SHARPER UTILITY BOUNDS FOR DIFFERENTIALLY PRIVATE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, by introducing Generalized Bernstein condition, we propose the first $\mathcal{O}\big(\frac{\sqrt{p}}{n\epsilon}\big)$ high probability excess population risk bound for differentially private algorithms under the assumptions G-Lipschitz, L-smooth, and Polyak-Łojasiewicz condition, based on gradient perturbation method. If we replace the properties G-Lipschitz and L-smooth by α-Hölder smoothness (which can be used in non-smooth setting), the high probability bound comes to $\mathcal{O}\big(n^{-\frac{2\alpha}{1+2\alpha}}\big)$ w.r.t n, which cannot achieve $\mathcal{O}\left(1/n\right)$ when $\alpha \in (0, 1]$. To solve this problem, we propose a variant of gradient perturbation method, **max$\{1, g\}$-Normalized Gradient Perturbation** (m-NGP). We further show that by normalization, the high probability excess population risk bound under assumptions α-Hölder smooth and Polyak-Łojasiewicz condition can achieve $\mathcal{O}\big(\frac{\sqrt{p}}{n\epsilon}\big)$, which is the first $\mathcal{O}\left(1/n\right)$ high probability utility bound w.r.t n for differentially private algorithms under non-smooth conditions. Moreover, we evaluate the performance of the new proposed algorithm m-NGP, the experimental results show that m-NGP improves the performance (measured by accuracy) of the DP model over real datasets. It demonstrates that m-NGP improves the excess population risk bound and the accuracy of the DP model on real datasets simultaneously.

1 INTRODUCTION

Machine learning has been widely used and found effective in many fields in recent years (Singha et al., 2021; Swapna & Soman, 2021; Ponnusamy et al., 2021). When training machine learning models, tremendous data was collected, and the data often contains sensitive information of individuals, which may leakage personal privacy (Shokri et al., 2017; Carlini et al., 2019).

Differential Privacy (DP) (Dwork et al., 2006; Dwork & Lei, 2009; Dwork et al., 2014) is a theoretically rigorous tool to prevent sensitive information. It introduces random noise to the machine learning model and blocks adversaries from inferring any single individual included in the dataset by observing the model. The mathematical definition of DP is well accepted and relative technologies are performed by Google (Erlingsson et al., 2014), Apple (McMillan, 2016) and Microsoft (Ding et al., 2017). As such, DP has attracted attention from the researchers and has been applied to numerous machine learning problems (Ullman & Sealfon, 2019; Xu et al., 2019; Bernstein & Sheldon, 2019; Wang & Xu, 2019; Heikkilä et al., 2019; Kulkarni et al., 2021; Bun et al., 2021; Nguyen & Vullikanti, 2021).

There are mainly three approaches to guarantee differential privacy: output perturbation (Chaudhuri et al., 2011), objective perturbation (Chaudhuri et al., 2011), and gradient perturbation (Song et al., 2013). Considering that gradient descent is a widely used optimization method, the gradient perturbation method can be used for a wide range of applications, and adding random noise to the gradient allows the model to escape local minima (Raginsky et al., 2017), we focus on the gradient perturbation method to guarantee DP in this paper.

In this paper, we aim to minimize the population risk, and measure the utility of the DP model by the excess population risk. To get the excess population risk, an important step is to analyze the generalization error (the reason is demonstrated in Section 3). Complexity theory (Bartlett et al., 2002) and algorithm stability theory (Bousquet & Elisseeff, 2002) are popular tools to analyze the generalization error. On one hand, Chaudhuri et al. (2011) applied the complexity theory and achieved

an $\mathcal{O}\big(\max\{\frac{1}{\sqrt{n}}, \sqrt[2/3]{\frac{p}{n\epsilon}}\}\big)$ high probability excess population risk bound under the assumption of strongly convex; Kifer et al. (2012) achieved $\mathcal{O}\big(\frac{\sqrt{p}}{n\epsilon}\big)$ expected excess population risk bound via complexity theory. On the other hand, the sharpest known high probability generalization bounds for DP algorithms analyzed via stability theory under different assumptions (Wu et al., 2017; Bassily et al., 2019; Feldman et al., 2020; Bassily et al., 2020; Wang et al., 2021) are $\mathcal{O}\big(\frac{\sqrt{p}}{n\epsilon} + \frac{1}{\sqrt{n}}\big)$ or $\mathcal{O}\big(\frac{\sqrt[4]{p}}{\sqrt{n\epsilon}}\big)$, containing an inevitable $\mathcal{O}\big(\frac{1}{\sqrt{n}}\big)$ term, which is a bottleneck on the utility analysis. Thus, we are focusing on the following question, which is still an open problem:

Can we achieve the high probability excess risk bounds with rate $\mathcal{O}(\frac{\sqrt{p}}{n\epsilon})$ for differentially private models via uniform stability?

This paper answers the question positively under more (or different) assumptions and provides the first high probability bound allowing an $\mathcal{O}\big(\frac{\sqrt{p}}{n\epsilon}\big)$ rate of convergence in the setting of DP. By introducing *Generalized Bernstein condition* (Koltchinskii, 2006), we remove the $\mathcal{O}\big(\frac{1}{\sqrt{n}}\big)$ term in the generalization error and furthermore improve the high probability excess population risk bound. Comparing with previous high probability bounds, the improvement is approximately up to $\mathcal{O}\left(\sqrt{n}\right)$.

CONTRIBUTIONS

We first prove that by introducing Generalized Bernstein condition (Koltchinskii, 2006), under the assumptions G-Lipschitz, L-smooth, and Polyak-Łojasiewicz (PL) condition, the high probability excess population risk bound can be improved to $\mathcal{O}\big(\frac{\sqrt{p}}{n\epsilon}\big)$. To the best of our knowledge, this is the first $\mathcal{O}\big(\frac{\sqrt{p}}{n\epsilon}\big)$ high probability excess population risk bound in the field of DP.

Then, we relax the assumptions G-Lipschitz and L-smooth, by introducing α-Hölder smooth. Under these assumptions, we prove that the high probability excess population risk bound comes to $\mathcal{O}\big(\frac{\sqrt{p}}{\epsilon} n^{\frac{-2\alpha}{1+2\alpha}}\big)$. Considering that $\alpha \in (0, 1]$, the result cannot achieve $\mathcal{O}\big(\frac{\sqrt{p}}{n\epsilon}\big)$.

To overcome the bottleneck, we design a variant of gradient perturbation method, called **max $\{1, g\}$-Normalized Gradient Perturbation** (m-NGP) algorithm. Via this new proposed algorithm, we prove that under the assumptions α-Hölder smooth, PL condition, and generalized Bernstein condition, the high probability excess population risk bound can be improved to $\mathcal{O}\big(\frac{\sqrt{p}}{n\epsilon}\big)$. To the best of our knowledge, this is the first $\mathcal{O}\big(\frac{\sqrt{p}}{n\epsilon}\big)$ high probability excess population risk bound for non-smooth loss in the field of DP.

Moreover, to evaluate the performance of our proposed max $\{1, g\}$-Normalized Gradient Perturbation algorithm, we perform experiments on real datasets, the experimental results show that m-NGP method also improves the accuracy of the DP model on real datasets.

The rest of the paper is organized as follows. We discuss some related work in Section 2. Some preliminaries are formally introduced in Section 3. In Section 4, we propose sharper utility bounds under different assumptions and design a variant of gradient perturbation method, **max $\{1, g\}$-Normalized Gradient Perturbation**. The experimental results are shown in Section 5. Finally, we conclude the paper in Section 6.

2 RELATED WORK

Dwork et al. (2006) proposed the mathematical definition of DP for the first time. Then, it was developed to protect the privacy in the field of machine learning (e.g. Empirical Risk Minimization (ERM)) via output perturbation, objective perturbation, and gradient perturbation methods. For DP-ERM formulations, Chaudhuri et al. (2011) first proposed output perturbation and objective perturbation methods, and Song et al. (2013) first proposed the gradient perturbation method. Based on these works, Kifer et al. (2012); Bassily et al. (2014); Abadi et al. (2016); Wang et al. (2017); Zhang et al. (2017); Wu et al. (2017); Bassily et al. (2019); Feldman et al. (2020); Bassily et al. (2020) further improved the results under different assumptions.

Table 1: Previous excess population risk bounds and ours under different assumptions

	Assumptions	Method	Utility Bound
Bassily et al. (2019)	Lipschitz, smooth, convex	Gradient	$\mathcal{O}\left(\frac{\sqrt{p}}{n\epsilon}+\frac{1}{\sqrt{n}}\right)$
Feldman et al. (2020)	Lipschitz, convex	Gradient	$\mathcal{O}\left(\frac{\sqrt{p}}{n\epsilon}+\frac{1}{\sqrt{n}}\right)$
Bassily et al. (2020)	Lipschitz, convex	Gradient	$\mathcal{O}\left(\frac{\sqrt{p}}{n\epsilon}+\frac{1}{\sqrt{n}}\right)$
Wang et al. (2021)	α-Hölder smooth, convex	Gradient	$\mathcal{O}\left(\frac{\sqrt{p}}{n\epsilon}+\frac{1}{\sqrt{n}}\right)$
Wang et al. (2021)	α-Hölder smooth, convex	Output	$\mathcal{O}\left(\frac{\sqrt[4]{p}}{\sqrt{n\epsilon}}\right)$
Ours	Lipschitz, smooth, PL condition	Gradient	$\mathcal{O}\left(\frac{\sqrt{p}}{n\epsilon}\right)$
Ours	α-Hölder smooth, PL condition	Gradient	$\mathcal{O}\left(\frac{\sqrt{p}}{n^{\frac{2\alpha}{1+2\alpha}}\epsilon}\right)$
Ours (m-NGP)	α-Hölder smooth, PL condition	Gradient	$\mathcal{O}\left(\frac{\sqrt{p}}{n\epsilon}\right)$

[1] In Table 1, n is the size of the dataset, ϵ is the privacy budget, and p is the dimension of the data.

Among the works mentioned above, some of them only analyzed the privacy guarantees (Song et al., 2013; Abadi et al., 2016), some of them only discussed the excess empirical risk bound (Wang et al., 2017; Zhang et al., 2017; Wu et al., 2017). Some works discussed the excess population risk under expectation, from different points of view, such as complexity theory, optimization theory, and stability theory: Kifer et al. (2012) achieved an $\mathcal{O}\big(\frac{\sqrt{p}}{n\epsilon}\big)$ expected excess population risk bound via complexity theory; Bassily et al. (2014) achieved similar expected bound under convexity assumption, via optimization theory; and Wang et al. (2019) proposed an $\mathcal{O}\big(\frac{p}{\log(n)\epsilon^2}\big)$ expected excess population risk bound under non-convex condition, via Langevin Dynamics (Gelfand & Mitter, 1991) and the stability of Gibbs algorithm.

Considering that the high probability bound is more concerned by researchers, we focus on the high probability utility bound. Meanwhile, we concentrate on the stability theory in this paper. Among many notions of stability, uniform stability is arguably the most popular one, which yields exponential generalization bounds. Via uniform stability, the high probability excess population risk bounds under different assumptions given by previous works all contain an $\mathcal{O}\big(\frac{1}{\sqrt{n}}\big)$ term, details can be found in Table 1. The reason is that when analyzing the generalization error, the technical routes followed works Bousquet & Elisseeff (2002); Hardt et al. (2016).

In this paper, by introducing *Generalized Bernstein condition* (Koltchinskii, 2006), we remove the $\mathcal{O}\big(\frac{1}{\sqrt{n}}\big)$ term from the generalization error, and further improve the excess population risk bound of DP models. The improved convergence rate is up to $\mathcal{O}\big(\frac{\sqrt{p}}{n\epsilon}\big)$, which positively answers the question: Can the high probability excess population risk bound achieve $\mathcal{O}\left(1/n\right)$ w.r.t n. The improvements are shown in Table 1.

Table 1 first shows that by adding more assumptions (we assume the loss function to be Lipschitz, smooth, and satisfy Polyak-Łojasiewicz (PL) condition, while previous results require α-Hölder smoothness and convexity), we achieve a better high probability excess population risk bound, $\mathcal{O}\big(\frac{\sqrt{p}}{n\epsilon}\big)$, which is state-of-the-art to the best of our knowledge. Then, we replace the Lipschitz and smooth property by α-Hölder smoothness and achieve $\mathcal{O}\big(\frac{\sqrt{p}}{n^{\frac{2\alpha}{1+2\alpha}}\epsilon}\big)$ high probability excess population risk bound, when $\alpha \in [\frac{1}{2}, 1]$, our result is better than previous ones, but it cannot achieve

the same bound ($\mathcal{O}(1/n)$ w.r.t n) under the condition that the loss function is Lipschitz, smooth, and satisfies PL condition. To overcome it, we propose an algorithm called m-NGP, and achieve the $\mathcal{O}\big(\frac{\sqrt{p}}{n\epsilon}\big)$ result under the same assumptions: α-Hölder smooth and PL condition.

Moreover, although it is hard to directly compare PL condition with convexity, PL condition can be applied to many non-convex conditions (more information can be found in Section 4.2). So, in this paper, we analyze the utility bound of DP algorithm under cases different from previous scenarios.

3 PRELIMINARIES

In this paper, we assume that there are n data instances in dataset D, i.e. $D = \{z_1, \cdots, z_n\}$ where $z = (x, y)$ with input $x \in \mathcal{X}$ and label $y \in \mathcal{Y}$, and $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. The data space is denoted by $\mathcal{D}$ and the parameter space is denoted by $\mathcal{C}$, the loss function ℓ is defined as $\ell(\cdot, \cdot) : \mathcal{D} \times \mathcal{C} \to \mathbb{R}$. Databases $D, D' \in \mathcal{D}^n$ differing by one data instance are denoted as $D \sim D'$, called *adjacent databases*. For a given vector $\boldsymbol{x} = [x_1, \cdots, x_d]^T$, its ℓ_2-norm is $\|\boldsymbol{x}\|_2 = (\sum_{i=1}^{d} |x_i|^2)^{\frac{1}{2}}$. And $A \lesssim B$ represents that there exists $c > 0$, $A \leq cB$.

Definition 1 (Differential Privacy (Dwork et al., 2006)). *A randomized algorithm: $\mathcal{A} : \mathcal{D}^n \to \mathbb{R}^p$ is (ϵ, δ)-differential privacy (DP) if for all $D \sim D'$ and events $S \in range(\mathcal{A})$:*

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

Definition 1 implies that the adversaries cannot infer whether an individual participates when training the machine learning model, because essentially the same distributions will be drawn over any adjacent datasets. Some kind of attacks, such as membership inference attack, attribute inference attack, and memorization attack, can be thwarted by DP (Backes et al., 2016; Jayaraman & Evans, 2019; Carlini et al., 2019).

Throughout this paper, we focus on gradient perturbation method to guarantee (ϵ, δ)-DP, the paradigm is based on gradient descent: at iteration t,

$$\hat{\theta}_t \leftarrow \hat{\theta}_{t-1} - \eta_t \left(\nabla_\theta R_n(\hat{\theta}_{t-1}) + b\right), \tag{1}$$

where η_t is the learning rate, b is the random noise injected into the gradient, $\hat{\theta}$ is corresponding model with privacy, and $R_n(\theta)$ is the empirical risk, defined as $R_n(\theta) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \ell(z_i, \theta)$.

In this paper, we focus on minimizing the population risk: $R(\theta) = \mathbb{E}_{z\sim\mathcal{D}}\left[\ell(z, \theta)\right]$. In the setting of DP, the excess population risk is defined by $R(\hat{\theta}) - \min_{\theta\in\mathcal{C}} R(\theta)$, which can be decomposed into:

$$\begin{aligned} R(\hat{\theta}_n) - R(\theta^*) &= R(\hat{\theta}_n) - R_n(\hat{\theta}_n) + R_n(\hat{\theta}_n) - R_n(\theta^*) + R_n(\theta^*) - R(\theta^*) \\ &\leq \underbrace{R(\hat{\theta}_n) - R_n(\hat{\theta}_n)}_{\text{GE}} + \underbrace{R_n(\hat{\theta}_n) - R_n(\theta^*_{\mathbf{n}})}_{\text{OE}} + R_n(\theta^*) - R(\theta^*), \end{aligned} \tag{2}$$

where $\theta^* = \arg\min_{\theta\in\mathcal{C}} R(\theta), \theta^*_n = \arg\min_{\theta\in\mathcal{C}} R_n(\theta)$, and the last inequality is becasuse of the definition of θ^*_n. In (2), GE, OE mean the generalization error and the optimization error (also called the excess empirical risk), respectively. Inequality (2) answers the question mentioned in Section 1: Why generalization error is an important step towards excess population risk.

To get the generalization error, algorithm stability theory is a popular tool, in which uniform stability yields exponential generalization bounds and is commonly used.

Definition 2 (Uniform Stability (Bousquet & Elisseeff, 2002)). *An algorithm θ_n is γ-uniformly stable if for any $z, z_1, \cdots, z_i, \cdots, z_n, z'_i \in \mathcal{Z}$ and $i = 1, \cdots, n$, it holds that*

$$|\ell(z, \theta_n(z_1, \cdots, z_n)) - \ell(z, \theta_n(z_1, \cdots, z_{i-1}, z'_i, z_{i+1}, \cdots, z_n))| \leq \gamma.$$

In this paper, we use notation θ_n for both algorithm and model parameter. By Definition 2, it is easy to follow that the uniform stability measures the upper bound of the difference (on the loss function) between the models derived from adjacent datasets.

Assumption 1 (G-Lipschitz). *The loss function $\ell : \mathcal{D} \times \mathcal{C} \to \mathbb{R}$ is G-Lipschitz over θ if for any $z \in \mathcal{D}$ and $\theta_1, \theta_2 \in \mathcal{C}$, we have: $|\ell(z, \theta_1) - \ell(z, \theta_2)| \leq G\|\theta_1 - \theta_2\|_2$.*

Assumption 2 (L-smooth)**.** *The loss function $\ell : \mathcal{D} \times \mathcal{C} \to \mathbb{R}$ is L-smooth over θ if for any $z \in \mathcal{D}$ and $\theta_1, \theta_2 \in \mathcal{C}$, we have: $\|\nabla_\theta \ell(z, \theta_1) - \nabla_\theta \ell(z, \theta_2)\|_2 \leq L\|\theta_1 - \theta_2\|_2$.*

If ℓ is differentiable, smoothness yields: $\ell(z, \theta_1) - \ell(z, \theta_2) \leq \langle \nabla_\theta \ell(z, \theta_2), \theta_1 - \theta_2 \rangle + \frac{L}{2} \|\theta_1 - \theta_2\|_2^2$.

Assumptions G-Lipschitz and L-smooth are commonly used in the utility analysis of DP machine learning (Chaudhuri et al., 2011; Kifer et al., 2012; Abadi et al., 2016; Bassily et al., 2019; Feldman et al., 2020; Bassily et al., 2020). To relax the Lipschitz and smoothness assumptions, we introduce the α-Hölder smoothness of the loss function:

Assumption 3 (α-Hölder smooth)**.** *Let $\alpha \in (0, 1]$. The loss function $\ell : \mathcal{D} \times \mathcal{C} \to \mathbb{R}$ is α-Hölder smooth over θ with parameter H if for any $z \in \mathcal{D}$ and $\theta_1, \theta_2 \in \mathcal{C}$, we have: $\|\nabla_\theta \ell(z, \theta_1) - \nabla_\theta \ell(z, \theta_2)\|_2 \leq H\|\theta_1 - \theta_2\|_2^\alpha$.*

Lemma 1. *If the loss function $\ell(\cdot, \cdot)$ is differentiable, then Assumption 3 yields $\ell(z, \theta_1) - \ell(z, \theta_2) \leq \langle \nabla_\theta \ell(z, \theta_2), \theta_1 - \theta_2 \rangle + \frac{H}{2} \|\theta_1 - \theta_2\|_2^{\alpha+1}$.*

By the definition, it is easy to follow that if $\alpha = 1$, it is equivalent to H-smooth; and if $\alpha \to 0$, it satisfies the Lipschitz property given in Assumption 1. Besides, with bounded parameter space, i.e. $\|\mathcal{C}\|_2 \leq M_{\mathcal{C}}$, α-Hölder smoothness immediately implies $\max\{2HM_{\mathcal{C}}, H\}$-Lipschitz. Moreover, Assumption 3 instantiates many non-smooth loss functions. For example, the q-norm hinge loss $\ell(z, \theta) = (\max(0, 1 - y\langle \theta, z \rangle))^q$ for classification and the q-th power absolute distance loss $\ell(z, \theta) = |y - \langle \theta, z \rangle|^q$ for regression (Lei & Ying, 2020a), whose ℓ are $(q-1)$-Hölder smooth if $q \in (1, 2]$ (Li & Liu, 2021). Lemma 1 shows that Hölder smoothness shares similar property with smoothness defined in Assumption 2, details of the proof can be found in Appendix A.1.

4 Sharper utility bounds for differentially private models

4.1 Privacy Guarantees

Before analyzing the excess population risk bound, we first discuss the privacy guarantees in this section. Abadi et al. (2016) proposed the moments accountant method to measure the privacy costs of DP model training by stochastic gradient descent (SGD), Wang et al. (2017) further analyzed it under the setting of gradient descent (GD). In this paper, we focus more on the utility analysis, to improve the excess population risk, so we directly apply it to the gradient perturbation method.

Lemma 2 (Wang et al. (2017))**.** *In gradient perturbation method in (1), for $\epsilon, \delta > 0$, it is (ϵ, δ)-DP if the random noise b is zero mean Gaussian noise, i.e. $b \sim \mathcal{N}(0, \sigma^2 I_p)$, and for some constant c,*

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

Remark 1. *(3) assumes the loss function to be G-Lipschitz. If we only assume that $\ell(\cdot, \cdot)$ is α-Hölder smooth with parameter H, then G can be replaced by $\max\{2HM_{\mathcal{C}}, H\}$ as discussed above.*

4.2 Analysis of the excess population risk

To remove the $\mathcal{O}(1/\sqrt{n})$ term in previous results, we further need the Generalized Bernstein condition when analyzing the excess population risk.

Assumption 4 (Generalized Bernstein condition (Koltchinskii, 2006))**.** *We say the loss function ℓ satisfies the generalized Bernstein condition if for some $B > 0$ for any $\theta \in \mathcal{C}$, we have:*

$$\mathbb{E}\left[\left(\ell(z, \theta) - \ell(z, \theta^*)\right)^2\right] \leq B\left(R(\theta) - R(\theta^*)\right).$$

Assumption 4 is a general condition, if the loss function $\ell(\cdot, \cdot)$ is G-Lipschitz and bounded by M_ℓ, then many loss functions satisfy the generalized Bernstein condition, such as exponential loss function, logistic loss function, quadratic loss function, truncated quadratic loss, and hinge loss (Bartlett et al., 2006; Steinwart & Christmann, 2008).

Most of the previous works assumed that the loss function is convex (or strongly convex) when analyzing the optimization error (the excess empirical risk) $R_n(\hat{\theta}_n) - R_n(\theta_n^*)$. In this paper, we use the Polyak-Łojasiewicz (PL) condition to replace the convexity assumption.

Assumption 5 (Polyak-Łojasiewicz condition)**.** *The empirical risk $R_n(\theta)$ satisfies the Polyak-Łojasiewicz (PL) condition if there exists $\mu > 0$ and for every θ,*

$$\|\nabla_\theta R_n(\theta)\|_2^2 \geq 2\mu \left(R_n(\theta) - R_n(\theta_n^*)\right).$$

The Polyak-Łojasiewicz condition is one of the weakest curvature conditions, so all the results given in this paper can be expanded to strongly convex conditions. (Karimi et al., 2016; Li & Liu, 2021), weaker than 'one-point convexity' (Kleinberg et al., 2018), 'star convexity' (Zhou et al., 2019), and 'quasar convexity' (Hinder et al., 2020). It is widely used in the analysis of non-convex learning (Wang et al., 2017; Charles & Papailiopoulos, 2018; Lei & Ying, 2020b; Lei & Tang, 2021) and many popular non-convex objective functions satisfy the PL condition, such as: matrix factorization (Liu et al., 2016), robust regression (Liu et al., 2016), neural networks with one hidden layer (Li & Yuan, 2017), mixture of two Gaussians (Balakrishnan et al., 2017), ResNets with linear activations (Hardt & Ma, 2017), linear dynamical systems (Hardt et al., 2018), phase retrieval (Sun et al., 2018), and blind deconvolution (Li et al., 2019).

Remark 2. *With G-Lipschitz and λ-strongly convex, we have $\mathbb{E}\big[\big(\ell(z,\theta) - \ell(z,\theta^*)\big)^2\big] \leq G^2\|\theta - \theta^*\|_2^2$, and $R(\theta) - R(\theta^*) \geq \frac{\lambda}{2}\|\theta - \theta^*\|_2^2$, which implies $\mathbb{E}\big[\big(\ell(z,\theta) - \ell(z,\theta^*)\big)^2\big] \leq \big(2G^2/\lambda\big)\big(R(\theta) - R(\theta^*)\big)$, Assumption 4 is naturally satisfied. And PL condition can be directly derived from strongly convex (Karimi et al., 2016), so all strongly convex loss functions satisfy Assumptions 4 and 5 simultaneously and all the results given in this paper can be directly extended to strongly convex condition. Expect for strongly convex functions, several interesting machine learning setups also satisfy Assumptions 4 and 5. (1) 1-layer neural networks with a squared error loss and leaky ReLU activations. Charles & Papailiopoulos (2018) shows that 1-layer neural networks with a squared error loss and leaky ReLU activations satisfy Assumption 5, and Bartlett et al. (2006) shows that quadratic functions satisfy Assumption 4, so (1) holds. (2) Loss functions of least squares minimizations. Charles & Papailiopoulos (2018) shows that least squares minimization satisfy Assumption 5 and Bartlett et al. (2006) shows that the quadratic functions satisfy Assumption 4, so (2) holds. (3) Squared piecewise-linear functions with regularized term. Bartlett et al. (2006) shows that the composition of strongly convex functions with piecewise-linear functions satisfy Assumption 5, and Bartlett et al. (2006) shows that squared piecewise-linear functions satisfy Assumption 4. We prove that if a function satisfies Assumption 4, then with regularized term $\lambda\|\theta\|_2^2$, it also satisties Assumption 4 (details can be found in Appendix A.5). Thus, (3) holds.*

Theorem 1. *If Assumptions 1, 2, 4 and 5 hold, the loss function is bounded, i.e. $0 \leq \ell(\cdot,\cdot) \leq M_\ell$, taking σ given by Lemma 2, $T = \mathcal{O}\left(\log(n)\right)$, $\eta_1 = \cdots = \eta_T = \frac{1}{L}$, if $\zeta \in (\exp(-p/8), 1)$, then with probability at least $1 - \zeta$:*

$$\begin{aligned} R(\hat{\theta}_n) - R(\theta^*) \leq\ & c_1 \frac{G^2 p \log(n)\log(1/\delta)}{n^2\epsilon^2}\left(1 + \left(\frac{8\log(T/\zeta)}{p}\right)^{1/4}\right)^2 \\ & + c_2\left(\frac{G^2\log^2(n)}{n} + \frac{B + M_\ell}{n}\right) \\ & + c_3 \frac{G^2\log^{2.5}(n)\sqrt{p\log(1/\delta)}}{n\epsilon}\left(1 + \left(\frac{8\log(1/\zeta)}{p}\right)^{1/4}\right). \end{aligned}$$

for some constants $c_1, c_2, c_3 > 0$.

Detailed proof can be found in Appendix A.2, we give a proof sketch here. First, we discuss the stability of the gradient perturbation based DP algorithm and show that it is $\mathcal{O}\left(T\eta/n\right)$ uniformly stable w.r.t n with high probability. Then, we analyze the generalization error via stability theory. Meanwhile, via Assumption 4 and its moments bound, we couple term $R(\hat{\theta}_n) - R_n(\hat{\theta}_n)$ (the generalization error of $\hat{\theta}$) and term $R_n(\theta^*) - R(\theta^*)$ in (2) together, to remove the $\mathcal{O}\left(1/\sqrt{n}\right)$ term in the generalization error. In this way, a better excess population risk bound is achieved by combining the optimization error together.

The proof is motivated by Klochkov & Zhivotovskiy (2021) in the non-private case. The key challenges include that in the setting of DP, the random noise is injected into the algorithm. In Klochkov & Zhivotovskiy (2021), a key step to analyze the generalization error is summing

$X_i = \mathbb{E}'\left[\ell(z_i, \theta_n') - \ell(z_i, \theta^*)\right]$ for $i = 1, \cdots, n$, where θ_n' is derived from an independent copy of the original dataset and $\mathbb{E}'$ means the expectation taken over the independent copy. When summing, X_i is required to be zero mean. However, in the cases of DP, if we replace θ_n' by $\hat{\theta}_n'$, then X_i are not zero mean. Besides, for output perturbation, a common way to decompose the excess population risk is $R(\hat{\theta}_n) - R(\theta^*) \leq R(\hat{\theta}_n) - R(\theta_n) + R(\theta_n) - R_n(\theta_n) + R_n(\theta_n) - R_n(\theta_n^*) + R_n(\theta^*) - R(\theta^*)$, which naturally solves the problem mentioned above (because the generalization error is discussed over the non-private model). However, when it comes to the gradient perturbation method, we cannot solve the problem easily in this way, because the random noise is coupled with the gradient. So, we decouple the noise terms and overcome the challenge by the moment Bernstein inequality.

By Theorem 1, it is easy to follow that with high probability, $R(\hat{\theta}_n) - R(\theta^*) = \mathcal{O}\big(\frac{\sqrt{p}}{n\epsilon}\big)$, which is the first $\mathcal{O}\left(1/n\right)$ high probability excess population risk bound over DP algorithm w.r.t n, to the best of our knowledge.

Theorem 2. *If Assumptions 3, 4, 5 hold, the loss function and the parameter space are bounded, i.e.* $0 \leq \ell(\cdot, \cdot) \leq M_\ell$, $\|\mathcal{C}\|_2 \leq M_\mathcal{C}$. *Taking* σ *given by Lemma 2,* $T = \mathcal{O}\left(n^{\frac{2}{1+2\alpha}}\right)$, *and* $\eta_t = \frac{2}{\mu(t+\kappa)}$, *where* $\kappa \geq \frac{2H^{1/\alpha}}{\mu}$, *if* $\zeta \in (\exp(-p/8), 1)$, *then with probability at least* $1 - \zeta$*:*

$$
\begin{aligned}
R(\hat{\theta}_n) - R(\theta^*) \leq c_1 \frac{G'^2 \sqrt{p \log(1/\delta)}}{n^{\frac{2\alpha}{1+2\alpha}} \epsilon} \left(1 + \left(\frac{8 \log(T/\zeta)}{p}\right)^{1/4}\right) \\
+ c_2 \left(\frac{G'^2 \log^2(n)}{n} + \frac{B + M_\ell}{n}\right) \\
+ c_3 \frac{G'^2 \log^2(n) \sqrt{p \log(1/\delta)}}{n^{\frac{2\alpha}{1+2\alpha}} \epsilon} \left(1 + \left(\frac{8 \log(1/\zeta)}{p}\right)^{1/4}\right).
\end{aligned}
$$

for some constants $c_1, c_2, c_3 > 0$, *where* $G' = \max\{2HM_\mathcal{C}, H\}$.

Detailed proof can be found in Appendix A.3. The proof is similar to Theorem 1, the challenge is that the properties G-Lipschitz and L-smooth are replaced by the assumption α-Hölder smooth when analyzing the optimization error (the excess empirical risk). To overcome the challenge, we use Lemma 1 to bound the optimization error and Young's inequality is used to normalize the exponential rate, details are shown in the proof of Lemma 8.

By Theorem 2, it is easy to follow that with high probability,

$$
R(\hat{\theta}_n) - R(\theta^*) = \mathcal{O}\left(\frac{\sqrt{p}}{\epsilon} n^{\frac{-2\alpha}{1+2\alpha}}\right).
$$

By the definition of α-Hölder smooth, $\alpha \in (0, 1]$, so if $\alpha \in [\frac{1}{2}, 1]$,

$$
R(\hat{\theta}_n) - R(\theta^*) = \mathcal{O}\left(n^{\frac{-2\alpha}{1+2\alpha}}\right) \leq \mathcal{O}\left(n^{-\frac{1}{2}}\right)
$$

w.r.t n, which implies that our result is better than previous results when $\alpha \in [\frac{1}{2}, 1]$.

Via the discussion mentioned above, we observe that under the assumption α-Hölder smooth, our result is better than $\mathcal{O}(1/\sqrt{n})$ w.r.t n only in the case that $\alpha \in [\frac{1}{2}, 1]$. Besides, the best result is $\mathcal{O}\left(n^{-2/3}\right)$, which comes when $\alpha = 1$. And it cannot achieve the convergence rate $\mathcal{O}(\frac{\sqrt{p}}{n\epsilon})$. The reason is that when applying Young's inequality in the optimization error analysis, an additional term $\frac{H\eta_t^{\alpha+1}(1-\alpha)}{2(\alpha+1)}$ appears, leading a loose excess population risk bound.

Motivated by this, we design a variant of gradient perturbation method given in (1), called **max$\{1, g\}$-Normalized Gradient Perturbation** DP algorithm, to overcome the loose excess population risk bound. Details are shown in Algorithm 1.

Remark 3. *The difference between Algorithm 1 and (1) is that in lines 4 and 5, we normalize the* ℓ_2*-norm of the gradient to* 1 *if it is less than* 1. *In this way, we can 'bypass' the Young's inequality when scaling* $\|\theta_t - \theta_n^*\|_2^{1+\alpha}$ *(derived from Lemma 1), further remove term* $\frac{H\eta_t^{\alpha+1}(1-\alpha)}{2(\alpha+1)}$ *in the theoretical analysis. Details can be found in Appendix A.4.*

Algorithm 1 $\max\{\mathbf{1}, \mathbf{g}\}$-Normalized Gradient Perturbation

Require: dataset D, learning rate at iteration t: η_t, the variance of the Gaussian noise injected to the gradient: σ.

function M-NGP(D, η_t, σ)
  Initialize θ_0.
  for $t = 0$ to $T-1$ **do**
    if $\left\|\nabla_\theta R_n(\hat{\theta}_t)\right\|_2 < 1$ **then**
      $\nabla_\theta R_n(\hat{\theta}_t) \leftarrow \nabla_\theta R_n(\hat{\theta}_t) / \left\|\nabla_\theta R_n(\hat{\theta}_t)\right\|_2$.
    endif
    $\hat{\theta}_{t+1} \leftarrow \hat{\theta}_t - \eta_t \left(\nabla_\theta R_n(\hat{\theta}_t) + b\right)$, where $b \sim \mathcal{N}\left(0, \sigma^2 I_p\right)$.
  endfor
  return $\hat{\theta}_n = \hat{\theta}_T$.
end function

Then, via Algorithm 1, we can improve the excess population risk bound as shown below.

Theorem 3. *If Assumptions 3, 4, 5 hold, the loss function and the parameter space are bounded, i.e.* $0 \leq \ell(\cdot,\cdot) \leq M_\ell$, $\|\mathcal{C}\|_2 \leq M_\mathcal{C}$. *Taking* σ *given by Lemma 2,* $T = \mathcal{O}(\log(n))$, *and* $\eta_1 = \cdots = \eta_T = \eta$, *where* $\left(\frac{2}{H} - \frac{2^{-1/\alpha}}{\mu H^{(\alpha-1)/\alpha}}\right)^{1/\alpha} < \eta < \left(\frac{2}{H}\right)^{1/\alpha}$, *if* $\zeta \in (\exp(-p/8), 1)$, *then with probability at least* $1-\zeta$,

$$
\begin{aligned}
R(\hat{\theta}_n) - R(\theta^*) \leq\ & c_1 \frac{G' \sqrt{p \log(n) \log(1/\delta)}}{n\epsilon} \left(1 + \left(\frac{8\log(T/\zeta)}{p}\right)^{1/4}\right) \\
& + c_2 \left(\frac{G'^2 \log^2(n)}{n} + \frac{B + M_\ell}{n}\right) \\
& + c_3 \frac{G'^2 \log^{2.5}(n) \sqrt{p \log(1/\delta)}}{n\epsilon} \left(1 + \left(\frac{8\log(1/\zeta)}{p}\right)^{1/4}\right),
\end{aligned}
$$

for some constants $c_1, c_2, c_3 > 0$, *where* $G' = \max\{2HM_\mathcal{C}, H\}$.

Detailed proof can be found in Appendix A.4. The proof is similar to Theorems 1 and 2, the key difference is that by *gradient normalization* in Algorithm 1, Young's inequality is abandoned in the theoretical analysis (as discussed in Remark 3), which implies a better excess population risk bound.

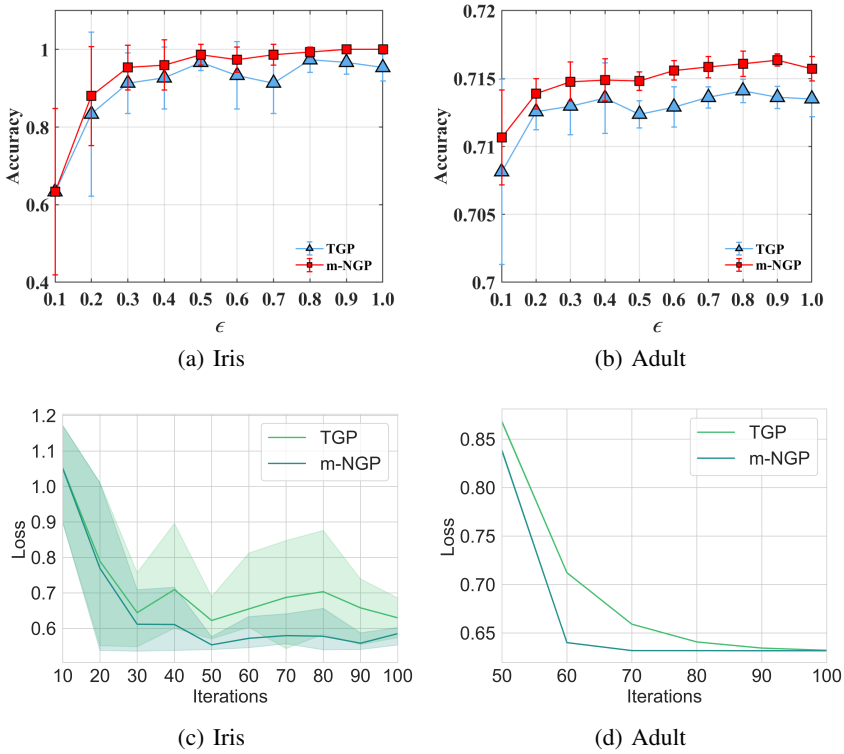
By Theorem 3, it is easy to follow that with high probability, $R(\hat{\theta}_n) - R(\theta^*) = \mathcal{O}\left(\frac{\sqrt{p}}{n\epsilon}\right)$. The bound is of the same order as the result given in Theorem 1.This is also the first $\mathcal{O}(1/n)$ high probability excess population risk bound over DP algorithm w.r.t n without smoothness assumption.

5 Experiments

In this section, we perform experiments on real datasets to evaluate the difference between our proposed m-NGP algorithm and the traditional gradient perturbation (TGP), like (1).

The experiments are performed on classification task over datasets Iris (Dua & Graff, 2017), Breast Cancer (Mangasarian & Wolberg, 1990), Credit Card Fraud (Bontempi & Worldline, 2018), Bank (Moro et al., 2014), and Adult (Dua & Graff, 2017), the number of total data instances are 150, 699, 984, 41188, and 45222, respectively. We split the training and testing sets randomly and evaluate the accuracy on the testing set and the convengence rate on the training set. In all the experiments, the privacy budget δ is set $\frac{1}{n}$ and we choose $\epsilon = 0.1$ to 1.0.

We apply the regularized logistic regression method to the classification task, the loss function satisfies the assumptions mentioned before, and the experimental results are shown in Figure 1. We show the experimental results over datasets Iris and Adult in this section and experiments on other datasets are shown in Appendices B.1 and B.2. For convergence rate, the shadow area represents the



(a) Iris (b) Adult

(c) Iris (d) Adult

Figure 1: Comparisons between Traditional Gradient Perturbation (TGP) method and $\max\{\mathbf{1}, \mathbf{g}\}$-Normalized Gradient Perturbation (m-NGP) method.

maximum and minimum loss over mutiple experiments, reflecting the variance. The shadow area in part (d) of Figure 1 is not obvious, the reason is that the variances are small. Over most datasets, the accuracy and the convergence rate of $\max\{\mathbf{1}, \mathbf{g}\}$-Normalized Gradient Perturbation method is better than traditional gradient perturbation method. Besides, the accuracy of the DP model increases with the increasing of the privacy budget ϵ, which is in line with the theoretical analysis.

6 Conclusions

In this paper, we first propose a state-of-the-art $\mathcal{O}\big(\frac{\sqrt{p}}{n\epsilon}\big)$ high probability excess population risk bound for gradient perturbation based DP algorithms, under the assumptions of G-Lipschitz, L-smooth, Polyak-Łojasiewicz condition, and generalized Bernstein condition. The result positively answers the open problem: *Can we achieve high probability excess risk bound with rate* $\mathcal{O}(1/n)$ *w.r.t* n *for DP models via uniform stability?* Then, we extend the result to a more general case, requiring α-Hölder smoothness, Polyak-Łojasiewicz condition, and generalized Bernstein condition. However, the result is not as satisfactory as before, we achieve an $\mathcal{O}\big(n^{\frac{-2\alpha}{1+2\alpha}}\big)$ high probability utility bound, which is better than previous results when $\alpha \in [\frac{1}{2}, 1]$ and cannot achieve an $\mathcal{O}\left(1/n\right)$ bound. To get a better result, we further propose a new algorithm: $\max\{1, g\}$-Normalized Gradient Perturbation (m-NGP). Detailed theoretical analysis shows that m-NGP can achieve $\mathcal{O}\big(\frac{\sqrt{p}}{n\epsilon}\big)$ high probability excess population risk bound, under the assumptions of α-Hölder smoothness, Polyak-Łojasiewicz condition, and generalized Bernstein condition, which is the first $\mathcal{O}\left(1/n\right)$ high probability bound w.r.t n under non-smoothness cases. Experimental results show that the accuracy of m-NGP algorithm is better than traditional gradient perturbation method. Thus, our proposed $\max\{1, g\}$-Normalized Gradient Perturbation method improves the excess population risk bound and the accuracy of the DP model over real datasets, simultaneously.

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 308–318, 2016.

Michael Backes, Pascal Berrang, Mathias Humbert, and Praveen Manoharan. Membership privacy in microrna-based studies. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 319–330, 2016.

Sivaraman Balakrishnan, Martin J. Wainwright, and Bin Yu. Statistical guarantees for the em algorithm: From population to sample-based analysis. *The Annals of Statistics*, pp. 77 – 120, 2017.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Localized rademacher complexities. In *Computational Learning Theory, 15th Annual Conference on Computational Learning Theory, COLT 2002*, pp. 44–58, 2002.

Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, pp. 138–156, 2006.

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

Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Guha Thakurta. Private stochastic convex optimization with optimal rates. In *Advances in Neural Information Processing Systems*, pp. 11279–11288, 2019.

Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. In *Advances in Neural Information Processing Systems*, pp. 4381–4391, 2020.

Garrett Bernstein and Daniel R Sheldon. Differentially private bayesian linear regression. In *Advances in Neural Information Processing Systems*, pp. 523–533, 2019.

Gianluca Bontempi and Worldline. ULB the machine learning group, 2018.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, pp. 499–526, 2002.

Olivier Bousquet, Yegor Klochkov, and Nikita Zhivotovskiy. Sharper bounds for uniformly stable algorithms. In *Conference on Learning Theory*, pp. 610–626, 2020.

Mark Bun, Marek Elias, and Janardhan Kulkarni. Differentially private correlation clustering. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 1136–1146, 2021.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pp. 267–284, 2019.

Zachary Charles and Dimitris Papailiopoulos. Stability and generalization of learning algorithms that converge to global optima. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 745–754, 2018.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, pp. 1069–1109, 2011.

Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. Collecting telemetry data privately. In *Advances in Neural Information Processing Systems*, pp. 3571–3580, 2017.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing, 2009*, pp. 371–380, 2009.

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

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, pp. 211–407, 2014.

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

Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: Optimal rates in linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 439–449, 2020.

Saul B Gelfand and Sanjoy K Mitter. Recursive stochastic algorithms for global optimization in r^d. *SIAM Journal on Control and Optimization*, pp. 999–1018, 1991.

Moritz Hardt and Tengyu Ma. Identity matters in deep learning. In *5th International Conference on Learning Representations, 2017*, 2017.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of The 33rd International Conference on Machine Learning*, pp. 1225–1234, 2016.

Moritz Hardt, Tengyu Ma, and Benjamin Recht. Gradient descent learns linear dynamical systems. *Journal of Machine Learning Research*, pp. 29:1–29:44, 2018.

Mikko Heikkilä, Joonas Jälkö, Onur Dikmen, and Antti Honkela. Differentially private markov chain monte carlo. In *Advances in Neural Information Processing Systems 32*, pp. 4115–4125. 2019.

S. Hettich and S. D. Bay. The uci kdd archive, 1999.

Oliver Hinder, Aaron Sidford, and Nimit Sohoni. Near-optimal methods for minimizing star-convex functions and beyond. In *Proceedings of Thirty Third Conference on Learning Theory*, pp. 1894–1938, 2020.

Bargav Jayaraman and David Evans. Evaluating differentially private machine learning in practice. In *Proceedings of the 28th USENIX Conference on Security Symposium*, pp. 1895–1912, 2019.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 795–811, 2016.

Daniel Kifer, Adam Smith, and Abhradeep Thakurta. Private convex empirical risk minimization and high-dimensional regression. In *Conference on Learning Theory*, pp. 25–1, 2012.

Bobby Kleinberg, Yuanzhi Li, and Yang Yuan. An alternative view: When does SGD escape local minima? In *Proceedings of the 35th International Conference on Machine Learning*, pp. 2698–2707, 2018.

Yegor Klochkov and Nikita Zhivotovskiy. Stability and deviation optimal risk bounds with convergence rate $o(1/n)$. *arXiv preprint arXiv:2103.12024*, 2021.

Vladimir Koltchinskii. Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, pp. 2593–2656, 2006.

Tejas Kulkarni, Joonas Jälkö, Antti Koskela, Samuel Kaski, and Antti Honkela. Differentially private bayesian inference for generalized linear models. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 5838–5849, 2021.

Yunwen Lei and Ke Tang. Learning rates for stochastic gradient descent with nonconvex objectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Yunwen Lei and Yiming Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 5809–5819, 2020a.

Yunwen Lei and Yiming Ying. Sharper generalization bounds for learning with gradient-dominated objective functions. In *International Conference on Learning Representations*, 2020b.

Shaojie Li and Yong Liu. Improved learning rates for stochastic optimization: Two theoretical viewpoints, 2021.

Xiaodong Li, Shuyang Ling, Thomas Strohmer, and Ke Wei. Rapid, robust, and reliable blind deconvolution via nonconvex optimization. *Applied and computational harmonic analysis*, pp. 893–934, 2019.

Yuanzhi Li and Yang Yuan. Convergence analysis of two-layer neural networks with relu activation. In *Advances in Neural Information Processing Systems 30, 2017*, pp. 597–607, 2017.

Huikang Liu, Weijie Wu, and Anthony Man-Cho So. Quadratic optimization with orthogonality constraints: Explicit lojasiewicz exponent and linear convergence of line-search methods. In *Proceedings of the 33nd International Conference on Machine Learning, 2016*, pp. 1158–1167, 2016.

Olvi L Mangasarian and William H Wolberg. Cancer diagnosis via linear programming. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 1990.

Robert McMillan. Apple tries to peek at user habits without violating privacy. *The Wall Street Journal*, 2016.

Sérgio Moro, Paulo Cortez, and Paulo Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, pp. 22–31, 2014.

Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

Dung Nguyen and Anil Vullikanti. Differentially private densest subgraph detection. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8140–8151, 2021.

Vijayakumar Ponnusamy, J. Christopher Clement, K. C. Sriharipriya, and Sowmya Natarajan. *Smart Healthcare Technologies for Massive Internet of Medical Things*, pp. 71–101. Springer International Publishing, 2021.

Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. In *Proceedings of the 2017 Conference on Learning Theory*, pp. 1674–1703, 2017.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy, 2017*, pp. 3–18, 2017.

Monoj Kumar Singha, Priyanka Dwivedi, Gaurav Sankhe, Aniket Patra, and Vineet Rojwal. *Role of Sensors, Devices and Technology for Detection of COVID-19 Virus*, pp. 293–312. Springer International Publishing, 2021.

Shuang Song, Kamalika Chaudhuri, and Anand D Sarwate. Stochastic gradient descent with differentially private updates. In *2013 IEEE Global Conference on Signal and Information Processing*, pp. 245–248, 2013.

Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

Ju Sun, Qing Qu, and John Wright. A geometric analysis of phase retrieval. *Foundations of Computational Mathematics*, pp. 1131–1198, 2018.

G. Swapna and K. P. Soman. *Diabetes Detection and Sensor-Based Continuous Glucose Monitoring – A Deep Learning Approach*, pp. 245–268. Springer International Publishing, 2021.

Jonathan Ullman and Adam Sealfon. Efficiently estimating erdos-renyi graphs with node differential privacy. In *Advances in Neural Information Processing Systems*, pp. 3765–3775, 2019.

Di Wang and Jinhui Xu. Principal component analysis in the local differential privacy model. *Theoretical Computer Science*, 2019.

Di Wang, Minwei Ye, and Jinhui Xu. Differentially private empirical risk minimization revisited: Faster and more general. In *Advances in Neural Information Processing Systems*, pp. 2722–2731, 2017.

Di Wang, Changyou Chen, and Jinhui Xu. Differentially private empirical risk minimization with non-convex loss functions. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, pp. 6526–6535, 2019.

Puyu Wang, Yunwen Lei, Yiming Ying, and Hai Zhang. Differentially private sgd with non-smooth loss. *arXiv preprint arXiv:2101.08925*, 2021.

Xi Wu, Fengan Li, Arun Kumar, Kamalika Chaudhuri, Somesh Jha, and Jeffrey Naughton. Bolt-on differential privacy for scalable stochastic gradient descent-based analytics. In *Proceedings of the 2017 ACM International Conference on Management of Data*, pp. 1307–1322, 2017.

Chugui Xu, Ju Ren, Deyu Zhang, Yaoxue Zhang, Zhan Qin, and Kui Ren. Ganobfuscator: Mitigating information leakage under gan via differential privacy. *IEEE Transactions on Information Forensics and Security*, pp. 2358–2371, 2019.

Zhenhuan Yang, Yunwen Lei, Siwei Lyu, and Yiming Ying. Stability and differential privacy of stochastic gradient descent for pairwise learning with non-smooth loss. In *International Conference on Artificial Intelligence and Statistics*, pp. 2026–2034, 2021.

Jiaqi Zhang, Kai Zheng, Wenlong Mou, and Liwei Wang. Efficient private erm for smooth objectives. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pp. 3922–3928, 2017.

Yi Zhou, Junjie Yang, Huishuai Zhang, Yingbin Liang, and Vahid Tarokh. SGD converges to global minimum in deep learning via star-convex path. In *7th International Conference on Learning Representations, 2019*, 2019.

A Details of Proofs

A.1 Proof of Lemma 1

Lemma. *If $\ell(\cdot,\cdot)$ is differentiable, α-Hölder Smoothness with parameter H yields*

$$\ell(z,\theta_1)-\ell(z,\theta_2)\leq\langle\nabla_\theta\ell(z,\theta_2),\theta_1-\theta_2\rangle+\frac{H}{2}\left\|\theta_1-\theta_2\right\|_2^{\alpha+1}.$$

Proof. First, following Nesterov et al. (2018), for any θ_1,θ_2, and data instance z, we have

$$\begin{aligned}\ell(z,\theta_1)&=\ell(z,\theta_2)+\int_0^1\langle\nabla_\theta\ell(z,\theta_2+\tau\left(\theta_1-\theta_2\right)),\theta_1-\theta_2\rangle\mathrm{d}\tau\\&=\ell(z,\theta_2)+\langle\nabla_\theta\ell(z,\theta_2),\theta_1-\theta_2\rangle\\&\quad+\int_0^1\langle\nabla_\theta\ell(z,\theta_2+\tau\left(\theta_1-\theta_2\right))-\nabla_\theta\ell(z,\theta_2),\theta_1-\theta_2\rangle\mathrm{d}\tau.\end{aligned}$$

By Cauthy-Schwarz inequality,

$$\begin{aligned}\ell(z,\theta_1) &\leq \ell(z,\theta_2) + \langle \nabla_\theta \ell(z,\theta_2), \theta_1 - \theta_2 \rangle \\ &\quad + \int_0^1 \left\| \nabla_\theta \ell(z, \theta_2 + \tau(\theta_1 - \theta_2)) - \nabla_\theta \ell(z,\theta_2) \right\|_2 \left\| \theta_1 - \theta_2 \right\|_2 \mathrm{d}\tau \\ &\leq \ell(z,\theta_2) + \langle \nabla_\theta \ell(z,\theta_2), \theta_1 - \theta_2 \rangle + \int_0^1 \tau H \left\| \theta_1 - \theta_2 \right\|_2^{\alpha+1} \mathrm{d}\tau \\ &= \ell(z,\theta_2) + \langle \nabla_\theta \ell(z,\theta_2), \theta_1 - \theta_2 \rangle + \frac{H}{2} \left\| \theta_1 - \theta_2 \right\|_2^{\alpha+1},\end{aligned}$$

where the second inequality holds because of the definition of α-Hölder smooth (Assumption 3).

The result holds. □

A.2 Proof of Theorem 1

Before the detailed proof, we first prove the following lemma 5. To get Lemma 5, we need the following lemmas given in Bousquet et al. (2020).

Lemma 3 (Bousquet et al. (2020))**.** *Assume that $z_1, \cdots, z_n$ are independent variables and the function $g_i : \mathcal{Z}^n \to \mathbb{R}$ satisfy the following properties for $i = 1, \cdots, n$,*

- $\mathbb{E}_{z_i} g_i(z_1, \cdots, z_n) = 0$ *almost surely;*
- $|\mathbb{E}[g_i(z_1, \cdots, z_n)|z_i]| \leq K$ *almost surely;*
- $\left| g_i(z_1, \cdots, z_n) - g_i(z_1, \cdots, z_{j-1}, z_j', z_{j+1}, \cdots, z_n) \right| \leq \beta$.

Then the following inequality holds for all $q \geq 2$,

$$\left\| \sum_{i=1}^n g_i \right\|_q \leq 12\sqrt{2}\beta q n \log(n) + 4K\sqrt{qn}.$$

Lemma 4 (Bousquet et al. (2020))**.** *Under the uniform stability condition with parameter γ and uniformly bounded loss function $\ell(\cdot,\cdot) \leq M_\ell$, we have for $g_i = \mathbb{E}_{z_i'}\left(\ell(z_i, \theta_n^{(i)}) - \mathbb{E}_z \ell(z, \theta_n^{(i)})\right)$,*

$$\left| n\left(R_n(\theta_n) - R(\theta_n)\right) - \sum_{i=1}^n g_i \right| \leq 2\gamma n.$$

Lemma 5. *Defining the DP algorithm (model) training by T-iterations gradient perturbation method (like (1)) $\hat{\theta}_n = \hat{\theta}(z_1, \cdots, z_n)$ and its independent copy $\hat{\theta}_n' = \hat{\theta}(z_1', \cdots, z_n')$. Then for all $q \geq 2$,*

$$\left\| R_n(\hat{\theta}_n) - R(\hat{\theta}_n) - \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\ell(z_i, \hat{\theta}_n')|z_i\right] + \mathbb{E}R(\hat{\theta}_n) \right\|_q \lesssim \left(\frac{G}{n} + \|b\|_2\right) G q \log(n) \sum_{t=1}^T \eta_t,$$

where $b \sim \mathcal{N}\left(0, \sigma^2 I_p\right)$ and σ is the same as in Lemma 2.

Proof. First, we discuss the stability of the DP algorithm.

Recalling the definition of γ-uniformly stability: If for any $z, z', z_1, \cdots, z_n \in \mathcal{Z}$ and $i = 1, \cdots, n$, it holds that

$$|\ell(z, \theta_n(z_1, \cdots, z_n)) - \ell(z, \theta_n(z_1, \cdots, z_{i-1}, z', z_{i+1}, \cdots, z_n))| \leq \gamma.$$

In the following, $\theta_n^{(i)}, \hat{\theta}_n^{(i)}$ represent $\theta_n(z_1, \cdots, z_{i-1}, z', z_{i+1}, \cdots, z_n), \hat{\theta}_n(z_1, \cdots, z_{i-1}, z', z_{i+1}, \cdots, z_n)$, respectively.

Gradient Descent:

$$\theta_t \leftarrow \theta_{t-1} - \eta_t \left(\frac{1}{n} \sum_{i=1}^n \nabla_\theta \ell(z_i, \theta_{t-1}) \right).$$

Private Gradient Descent (gradient perturbation): with $b \sim \mathcal{N}\left(0, \sigma^2 I_p\right)$,

$$\hat{\theta}_t \leftarrow \hat{\theta}_{t-1} - \eta_t \left(\frac{1}{n} \sum_{i=1}^{n} \nabla_\theta \ell(z_i, \hat{\theta}_{t-1}) + b \right).$$

Bounding the stability of DP model, i.e. $\left| \ell(z, \hat{\theta}_n) - \ell(z, \hat{\theta}_n^{(i)}) \right|$:

At the first iteration:

$$\hat{\theta}_1 = \theta_0 - \eta_1 \left(\frac{1}{n} \sum_{j=1}^{i-1} \nabla_\theta \ell(z_j, \theta_0) + \frac{1}{n} \nabla_\theta \ell(z_i, \theta_0) + \frac{1}{n} \sum_{j=i+1}^{n} \nabla_\theta \ell(z_j, \theta_0) + b \right),$$

$$\hat{\theta}_1^{(i)} = \theta_0 - \eta_1 \left(\frac{1}{n} \sum_{j=1}^{i-1} \nabla_\theta \ell(z_j, \theta_0) + \frac{1}{n} \nabla_\theta \ell(z_i', \theta_0) + \frac{1}{n} \sum_{j=i+1}^{n} \nabla_\theta \ell(z_j, \theta_0) + b \right).$$

Then, considering that $\ell(\cdot, \cdot)$ is G-Lipschitz (denoted by G),

$$\begin{aligned} \left\| \hat{\theta}_1 - \hat{\theta}_1^{(i)} \right\|_2 &\leq \left\| \frac{\eta_1}{n} \left(\nabla_\theta \ell(z_i', \theta_0) - \nabla_\theta \ell(z_i, \theta_0) \right) + 2\eta_1 b \right\|_2 \\ &\overset{(G)}{\leq} \frac{2\eta_1 G}{n} + 2\eta_1 \left\| b \right\|_2. \end{aligned}$$

After T iterations,

$$\left\| \hat{\theta}_n - \hat{\theta}_n^{(i)} \right\|_2 \leq \left(\frac{2G}{n} + 2 \left\| b \right\|_2 \right) \sum_{t=1}^{T} \eta_t.$$

So,

$$\left| \ell(z, \hat{\theta}_n) - \ell(z, \hat{\theta}_n^{(i)}) \right| \overset{(G)}{\leq} G \left\| \hat{\theta}_n - \hat{\theta}_n^{(i)} \right\|_2 \leq \left(\frac{2G^2}{n} + 2G \left\| b \right\|_2 \right) \sum_{t=1}^{T} \eta_t. \tag{4}$$

Considering the function $g_i(z_1, \cdots, z_n) = \mathbb{E}_{z_i'}[\ell(z_i, \hat{\theta}_n^{(i)})] - \mathbb{E}_{z_i'}[R(\hat{\theta}_n^{(i)})]$, via the definition of $R(\hat{\theta}^{(i)})$, we have: $\mathbb{E}_{z_i} g_i(z_1, \cdots, z_n) = 0$.

Via (4), with the stability of the DP model, we have:

$$\left| g_i(z_1, \cdots, z_n) - g_i(z_1, \cdots, z_{j-1}, z_j', z_{j+1}, \cdots, z_n) \right| \leq \beta \coloneqq 2 \left(\frac{2G^2}{n} + 2G \left\| b \right\|_2 \right) \sum_{t=1}^{T} \eta_t.$$

If considering $h_i(z_1, \cdots, z_n) = g_i(z_1, \cdots, z_n) - \mathbb{E}[g_i(z_1, \cdots, z_n) | z_i]$, we have:

$$\mathbb{E}_{z_i} h_i(z_1, \cdots, z_n) = 0$$

almost surely, and

$$\left| h_i(z_1, \cdots, z_n) - h_i(z_1, \cdots, z_{j-1}, z_j', z_{j+1}, \cdots, z_n) \right| \leq 2\beta = 4 \left(\frac{2G^2}{n} + 2G \left\| b \right\|_2 \right) \sum_{t=1}^{T} \eta_t.$$

Via the definition of h_i, we observe that $\mathbb{E}\left[h_i | z_i \right] = 0$ almost surely, which further implies $K = 0$ in Lemma 3, so we have for $q \geq 2$:

$$\left\| \sum_{i=1}^{n} h_i \right\|_q = \left\| \sum_{i=1}^{n} \left(g_i - \mathbb{E}\left[g_i | z_i \right] \right) \right\|_q \leq 96\sqrt{2} G q n \log(n) \left(\frac{G}{n} + \left\| b \right\|_2 \right) \sum_{t=1}^{T} \eta_t.$$

Via Lemma 4, we have:

$$\left| n \left(R_n(\hat{\theta}_n) - R(\hat{\theta}_n) \right) - \sum_{i=1}^{n} g_i \right| \leq \beta n = 4Gn \left(\frac{G}{n} + \left\| b \right\|_2 \right) \sum_{t=1}^{T} \eta_t.$$

Noting that

$$\mathbb{E}\left[g_i|z_i\right] = \mathbb{E}\left[\ell(z_i, \hat{\theta}'_n)|z_i\right] - \mathbb{E}R(\hat{\theta}'_n),$$

we have:

$$\left\| R_n(\hat{\theta}_n) - R(\hat{\theta}_n) - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\ell(z_i, \hat{\theta}'_n)|z_i\right] + \mathbb{E}R(\hat{\theta}_n) \right\|_q \lesssim Gq\log(n)\left(\frac{G}{n} + \|b\|_2\right)\sum_{t=1}^{T}\eta_t.$$

The result follows.

□

As discussed before, the excess population risk can be decomposed into:

$$\begin{aligned} R(\hat{\theta}_n) - R(\theta^*) &= R(\hat{\theta}_n) - R_n(\hat{\theta}_n) + R_n(\hat{\theta}_n) - R_n(\theta^*) + R_n(\theta^*) - R(\theta^*) \\ &\leq R(\hat{\theta}_n) - R_n(\hat{\theta}_n) + R_n(\hat{\theta}_n) - R_n(\theta^*_n) + R_n(\theta^*) - R(\theta^*). \end{aligned} \tag{5}$$

We next discuss the optimization error (excess empirical risk) of the private model $\hat{\theta}_n$, i.e. $R_n(\hat{\theta}_n) - R_n(\theta^*_n)$, under different assumptions.

To get the optimization error bound, we need the following lemma given in (Yang et al., 2021).

Lemma 6 (Yang et al. (2021))**.** *If Gaussian random noise $b \sim \mathcal{N}(0, \sigma^2 I_p)$, then for $\zeta \in (\exp(-p/8), 1)$, we have with probability $1-\zeta$,*

$$\|b\|_2 \leq \sigma\sqrt{p}\left(1 + \left(\frac{8\log(1/\zeta)}{p}\right)^{1/4}\right).$$

Lemma 7. *If the Assumptions 1, 2, 5 hold and the DP model is trained by T-iterations gradient perturbation method (1), then taking $T = \mathcal{O}(\log(n))$, $\eta_1 = \cdots = \eta_T = \frac{1}{L}$, if $\zeta \in (\exp(-p/8), 1)$, with probability at least $1-\zeta$,*

$$R_n(\hat{\theta}_n) - R_n(\theta^*_n) \lesssim \frac{G^2 p\log(n)\log(1/\delta)}{n^2\epsilon^2}\left(1 + \left(\frac{8\log(T/\zeta)}{p}\right)^{1/4}\right)^2.$$

Proof. Note that we assume the loss function is L-smooth (Assumption 2, denoted by L) and satisfies the PL condition (Assumption 5, denoted by PL), at iteration t, taking $\eta_t = \frac{1}{L}$, we have:

$$\begin{aligned} R_n(\hat{\theta}_{t+1}) - R_n(\hat{\theta}_t) &\overset{(L)}{\leq} \langle \nabla_\theta R_n(\hat{\theta}_t), \hat{\theta}_{t+1} - \hat{\theta}_t\rangle + \frac{L}{2}\left\|\hat{\theta}_{t+1} - \hat{\theta}_t\right\|_2^2 \\ &= -\eta_t\langle \nabla_\theta R_n(\hat{\theta}_t), \nabla_\theta R_n(\hat{\theta}_t) + b\rangle + \frac{L\eta_t^2}{2}\left\|\nabla_\theta R_n(\hat{\theta}_t) + b\right\|_2^2 \\ &= -\frac{1}{L}\left\|\nabla_\theta R_n(\hat{\theta}_t)\right\|_2^2 + \frac{1}{2L}\left\|\nabla_\theta R_n(\hat{\theta}_t)\right\|_2^2 + \frac{1}{2L}\|b\|_2^2 \\ &= -\frac{1}{2L}\left\|\nabla_\theta R_n(\hat{\theta}_t)\right\|_2^2 + \frac{1}{2L}\|b\|_2^2 \\ &\overset{(PL)}{\leq} -\frac{\mu}{L}\left(R_n(\hat{\theta}_t) - R_n(\theta^*_n)\right) + \frac{1}{2L}\|b\|_2^2. \end{aligned} \tag{6}$$

With Lemma 6, with probability at least $1-\xi$, we have:

$$\|b\|_2^2 \leq \sigma^2 p\left(1 + \left(\frac{8\log(1/\xi)}{p}\right)^{1/4}\right)^2.$$

Taking over T iterations and setting $\xi = \zeta/T$, then with probability at least $1-\zeta$ we have:

$$\begin{aligned} R_n(\hat{\theta}_n) - R_n(\theta_n^*) &\leq \left(1-\frac{\mu}{L}\right)^T \left(R_n(\hat{\theta}_0) - R_n(\theta_n^*)\right) + \sum_{t=0}^{T-1} \left(1-\frac{\mu}{L}\right)^t \frac{\sigma^2 p}{2L} \left(1 + \left(\frac{8\log(T/\zeta)}{p}\right)^{1/4}\right)^2 \\ &\leq \left(1-\frac{\mu}{L}\right)^T M_\ell + \frac{\left(1-\left(1-\frac{\mu}{L}\right)^{T-1}\right)\sigma^2 p}{2\mu} \left(1 + \left(\frac{8\log(T/\zeta)}{p}\right)^{1/4}\right)^2 \\ &\leq \left(1-\frac{\mu}{L}\right)^T M_\ell + \frac{\sigma^2 p}{2\mu} \left(1 + \left(\frac{8\log(T/\zeta)}{p}\right)^{1/4}\right)^2, \end{aligned} \tag{7}$$

where the second inequality holds because $0 \leq \ell(\cdot,\cdot) \leq M_\ell$.

Taking $\sigma = c\frac{G\sqrt{T\log(1/\delta)}}{n\epsilon}$ given in Lemma 2, and taking $T = \mathcal{O}(\log(n))$, then if $\zeta \in (\exp(-p/8), 1)$, with probability at least $1-\zeta$, we have:

$$R_n(\hat{\theta}_n) - R_n(\theta_n^*) \lesssim \frac{G^2 p \log(n)\log(1/\delta)}{2\mu n^2 \epsilon^2} \left(1 + \left(\frac{8\log(T/\zeta)}{p}\right)^{1/4}\right)^2.$$

The result follows.

□

Lemma 8. *If the loss function is α-Hölder smooth with parameter H, satisfies the PL inequality with parameter 2μ and the DP model is trained by T-iterations gradient perturbation method (1), then taking $T = \mathcal{O}\left(n^{\frac{2}{1+2\alpha}}\right)$, $\eta_t = \frac{2}{\mu(t+\kappa)}$, where $\kappa \geq \frac{2H^{1/\alpha}}{\mu}$, if $\zeta \in (\exp(-p/8), 1)$, with probability at least $1-\zeta$,*

$$R_n(\hat{\theta}_n) - R_n(\theta_n^*) \lesssim \frac{G'^2\sqrt{p\log(1/\delta)}}{n^{\frac{2\alpha}{1+2\alpha}}\epsilon} \left(1 + \left(\frac{8\log(T/\zeta)}{p}\right)^{1/4}\right).$$

Proof. The proof is motivated by (Li & Liu, 2021).

Like the proof of Lemma 7, by assuming that the loss function is α-Hölder smooth (Assumption 3, denoted by α), via Lemma 1, at iteration t,

$$\begin{aligned} R_n(\hat{\theta}_{t+1}) - R_n(\hat{\theta}_t) &\overset{(\alpha)}{\leq} \langle \nabla_\theta R_n(\hat{\theta}_t), \hat{\theta}_{t+1} - \hat{\theta}_t\rangle + \frac{H}{2}\left\|\hat{\theta}_{t+1} - \hat{\theta}_t\right\|_2^{\alpha+1} \\ &\leq \langle \nabla_\theta R_n(\hat{\theta}_t), \hat{\theta}_{t+1} - \hat{\theta}_t\rangle + \frac{H}{\alpha+1}\left\|\hat{\theta}_{t+1} - \hat{\theta}_t\right\|_2^{\alpha+1} \\ &= -\eta_t \langle \nabla_\theta R_n(\hat{\theta}_t), \nabla_\theta R_n(\hat{\theta}_t) + b\rangle + \frac{H\eta_t^{\alpha+1}}{\alpha+1}\left\|\nabla_\theta R_n(\hat{\theta}_t) + b\right\|_2^{\alpha+1} \\ &\leq -\eta_t\left(\left\|\nabla_\theta R_n(\hat{\theta}_t)\right\|_2^2 + \langle \nabla_\theta R_n(\hat{\theta}_t), b\rangle\right) \\ &\quad + \frac{H\eta_t^{\alpha+1}}{\alpha+1}\left(\frac{1-\alpha}{2} + \frac{\alpha+1}{2}\left(\left\|\nabla_\theta R_n(\hat{\theta}_t) + b\right\|_2^{\alpha+1}\right)^{\frac{2}{\alpha+1}}\right) \\ &\leq -\eta_t\left\|\nabla_\theta R_n(\hat{\theta}_t)\right\|_2^2 + \left(\eta_t + H\eta_t^{\alpha+1}\right)\|\nabla_\theta R_n(\hat{\theta}_t)\|_2\|b\|_2 \\ &\quad + \frac{H\eta_t^{\alpha+1}(1-\alpha)}{2(\alpha+1)} + \frac{H\eta_t^{\alpha+1}}{2}\left(\left\|\nabla_\theta R_n(\hat{\theta}_t)\right\|_2^2 + \|b\|_2^2\right), \end{aligned}$$

where the third inequality is because of Young's inequality: if $p^{-1} + q^{-1} = 1$ and $p > 0$, then $uv \leq p^{-1}|u|^p + q^{-1}|v|^q$. Here we set $p^{-1} = (1-\alpha)/2$, $q^{-1} = (\alpha+1)/2$. And the last inequality holds because of Cauthy-Schwarz inequality.

Noting that $\eta_t = \frac{2}{\mu(t+\kappa)}$ and $\kappa \geq \frac{2H^{1/\alpha}}{\mu}$, so we have: $\eta_t \leq \left(\frac{1}{H}\right)^{1/\alpha}$.

As a result, we have:

$$H\eta_t^{\alpha+1} \leq H\left[\left(\frac{1}{H}\right)^{1/\alpha}\right]^{\alpha}\eta_t \leq \eta_t. \tag{8}$$

As a result,

$$\begin{aligned}
R_n(\hat{\theta}_{t+1}) - R_n(\hat{\theta}_t) &\leq -\eta_t \left\|\nabla_\theta R_n(\hat{\theta}_t)\right\|_2^2 + \frac{H\eta_t^{\alpha+1}(1-\alpha)}{2(\alpha+1)} + \frac{H\eta^{\alpha+1}}{2}\left\|\nabla_\theta R_n(\hat{\theta}_t)\right\|_2^2 \\
&\quad + \left(\eta_t + H\eta_t^{\alpha+1}\right)\|\nabla_\theta R_n(\hat{\theta}_t)\|_2\|b\|_2 + \frac{H\eta_t^{\alpha+1}}{2}\|b\|_2^2 \\
&\leq -\frac{\eta_t}{2}\left\|\nabla_\theta R_n(\hat{\theta}_t)\right\|_2^2 + \frac{H\eta_t^{\alpha+1}(1-\alpha)}{2(\alpha+1)} + 2\eta_t\|\nabla_\theta R_n(\hat{\theta}_t)\|_2\|b\|_2 + \frac{\eta_t}{2}\|b\|_2^2 \\
&\overset{(PL)}{\leq} -\frac{\eta_t}{4}\left\|\nabla_\theta R_n(\hat{\theta}_t)\right\|_2^2 - \mu\eta_t\left(R_n(\hat{\theta}_t) - R_n(\theta_n^*)\right) + \frac{H\eta_t^{\alpha+1}(1-\alpha)}{2(\alpha+1)} \\
&\quad + 2\eta_t G'\|b\|_2 + \frac{\eta_t}{2}\|b\|_2^2,
\end{aligned} \tag{9}$$

where the second inequality is because of (8) and the last inequality holds because we assume that the loss function satisfies the PL condition with parameter 2μ, and $G' = \max\{2HM_{\mathcal{C}}, H\}$, as discussed in Lemma 2.

Adding $\frac{\eta_t}{4}\left\|\nabla_\theta R_n(\hat{\theta}_t)\right\|_2^2 - R_n(\theta_n^*)$ to both sides of (9), we have

$$\begin{aligned}
R_n(\hat{\theta}_{t+1}) - R_n(\theta_n^*) + \frac{\eta_t}{4}\left\|\nabla_\theta R_n(\hat{\theta}_t)\right\|_2^2 &\leq (1-\mu\eta_t)\left(R_n(\hat{\theta}_t) - R_n(\theta_n^*)\right) + \frac{H\eta_t^{\alpha+1}(1-\alpha)}{2(\alpha+1)} \\
&\quad + 2\eta_t G'\|b\|_2 + \frac{\eta_t}{2}\|b\|_2^2.
\end{aligned}$$

Taking $\eta_t = \frac{2}{\mu(t+\kappa)}$,

$$\begin{aligned}
R_n(\hat{\theta}_{t+1}) - R_n(\theta_n^*) + \frac{1}{2\mu(t+\kappa)}\left\|\nabla_\theta R_n(\hat{\theta}_t)\right\|_2^2 &\leq \frac{t+\kappa-2}{t+\kappa}\left(R_n(\hat{\theta}_t) - R_n(\theta_n^*)\right) + \frac{H\eta_t^{\alpha+1}(1-\alpha)}{2(\alpha+1)} \\
&\quad + 2\eta_t G'\|b\|_2 + \frac{\eta_t}{2}\|b\|_2^2.
\end{aligned}$$

Multiply both side by $(t+\kappa)(t+\kappa-1)$,

$$\begin{aligned}
&(t+\kappa)(t+\kappa-1)\left(R_n(\hat{\theta}_{t+1}) - R_n(\theta_n^*)\right) + \frac{t+\kappa-1}{2\mu}\left\|\nabla_\theta R_n(\hat{\theta}_t)\right\|_2^2 \\
&\leq (t+\kappa-1)(t+\kappa-2)\left(R_n(\hat{\theta}_t) - R_n(\theta_n^*)\right) + (t+\kappa)^{-\alpha}(t+\kappa-1)\frac{H(1-\alpha)}{2(\alpha+1)}\left(\frac{2}{\mu}\right)^{1+\alpha} \\
&\quad + \frac{4G'(t+\kappa-1)}{\mu}\|b\|_2 + \frac{t+\kappa-1}{\mu}\|b\|_2^2.
\end{aligned} \tag{10}$$

With Lemma 6, with probability at least $1-\xi$, we have:

$$\begin{aligned}
\|b\|_2 &\leq \sigma\sqrt{p}\left(1 + \left(\frac{8\log(1/\xi)}{p}\right)^{1/4}\right), \\
\|b\|_2^2 &\leq \sigma^2 p\left(1 + \left(\frac{8\log(1/\xi)}{p}\right)^{1/4}\right)^2.
\end{aligned}$$

So with probability at least $1-\xi$:

$$
\begin{aligned}
&(t+\kappa)(t+\kappa-1)\left(R_n(\hat{\theta}_{t+1})-R_n(\theta_n^*)\right)+\frac{t+\kappa-1}{2\mu}\left\|\nabla_\theta R_n(\hat{\theta}_t)\right\|_2^2 \\
&\leq (t+\kappa-1)(t+\kappa-2)\left(R_n(\hat{\theta}_t)-R_n(\theta_n^*)\right)+(t+\kappa)^{-\alpha}(t+\kappa-1)\frac{H(1-\alpha)}{2(\alpha+1)}\left(\frac{2}{\mu}\right)^{1+\alpha} \\
&\quad+\frac{4G'(t+\kappa-1)\sigma\sqrt{p}}{\mu}\left(1+\left(\frac{8\log(1/\xi)}{p}\right)^{1/4}\right)+\frac{(t+\kappa-1)\sigma^2 p}{\mu}\left(1+\left(\frac{8\log(1/\xi)}{p}\right)^{1/4}\right)^2.
\end{aligned}
$$

By summing over T iterations and taking $\xi=\zeta/T$, with probability at least $1-\zeta$, we have:

$$
\begin{aligned}
&(T+\kappa)(T+\kappa-1)\left(R_n(\hat{\theta}_{T+1})-R_n(\theta_n^*)\right)+\sum_{t=1}^{T}\frac{t+\kappa-1}{2\mu}\left\|\nabla_\theta R_n(\hat{\theta}_t)\right\|_2^2 \\
&\leq \kappa(\kappa-1)\left(R_n(\hat{\theta}_1)-R_n(\theta_n^*)\right)+\sum_{t=1}^{T}(t+\kappa)^{-\alpha}(t+\kappa-1)\frac{H(1-\alpha)}{2(\alpha+1)}\left(\frac{2}{\mu}\right)^{1+\alpha} \\
&\quad+\sum_{t=1}^{T}\frac{4G'(t+\kappa-1)\sigma\sqrt{p}}{\mu}\left(1+\left(\frac{8\log(T/\zeta)}{p}\right)^{1/4}\right) \\
&\quad+\sum_{t=1}^{T}\frac{(t+\kappa-1)\sigma^2 p}{\mu}\left(1+\left(\frac{8\log(T/\zeta)}{p}\right)^{1/4}\right)^2.
\end{aligned}
\tag{11}
$$

Here, for simplicity, we represent $t=0,\cdots,T-1$ by $t=1,\cdots,T$.

Then we bound term $\sum_{t=1}^{T}(t+\kappa)^{-\alpha}(t+\kappa-1)\frac{H(1-\alpha)}{2(\alpha+1)}\left(\frac{2}{\mu}\right)^{1+\alpha}$.

Note that:

$$
\sum_{t=1}^{T}(t+\kappa)^{-\alpha}(t+\kappa-1)\leq\sum_{t=1}^{T}(t+\kappa)^{1-\alpha}\leq\int_1^T(t+\kappa)^{1-\alpha}dt\leq\frac{(T+\kappa)^{2-\alpha}}{2-\alpha}.
$$

Plugging the result above back into (11), and note that $0\leq\ell(\cdot,\cdot)\leq M_\ell$, we have:

$$
\begin{aligned}
(T+\kappa)(T+\kappa-1)\left(R_n(\hat{\theta}_{T+1})-R_n(\theta_n^*)\right)\leq\kappa(\kappa-1)M_\ell&+\frac{(T+\kappa)^{2-\alpha}}{2-\alpha}\frac{H(1-\alpha)}{2(\alpha+1)}\left(\frac{2}{\mu}\right)^{1+\alpha} \\
&+\left(T\kappa+\frac{T(T-1)}{2}\right)\frac{4G'\sigma\sqrt{p}}{\mu}\left(1+\left(\frac{8\log(T/\zeta)}{p}\right)^{1/4}\right) \\
&+\left(T\kappa+\frac{T(T-1)}{2}\right)\frac{\sigma^2 p}{\mu}\left(1+\left(\frac{8\log(T/\zeta)}{p}\right)^{1/4}\right)^2.
\end{aligned}
$$

As a result, taking σ given in Lemma 2, with probability at least $1-\zeta$, we have:

$$
R_n(\hat{\theta}_{T+1})-R_n(\theta_n^*)\lesssim T^{-\alpha}+\frac{G'^2\sqrt{Tp\log(1/\delta)}}{n\epsilon}\left(1+\left(\frac{8\log(T/\zeta)}{p}\right)^{1/4}\right).
$$

Taking $T=\mathcal{O}\left(n^{\frac{2}{1+2\alpha}}\right)$, with probability at least $1-\zeta$, we have:

$$
R_n(\hat{\theta}_n)-R_n(\theta_n^*)\lesssim\frac{G'^2\sqrt{p\log(1/\delta)}}{n^{\frac{2\alpha}{1+2\alpha}}\epsilon}\left(1+\left(\frac{8\log(T/\zeta)}{p}\right)^{1/4}\right).
$$

The result follows.

□

To get Theorem 1, we further need the following lemma given in Boucheron et al. (2013).

Lemma 9 (Boucheron et al. (2013))**.** *If $X_1, \cdots, X_n$ are zero mean, independent and bounded $|X_i| \leq M$ almost surely, then for $q \geq 2$,*

$$\|X_1 + \cdots X_n\|_q \leq 6\sqrt{\left(\sum_{i=1}^{n} \mathbb{E}[X_i^2]\right) q} + 4qM.$$

Then, we can start our proof.

Theorem. *If Assumptions 1, 2, 4 and 5 hold, the loss function is bounded, i.e. $0 \leq \ell(\cdot,\cdot) \leq M_\ell$, taking σ given by Lemma 2, $T = \mathcal{O}(\log(n))$, $\eta_1 = \cdots = \eta_T = \frac{1}{L}$, if $\zeta \in (\exp(-p/8), 1)$, then with probability at least $1-\zeta$:*

$$\begin{aligned} R(\hat{\theta}_n) - R(\theta^*) \leq\ & c_1 \frac{G^2 p \log(n) \log(1/\delta)}{n^2 \epsilon^2} \left(1 + \left(\frac{8\log(T/\zeta)}{p}\right)^{1/4}\right)^2 \\ & + c_2 \left(\frac{G^2 \log^2(n)}{n} + \frac{B + M_\ell}{n}\right) \\ & + c_3 \frac{G^2 \log^{2.5}(n) \sqrt{p \log(1/\delta)}}{n\epsilon} \left(1 + \left(\frac{8\log(1/\zeta)}{p}\right)^{1/4}\right). \end{aligned}$$

for some constants $c_1, c_2, c_3 > 0$.

Proof. Via Lemma 5, we have:

$$R_n(\hat{\theta}_n) - R(\hat{\theta}_n) = \rho + \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}'\ell(z_i, \hat{\theta}'_n) - \mathbb{E}R(\hat{\theta}_n),$$

where $\|\rho\|_q \lesssim G\big(G/n + \|b\|_2\big) q \log(n) \sum_{t=1}^{T} \eta_t$ for $q \geq 2$ and $\mathbb{E}'$ denotes the expectation taken over the independent copy.

Plugging this back to (5), we have:

$$R(\hat{\theta}_n) - R(\theta^*) \leq \Big(R_n(\hat{\theta}_n) - R_n(\theta^*_n)\Big) + (R_n(\theta^*) - R(\theta^*)) - \rho - \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}'\ell(z_i, \hat{\theta}'_n) + \mathbb{E}R(\hat{\theta}_n).$$

Noting that $R_n(\theta^*) = \frac{1}{n}\sum_{i=1}^{n} \ell(z_i, \theta^*)$, we have:

$$R(\hat{\theta}_n) - R(\theta^*) \leq \Big(R_n(\hat{\theta}_n) - R_n(\theta^*_n)\Big) + \Big(\mathbb{E}R(\hat{\theta}_n) - R(\theta^*)\Big) - \rho - \frac{1}{n}\sum_{i=1}^{n} \Big(\mathbb{E}'\ell(z_i, \hat{\theta}'_n) - \ell(z_i, \theta^*)\Big). \tag{12}$$

Based on the definition of $R(\theta)$, Assumption 4 is equivalent to:

$$\mathbb{E}\left[\left(\ell(z,\theta) - \ell(z,\theta^*)\right)^2\right] \leq B\left(\mathbb{E}\ell(z,\theta) - \mathbb{E}\ell(z,\theta^*)\right). \tag{13}$$

So, via (13),

$$\begin{aligned} \mathbb{E}\left[\Big(\mathbb{E}'\ell(z_i, \hat{\theta}'_n) - \ell(z_i, \theta^*)\Big)^2\right] &\leq B\left(\mathbb{E}\mathbb{E}'\ell(z_i, \hat{\theta}'_n) - \mathbb{E}\ell(z_i, \theta^*)\right) \\ &= B\left(\mathbb{E}[R(\hat{\theta}'_n)] - R(\theta^*)\right), \end{aligned} \tag{14}$$

where the last equation holds because $\mathbb{E}\mathbb{E}'\ell(z_i, \hat{\theta}'_n) = \mathbb{E}[R(\hat{\theta}'_n)]$.

Note that term $\mathbb{E}'\ell(z_i, \hat{\theta}'_n) - \ell(z_i, \theta^*)$ can be decomposed as the following:

$$\underbrace{\mathbb{E}'\ell(z_i, \hat{\theta}'_n) - \ell(z_i, \theta^*)}_{X_i} = \underbrace{\mathbb{E}'\ell(z_i, \hat{\theta}'_n) - \mathbb{E}'\ell(z_i, \theta'_n)}_{X'_i} + \underbrace{\mathbb{E}'\ell(z_i, \theta'_n) - \ell(z_i, \theta^*)}_{X''_i}.$$

Via triangle inequality,

$$\|X_i\|_q \leq \|X_i'\|_q + \|X_i''\|_q .$$

Recalling the definition of $R_n(\hat{\theta}_n) - R_n(\theta_n^*)$, we have:

$$\left\| \frac{1}{n} \sum_{i=1}^{n} X_i' \right\|_q = R_n(\hat{\theta}_n) - R_n(\theta_n^*). \tag{15}$$

Via Lemma 9, since $\mathbb{E}[R(\theta_n')] - R(\theta^*)$ is exactly the expectation of each X_i'', we have for $q \geq 2$,

$$\begin{aligned} \left\| \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}' \left[\ell(z_i, \theta_n') \right] - \ell(z_i, \theta^*) - \mathbb{E}[R(\theta_n')] + R(\theta^*) \right\|_q &\lesssim \sqrt{\mathbb{E}\left[\left(\mathbb{E}' \ell(z_i, \hat{\theta}_n') - \ell(z_i, \theta^*) \right)^2 \right] \frac{q}{n}} \\ &\leq \sqrt{B \left(\mathbb{E}[R(\theta_n)] - R(\theta^*) \right) \frac{q}{n}} + \frac{q M_\ell}{n}, \end{aligned} \tag{16}$$

where the last inequality holds because of (14) and $\mathbb{E}[R(\theta_n)] = \mathbb{E}[R(\hat{\theta}_n')]$.

Plugging (15) and (16) back into (12), we obtain for each $q \geq 2$ and some constant $C > 0$,

$$\begin{aligned} &\left\| R(\hat{\theta}_n) - R(\theta^*) - \left(R_n(\hat{\theta}_n) - R_n(\theta_n^*) \right) \right\|_q \\ &\leq C \left(G \left(\frac{G}{n} + \|b\|_2 \right) q \log(n) \sum_{t=1}^{T} \eta_t + \left(R_n(\hat{\theta}_n) - R_n(\theta_n^*) \right) + \sqrt{B \left(\mathbb{E}[R(\theta_n)] - R(\theta^*) \right) \frac{q}{n}} + \frac{q M_\ell}{n} \right) \\ &\leq \varphi C \left(\mathbb{E}[R(\theta_n)] - R(\theta^*) \right) + C \left(G \left(\frac{G}{n} + \|b\|_2 \right) q \log(n) \sum_{t=1}^{T} \eta_t + \left(R_n(\hat{\theta}_n) - R_n(\theta_n^*) \right) + \left(\frac{B}{\varphi} + M_\ell \right) \frac{q}{n} \right), \end{aligned} \tag{17}$$

where the last inequality holds because for $a, b, \varphi > 0$, $\sqrt{ab} \leq \varphi a + b/\varphi$.

Taking $q = 2$, and via Cauchy-Schwarz inequality,

$$\begin{aligned} &\mathbb{E}[R(\hat{\theta}_n)] - R(\theta^*) - \mathbb{E}\left[R_n(\hat{\theta}_n) - R_n(\theta_n^*) \right] \\ &\leq \left\| R(\hat{\theta}_n) - R^* - \left(R_n(\hat{\theta}_n) - R_n(\theta_n^*) \right) \right\|_2 \\ &\leq \varphi C \left(\mathbb{E}[R(\hat{\theta}_n)] - R(\theta^*) \right) + C \left(2G \left(\frac{G}{n} + \|b\|_2 \right) \log(n) \sum_{t=1}^{T} \eta_t + \left(R_n(\hat{\theta}_n) - R_n(\theta_n^*) \right) + 2 \left(\frac{B}{\varphi} + M_\ell \right) / n \right). \end{aligned}$$

The inequality above can be rewritten as:

$$\begin{aligned} \mathbb{E}[R(\hat{\theta}_n)] - R(\theta^*) \leq \frac{1}{1 - \varphi C} \mathbb{E}[R_n(\hat{\theta}_n) - R_n(\theta_n^*)] + \frac{C}{1 - \varphi C} \left(R_n(\hat{\theta}_n) - R_n(\theta_n^*) \right) \\ + \frac{C}{1 - \varphi C} \left(2G \left(\frac{G}{n} + \|b\|_2 \right) \log(n) \sum_{t=1}^{T} \eta_t + 2 \left(\frac{B}{\varphi} + M_\ell \right) / n \right). \end{aligned}$$

Taking this back to (17), we have:

$$\begin{aligned} R(\hat{\theta}_n) - R(\theta^*) \leq c_1 \left(R_n(\hat{\theta}_n) - R_n(\theta_n^*) \right) + c_2 \mathbb{E}[R_n(\hat{\theta}_n) - R_n(\theta_n^*)] \\ + c_3 \left(G \left(\frac{G}{n} + \|b\|_2 \right) \log(n) \sum_{t=1}^{T} \eta_t + \frac{B + M_\ell}{n} \right), \end{aligned} \tag{18}$$

for some constants c_1, c_2 and c_3.

The first term $R_n(\hat{\theta}_n) - R_n(\theta_n^*)$ has been discussed in Lemma 7, so we analyze its expectation (i.e. the second term in 18) here.

Via (6), we have:

$$\begin{aligned}\mathbb{E}[R_n(\hat{\theta}_{t+1}) - R_n(\hat{\theta}_t)] &\leq -\frac{\mu}{L}\mathbb{E}\left[R_n(\hat{\theta}_t) - R_n(\theta_n^*)\right] + \frac{1}{2L}\mathbb{E}\left[\|b\|_2^2\right] \\ &= -\frac{\mu}{L}\mathbb{E}\left[R_n(\hat{\theta}_t) - R_n(\theta_n^*)\right] + \frac{1}{2L}\left(\mathbb{E}^2\left[\|b\|_2\right] + v(\|b\|_2)\right) \\ &= -\frac{\mu}{L}\mathbb{E}\left[R_n(\hat{\theta}_t) - R_n(\theta_n^*)\right] + \frac{p\sigma^2}{2L}.\end{aligned} \tag{19}$$

The first equation holds because for random variable X, $\mathbb{E}\left[X^2\right] = \mathbb{E}^2\left[X\right] + v(X)$, where $v(X)$ denotes the variance of X, and the last equation holds because the random noise b is zero mean.

Then, via similar steps given in (7), by summing over T iterations, we have:

$$\begin{aligned}\mathbb{E}[R_n(\hat{\theta}_n) - R_n(\theta_n^*)] &\leq \left(1 - \frac{\mu}{L}\right)^T M_\ell + \frac{p\sigma^2}{2\mu} \\ &\lesssim \left(1 - \frac{\mu}{L}\right)^T M_\ell + \frac{G^2 T p \log(1/\delta)}{n^2\epsilon^2},\end{aligned} \tag{20}$$

where the last inequality holds because $\sigma = c\frac{G\sqrt{T\log(1/\delta)}}{n\epsilon}$.

Then via Lemmas 6, 7 and inequality (20), considering that the random noise b in γ and $R(\hat{\theta}_n)$ are all derived from the gradient (the injected noise b), so if taking $T = \mathcal{O}(\log(n))$, with probability at least $1 - \zeta$,

$$\begin{aligned}R(\hat{\theta}_n) - R(\theta^*) \leq\ & c_1 \frac{G^2 p \log(n)\log(1/\delta)}{n^2\epsilon^2}\left(1 + \left(\frac{8\log(T/\zeta)}{p}\right)^{1/4}\right)^2 \\ &+ c_2\left(\frac{G^2\log^2(n)}{n} + \frac{B + M_\ell}{n}\right) \\ &+ c_3\frac{G^2\log^{2.5}(n)\sqrt{p\log(1/\delta)}}{n\epsilon}\left(1 + \left(\frac{8\log(1/\zeta)}{p}\right)^{1/4}\right).\end{aligned}$$

for some constants $c_1, c_2, c_3 > 0$.

The result follows. □

A.3 Proof of Theorem 2

Theorem. *If the loss function is α-Hölder smooth (Assumption 3) with parameter H, satisfies the generalized Bernstein condition with parameter B (Assumption 4), and satisfies the PL condition with parameter 2μ (Assumption 5), the loss function and the parameter space are bounded, i.e. $0 \leq \ell(\cdot,\cdot) \leq M_\ell$, $\|\mathcal{C}\|_2 \leq M_{\mathcal{C}}$. Taking σ given by Lemma 2, $T = \mathcal{O}\left(n^{\frac{2}{1+2\alpha}}\right)$, and $\eta_t = \frac{2}{\mu(t+\kappa)}$, where $\kappa \geq \frac{2H^{1/\alpha}}{\mu}$, if $\zeta \in (\exp(-p/8), 1)$, then with probability at least $1 - \zeta$:*

$$\begin{aligned}R(\hat{\theta}_n) - R(\theta^*) \leq\ & c_1 \frac{G'^2\sqrt{p\log(1/\delta)}}{n^{\frac{2\alpha}{1+2\alpha}}\epsilon}\left(1 + \left(\frac{8\log(T/\zeta)}{p}\right)^{1/4}\right) \\ &+ c_2\left(\frac{G'^2\log^2(n)}{n} + \frac{B + M_\ell}{n}\right) \\ &+ c_3\frac{G'^2\log^2(n)\sqrt{p\log(1/\delta)}}{n^{\frac{2\alpha}{1+2\alpha}}\epsilon}\left(1 + \left(\frac{8\log(1/\zeta)}{p}\right)^{1/4}\right).\end{aligned}$$

for some constants $c_1, c_2, c_3 > 0$, where $G' = \max\{2HM_{\mathcal{C}}, H\}$.

Proof. Like inequality (18) in the proof of Theorem 1 (Appendix A.2), we have:

$$R(\hat{\theta}_n) - R(\theta^*) \lesssim c_1 \left(R_n(\hat{\theta}_n) - R_n(\theta_n^*) \right) + c_2 \mathbb{E}[R_n(\hat{\theta}_n) - R_n(\theta_n^*)] + c_3 \left(G' \left(\frac{G'}{n} + \|b\|_2 \right) \log(n) \sum_{t=1}^{T} \eta_t + \frac{B + M_\ell}{n} \right) \tag{21}$$

for some constants c_1, c_2 and c_3, where $G' = \max\{2HM_{\mathcal{C}}, H\}$ as discussed in Remark 1.

Like (19), via inequality (10), we have:

$$\begin{aligned}
&(t+\kappa)(t+\kappa-1)\mathbb{E}\left[R_n(\hat{\theta}_{t+1}) - R_n(\theta_n^*)\right] \\
&\leq (t+\kappa-1)(t+\kappa-2)\mathbb{E}\left[R_n(\hat{\theta}_t) - R_n(\theta_n^*)\right] + (t+\kappa)^{-\alpha}(t+\kappa-1)\frac{H(1-\alpha)}{2(\alpha+1)}\left(\frac{2}{\mu}\right)^{1+\alpha} \\
&\quad + \frac{4G'(t+\kappa-1)}{\mu}\mathbb{E}\left[\|b\|_2\right] + \frac{t+\kappa-1}{\mu}\mathbb{E}\left[\|b\|_2^2\right] \\
&= (t+\kappa-1)(t+\kappa-2)\mathbb{E}\left[R_n(\hat{\theta}_t) - R_n(\theta_n^*)\right] + (t+\kappa)^{-\alpha}(t+\kappa-1)\frac{H(1-\alpha)}{2(\alpha+1)}\left(\frac{2}{\mu}\right)^{1+\alpha} \\
&\quad + \frac{(t+\kappa-1)p\sigma^2}{\mu},
\end{aligned}$$

where the last inequality holds because of the property of random noise b, like discussed in (19).

Following the steps in the proof of Lemma 8, by summing over T iterations, we have:

$$\mathbb{E}\left[R_n(\hat{\theta}_n) - R_n(\theta_n^*)\right] \lesssim T^{-\alpha} + \frac{G'^2 pT\log(1/\delta)}{n^2\epsilon^2}. \tag{22}$$

Via Lemma 6 and Lemma 8, taking σ given in Lemma 2, $T = \mathcal{O}\left(n^{\frac{2}{1+2\alpha}}\right)$, and $\eta_t = \frac{2}{\mu(t+\kappa)}$, where $\kappa \geq \frac{2H^{1/\alpha}}{\mu}$ analyzed in Lemma 8, then with probability at least $1-\zeta$, we have:

$$\begin{aligned}
R_n(\hat{\theta}_n) - R_n(\theta_n^*) &\lesssim \frac{G'^2\sqrt{p\log(1/\delta)}}{n^{\frac{2\alpha}{1+2\alpha}}\epsilon}\left(1 + \left(\frac{8\log(T/\zeta)}{p}\right)^{1/4}\right), \\
\|b\|_2 &\leq \frac{G'\sqrt{p\log(1/\delta)}}{n^{\frac{2\alpha}{1+2\alpha}}\epsilon}\left(1 + \left(\frac{8\log(1/\zeta)}{p}\right)^{1/4}\right),
\end{aligned}$$

for $\zeta \in (\exp(-p/8), 1)$.

Taking the results given above back into (21), noting that $\eta_t = \frac{2}{\mu(t+\kappa)}$, where $\kappa \geq \frac{2H^{1/\alpha}}{\mu}$, then if $\zeta \in (\exp(-p/8), 1)$, with probability at least $1-\zeta$, we have:

$$\begin{aligned}
R(\hat{\theta}_n) - R(\theta^*) \leq\ & c_1 \frac{G'^2\sqrt{p\log(1/\delta)}}{n^{\frac{2\alpha}{1+2\alpha}}\epsilon}\left(1 + \left(\frac{8\log(T/\zeta)}{p}\right)^{1/4}\right) \\
&+ c_2\left(\frac{G'^2\log^2(n)}{n} + \frac{B+M_\ell}{n}\right) \\
&+ c_3 \frac{G'^2\log^2(n)\sqrt{p\log(1/\delta)}}{n^{\frac{2\alpha}{1+2\alpha}}\epsilon}\left(1 + \left(\frac{8\log(1/\zeta)}{p}\right)^{1/4}\right).
\end{aligned}$$

The result holds. □

A.4 Proof of Theorem 3

Theorem. *If Assumptions 3, 5 hold, and Assumption 4 with parameter B holds, the loss function and the parameter space are bounded, i.e. $0 \leq \ell(\cdot,\cdot) \leq M_\ell$, $\|\mathcal{C}\|_2 \leq M_{\mathcal{C}}$. Taking σ given by*

Lemma 2, $T = \mathcal{O}(\log(n))$*, and* $\eta_1 = \cdots = \eta_T = \eta$*, where* $\left(\frac{2}{H} - \frac{2^{-1/\alpha}}{\mu H^{(\alpha-1)/\alpha}}\right)^{1/\alpha} < \eta < \left(\frac{2}{H}\right)^{1/\alpha}$*, if* $\zeta \in (\exp(-p/8), 1)$*, then with probability at least* $1-\zeta$*,*

$$
\begin{aligned}
R(\hat{\theta}_n) - R(\theta^*) \leq\ & c_1 \frac{G'\sqrt{p\log(n)\log(1/\delta)}}{n\epsilon}\left(1 + \left(\frac{8\log(T/\zeta)}{p}\right)^{1/4}\right) \\
& + c_2\left(\frac{G'^2\log^2(n)}{n} + \frac{B + M_\ell}{n}\right) \\
& + c_3 \frac{G'^2\log^{2.5}(n)\sqrt{p\log(1/\delta)}}{n\epsilon}\left(1 + \left(\frac{8\log(1/\zeta)}{p}\right)^{1/4}\right),
\end{aligned}
$$

for some constants $c_1, c_2, c_3 > 0$*, where* $G' = \max\{2HM_{\mathcal{C}}, H\}$*.*

Proof. The proof is similar to Theorems 1 and 2, we first analyze the optimization error $R_n(\hat{\theta}_n) - R_n(\theta_n^*)$.

For algorithm 1, with normalization, we have:

$$
\begin{aligned}
R_n(\hat{\theta}_{t+1}) - R_n(\hat{\theta}_t) &\overset{(\alpha)}{\leq} \langle \nabla_\theta R_n(\hat{\theta}_t), \hat{\theta}_{t+1} - \hat{\theta}_t\rangle + \frac{H}{2}\left\|\hat{\theta}_{t+1} - \hat{\theta}_t\right\|_2^{\alpha+1} \\
&= -\eta_t\langle \nabla_\theta R_n(\hat{\theta}_t), \nabla_\theta R_n(\hat{\theta}_t) + b\rangle + \frac{H\eta_t^{\alpha+1}}{2}\left(\left\|\nabla_\theta R_n(\hat{\theta}_t)\right\|_2 + \|b\|_2\right)^{\alpha+1} \\
&\leq -\eta_t\left\|\nabla_\theta R_n(\hat{\theta}_t)\right\|_2^2 + \frac{H\eta_t^{\alpha+1}}{2}\left(\left\|\nabla_\theta R_n(\hat{\theta}_t)\right\|_2 + \|b\|_2\right)^2 + \eta_t\left\|\nabla_\theta R_n(\hat{\theta}_t)\right\|_2\|b\|_2 \\
&\leq \left(\frac{H\eta_t^{\alpha+1}}{2} - \eta_t\right)\left\|\nabla_\theta R_n(\hat{\theta}_t)\right\|_2^2 + \frac{H\eta_t^{\alpha+1}}{2}\|b\|_2^2 + G'\left(\eta_t + H\eta_t^{\alpha+1}\right)\|b\|_2 \\
&\overset{(PL)}{\leq} \left(\mu H\eta_t^{\alpha+1} - 2\mu\eta_t\right)\left(R_n(\hat{\theta}_t) - R_n(\theta_n^*)\right) + \frac{H\eta_t^{\alpha+1}}{2}\|b\|_2^2 + G'\left(\eta_t + H\eta_t^{\alpha+1}\right)\|b\|_2,
\end{aligned}
$$

the second inequality holds because by normalization.

Summing $R_n(\theta_t) - R_n(\theta_n^*)$ to both sides, we have:

$$
R_n(\hat{\theta}_{t+1}) - R_n(\theta_n^*) \leq \left(1 + \mu H\eta_t^{\alpha+1} - 2\mu\eta_t\right)\left(R_n(\hat{\theta}_t) - R_n(\theta_n^*)\right) + \frac{H\eta_t^{\alpha+1}}{2}\|b\|_2^2 + G'\left(\eta_t + H\eta_t^{\alpha+1}\right)\|b\|_2.
$$

Note that b is zero mean, so if taking expectation of both sides, we have:

$$
\begin{aligned}
\mathbb{E}\left[R_n(\hat{\theta}_{t+1}) - R_n(\theta_n^*)\right] &\leq \left(1 + \mu H\eta_t^{\alpha+1} - 2\mu\eta_t\right)\mathbb{E}\left[R_n(\hat{\theta}_t) - R_n(\theta_n^*)\right] + \frac{H\eta_t^{\alpha+1}}{2}\mathbb{E}\left[\|b\|_2^2\right] \\
&= \left(1 + \mu H\eta_t^{\alpha+1} - 2\mu\eta_t\right)\mathbb{E}\left[R_n(\hat{\theta}_t) - R_n(\theta_n^*)\right] + \frac{H\eta_t^{\alpha+1}\sigma^2 p}{2}.
\end{aligned}
$$

With Lemma 6, with probability at least $1-\xi$,

$$
\begin{aligned}
R_n(\hat{\theta}_{t+1}) - R_n(\theta_n^*) \leq\ & \left(1 + \mu H\eta_t^{\alpha+1} - 2\mu\eta_t\right)\left(R_n(\hat{\theta}_t) - R_n(\theta_n^*)\right) \\
& + \frac{H\eta_t^{\alpha+1}\sigma^2 p}{2}\left(1 + \left(\frac{8\log(1/\xi)}{p}\right)^{1/4}\right)^2 \\
& + G'\left(\eta_t + H\eta_t^{\alpha+1}\right)\sigma\sqrt{p}\left(1 + \left(\frac{8\log(1/\xi)}{p}\right)^{1/4}\right).
\end{aligned}
$$

Then summing over T iterations and setting $\xi = \zeta/T$, we have:

$$\begin{aligned}
R_n(\hat{\theta}_n) - R_n(\theta_n^*) \leq & \left(1 + \mu H \eta_t^{\alpha+1} - 2\mu\eta_t\right)^T \left(R_n(\hat{\theta}_0) - R_n(\theta_n^*)\right) \\
& + \frac{\left(1 - \left(1 + \mu H \eta_t^{\alpha+1} - 2\mu\eta_t\right)^T\right)}{1 - \left(1 + \mu H \eta_t^{\alpha+1} - 2\mu\eta_t\right)} \frac{H\eta_t^{\alpha+1}\sigma^2 p}{2} \left(1 + \left(\frac{8\log(T/\zeta)}{p}\right)^{1/4}\right)^2 \\
& + \frac{\left(1 - \left(1 + \mu H \eta_t^{\alpha+1} - 2\mu\eta_t\right)^T\right)}{1 - \left(1 + \mu H \eta_t^{\alpha+1} - 2\mu\eta_t\right)} G'\left(\eta_t + H\eta_t^{\alpha+1}\right)\sigma\sqrt{p} \left(1 + \left(\frac{8\log(T/\zeta)}{p}\right)^{1/4}\right) \\
\leq & \left(1 + \mu H \eta_t^{\alpha+1} - 2\mu\eta_t\right)^T M_\ell + \frac{H\eta_t^{\alpha+1}\sigma^2 p}{2(2\mu\eta_t - \mu H\eta_t^{\alpha+1})} \left(1 + \left(\frac{8\log(T/\zeta)}{p}\right)^{1/4}\right)^2 \\
& + \frac{G'\left(\eta_t + H\eta_t^{\alpha+1}\right)\sigma\sqrt{p}}{2\mu\eta_t - \mu H\eta_t^{\alpha+1}} \left(1 + \left(\frac{8\log(T/\zeta)}{p}\right)^{1/4}\right).
\end{aligned}$$

For the expectation,

$$\mathbb{E}\left[R_n(\hat{\theta}_n) - R_n(\theta_n^*)\right] \leq \left(1 + \mu H \eta_t^{\alpha+1} - 2\mu\eta_t\right)^T M_\ell + \frac{H\eta_t^{\alpha+1}\sigma^2 p}{2(2\mu\eta_t - \mu H\eta_t^{\alpha+1})}.$$

Noting that $\eta_1 = \cdots = \eta_T = \eta$, where $\left(\frac{2}{H} - \frac{2^{-1/\alpha}}{\mu H^{(\alpha-1)/\alpha}}\right)^{1/\alpha} < \eta < \left(\frac{2}{H}\right)^{1/\alpha}$, we have:

$$0 < 1 + \mu H \eta_t^{\alpha+1} - 2\mu\eta_t < 1.$$

Taking $T = \mathcal{O}\left(\log(n)\right)$, with probability at least $1 - \zeta$, we have:

$$R_n(\hat{\theta}_n) - R_n(\theta_n^*) \lesssim \frac{G'\sqrt{p\log(n)\log(1/\delta)}}{n\epsilon} \left(1 + \left(\frac{8\log(T/\zeta)}{p}\right)^{1/4}\right).$$

Then, like in Theorem 1, we have:

$$\begin{aligned}
R(\hat{\theta}_n) - R(\theta^*) \leq & \, c_1 \frac{G'\sqrt{p\log(n)\log(1/\delta)}}{n\epsilon} \left(1 + \left(\frac{8\log(T/\zeta)}{p}\right)^{1/4}\right) \\
& + c_2 \left(\frac{G'^2\log^2(n)}{n} + \frac{B + M_\ell}{n}\right) \\
& + c_3 \frac{G'^2\log^{2.5}(n)\sqrt{p\log(1/\delta)}}{n\epsilon} \left(1 + \left(\frac{8\log(1/\zeta)}{p}\right)^{1/4}\right),
\end{aligned}$$

for some constants $c_1, c_2, c_3 > 0$.

The result follows.

□

A.5 Proof in Remark 2

In this section, we prove that if a loss function ℓ satisfies the Generalized Bernstein condition, then if regularized term $\lambda\|\theta\|_2^2$ is added to ℓ, it also satisfies the Generalized Bernstein condition.

Proof. With regularized term $\lambda\|\theta\|_2^2$, we have:

$$\ell_r(z, \theta) = \ell(z, \theta) + \lambda\|\theta\|_2^2, \quad R_r(\theta) = R(\theta) + \lambda\|\theta\|_2^2.$$

As a result,

$$\begin{aligned}
&\mathbb{E}\left[\left(\ell_r(z,\theta)-\ell_r(z,\theta^*)\right)^2\right] \\
&=\mathbb{E}\left[\left(\left(\ell(z,\theta)-\ell(z,\theta^*)\right)+\lambda\left(\|\theta\|_2^2-\|\theta^*\|_2^2\right)\right)^2\right] \\
&\leq \mathbb{E}\left[\left(\ell(z,\theta)-\ell(z,\theta^*)\right)^2\right]+\lambda^2\left(\|\theta\|_2^2-\|\theta^*\|_2^2\right)^2+2\lambda M_{\mathcal{C}}^2\mathbb{E}\left[\ell(z,\theta)-\ell(z,\theta^*)\right],
\end{aligned}$$

where the last inequality holds because $\mathbb{E}$ is taken over z.

Note that ℓ satisfies the Generalized Bernstein condition, we assume the parameter is B, so we have:

$$\begin{aligned}
\mathbb{E}\left[\left(\ell_r(z,\theta)-\ell_r(z,\theta^*)\right)^2\right] &\leq \left(B+2\lambda M_{\mathcal{C}}^2\right)\left(R(\theta)-R(\theta^*)\right)+\lambda M_{\mathcal{C}}^2\lambda\left(\|\theta\|_2^2-\|\theta^*\|_2^2\right) \\
&\leq \left(B+2\lambda M_{\mathcal{C}}^2\right)\left(R(\theta)-R(\theta^*)+\lambda\left(\|\theta\|_2^2-\|\theta^*\|_2^2\right)\right) \\
&= \left(B+2\lambda M_{\mathcal{C}}^2\right)\left(R_r(\theta)-R_r(\theta^*)\right).
\end{aligned}$$

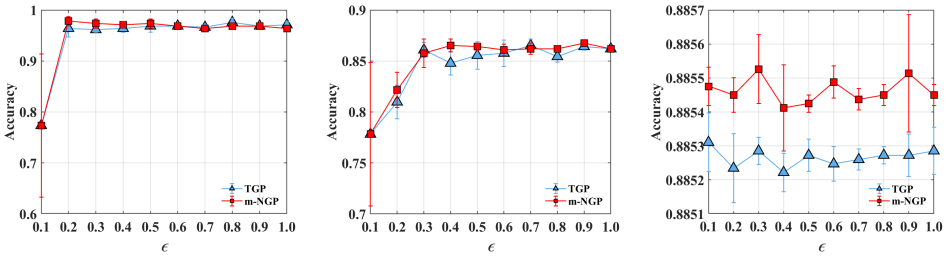
Bartlett et al. (2006) shows that squared piecewise-linear functions satisfy the Generalized Bernstein condition, so if a regularized term $\lambda\|\theta\|_2^2$ is added to it, the Generalized Bernstein condition is also guaranteed. And squared piecewise-linear functions with regularized term $\lambda\|\theta\|_2^2$ can also be seemed as a strongly convex composition to piecewise-linear functions, so along with the PL condition shown in Charles & Papailiopoulos (2018), claim (3) in Remark 2 holds. □

B More Experimental Results

B.1 Accuracies on More Datasets

In this section, we show the experimental results on datasets Breast Cancer, Credit Card Fraud, and Bank. Details are shown in Figure 2.

The results are similar to which given by Figure 1 in Section 5: although there are some fluctuations over some datasets (such as Bank), the performance of our proposed m-NGP method is similar to or better than traditional method on most datasets.



(a) Breast Cancer (b) Credit Card Fraud (c) Bank

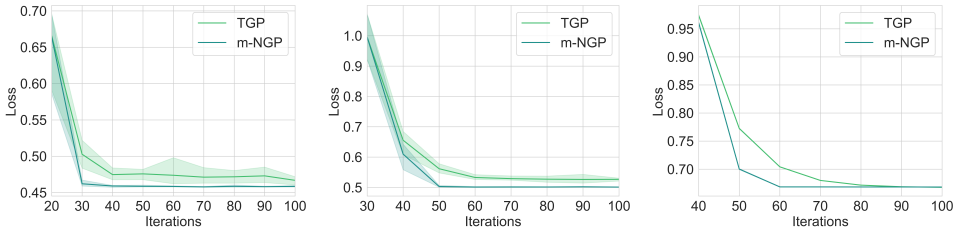
Figure 2: Comparisons between Traditional Gradient Perturbation (TGP) method and $\max\{\mathbf{1},\mathbf{g}\}$-Normalized Gradient Perturbation (m-NGP) method.

B.2 Convergence Rate and Normalization

In this section, we perform experiments to demonstrate the effects on the convergence rate caused by normalization when applying m-NGP. The privacy budget ϵ is set 0.5. Detailed results are shown in Figure 3.

In Figure 3, the lines with dark color and light color correspond to m-NGP and TGP, respectively, and the shadow area represents the maximum and minimum loss over mutiple experiments, reflecting the

variance. And the horizontal axis is iterations and the ordinate is the loss. The experimental results show that over most datasets, m-NGP (normalization) achieves faster convergence rate, comparing with TGP, which is in line with the theoretical analysis.



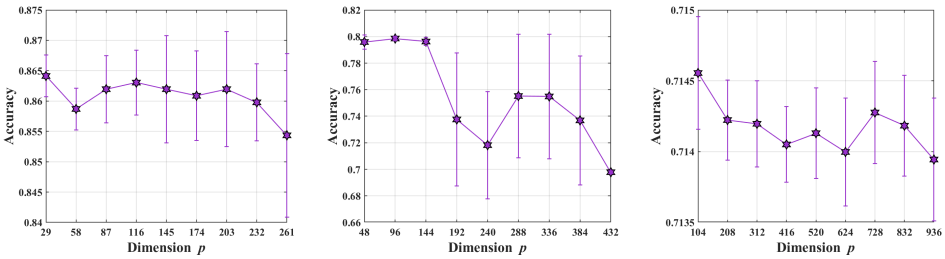
(a) Breast Cancer (b) Credit Card Fraud (c) Bank

Figure 3: Convergence Rates of TGP and m-NGP.

B.3 Accuracy and Dimension p

In this section, we perform experiments to demonstrate the effects on the accuracy brought by the dimensions of data instances. The experiments are performed on datasets Credit Card Fraud, Bank, and Adult, whose dimensions are 29, 48, and 104, respectively. And the privacy budget ϵ is set 0.5. The results are shown in Figure 4.

For abscissa, the first dimensions of parts (a), (b), and (c) are set $p = 29, 48, 104$, they are original features given by the datasets. And the dimensions more than them are all set 0, to evaluate the effects brought by the magnitude of p, without introducing new information.



(a) Credit Card Fraud (b) Bank (c) Adult

Figure 4: Effects of dimension p on m-NGP.

Experimental results show that although there may exist some fluctuations caused by the injected random noise, the accuracy decreases with the increasing of p overall, which is in line with the theoretical analysis given in Section 4.