research on deep neural network training [16, 23] and many other machine learning problems [15, 5] has shifted the attentions to obtaining local minima. It

[3] A point w of a function $F(\cdot)$ is called a first-order stationary point (critical point) if it satisfies the condition of $\|\nabla F(w)\| = 0$.

has been shown that fast convergence to a local minimum is actually sufficient for such tasks to have good generalization performance. This motivates us to investigate efficient techniques for finding local minima. However, as shown in [2], computing a local minimum could be quite challenging as it is actually NP-hard for general non-convex functions. Fortunately, many non-convex functions in machine learning are known to be strict saddle [15], meaning that a second-order stationary point (or approximate local minimum) is sufficient to obtain a close enough point to some local minimum. With this, the authors in [28] used a new way to measure the utility, based on the ℓ_2-gradient norm and the minimal eigenvalue of a Hessian matrix (see Preliminaries section for details), where the goal is to design an algorithm with the ability of escaping saddle points and approximating some second-order stationary point. Specifically, they showed that when n is large enough, the classical differentially private gradient descent method could escape saddle points and meanwhile output an α-second-order stationary point (α-SOSP). But their method has several issues, which hamper its applications in big data. Firstly, their sample complexity (or equivalently error bound) is relatively high. It is not clear whether it can be improved. Secondly, their method needs to calculate the gradient and Hessian matrix of the whole objective function in each iteration, which is prohibitive in large scale datasets. Finally, their result mainly focuses on theoretical development and does not provide any experimental study. Thus, it is not clear how practical it is.

Our Contributions: To address the aforementioned theoretical and practical issues, we propose in this paper a new method called Differentially Private Trust Region (DP-TR) which is capable of escaping saddle points privately. Particularly, we first show that our algorithm can output an α-SOSP with high probability and less sample complexity (compared to the one in [28]). To make our method scalable, we then present a stochastic version of DP-TR called Differentially Private Stochastic Trust Region (DP-STR) with the same functionality. We show that DP-STR is much faster and has asymptotically the same sample complexity as DP-TR. Finally, we provide comprehensive experimental studies on the practical performance of our methods in escaping saddle point under differential privacy model.

Due to space limit, all proofs are left to the Supplementary Material.

2 Related Work

DP-ERM is a fundamental problem in both machine learning and differential privacy communities. There are quite a number of results on DP-ERM with convex loss functions, which investigate the problem from different perspectives. For example, [29, 30, 35] considered ERM in the non-interactive local model. [21, 27] and [1] investigated the regret bound in online settings. [36] explored the problem from the perspective of learnability and stability. The problem has also been well-studied in the offline central model [7, 8, 4, 34, 20], as well as in high dimensional space [26, 22].

For general non-convex loss functions, as mentioned earlier, existing results have used three different ways to measure the utility of the private estimator [37, 34, 31, 28]. For ℓ_2-gradient norm based utility, [31] provided a comprehensive study following the work of [37]. However, since the first order stationary points often have inferior performance to the second order stationary points in practice, which is the focus of this paper, their results are incomparable with ours. For the expected excess empirical (or population) risk based utility, it has been applied only to some special non-convex loss functions before the work of [28]. For example, [34] showed a near optimal bound for some special non-convex loss functions satisfying the Polyak-Lojasiewicz condition. [3] studied the problem of optimizing privately piecewise Lipschitz functions which satisfy the dispersion condition in online settings. Recently, the authors of [28] provided the first result for general non-convex functions. However, their bound is loose compared to the convex case and needs to assume that $n \geq O(\exp(p))$, which may not hold in practice.

For the third type of utility (based on the ℓ_2-gradient norm and the minimal eigenvalue of a Hessian matrix), [28] was the first to use it to measure the closeness of the private estimator to some second order stationary points. Compared with theirs, our proposed methods improve considerably the sample complexity. More precisely, we show that to achieve an α-SOSP (see Definition 7 for details), the sample complexity of DP-TR and DP-STR is $O(\frac{p\sqrt{\ln\frac{1}{\delta}}}{\alpha^{1.75}\epsilon})$ (omitting other terms), while it is $O(\frac{p\sqrt{\ln\frac{1}{\delta}}}{\alpha^{2}\epsilon})$ in [28]. Equivalently, with a fixed data size n, our algorithms yield an $O\left((\frac{p\sqrt{\ln\frac{1}{\delta}}}{n\epsilon})^{\frac{4}{7}}\right)$-SOSP, while [28] can output only an $O\left((\frac{p\sqrt{\ln\frac{1}{\delta}}}{n\epsilon})^{\frac{1}{2}}\right)$-SOSP (with high probability). Moreover, we also show in experiments that our methods are more efficient and scalable.

3 Preliminaries

In this section, we review some definitions related to differential privacy and some terminologies and lemmas in optimization.

3.1 Differential Privacy

Informally speaking, DP ensures that an adversary cannot infer whether or not a particular individual is participating in the database query, even with unbounded computational power and access to every entry in the database except for that particular individual's data. DP considers a centralized setting that includes a trusted data curator, who generates the perturbed statistical information (e.g., counts and histograms) by using some randomized mechanism. Formally, it can be defined as follows.

Definition 2 (Differential Privacy [12]). *Given a data universe $\mathcal{X}$, we say that two datasets $D, D' \subseteq \mathcal{X}$ are neighbors if they differ by only one entry, which is denoted as $D \sim D'$. A randomized algorithm $\mathcal{A}$ is (ϵ, δ)-differentially private (DP) if for all neighboring datasets D, D' and for all events S in the output space of $\mathcal{A}$, the following holds*

$$\mathbb{P}(\mathcal{A}(D) \in S) \leq e^{\epsilon}\mathbb{P}(\mathcal{A}(D') \in S) + \delta.$$

When $\delta = 0$, $\mathcal{A}$ is ϵ-differentially private.

In this paper, we will study only (ϵ, δ)-DP and use the Gaussian Mechanism [12] to guarantee (ϵ, δ)-DP.

Definition 3 (Gaussian Mechanism). *Given any 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 Gaussian Distribution $\mathcal{N}(0, \sigma^2 I_p)$ 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 ensures (ϵ, δ)-differential privacy.*

We will use the sub-sampling property and the advanced composition theorem to ensure (ϵ, δ)-DP for the DP-STR algorithm.

Lemma 1 (Advanced Composition Theorem [13]). *Given target privacy parameters $0 < \epsilon, \delta \leq 1$, to ensure (ϵ, δ)-DP over T mechanisms, it suffices that each mechanism is $(\epsilon', \frac{\delta}{2T})$-DP, where $\epsilon' = \frac{\epsilon}{2\sqrt{2T\ln(2/\delta)}}$.*

Lemma 2 ([4]). *Over a domain of datasets $\mathcal{X}^n$, if an algorithm $\mathcal{A}$ is (ϵ, δ)-DP, then for any n-size dataset D, executing $\mathcal{A}$ on uniformly random γn entries of D ensures $(2\gamma\epsilon, \delta)$-DP.*

We will use a relaxation of DP called zero-Concentrated Differential Privacy (zCDP) [6] to guarantee (ϵ, δ)-DP for our DP-TR method. Due to its optimal composition proposition, zCDP is easier to analyze and can achieve a tighter bound, compared to those using the advanced composition theorem to ensure (ϵ, δ)-DP (Lemma 1)[31].

Definition 4 (zCDP[6]). *A randomized mechanism $\mathcal{A}$ is ρ-zero concentrated differentially private if for all $D \sim D'$ and all $\alpha \in (1, \infty)$,*

$$D_\alpha(\mathcal{A}(D) \| \mathcal{A}(D')) \leq \rho\alpha,$$

where $D_\alpha(\mathcal{A}(D) \| \mathcal{A}(D'))$ is the α-Rényi divergence [4] between the distribution of $\mathcal{A}(D)$ and $\mathcal{A}(D')$.

The following lemma shows the connection between zCDP and (ϵ, δ)-DP.

Lemma 3 ([6]). *If $\mathcal{A}$ is ρ-zCDP, then $\mathcal{A}$ is $(\rho + 2\sqrt{\rho \ln \frac{1}{\delta}}, \delta)$-DP for any $\delta > 0$.*

The following lemma says that adding Gaussian noise could also achieve zCDP.

Lemma 4 ([6]). *Given any function $q : \mathcal{X}^n \mapsto \mathbb{R}^p$, the Gaussian Mechanism $\mathcal{M}_G(D, q, \epsilon) = q(D) + Y$, where Y is drawn from Gaussian Distribution $\mathcal{N}(0, \sigma^2 I_p)$ with $\sigma \geq \frac{\Delta_2(q)}{\sqrt{2\rho}}$, is ρ-zCDP.*

Similar to DP, zCDP also has the composition property.

Lemma 5 ([6]). *Let $\mathcal{A}$ be ρ-zCDP and $\mathcal{A}'$ be ρ'-zCDP, then their composition $\mathcal{A}''(D) = (\mathcal{A}(D), \mathcal{A}'(D))$ is $(\rho + \rho')$-zCDP.*

[4] Generally, $D_\alpha(P\|Q)$ is the Rényi divergence between P and Q which is defined as

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1} \log \mathbb{E}_{x\sim Q}(\frac{P(x)}{Q(x)})^\alpha.$$

3.2 Optimization

We first specify the necessary assumptions on our loss functions that are commonly used in other related work such as [28, 38, 24].

Assumption 1 We assume that $L(\cdot, D)$ is bounded from below and its global minimum is achieved at w^*. We let Δ denote

$$\Delta = L(w^0, D) - L(w^*, D),$$

where w^0 is the initial vector of our algorithms.

Assumption 2 We assume that for each $x \in \mathcal{X}$, the loss function $\ell(\cdot, x)$ is G-Lipschitz, that is, for all $w, w' \in \mathbb{R}^d$

$$|\ell(w, x) - \ell(w', x)| \leq G\|w - w'\|_2.$$

We also assume that $\ell(\cdot, x)$ is M-smooth, that is, for all $w, w' \in \mathbb{R}^p$

$$\|\nabla\ell(w, x) - \nabla\ell(w', x)\|_2 \leq M\|w - w'\|_2.$$

Finally, we assume that $\ell(\cdot, x)$ is twice differentiable and ρ-Hessian Lipschitz, that is, for all $w, w' \in \mathbb{R}^p$

$$\|\nabla^2\ell(w, x) - \nabla^2\ell(w', x)\|_2 \leq \rho\|w - w'\|_2,$$

where $\|A\|_2$ is the spectral norm of a matrix A.

Note that the above assumption indicates that for any $w, h \in \mathbb{R}^p$

$$L(w + h, D) \leq L(w, D) + \langle\nabla L(w, D), h\rangle + \frac{1}{2}h^T\nabla^2 L(w, D)h + \frac{\rho}{6}\|h\|_2^3.$$

In this paper, we focus on approximating a second order stationary point, which is defined as follows.

Definition 5. *A point 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 minimal eigenvalue.

Since it is extremely challenging to find an exact SOSP [15], we turn to its approximation. The following defintion of α-approximate SOSP relaxes the first- and second-order optimality conditions.

Definition 6 ([15]). *w is an α-second-order stationary point (α-SOSP) or α-approximate local minimum [5] of a twice differentiable function F which is ρ-Hessian Lipschitz, if*

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

[5] This is a special version of (ϵ, γ)-SOSP [15]. Our results can be easily extended to the general definition. The same applies to the constrained case.

Based on this, we now formally define our problem of DP-SOSP.

Definition 7 (DP-SOSP). *Given $\alpha, \epsilon, \delta > 0$, DP-SOSP is to identify the smallest sample complexity $n(\alpha, p, \epsilon, \delta)$ such that when $n \geq n(\alpha, p, \epsilon, \delta)$, for any dataset D of size n, there is an (ϵ, δ)-DP algorithm which outputs an α-SOSP of the empirical risk (1) with high probability.*

Since our ideas are derived from the trust region method proposed in [9], we now briefly introduce the trust region method. In each step of the trust region method for a function $F(\cdot)$, it solves a Quadratic Constraint Quadratic Program (QCQP):

$$h^k = \arg \min_{h \in \mathbb{R}^d, \|h\|_2 \leq r} \langle \nabla F(w^k), h \rangle + \frac{1}{2} \langle \nabla^2 F(w^k) h, h \rangle, \tag{3}$$

where r is called the *trust-region radius*. Then, it updates in the following way

$$w^{k+1} = w^k + h^k.$$

Since the function $F(w)$ is non-convex, this indicates that the sub-problem (3) is non-convex. However, its global minimum can be characterized by the following lemma.

Lemma 6 (Corollary 7.2.2 in [9]). *Any global minimum of the problem (3) should satisfy*

$$(\nabla^2 F(w^k) + \lambda I) h^k = -\nabla F(w^k), \tag{4}$$

where the dual variable $\lambda \geq 0$ should satisfies the conditions of $\nabla^2 F(x^k) + \lambda I \succ 0$ and $\lambda(\|h^k\|_2 - r) = 0$.

It is worth noting that in practice sub-problem (3) can be solved by the Lanczos method efficiently (see [18] for details). For the dual variable λ in Lemma 6, it can be solved by almost any QCQP solver such as CVX [19].

4 Methodology

In this section, we first introduce our main method, DP-TR, and then extend to its stochastic version, *i.e.,* DP-STR.

4.1 Differentially Private Trust Region Method

The key idea of our DP-TR is the following. In each iteration, instead of using the gradient and Hessian of the empirical risk (1) directly to the sub-problem (3), we use their perturbed versions to ensure DP. That is, we use $\tilde{\nabla} L(w^k, D) = \nabla L(w^k, D) + \epsilon_k$ and $\tilde{\nabla}^2 L(w^k, D) = \nabla^2 L(w^k, D) + H_k$, where ϵ_t is a Gaussian vector and H_t is a randomized symmetric Gaussian matrix (since a Hessian matrix is symmetric, we need to add a symmetric random matrix). The main steps of DP-TR are given in Algorithm 1.

For the stopping criteria, we use the dual variable λ^k and see whether the value is greater or less than some threshold. This criteria enable the last-term convergence analysis in Theorem 2.

The following theorem shows that Algorithm 1 is (ϵ, δ)-DP.

Algorithm 1 DP-TR

```
Input: Privacy parameters $\epsilon, \delta$, trust-region radius $r$, iteration number $T$ (to be specified later), initial vector $w^0$ and error term $\alpha$
Let $\phi = (\sqrt{\epsilon + \ln \frac{1}{\delta}} - \sqrt{\ln \frac{1}{\delta}})^2$.
for $k = 0, \cdots, T-1$ do
    Denote $\tilde{\nabla} L(w^k, D) = \nabla L(w^k, D) + \epsilon_k$, where $\epsilon_t \sim \mathcal{N}(0, \sigma^2 I_d)$ with $\sigma^2 = \frac{4G^2T}{n^2\phi}$.
    Denote $\tilde{\nabla}^2 L(w^k, D) = \nabla^2 L(w^k, D) + H_k$, where $H_t$ is a symmetric matrix with its upper triangle (including the diagonal) being i.i.d samples from $\mathcal{N}(0, \sigma_2^2)$, $\sigma_2^2 = \frac{4pM^2T}{n^2\phi}$, and each lower triangle entry is copied from its upper triangle counterpart.
    Solve the following QCQP and get $h^k$ and dual variable $\lambda^k$,
        $$h^k = \arg \min_{h \in \mathbb{R}^d, \|h\|_2 \leq r} \langle \tilde{\nabla} L(w^k, D), h \rangle + \frac{1}{2} \langle \tilde{\nabla}^2 L(w^k, D) h, h \rangle,$$
    Let $w^{k+1} = w^k + h^k$.
    if $\lambda^k \leq \sqrt{\alpha\rho}$ then
        Output $w_\alpha = w^{k+1}$.
    end if
end for
```

Theorem 1. *For any $\epsilon, \delta > 0$, Algorithm 1 is (ϵ, δ)-differentially private under Assumption 2.*

The following theorem shows that when the data size n is large enough, then with high probability the output of Algorithm 1 will be an α-SOSP.

Theorem 2. *Under Assumptions 1 and 2, for any given α, if we take $r = \sqrt{\frac{\alpha}{\rho}}$, $T = \frac{6\sqrt{\rho}\Delta}{\alpha^{1.5}}$, then with probability at least $1 - \zeta - \frac{T}{p^c}$ for some universal constant $c > 0$ and $\zeta > 0$, the algorithm outputs a point which is an $O(\alpha)$-SOSP if n satisfies*

$$n \geq \Omega(\frac{p \ln \frac{1}{\zeta} \sqrt{\ln \frac{1}{\delta}}}{\alpha^{1.75}\epsilon}), \tag{5}$$

where the Big-Ω notation omits the terms of $G, M, \rho, \Delta, \ln \frac{1}{\alpha}$.

Remark 1. We note that in the previous work [28], to output an $O(\alpha)$-SOSP with high probability, the data size n needs to satisfy $n \geq \Omega(\frac{p\sqrt{\ln \frac{1}{\delta}}}{\alpha^2\epsilon})$, while the dependency on α in (5) is $\frac{1}{\alpha^{1.75}}$. Thus, we improve the sample size by a factor of $O(\frac{1}{\alpha^{0.25}})$. Equivalently, if we fix n, Theorem 2 ensures that Algorithm 1 outputs a point which is $O\left((\frac{p\sqrt{\ln \frac{1}{\delta}}}{n\epsilon})^{\frac{4}{7}}\right)$-SOSP, while the previous work in [28] outputs a point which is $O\left((\frac{p\sqrt{\ln \frac{1}{\delta}}}{n\epsilon})^{\frac{1}{2}}\right)$-SOSP. We can see that our algorithm yields better approximate SOSP than the previous one. We leave as open problems to determine whether the sample complexity in (5) can be further improved and what is the optimal bound of the sample complexity.

Also, in [28] the number of iterations is $T = \tilde{O}(\frac{1}{\alpha^2})$, while Algorithm 1 needs only $O(\frac{1}{\alpha^{1.5}})$ iterations. This means that the running time of Algorithm 1 is $O(\frac{n\text{Poly}(p)}{\alpha^{1.5}})$, while it is $O(\frac{n\text{Poly}(p)}{\alpha^2})$ in [28]. Thus, our algorithm has an improved time complexity for the term of $\frac{1}{\alpha}$ compared with the previous one. Moreover, as we will see in the experiment section, our algorithms is indeed faster than the previous one.

Theorem 2 shows the explicit step size control of the DP-TR method: Since the dual variable satisfies $\lambda^k > \sqrt{\alpha\rho}$ for all but the last iteration. Thus we can always find a solution to the trust-region sub-problem (3) in the boundary, *i.e.,* $\|h^k\|_2 = r$, according to Lemma 6.

4.2 Differentially Private Stochastic Trust Region Method

In the previous section we show that our method DP-TR needs less samples and is faster than DP-GD proposed in [28]. However, as mentioned in Remark 1, the time complexities of both algorithms are linearly dependent on the sample size n, which is prohibitive in large scale datasets. Thus, a natural question is to determine whether it is possible to design an algorithm that shares the advantages of DP-TR and meanwhile is scalable. In this section we give an affirmative answer to this question by providing a stochastic version of DP-TR called Differentially Private Stochastic Trust Region method (DP-STR).

The key idea of DP-STR is that, instead of evaluating the gradient and Hessian matrix of the whole function $L(w, D)$ in each iteration, we will uniformly sub-sample two sets of indices $\mathcal{S}, \mathcal{T} \subseteq [n]$ and calculate the gradients and Hessian matrix of the loss function with the samples corresponding to the set $\mathcal{S}$ and $\mathcal{T}$, respectively. That is

$$\nabla L(w^k, \mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \nabla \ell(w^k, x_i), \tag{6}$$

$$\nabla^2 L(w^k, \mathcal{T}) = \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} \nabla^2 \ell(w^k, x_i). \tag{7}$$

Then, similar to DP-TR, we add some Gaussian noise and random Gaussian matrix to $\nabla L(w^k, \mathcal{S})$ and $\nabla^2 L(w^k, \mathcal{T})$, respectively, to ensure (ϵ, δ)-DP. See Algorithm 2 for details. Note that since zCDP can not be guaranteed by sub-sampling, we use the traditional advanced composition theorem Lemma 1 and sub-sampling property Lemma 2 to guarantee (ϵ, δ)-DP.

Theorem 3. *For any* $0 < \epsilon, \delta < 1$*, Algorithm 2 is* (ϵ, δ)*-differentially private.*

Theorem 4. *Under Assumptions 1 and 2, for a given* α*, if we take* $r = \sqrt{\frac{\alpha}{\rho}}$, $T = \frac{6\sqrt{\rho}\Delta}{\alpha^{1.5}}$, $|\mathcal{S}| \geq \Omega(\frac{L^2 \ln \frac{p}{\zeta}}{\alpha^2})$ *and* $|\mathcal{T}| \geq \Omega(\frac{M^2 \ln \frac{p}{\zeta}}{\alpha\rho})$ *in Algorithm 2, then with probability at least* $1 - 3\zeta - \frac{T}{p^c}$ *for some universal constant* $c > 0$ *and* $\zeta > 0$*, the algorithm outputs a point that is an* $O(\alpha)$*-SOSP if* n *satisfies (5), which is the same as in Theorem 2.*

Algorithm 2 DP-STR

```
Input: Privacy parameters $\epsilon, \delta$, trust-region radius $r$, iteration number $T$, sub-sampling size $|\mathcal{S}|, |\mathcal{T}|$ (to be specified later), initial vector $w^0$ and error term $\alpha$.
for $k = 0, \cdots, T-1$ do
    Uniformly sub-sample two independent indices sets $\mathcal{S}, \mathcal{T} \subseteq [n]$ with size $|\mathcal{S}|$ and $|\mathcal{T}|$, respectively.
    Denote $\tilde{\nabla} L(w^k, \mathcal{S}) = \nabla L(w^k, \mathcal{S}) + \epsilon_k$, where $\epsilon_t \sim \mathcal{N}(0, \sigma^2 I_d)$ with $\sigma^2 = \frac{256G^2 \ln\frac{5T}{\delta} \ln\frac{2}{\delta}}{n^2\epsilon^2}$ and $\nabla L(w^k, \mathcal{S})$ is given in (6)
    Denote $\tilde{\nabla}^2 L(w^k, \mathcal{T}) = \nabla^2 L(w^k, \mathcal{T}) + H_k$, where $H_t$ is a symmetric matrix with its upper triangle (including the diagonal) being i.i.d samples from $\mathcal{N}(0, \sigma_2^2)$, $\sigma_2^2 = \frac{256pM^2T \ln\frac{2}{\delta} \ln\frac{5T}{\delta}}{n^2\epsilon^2}$, and each lower triangle entry is copied from its upper triangle counterpart. $\nabla^2 L(w^k, \mathcal{T})$ is given in (7).
    Solve the following QCQP and get $h^k$ and dual variable $\lambda^k$,
        $$h^k = \arg\min_{h \in \mathbb{R}^d, \|h\|_2 \leq r} \langle \tilde{\nabla} L(w^k, \mathcal{S}), h \rangle + \frac{1}{2} \langle \tilde{\nabla}^2 L(w^k, \mathcal{T}) h, h \rangle,$$
    Let $w^{k+1} = w^k + h^k$.
    if $\lambda^k \leq \sqrt{\alpha\rho}$ then
        Output $w_\alpha = w^{k+1}$.
    end if
end for
```

Comparing with Theorem 2, we can see that the sample complexity of Theorem 4 is the same while the time complexity of Algorithm 2 is $O\left(T(|\mathcal{S}| + |\mathcal{T}|)\text{Poly}(p)\right) = O(\frac{\text{Poly}(p)}{\alpha^{3.5}})$, which is independent of the sample size n. This means that DP-STR is faster and scalable to large scale datasets.

Remark 2. We note that it is unknown whether the algorithm in [28] can be extended to a stochastic version whose time complexity is independent of the size n. The algorithm in [28] consists of two routines, one is the Differentially Private Gradient Descent method and the other one is the procedure of selecting an α-SOSP. The first one can be easily extend to a stochastic version, which is similar as the one in[37]. However, for the second one, it needs to calculate the whole Hessian matrix and verify some conditions as stopping criteria, but it is unknown whether we can extend it to a stochastic version. Compared with their algorithm, in Algorithm 1 we use the Hessian matrix for Trust-Region sub-problem and use the dual variable λ^k as our stopping criteria. Thus, this is why we can extend Algorithm 1 to a stochastic version.

Note that in Algorithm 2 we use the basic subsampling technique for DP-STR to improve the time complexity. In [34], the authors proposed the Stochastic Variance Reduction Gradient method to improve the gradient complexity for DP-ERM with convex less functions and show it is superior to the DP-SGD method. Thus, it is unknown whether we can use the same idea to our problem to further improve the time complexity or gradient complexity. Moreover, in both of Algorithm 1 and 2, we assume that we can exactly solve the Trust-Region sub-problem (3). However, in most cases, exactly solving the problem is quite hard and costy. Thus whether we can relax this assumption is still an open problem. We leave these as further research.

5 Experiments

In this section, we present numerical experiments for different non-convex Empirical Risk Minimization problems on different datasets to demonstrate the advantage of our DP-TR and DP-STR algorithms in finding SOSP under differential privacy.

5.1 Experimental Settings

Baselines As mentioned in previous section, the only known method for this problem is DP-GD given in [28]. Thus, we compare it with our methods (DP-TR and DP-STR) after carefully tuning the algorithms for a fair comparison. For the QCQP sub-problem in Algorithm 1 and 2, we use the CVX package [19] to solve it.

Datasets We evaluate the algorithms on real-world datasets with $n \gg p$. Specifically, we use the datasets, Covertype and IJCNN, which are commonly used in the study of DP-ERM such as [34, 33, 32]. More information about these datasets is listed in Table 1. We normalize each row of the datasets as preprocessing.

Table 1: Summary of Datasets used in the experiments.

Dataset	Sample size n	dimension p
Covertype	$581,012$	54
IJCNN	$35,000$	22

Evaluated Problems For the loss functions we will follow the studies in [24, 38, 31]. The first non-convex problem that will be investigated is logistic regression with a non-convex regularizer $r(w) = \sum_{i=1}^{p} \frac{\lambda w_i^2}{1+w_i^2}$. Specifically, suppose that we are given training data $\{(x_i, y_i)\}_{i=1}^n$, where $x_i \in \mathbb{R}^p$ and $y \in \{-1, 1\}$ are, respectively, the feature vector and label of the i-th data record. The corresponding ERM is

$$\min_{w\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \log(1+\exp(-y_i\langle x_i, w\rangle)) + r(w).$$

In the experiment, we set $\lambda = 10^{-3}$.

The second problem that will be considered is the sigmoid regression with ℓ_2 norm regularizer. Given training dataset $\{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathbb{R}^p$ and $y \in \{-1, 1\}$ are, respectively, the feature vector and label of the i-th data record. Then, minimization problem is

$$\min_{w\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \frac{1}{1+\exp(-y_i\langle x_i, w\rangle)} + \frac{\lambda}{2}\|w\|_2^2.$$

In the experiment, we set $\lambda = 10^{-3}$.

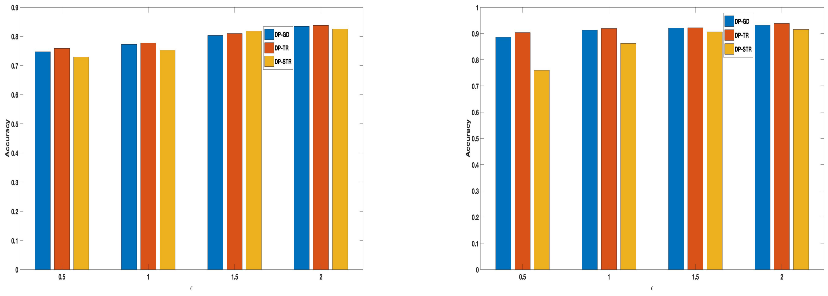
Measurements We first study how the optimaliy gap, *i.e.,* $L(w_\alpha, D) - \min_{w\in\mathbb{R}^p} L(w, D)$, changes w.r.t the privacy level ϵ or time (second). For the optimal solution of the problem $\min_{w\in\mathbb{R}^p} L(w, D)$, we obtain it through multiple runs of the classical trust region method and taking the best one. Besides the expected excess empirical risk, we also use the gradient norm, *i.e.,* $\|\nabla L(w_\alpha, D)\|_2$, to measure the utility. For logistic regression, we also consider its classification accuracy w.r.t privacy level, where the non-private case is obtained by running the trust region method and taking the best. For each experiment, we run 10 times and take the average as the final output. In all experiments, we set $\delta = \frac{1}{n}$ and $\alpha = 10^{-1}$.

5.2 Experimental Results

Figure 2 shows the classification accuracy of the private classifier given by the sigmoid regression on the Covertype and IJCNN datasets w.r.t different privacy levels. We can see that the accuracy increases when ϵ becomes larger, which means that the algorithm will be non-private. From Remark 1 we can see that this is due to the fact that when ϵ is larger, we can outpout an SOSP which is closer to the local minimum. Also, the accuracy of the non-private case is 86% and 95% for Covertype and IJCNN dataset, respectively. This indicates that the accuracy is comparable to the non-private case when $\epsilon \geq 1.5$.

The first and second subfigures of Figure 4, 6, 8 and 10 depict the optimality gap and the gradient norm w.r.t different privacy level ϵ of the two non-convex problems on Covertype and IJCNN datasets. For Covertype, we set the batchsize as 50000, while for IJCNN we set it as 5000. From the figures, we can see that compared with DP-GD, our DP-TR method has better performance on both the optimality gap and the gradient norm. This is due to the fact that DP-TR has improved the bound of SOSP (see Remark 1). However, the results of DP-STR are worse than that of DP-GD and DP-TR. We attribute this to the fact that the noise level of DP-STR added in each iteration (steps 2 and 3) is higher than that of DP-TR and DP-GD. For example, in Step 2 of Algorithm 2 we add a Gaussian noise with variance $\sigma^2 = \frac{256L^2 \ln \frac{5T}{\delta} \ln \frac{2}{\delta}}{n^2\epsilon^2}$ to each coordinate, while in step 3 of Algorithm 1 we only need to add a Gaussian noise with variance $\sigma^2 = \frac{4L^2T}{n^2\phi} \approx \frac{64L^2T \log \frac{1}{\delta}}{n^2\epsilon^2}$. Equivalently, the sub-optimality of DP-STR is due to the higher level of noise that needs to be added, which is required by the Advanced Composition Theorem to ensure (ϵ, δ)-DP. We leave it as an open problem to determine how to improve the practical performance of DP-STR.

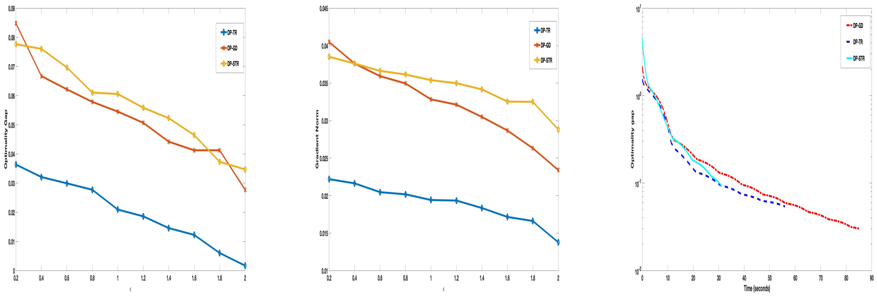
The third subfigures of Figure 4, 6, 8 and 10 show the results on the optimality gap w.r.t time of the two non-convex problems on the datasets of Covertype and IJCNN. Here we fix ϵ to be 1 in all the experiments. We can see that although the gap of DP-STR is worse than that of DP-GD and DP-TR, its running time is the least one. This is due to the fact that DP-STR needs only to evaluate a subset of the gradient and Hessian matrix, instead of the full ones as in DP-TR and DP-GD.



Covertype

IJCNN

Fig. 2: Accuracy w.r.t privacy level on Covertype and IJCNN datasets

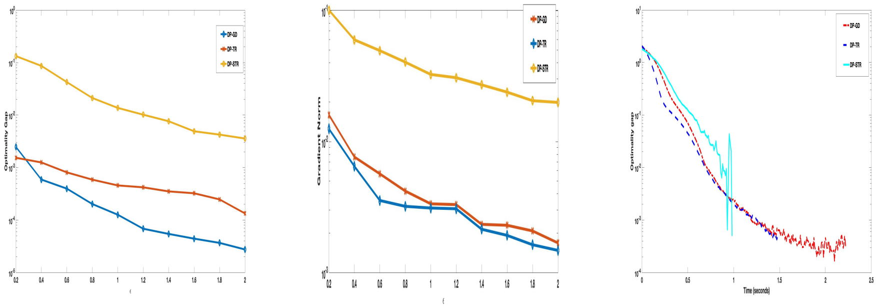


Optimality gap w.r.t privacy level ϵ

Gradient norm w.r.t privacy level ϵ

Optimality gap w.r.t time (seconds)

Fig. 4: Results of logistic regression with non-convex regularizer on Covertype dataset

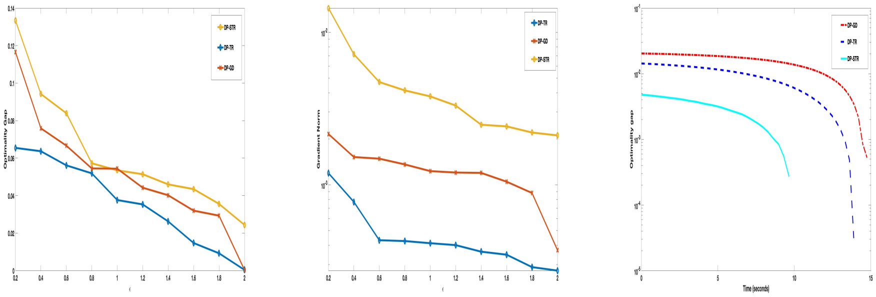


Optimality gap w.r.t privacy level ϵ

Gradient norm w.r.t privacy level ϵ

Optimality gap w.r.t time (seconds)

Fig. 6: Results of logistic regression with non-convex regularizer on IJCNN dataset

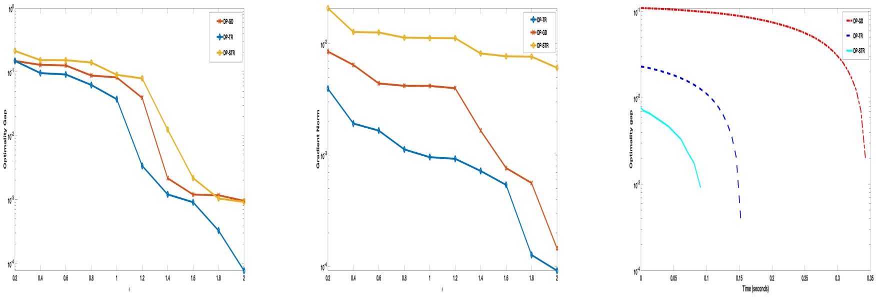


Optimality gap w.r.t privacy level ϵ

Gradient norm w.r.t privacy level ϵ

Optimality gap w.r.t time (seconds)

Fig. 8: Results of sigmoid regression with ℓ_2 norm regularizer on Covertype dataset



Optimality gap w.r.t privacy level ϵ

Gradient norm w.r.t privacy level ϵ

Optimality gap w.r.t time (seconds)

Fig. 10: Results of sigmoid regression with ℓ_2 norm regularizer on IJCNN dataset

6 Conclusion

In this paper we study the problem of escaping saddle points of empirical risk in the differential privacy model and propose a new method called DP-Trust Region (DP-TR) along with its stochastic version called DP-STR. Particularly, we show that to achieve an α-SOSP with high probability, DP-TR and DP-STR have lower sample complexities compared with the existing algorithm DP-GD. We also show that DP-TR is faster than DP-GD; DP-STR is more scalable and much faster than DP-TR. Experimental results on benchmark datasets confirm our theoretical claims.

References

1. Agarwal, N., Singh, K.: The price of differential privacy for online learning. In: Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017. pp. 32–40 (2017)
2. Anandkumar, A., Ge, R.: Efficient approaches for escaping higher order saddle points in non-convex optimization. In: Conference on Learning Theory. pp. 81–102 (2016)
3. Balcan, M.F., Dick, T., 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)
4. Bassily, R., Smith, A., 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)
5. Bhojanapalli, S., Neyshabur, B., Srebro, N.: Global optimality of local search for low rank matrix recovery. In: Advances in Neural Information Processing Systems. pp. 3873–3881 (2016)
6. Bun, M., Steinke, T.: Concentrated differential privacy: Simplifications, extensions, and lower bounds. In: Theory of Cryptography Conference. pp. 635–658. Springer (2016)
7. Chaudhuri, K., Monteleoni, C.: Privacy-preserving logistic regression. In: Advances in Neural Information Processing Systems. pp. 289–296 (2009)
8. Chaudhuri, K., Monteleoni, C., Sarwate, A.D.: Differentially private empirical risk minimization. Journal of Machine Learning Research **12**(Mar), 1069–1109 (2011)
9. Conn, A.R., Gould, N.I., Toint, P.L.: Trust region methods, vol. 1. Siam (2000)
10. Dauphin, Y.N., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S., Bengio, Y.: Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In: Advances in neural information processing systems. pp. 2933–2941 (2014)
11. Dwork, C.: Differential privacy: A survey of results. In: International Conference on Theory and Applications of Models of Computation. pp. 1–19. Springer (2008)
12. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Theory of cryptography conference. pp. 265–284. Springer (2006)
13. Dwork, C., Roth, A., et al.: The algorithmic foundations of differential privacy. Foundations and Trends® in Theoretical Computer Science **9**(3–4), 211–407 (2014)
14. Erlingsson, Ú., Pihur, V., 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)
15. Ge, R., Huang, F., Jin, C., Yuan, Y.: Escaping from saddle points-online stochastic gradient for tensor decomposition. In: Conference on Learning Theory. pp. 797–842 (2015)
16. Ge, R., Lee, J.D., Ma, T.: Learning one-hidden-layer neural networks with landscape design. In: International Conference on Learning Representations (2018)

17. Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y.: Deep learning, vol. 1. MIT press Cambridge (2016)
18. Gould, N.I., Lucidi, S., Roma, M., Toint, P.L.: Solving the trust-region subproblem using the lanczos method. SIAM Journal on Optimization **9**(2), 504–525 (1999)
19. Grant, M., Boyd, S.: CVX: Matlab software for disciplined convex programming, version 2.1. `http://cvxr.com/cvx` (Mar 2014)
20. Huai, M., Wang, D., Miao, C., Xu, J., Zhang, A.: Pairwise learning with differential privacy guarantees. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, New York City, New York, USA, February 7-12, 2020. (2020)
21. Jain, P., Kothari, P., Thakurta, A.: Differentially private online learning. In: Conference on Learning Theory. pp. 24–1 (2012)
22. Kasiviswanathan, S.P., Jin, H.: Efficient private empirical risk minimization for high-dimensional learning. In: International Conference on Machine Learning. pp. 488–497 (2016)
23. Kawaguchi, K.: Deep learning without poor local minima. In: Advances in Neural Information Processing Systems. pp. 586–594 (2016)
24. Kohler, J.M., Lucchi, A.: Sub-sampled cubic regularization for non-convex optimization. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. pp. 1895–1904. JMLR. org (2017)
25. 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)
26. Talwar, K., Thakurta, A.G., Zhang, L.: Nearly optimal private lasso. In: Advances in Neural Information Processing Systems. pp. 3025–3033 (2015)
27. Thakurta, A.G., Smith, A.: (nearly) optimal algorithms for private online learning in full-information and bandit settings. In: Advances in Neural Information Processing Systems. pp. 2733–2741 (2013)
28. Wang, D., Chen, C., Xu, J.: Differentially private empirical risk minimization with non-convex loss functions. In: International Conference on Machine Learning. pp. 6526–6535 (2019)
29. Wang, D., Gaboardi, M., Xu, J.: Empirical risk minimization in non-interactive local differential privacy revisited (2018)
30. Wang, D., Smith, A., Xu, J.: Noninteractive locally private learning of linear models via polynomial approximations. In: Algorithmic Learning Theory. pp. 897–902 (2019)
31. Wang, D., Xu, J.: Differentially private empirical risk minimization with smooth non-convex loss functions: A non-stationary view (2019)
32. Wang, D., Xu, J.: Differentially private empirical risk minimization with smooth non-convex loss functions: A non-stationary view. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 1182–1189 (2019)
33. Wang, D., Xu, J.: On sparse linear regression in the local differential privacy model. In: International Conference on Machine Learning. pp. 6628–6637 (2019)
34. Wang, D., Ye, M., Xu, J.: Differentially private empirical risk minimization revisited: Faster and more general. In: Advances in Neural Information Processing Systems. pp. 2722–2731 (2017)
35. Wang, D., Zhang, H., Gaboardi, M., Xu, J.: Estimating smooth glm in non-interactive local differential privacy model with public unlabeled data. arXiv preprint arXiv:1910.00482 (2019)
36. Wang, Y.X., Lei, J., Fienberg, S.E.: Learning with differential privacy: Stability, learnability and the sufficiency and necessity of erm principle. Journal of Machine Learning Research **17**(183), 1–40 (2016)
37. Zhang, J., Zheng, K., Mou, W., 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)
38. Zhou, D., Xu, P., Gu, Q.: Stochastic variance-reduced cubic regularized newton method. In: International Conference on Machine Learning. pp. 5985–5994 (2018)