Abstract

In this paper, by introducing Generalized Bernstein condition, we propose the first $O(\sqrt{\frac{p}{n}})$ high probability excess population risk bound for differentially private algorithms under the assumptions $L$-Lipschitz, $L$-smooth, and Polyak-Łojasiewicz condition, based on gradient perturbation method. If we replace the properties $G$-Lipschitz and $L$-smooth by $\alpha$-Hölder smoothness (which can be used in non-smooth setting), the high probability bound comes to $O(n^{-\frac{1}{1+\alpha}})$ w.r.t $n$, which cannot achieve $O(1/n)$ 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 normalizing, the high probability excess population risk bound under assumptions $\alpha$-Hölder smooth and Polyak-Łojasiewicz condition can achieve $O(\frac{\sqrt{p}}{n})$, which is the first $O(1/n)$ high probability excess population risk 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 of the differentially private model over real datasets. It demonstrates that m-NGP improves the utility 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 and 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). Under these circumstances, the privacy of machine learning models is of great importance.

Differential Privacy (DP) (Dwork et al. 2006; Dwork, Roth, and others 2014) is a theoretically rigorous tool to prevent sensitive information leakage. It introduces random noise when training 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, Pihur, and Korolova 2014), Apple (McMillan 2016) and Microsoft (Ding, Kulkarni, and Yekhanin 2017). As such, DP has attracted attentions from researchers and has been applied to numerous machine learning problems (Ullman and Sealfon 2019; Xu et al. 2019; Bernstein and Sheldon 2019; Wang and Xu 2019; Heikila et al. 2019; Kulkarni et al. 2021; Bun, Elias, and Kulkarni 2021; Nguyen and Vullikanti 2021).

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

In this paper, we aim to minimize the population risk, and measure the utility of the differentially private 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, Bousquet, and Mendelson 2002) and algorithm stability theory (Bousquet and Elisseeff 2002) are popular tools to analyze the generalization error. On one hand, (Chaudhuri, Monteleoni, and Sarwate 2011) applied the complexity theory and achieved an $O\left(\max\left(\frac{1}{\sqrt{n}}, \frac{\sqrt{p}}{\sqrt{n}}\right)\right)$ high probability excess population risk bound under the assumption of strongly convex; (Kifer, Smith, and Thakurta 2012) achieved $O\left(\frac{\sqrt{p}}{\sqrt{n}}\right)$ 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, Koren, and Talwar 2020; Bassily et al. 2020; Wang et al. 2021) are $O\left(\frac{\sqrt{p}}{\sqrt{nc}} + \frac{1}{\sqrt{n}}\right)$ or $O\left(\frac{\sqrt{p}}{\sqrt{nc}}\right)$, containing an inevitable $O\left(\frac{1}{\sqrt{n}}\right)$ 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 $O\left(\frac{\sqrt{p}}{n}\right)$ for DP models via uniform stability?*

By introducing Generalized Bernstein condition
(Koltchinskii 2006), this paper answers the question positively. We remove the $O\left(\frac{1}{\sqrt{n}}\right)$ term in the generalization error and provide the first high probability excess population risk bound with order $O\left(\frac{\sqrt{p}}{n}\right)$ in the setting of DP. Comparing with previous high probability bounds, the improvement is approximately up to $O\left(\sqrt{n}\right)$. The contributions of this paper include: (1) We prove that by introducing Generalized Bernstein condition (Koltchinskii 2006), the high probability excess population risk can be improved to $O\left(\frac{\sqrt{p}}{n}\right)$, under Lipschitz and smooth assumptions. To our knowledge, this is the first $O\left(\frac{\sqrt{p}}{n}\right)$ high probability excess population risk bound for DP model. (2) We relax the assumptions $G$-Lipschitz and $L$-smooth, by introducing $\alpha$-Hölder smooth. Under these assumptions, we prove that the high probability excess population risk bound comes to $O\left(\frac{\sqrt{p}}{n^{1+2/\alpha}}\right)$. Considering that $\alpha \in (0, 1]$, the result cannot achieve $O\left(\frac{1}{n}\right)$ w.r.t $n$, but better than previous one w.r.t $p$ and $\epsilon$. (3) 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 $\alpha$-Hölder smooth and PL condition, the high probability excess population risk bound can be improved to $O\left(\frac{\sqrt{p}}{n}\right)$. To the best of our knowledge, this is the first $O\left(\frac{\sqrt{p}}{n}\right)$ high probability excess population risk bound for non-smooth loss in the field of differential privacy. (4) To evaluate the performance of our proposed max $\{1, g\}$-Normalized Gradient Perturbation algorithm, we perform experiments on several real datasets. The experimental results show that m-NGP improves the accuracy and the convergence rate of the differentially private model on real datasets. The rest of the paper is organized as follows. Related work is given in Section 2. Preliminaries are 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 differential privacy 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, Monteleoni, and Sarwate 2011) first proposed output perturbation and objective perturbation methods, and (Song, Chaudhuri, and Sarwate 2013) first proposed the gradient perturbation method. Based on these works, (Kifer, Smith, and Thakurta 2012; Bassily, Smith, and Thakurta 2014; Abadi et al. 2016; Wang, Ye, and Xu 2017; Zhang et al. 2017; Wu et al. 2017; Bassily et al. 2019; Feldman, Koren, and Talwar 2020) further improved the results under different assumptions.

Among the works mentioned above, some of them analyzed the privacy guarantees (Song, Chaudhuri, and Sarwate 2013; Abadi et al. 2016), some of them discussed the excess empirical risk bound (Wang, Ye, and Xu 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, Smith, and Thakurta 2012) achieved an $O\left(\frac{\sqrt{p}}{n}\right)$ excess population risk bound via complexity theory under expectation condition; (Bassily, Smith, and Thakurta 2014) achieved similar expected excess population risk bound under convexity assumption, via optimization theory; (Wang, Chen, and Xu 2019) proposed an $O\left(\frac{p}{\log(n)\sqrt{n}}\right)$ excess population risk bound under non-convex condition in expectation, via Langevin Dynamics method (Gelfand and Mitter 1991) and the stability of Gibbs algorithm; and (Feldman, Koren, and Talwar 2020) gives expected population risk bound of the order $O\left(\frac{1}{n} + \frac{\sqrt{p}}{n}\right)$, under strongly convex condition.

However, the behavior of the algorithm within a single or few runs cannot be well captured by expectation bounds, which is related to the probabilistic nature. In addition, in practical applications such as deep learning, it is often the case that the algorithm runs only once since the training process may take a long time. Therefore, obtaining a high probability bound is essential to ensure the performance of the algorithm on a single or few runs. So we focus on the high probability bound in this paper. Meanwhile, we concentrate on stability theory. 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 $O\left(\frac{1}{\sqrt{n}}\right)$ term, details can be found in Table 1. The reason is that when analyzing the generalization error, the technical routes follow works (Bousquet and Elisseeff 2002; Hardt, Recht, and Singer 2016). Besides, when analyzing the stability, previous works always do not consider the injected noise (e.g. (Wang et al. 2021)), or assume the random noise injected into adjacent datasets is the same. However, this is not reasonable, because ‘adjacent dataset’ is also the basis of DP. With the same random noise, it is hard to say that ‘DP is guaranteed’.

In this paper, we consider the random noise injected into adjacent datasets and analyze the stability under the noisy version. By introducing the Generalized Bernstein condition (Koltchinskii 2006), we remove the $O\left(\frac{1}{\sqrt{n}}\right)$ term when combining the stability and the generalization error, and further improve the excess population risk bound of differentially private models. The improved convergence rate is up to $O\left(\frac{\sqrt{p}}{n}\right)$, which positively answers the question given in Section 1: Can the high probability excess population risk bound achieve $O\left(\frac{1}{n}\right)$ w.r.t $n$. The improvements are shown in Table 1. For ‘TYPE’, E. means expectation bound and H.P. means the high probability bound.

Table 1 first shows that by adding more assumptions, we achieve a better high probability excess population risk
A

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

| Assumptions                  | Method                     | Utility Bound          | Type |
|------------------------------|----------------------------|------------------------|------|
| Bassily et al. 2019          | Lipschitz, smooth, convex  | $O \left( \frac{\sqrt{p}}{n\epsilon} + \frac{1}{\sqrt{n}} \right)$ | E.   |
| (Feldman, Koren, and Talwar 2020) | Lipschitz, convex          | $O \left( \frac{\sqrt{p}}{n\epsilon} + \frac{1}{\sqrt{n}} \right)$ | E.   |
| (Feldman, Koren, and Talwar 2020) | Lipschitz, strongly convex | $O \left( \frac{\sqrt{p}}{n\epsilon} + \frac{1}{\sqrt{n}} \right)$ | E.   |
| Bassily et al. 2020          | Lipschitz, convex          | $O \left( \frac{\sqrt{p}}{n\epsilon} + \frac{1}{\sqrt{n}} \right)$ | H.P. |
| (Wang et al. 2021)           | $\alpha$-Hölder smooth, convex | $O \left( \frac{\sqrt{p}}{n\epsilon} + \frac{1}{\sqrt{n}} \right)$ | H.P. |
| (Wang et al. 2021)           | $\alpha$-Hölder smooth, convex | $O \left( \frac{\sqrt{p}}{n\epsilon} + \frac{1}{\sqrt{n}} \right)$ | H.P. |
| Ours                        | Lipschitz, smooth, PL condition | $O \left( \frac{\sqrt{p}}{n\epsilon} \right)$ | H.P. |
| Ours                        | $\alpha$-Hölder smooth, PL condition | $O \left( \frac{\sqrt{p}}{n\epsilon} \right)$ | H.P. |

bound, $O \left( \frac{\sqrt{p}}{n\epsilon} \right)$, which is state-of-the-art to the best of our knowledge. Then, we relax the assumptions and achieve $O \left( \frac{\sqrt{p}}{n\epsilon} \right)$ high probability bound, but it cannot achieve the same bound ($O \left( 1/n \right)$ w.r.t. $n$) under the condition that the loss function is Lipschitz, smooth, and satisfies PL condition. To overcome this problem, we propose an algorithm called m-NGP, and achieve the $O \left( \frac{\sqrt{p}}{n\epsilon} \right)$ result under the same assumptions: $\alpha$-Hölder smooth and PL condition.

Moreover, although it is hard to directly compare the PL condition with convexity, PL condition can be applied to many non-convex conditions (more information can be found in Section 4.2). Besides, PL condition is weaker compared with strongly convex condition, and one of the best population risks under strongly convex condition is $O \left( \frac{1}{n} + \frac{p}{\epsilon^2} \right)$ (Feldman, Koren, and Talwar 2020) (line 2 in Table 1). However, the result is an expectation one, different from ours. In this paper, we analyze the excess population risk bound of DP algorithm under high probability and PL cases, different from previous scenarios.

3. Preliminaries

In this paper, we assume that there are $n$ data instances in dataset $D$, i.e., $D = \{z_1, \ldots, z_n\}$ 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 $\ell$ 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 $x = [x_1, \ldots, x_d]^T$, its $l_p$-norm is $\|x\|_p = (\sum_{i=1}^d |x_i|^p)^{\frac{1}{p}}$. And $A \leq B$ represents that there exists some constant $c > 0$, $A \leq cB$.

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

$$
\Pr[A(D) \in S] \leq e^\epsilon \Pr[A(D') \in S] + \delta.
$$

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 and Evans 2019; Carlini et al. 2019).

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

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

where $\eta_t$ is the learning rate at iteration $t$, $b_t$ is the random noise injected into the gradient, $\hat{\theta}$ is corresponding model with privacy, and $R_n(\hat{\theta})$ is the empirical risk, defined as $R_n(\hat{\theta}) := \frac{1}{n} \sum_{i=1}^n \ell(\hat{z}_i, \hat{\theta})$.

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

$$
R(\hat{\theta}_n) - \min_{\theta \in \mathcal{C}} R(\theta) = R(\hat{\theta}_n) - R_n(\hat{\theta}_n) + R_n(\hat{\theta}_n) - R_n(\theta^*) + R_n(\theta^*) - R(\theta^*), \tag{2}
$$
where \( \theta^* = \arg\min_{\theta \in \mathcal{C}} R(\theta), \theta_n^* = \arg\min_{\theta_n \in \mathcal{C}} R_n(\theta). \) In (2), part A is exactly the generalization. Via the definition of \( \theta_n^* \), we have \( R_n(\theta_n^*) = R_n(\theta^*) \leq R_n(\theta) - R_n(\theta_n^*) \), which bounds part B by the optimization error (also called the excess empirical risk). In this way, we answer the question mentioned in Section 1: Why generalization error is an important step towards the excess population risk.

To get the generalization error, algorithm stability theory is a popular tool, we introduce uniform stability and uniform argument stability here.

**Definition 2 (Uniform Stability (Bousquet and Elisseeff 2002)).** An algorithm \( \theta_n \) is \( \gamma \)-uniformly stable if for any \( S = \{z_1, \ldots, z_i, \ldots, z_n\} \) and \( S' = \{z_1, \ldots, z_i', \ldots, z_n\} \), where \( i = 1, \ldots, n \), it holds that

\[
|\ell(z, \theta_n(S)) - \ell(z, \theta_n(S'))| \leq \gamma.
\]

In this paper, we use notation \( \ell_n(S) \) 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.

**Definition 3 (Uniform Argument Stability (Bassily et al. 2020)).** Algorithm \( \theta_n \) is \( \gamma \)-uniformly argument stable if for any \( S = \{z_1, \ldots, z_i, \ldots, z_n\} \) and \( S' = \{z_1, \ldots, z_i', \ldots, z_n\} \), where \( i = 1, \ldots, n \), it holds that

\[
\|\ell_n(S) - \ell_n(S')\|_2 \leq \gamma.
\]

Definition 3 shows that the uniform argument stability measures the upper bound of the difference (on the model parameter) between the models derived from adjacent datasets.

Furthermore, we introduce some assumptions.

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

\[
|\ell(z, \theta_1) - \ell(z, \theta_2)| \leq G \|\theta_1 - \theta_2\|_2.
\]

With Assumption 1, one can easily get that if the loss function is G-Lipschitz, then \( \gamma \)-uniformly argument stability implies \( G \gamma \)-uniformly stability.

**Assumption 2 (L-smooth).** The loss function \( \ell : \mathcal{D} \times \mathcal{C} \to \mathbb{R} \) is L-smooth over \( \theta \) 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 loss function \( \ell(\cdot, \cdot) \) 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, Monteleoni, and Sarwate 2011; Kifer, Smith, and Thakurta 2012; Abadi et al. 2016; Bassily et al. 2018; Feldman, Koren, and Talwar 2020; Bassily et al. 2020). To relax the Lipschitz and smoothness assumptions, we introduce the \( \alpha \)-Hölder smoothness of the loss function:

**Assumption 3 (\( \alpha \)-Hölder smooth).** Let \( \alpha \in (0, 1] \), The loss function \( \ell : \mathcal{D} \times \mathcal{C} \to \mathbb{R} \) is \( \alpha \)-Hölder smooth over \( \theta \) 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 (Ying and Zhou 2017).** If the loss function \( \ell(\cdot, \cdot) \) is differentiable, then Assumption 3 yields

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

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., \(|\theta|_2 \leq M_\mathcal{C} \), \( \alpha \)-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 \( \ell(z, \theta) = (\max(0, 1 - y(z, \theta)))^q \) for classification and the \( q \)-th power absolute distance loss \( \ell(z, \theta) = |y - \theta(z, \theta)|^q \) for regression (Lei and Ying 2020a), whose \( \ell \) are \((q - 1)\)-Hölder smooth if \( q \in (1, 2] \) (Li and Liu 2021). Lemma 1 shows that Hölder smoothness shares similar property with smoothness defined in Assumption 2.

### 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, Ye, and Xu 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, Ye, and Xu 2017).** In gradient perturbation method in (1), if Assumption 1 holds, then for \( \epsilon, \delta > 0 \), it is \((\epsilon, \delta)\)-DP if the Gaussian random noise \( \eta_t \sim \mathcal{N}(0, \sigma^2 I_p) \), and for some constant \( c \),

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

**Remark 1.** Lemma 2 only assumes the loss function to be G-Lipschitz. If we assume that \( \ell(\cdot, \cdot) \) is \( \alpha \)-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

Before analyzing the excess population risk, we first introduce some assumptions. Most of the previous works assumed that the loss function is convex (or strongly convex) when analyzing the empirical and population risks. In this paper, we use the Polyak-Łojasiewicz (PL) condition to replace convexity.

**Assumption 4 (Polyak-Łojasiewicz condition).** Function \( f(\theta) \) satisfies the Polyak-Łojasiewicz (PL) condition if there exists \( \mu > 0 \) and for every \( \theta \),

\[
\|\nabla_\theta f(\theta)\|_2^2 \geq 2\mu (f(\theta) - f(\theta^*)),
\]

where \( \theta^* = \arg\min_{\theta \in \mathcal{C}} f(\theta) \).

In this paper, we assume the empirical risk and the population risk both satisfy the PL condition.

---

1PL condition can be directly derived from strongly convex (Karimi, Nutini, and Schmidt 2016), so all the results given in this paper hold when it comes to the strongly convex conditions.
The Polyak-Łojasiewicz condition is one of the weakest curvature conditions [Karimi, Nutini, and Schmidt 2016; Li and Liu 2021], weaker than ‘one-point convexity’ [Kleinberg, Li, and Yuan 2018], ‘star convexity’ [Zhou et al. 2019], and ‘quasar convexity’ [Hinder, Sidford, and Sohoni 2020]. It is widely used in the analysis of non-convex learning (Wang, Ye, and Xu 2017), Charles and Papailiopoulos 2018 [Lei and Ying 2020b; Lei and Tang 2021] and many popular non-convex objective functions satisfy the PL condition, such as: matrix factorization (Liu, Wu, and So 2016), robust regression [Liu, Wu, and So 2016], neural networks with one hidden layer (Li and Yuan 2017), mixture of two Gaussians (Balakrishnan, Wainwright, and Yu 2017), ResNets with linear activations (Hardt and Ma 2017), linear dynamical systems (Hardt, Ma, and Recht 2018), phase retrieval (Sun, Qu, and Wright 2018), and blind deconvolution (Li et al. 2019).

**Remark 2.** [Karimi, Nutini, and Schmidt 2016] shows that the Polyak-Łojasiewicz inequality directly implies the following inequality: \( R_n(\theta) - R_n(\theta^*_n) \geq \frac{\eta}{2} \| \theta - \theta^*_n \|^2 \), which is called the Quadratic Growth (QG) condition. In the following, we use QG to bound the argument stability of the gradient perturbation algorithm.

Then, with Assumption \( \mathbb{3} \) we discuss the uniform argument stability of the private model.

**Lemma 3.** In gradient perturbation method \( (\mathbb{7}) \), if Assumptions \( \mathbb{2}, \mathbb{4} \) hold, then

\[
\left\| \hat{\theta}_n - \hat{\theta}_n^* \right\|_2 \leq 2\sqrt{2} \sqrt{\frac{R_n(\hat{\theta}_n) - R_n(\theta^*_n)}{\mu} + 4G \frac{p}{\mu n^2}},
\]

where \( \hat{\theta}_n \) and \( \hat{\theta}_n^* \) denote the models derived from adjacent datasets \( S \) and \( S' \).

The proof is given in Appendix A.1. By the QG condition (implied by PL inequality) discussed in Remark 2, Lemma 3 connects the uniform argument stability with the empirical risk. And as a result, we only need to consider the noise when analyzing the empirical risk. Besides, when it comes to the stability of DP models, previous results often assume that the noise added to the gradient in each iteration is the same for adjacent datasets \( S \) and \( S' \) (e.g. [Wang et al. 2021]). This is not reasonable because noise injection is an independent process, so we expand it in this paper.

And to remove the \( O(\sqrt{n}) \) term in previous results, we further need the Generalized Bernstein condition when analyzing the excess population risk.

**Assumption 5** (Generalized Bernstein condition [Koltchinskii 2006]). We say the loss function \( \ell \) satisfies the generalized Bernstein condition if for some \( B > 0 \) for any \( \theta \in \mathcal{C} \), we have:

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

**Remark 3.** Here, we discuss the connections between Assumptions \( \mathbb{2}, \mathbb{4} \) and \( \mathbb{5} \). Via Assumption \( \mathbb{2} \), we have

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

And Remark 2 shows that if \( R(\theta) \) satisfies the Polyak-Łojasiewicz condition, we have \( \frac{\eta}{2} \| \theta - \theta^* \|^2 \leq R(\theta) - R(\theta^*) \). Combining these inequalities together, we observe that if Assumptions \( \mathbb{2}, \mathbb{4}, \mathbb{5} \) hold, then the loss function \( \ell(\cdot) \) satisfies Assumption \( \mathbb{5} \) with parameter \( B = \frac{2G^2}{\mu} \).

With the stability of the private model and Assumption \( \mathbb{5} \) now we come to the excess population risk.

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

\[
R(\theta_n) - R(\theta^*) \leq c_1 G \log^{1.5}(n) \sqrt{p \log(1/\delta)} \left( 1 + \left( \frac{8 \log(T/\zeta)}{p} \right)^{1/4} \right)
+ c_2 G^2 p \log(n) \log(1/\delta) \left( 1 + \left( \frac{8 \log(T/\zeta)}{p} \right)^{1/4} \right)^2
+ c_3 \log(n) \sqrt{n/\delta},
\]

for some constants \( c_1, c_2, c_3 > 0 \).

Detailed proof can be found in Appendix A.3, we give a proof sketch here. First, by Lemma 3 and Lipschitzness, we get the uniform stability of gradient perturbation \( (\mathbb{7}) \). Then, we analyze the generalization error via stability theory. By novel decomposition method, via Assumption 5 and its moments bound, we couple term \( R(\theta_n) - R(\theta_n^*) \) and term \( R_n(\theta^*) - R(\theta^*) \) in \( (\mathbb{2}) \) together, to remove the \( O(1/\sqrt{n}) \) term in the generalization error. In this way, a better excess population risk bound is achieved. The proof is motivated by [Klochkov and Zhivotovskiy 2021] in the non-private case. The key difference is that in the setting of DP, random noise is injected into the algorithm. In [Klochkov and Zhivotovskiy 2021], a key step to analyze the generalization error is summing \( X_i = \mathbb{E}' [\ell(z_i, \theta_n^*) - \ell(z_i, \theta^*)] \) for \( i = 1, \cdots, n \), where \( \theta_n^* \) is derived from an independent copy of the original dataset and \( \mathbb{E}' \) means the expectation over the independent copy. When summing, \( X_i \) is required to be zero mean. However, in the cases of DP, if we replace \( \theta_n^* \) 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 \( R(\theta_n) - R(\theta^*) \geq R(\theta_n) + R(\theta^*) - R_n(\theta_n) + R_n(\theta_n) - R_n(\theta_n^*) + R_n(\theta_n^*) - R(\theta^*) \), which naturally solves the problem mentioned above (the generalization error is discussed over the non-private model). However, when it comes to the gradient perturbation method, we cannot solve the problem 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.

**Remark 4.** If omitting \( \log(\cdot) \) and constant terms, the excess population risk bound comes to:

\[
O \left( \frac{p}{n^2 \epsilon^2} + \sqrt{\frac{p}{nc} + \frac{1}{n}} \right) = O \left( \frac{\sqrt{p}}{nc} \right).
\]

The result is better than previous ones containing \( O(\frac{1}{\sqrt{n}}) \) terms. And it is the first high probability excess population
risk bound over DP algorithm overcoming the $O(n^{-1/2})$ bottleneck, to the best of our knowledge.

Then, we replace Assumptions (L1-G-Lipschitz) and (L-smooth) by Assumption (3-α-Hölder smooth).

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

$$R(\hat{\theta}_n) - R(\theta^*) \leq c_1 \frac{G' \log(n)^{1/2} p \log(1/\delta)}{n^{\frac{1}{2\alpha+1}}} \left(1 + \left(\frac{8 \log(T/\zeta)}{p}\right)^{1/4}\right)$$

$$+ c_2 \frac{G'^2 \log(n)^{1/2}}{n^{2\alpha+1}} \left(1 + \left(\frac{8 \log(T/\zeta)}{p}\right)^{1/4}\right)$$

$$+ c_3 \frac{\log(n)}{n},$$

for some constants $c_1, c_2$ and $c_3$, where $G' = \max\{2HM_C, H\}$.

Detailed proof can be found in Appendix A.4. The proof is similar to Theorem (1). Considering that the stability parameter given by Lemma (3) is related to the optimization error, so the key is to obtain the optimization error. The challenge is that the properties $G$-Lipschitz and $L$-smooth are replaced by the assumption $\alpha$-Hölder smooth when analyzing it. To overcome the challenge, we use Lemma (4) to bound the optimization error and Young’s inequality is used to normalize the exponential rate. By choosing proper learning rate, we get an acceptable excess population risk bound. Details are shown in the proof of Lemma (7).

By Theorem (2) it is easy to follow that with high probability, the excess population risk satisfies:

$$R(\hat{\theta}_n) - R(\theta^*) = O\left(\frac{p^{1/4}}{\epsilon^{2/3} n^{\frac{1}{2\alpha+1}}}\right).$$

**Remark 5.** By the definition of $\alpha$-Hölder smooth, we have $\alpha \in (0, 1]$, so our result is worse than the gradient based result given by (Wang et al. 2021) ($O(1/\sqrt{n})$ w.r.t $n$). One of the reasons is that (Wang et al. 2021) does not consider the noise injected into the model when analyzing stability. However, as discussed before, the noise addition is independent to datasets, so we expand the stability to the noisy version in this paper, which generalize previous settings. Besides, our result is of $O(\epsilon^{-1/2})$ w.r.t $\epsilon$, and previous results are of the order $O(\epsilon^{-1})$. In practice, $\epsilon$ is always set less than 1 to guarantee meaningful privacy, so our result is more superior when it comes to conditions that privacy requirements are strict (low $\epsilon$ conditions).

Via the discussion mentioned above, we observe that the result given in Theorem (2) is far worse than which given in Theorem (1) if we replace Lipschitzness and smoothness by $\alpha$-Hölder smoothness. The reason is that when applying

**Algorithm 1** max{1, g}-Normalized Gradient Perturbation

1. **Input:** Dataset $D$, learning rate at iteration $t$: $\eta_t$, the variance of the Gaussian noise injected to the gradient: $\sigma$.
2. Initialize $\theta_0$.
3. for $t = 0$ to $T - 1$
   4. $G_t \leftarrow R_n(\theta_t) + b_t, b_t \sim N(0, \sigma^2 I_p)$.
   5. if $\|G_t\|_2 < 1$ then
      6. $G_t \leftarrow G_t / \|G_t\|_2$.
   7. end if
   8. $\hat{\theta}_{t+1} \leftarrow \hat{\theta}_t - \eta_t G_t$.
9. end for
10. Return $\hat{\theta}_n = \hat{\theta}_T$.

Young’s inequality in the optimization error analysis, an additional term $\eta^{1+\alpha}/2(1+\alpha)$ appears, leading a loose excess population risk bound.

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

**Remark 6.** The difference between Algorithm (1) and (1) is that in lines 5 and 6, 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 $\|\theta_t - \theta_n\|^{1+\alpha}$ (derived from Lemma (1)), further remove term $\frac{H n^{\frac{1}{1+\alpha}}}{2(1+\alpha)}$ in the theoretical analysis. Details can be found in Appendix A.4.

Then, we improve the excess population risk bound given in Theorem (2).

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

$$R(\hat{\theta}_n) - R(\theta^*) \leq c_1 \frac{G' \log(n)^{1.5} p \log(1/\delta)}{n^{\frac{1}{2\alpha+1}}} \left(1 + \left(\frac{8 \log(T/\zeta)}{p}\right)^{1/4}\right)$$

$$+ c_2 \frac{G'^2 \log(n) p \log(1/\delta)}{n^{2\alpha+1}} \left(1 + \left(\frac{8 \log(T/\zeta)}{p}\right)^{1/4}\right)^2$$

$$+ c_3 \frac{\log(n)}{n},$$

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

Detailed proof is given in Appendix A.5. The proof is similar to Theorems (1) and (2) the key difference is that by gradient normalization, Young’s inequality is abandoned in the theoretical analysis (as discussed in Remark (6)), which implies a better excess population risk bound.

**Remark 7.** We introduce normalization from theoretical view, and the experiments show that it works in practice (de-
tails can be found in Section 5). Here, we provide some intu-
tions on why normalization works. For gradient perturba-
tion method, it is easy to observe that the random noise $b_t$ is
sampled independently at each iteration $t$, so it is reasonable
to suppose $b_t$ does not change too much when it comes to
different iterations. As a result, if the $L_2$-norm of the gradient
is too small, the random noise will play a more important
role and the gradient property will be concealed beneath the
noise. Besides, the random noise itself would help the gen-
eralization error (Li, Luo, and Qiao 2020). So in m-NGP, we
apply normalization to scale the $L_2$-norm of the gradient to
1 if it is less than 1, strengthens the gradient along with the
random noise.

By Theorem 5, it is easy to follow that with high proba-
bility, $R(\hat{\theta}_n) - R(\theta^*) = O(\frac{\sqrt{p}}{n})$. The bound is of the same
order as the result given in Theorem 1. This is also the first
$O(1/n)$ high probability excess population risk bound over
DP algorithm w.r.t $n$ without smoothness.

**Remark 8.** Theorems 2 and 3 require Assumptions 3 and
4 in this remark, we give some examples satisfy these as-
dumptions. As discussed in Section 3, the $q$-norm hinge loss
and $q$-th power absolute distance loss can be seemed as
squared piecewise-linear loss functions when $q = 2$, so they
satisfy Hölder smoothness. For one-layer neural networks
with squared error loss and leaky ReLU activations, the
same phenomenon holds because the neural network can be
seemed as a matrix multiplication. Beside, (Charles and Pa-
palliopoulos 2013) shows that several interesting machine
learning setups satisfy Assumption 4 such as 1-layer neural
networks with a squared error loss and leaky ReLU activa-
tions, loss functions of least squares minimizations, squared
piecewise-linear functions with regularized term, etc. As a
result, loss functions satisfy those assumptions including
but not limited to: (1) logistic regression and least squared
minimization; (2) some of the squared piecewise linear func-
tions; (3) some of the neural networks such as one-layer neu-
ral networks with squared error loss and leaky ReLU activa-
tions. The examples listed above are only part of the loss
functions who satisfy those assumptions. We do not only fo-
cus on least squared minimization and logistic regression
models, but extend the condition from smoothness to Hölder
smoothness, and from strongly convex (convex) to the PL
condition (some of the non-convex cases).

5. Experiments

In this section, we perform experiments on real datasets to
evaluate our proposed m-NGP algorithm.

The experiments are performed on classification task over
datasets Iris (Dua and Graff 2017), Breast Cancer (Man-
gasarian and Wolberg 1990), Credit Card Fraud (Bontempi
and Worldline 2018), Bank (Moro, Cortez, and Rita 2014),
and Adult (Dua and 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 convergence rate on
the training set. In all the experiments, the privacy budget $\delta$
is set $\frac{1}{n}$ and we choose $\epsilon = 0.1$ to 1.0.

We apply regularized logistic regression to the classifica-
tion task, which satisfies the assumptions mentioned before,
the experimental results are shown in Figure 1. We show the
results over datasets Iris and Adult in this section, experi-
ments on other datasets are shown in Appendices B.1 and
B.2. For convergence rate, the shadow area represents the
maximum and minimum loss over mutiple experiments, ref-
lecting 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 m-NGP is better than traditional gradient perturbation
method, which is in line with the theoretical analysis. More-
over, in Appendix B.3, we perform experiments to demon-
strate the effects brought by the dimension parameter $p$, the
experimental results follow the theoretical results: with in-
creasing $p$, the accuracy becomes worse in general.

6. Conclusions

In this paper, we first propose a state-of-the-art $O(\frac{\sqrt{p}}{n})$ high
probability excess population risk bound for gradient per-
turbation based DP algorithms, under the assumptions of
$G$-Lipschitz, $L$-smooth, and Polyak-Łojasiewicz condition.
The result positively answers the open problem: Can we
achieve high probability excess risk bound with rate $O(1/n)$ w.r.t $n$ for DP models via uniform stability? Then, we ex-
tend the result to a more general case, requiring $\alpha$-Hölder
smoothness and Polyak-Łojasiewicz condition. However,
the result is not as satisfactory as before, we achieve an
$O((n \rightarrow n)\alpha)$ high probability excess population risk bound,
which cannot achieve an $O(1/n)$ bound. To get a bet-
ter result, we further propose a new algorithm: max$\{1, g\}$-
Normalized Gradient Perturbation (m-NGP). Detailed theo-
retical analysis shows that m-NGP can achieve $O(\frac{\sqrt{p}}{n\alpha})$ high
probability excess population risk bound, under the assump-
tions of $\alpha$-Hölder smoothness and Polyak-Łojasiewicz condition,
which is the first $O(1/n)$ high probability bound w.r.t $n$ under non-smoothness differentially private cases. Exper-
imental 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

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

[Backes et al. 2016] Backes, M.; Berrang, P.; Humbert, M.;
and Manoharan, P. 2016. Membership privacy in microrna-
based studies. In Proceedings of the 2016 ACM SIGSAC
Conference on Computer and Communications Security, 319–330.

[Balakrishnan, Wainwright, and Yu 2017] Balakrishnan, S.;
Wainwright, M. J.; and Yu, B. 2017. Statistical guaran-
Figure 1: Comparisons between Traditional Gradient Perturbation (TGP) and max\{1, g\}-Normalized Gradient Perturbation (m-NGP).

[Bartlett, Bousquet, and Mendelson 2002] Bartlett, P. L.; Bousquet, O.; and Mendelson, S. 2002. Localized rademacher complexities. In *Computational Learning Theory, 15th Annual Conference on Computational Learning Theory, COLT 2002*, 44–58.

[Bassily et al. 2020] Bassily, R.; Feldman, V.; Guzmán, C.; and Talwar, K. 2020. Stability of stochastic gradient descent on nonsmooth convex losses. In *Advances in Neural Information Processing Systems*, 4381–4391.

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

[Bernstein and Sheldon 2019] Bernstein, G., and Sheldon, D. R. 2019. Differentially private bayesian linear regression. In *Advances in Neural Information Processing Systems*, 523–533.

[Bontempi and Worldline 2018] Bontempi, G., and Worldline. 2018. ULB the machine learning group.

[Boucheron, Lugosi, and Massart 2013] Boucheron, S.; Lugosi, G.; and Massart, P. 2013. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

[Boucheron, Lugosi, and Massart 2013] Boucheron, S.; Lugosi, G.; and Massart, P. 2013. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

[Bousquet and Elisseeff 2002] Bousquet, O., and Elisseeff, A. 2002. Stability and generalization. *The Journal of Machine Learning Research* 499–526.

[Bousquet, Klochkov, and Zhivotovskiy 2020] Bousquet, O.; Klochkov, Y.; and Zhivotovskiy, N. 2020. Sharper bounds for uniformly stable algorithms. In *Conference on Learning Theory*, 610–626.

[Bun, Elias, and Kulkarni 2021] Bun, M.; Elias, M.; and
A. Details of Proofs

A.1. Proof of Lemma 3

Lemma 4. In gradient perturbation method (1), if Assumptions 1, 4 hold, then

\[ \| \hat{\theta}_n - \tilde{\theta}_n \|_2 \leq 2 \sqrt{2 \left( \frac{R_n(\hat{\theta}_n) - R_n(\hat{\theta}_n^\ast)}{\mu} \right) + \frac{4G}{\mu n}} \]

where \( \hat{\theta}_n \) and \( \tilde{\theta}_n \) denote the models derived from adjacent datasets \( S \) and \( S' \).

Proof. By triangle inequality, we have:

\[ \| \hat{\theta}_n - \tilde{\theta}_n \|_2 \leq \| \hat{\theta}_n - \theta^\ast_n(S) \|_2 + \| \theta^\ast_n(S) - \theta^\ast_n(S') \|_2 + \| \theta^\ast_n(S') - \tilde{\theta}_n \|_2. \]  \hspace{1cm} (3)

Recalling that PL condition implies QG condition:

\[ R_n(\theta) - R_n(\theta^\ast_n) \geq \frac{\mu}{2} \| \theta - \theta^\ast_n \|_2^2. \]  \hspace{1cm} (4)

So we have:

\[ \| \hat{\theta}_n - \theta^\ast_n(S) \|_2 \leq \sqrt{\frac{2}{\mu} \left( R_n(\hat{\theta}_n) - R_n(\hat{\theta}_n^\ast) \right)}, \]

\[ \| \theta^\ast_n(S') - \tilde{\theta}_n \|_2 \leq \sqrt{\frac{2}{\mu} \left( R_n(\hat{\theta}_n) - R_n(\hat{\theta}_n^\ast(S')) \right)}. \]

Due to the symmetry, the two inequalities can be integrated into one, i.e.

\[ \| \hat{\theta}_n - \theta^\ast_n(S) \|_2 + \| \theta^\ast_n(S') - \tilde{\theta}_n \|_2 \leq 2 \sqrt{2 \left( \frac{R_n(\hat{\theta}_n) - R_n(\hat{\theta}_n^\ast)}{\mu} \right) + \frac{4G}{\mu n}}. \]  \hspace{1cm} (5)

Now we turn to term \( \| \theta^\ast_n(S) - \theta^\ast_n(S') \|_2 \).

For \( R_n(\theta^\ast_n(S')) - R_n(\theta^\ast_n(S)) \), if assuming the different data instance is \( z_j \), we have:

\[ R_n(\theta^\ast_n(S')) - R_n(\theta^\ast_n(S)) = \frac{1}{n} \left( \ell(z_j, \theta^\ast_n(S')) - \ell(z_j, \theta^\ast_n(S)) \right) + \frac{1}{n} \sum_{i \neq j} \ell(z_i, \theta^\ast_n(S')) - \ell(z_i, \theta^\ast_n(S)) \]

\[ = \frac{1}{n} \left( \ell(z_j, \theta^\ast_n(S')) - \ell(z_j, \theta^\ast_n(S)) \right) + \frac{1}{n} \left( \ell(z'_j, \theta^\ast_n(S')) - \ell(z'_j, \theta^\ast_n(S)) \right) \]

\[ + R_n(\theta^\ast_n(S')) - R_n(\theta^\ast_n(S)) \]

\[ \leq \frac{2G}{n} \| \theta^\ast_n(S') - \theta^\ast_n(S) \|_2 + R_n(\theta^\ast_n(S')) - R_n(\theta^\ast_n(S)) \]

\[ \leq \frac{2G}{\mu} \| \theta^\ast_n(S') - \theta^\ast_n(S) \|_2. \]  \hspace{1cm} (6)

where \( R_n(\theta) \) is the empirical risk over dataset \( S' \), the first inequality holds because \( G \)-Lipschitzness and the last inequality holds because \( R_n(\theta^\ast_n(S')) - R_n(\theta^\ast_n(S)) \leq 0 \) due to the definition of \( \theta^\ast_n(S') \).

By inequality (4),

\[ R_n(\theta^\ast_n(S')) - R_n(\theta^\ast_n(S)) \geq \frac{\mu}{2} \| \theta^\ast_n(S') - \theta^\ast_n(S) \|_2. \]  \hspace{1cm} (7)

Combining inequalities (6) and (7), we have:

\[ \frac{\mu}{2} \| \theta^\ast_n(S') - \theta^\ast_n(S) \|_2^2 \leq \frac{2G}{n} \| \theta^\ast_n(S') - \theta^\ast_n(S) \|_2, \]

which implies

\[ \| \theta^\ast_n(S') - \theta^\ast_n(S) \|_2 \leq \frac{4G}{\mu n}. \]  \hspace{1cm} (8)

Plugging inequalities (5) and (8) back into (3), we have:

\[ \| \hat{\theta}_n - \tilde{\theta}_n \|_2 \leq 2 \sqrt{2 \left( \frac{R_n(\hat{\theta}_n) - R_n(\hat{\theta}_n^\ast)}{\mu} \right) + \frac{4G}{\mu n}}, \]

which completes the proof. \( \square \)
A.2. The Optimization Error

As discussed before, the excess population risk can be decomposed into:
\[
R(\hat{\theta}_n) - R(\theta^*) = R(\hat{\theta}_n) - R_n(\hat{\theta}_n) + R_n(\hat{\theta}_n) - R_n(\theta^*) + R_n(\theta^*) - R(\theta^*)
\]
\[
\leq R(\hat{\theta}_n) - R_n(\hat{\theta}_n) + R_n(\hat{\theta}_n) - R_n(\theta^*_n) + R_n(\theta^*_n) - R(\theta^*).
\]

(9)

In Lemma 5, the stability is also related to the optimization error (excess empirical risk), so we first discuss the optimization error 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 5 (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 \( 1 - \zeta \),
\[
\|b\|_2 \leq \sigma \sqrt{p} \left( 1 + \frac{8 \log(1/\zeta)}{p} \right)^{1/4}.
\]

**Lemma 6.** If the Assumptions hold and the DP model is trained by T-iterations gradient perturbation method (7), then taking \( T = \mathcal{O} \left( \log(n) \right) \), \( \eta_1 = \cdots = \eta_T = \frac{1}{L} \), if \( \zeta \in (\exp(-p/8), 1) \), with probability at least \( 1 - \zeta \),
\[
R_n(\hat{\theta}_n) - R_n(\theta^*_n) \leq \frac{G^2 p \log(n) \log(1/\delta)}{n^2 \epsilon^2} \left( 1 + \frac{8 \log(T/\zeta)}{p} \right)^{1/4}.
\]

**Proof.** Note that we assume the loss function is \( L \)-smooth (Assumption 2 denoted by \( L \)) and satisfies the PL condition (Assumption 4 denoted by PL), at iteration \( t \), taking \( \eta_t = \frac{1}{L} \), we have:
\[
R_n(\hat{\theta}_{t+1}) - R_n(\theta^*_n) \leq \left( \frac{L}{\mu} \right) (R_n(\hat{\theta}_t) - R_n(\theta^*_n)) + \frac{1}{2L} \|b_{t+1}\|^2.
\]

Adding \( R_n(\hat{\theta}_t) - R_n(\theta^*_n) \) to both sides, we have:
\[
R_n(\hat{\theta}_{t+1}) - R_n(\theta^*_n) \leq \left( 1 - \frac{\mu}{L} \right) (R_n(\hat{\theta}_t) - R_n(\theta^*_n)) + \frac{1}{2L} \|b_{t+1}\|^2.
\]

Summing over \( T \) iterations, we have:
\[
R_n(\hat{\theta}_n) - R_n(\theta^*_n) \leq \left( 1 - \frac{\mu}{L} \right)^T (R_n(\hat{\theta}_0) - R_n(\theta^*_n)) + \frac{1}{2L} \sum_{t=0}^{T-1} \left( 1 - \frac{\mu}{L} \right)^t \|b_{t+1}\|^2.
\]

(11)

With Lemma 5 with probability at least \( 1 - \zeta \), we have:
\[
\|b_t\|^2 \leq \sigma^2 p \left( 1 + \frac{8 \log(T/\zeta)}{p} \right)^{1/4},
\]
for all \( t = 1, \cdots, T \).

Then, with the high probability upper bound of \( b_t \), inequality (11) comes to:
\[
R_n(\hat{\theta}_n) - R_n(\theta^*_n) \leq \left( 1 - \frac{\mu}{L} \right)^T (R_n(\hat{\theta}_0) - R_n(\theta^*_n)) + \frac{1}{2L} \sum_{t=0}^{T-1} \left( 1 - \frac{\mu}{L} \right)^t \sigma^2 p \left( 1 + \frac{8 \log(T/\zeta)}{p} \right)^{1/4}.
\]

(12)
where the second inequality holds because $0 \leq \ell(\cdot, \cdot) \leq M_\ell$, and $0 < \frac{\mu}{\ell} < 1$ because of the definitions of $\mu$ and $L$ (Wang, Ye, and Xu 2017).

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

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

The result follows.

**Lemma 7.** If the loss function is $\alpha$-Hölder smooth with parameter $H$, satisfies the PL inequality with parameter $2\mu$ and the DP model is trained by $T$-iterations gradient perturbation method (1), then taking $T = O\left(n^{1-2/\alpha}\right)$, $\eta_t = \frac{2}{\mu (t+\alpha)}$, where $\kappa \geq \frac{2H^{1/\alpha}}{\mu}$, if $\zeta \in (\exp(-p/8), 1)$, with probability at least $1 - \zeta$,

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

**Proof.** The proof is motivated by (Li and Liu 2021).

Like the proof of Lemma 6 by assuming that the loss function is $\alpha$-Hölder smooth (Assumption 3) denoted by $\alpha$, via Lemma 1 at iteration $t$,

$$R_n(\hat{\theta}_{t+1}) - R_n(\hat{\theta}_t) \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_{t+1} \rangle + \frac{H \eta_t^{\alpha + 1}}{\alpha + 1} \left\| \nabla_\theta R_n(\hat{\theta}_t) + b_{t+1} \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_{t+1} \rangle \right)

+ \frac{H \eta_t^{\alpha + 1}}{\alpha + 1} \left( 1 - \frac{\alpha + 1}{2} \left( \left\| \nabla_\theta R_n(\hat{\theta}_t) + b_{t+1} \right\|_2^{\alpha + 1} \right) \right)

\leq -\eta_t \left\| \nabla_\theta R_n(\hat{\theta}_t) \right\|_2^2 + (H \eta_t^{\alpha + 1} - \eta_t) \langle \nabla_\theta R_n(\hat{\theta}_t), b_{t+1} \rangle

+ \frac{H \eta_t^{\alpha + 1}}{2(\alpha + 1)} \left( \left\| \nabla_\theta R_n(\hat{\theta}_t) \right\|_2^2 + \left\| b_{t+1} \right\|_2^2 \right) ,$$

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 Cauhby-Schwarz inequality.

Noting that $\eta_t = \frac{2}{\mu (t+\alpha)}$ 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:

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

As a result,

$$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_t^{\alpha + 1}}{2} \left\| \nabla_\theta R_n(\hat{\theta}_t) \right\|_2^2

+ \eta_t \left\| \nabla_\theta R_n(\hat{\theta}_t) \right\|_2 \left\| b_{t+1} \right\|_2 + \frac{H \eta_t^{\alpha + 1}}{2} \left\| b_{t+1} \right\|_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 \left\| \nabla_\theta R_n(\hat{\theta}_t) \right\|_2 \left\| b_{t+1} \right\|_2 + \frac{\eta_t}{2} \left\| b_{t+1} \right\|_2^2

\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)}

+ \eta_t G^2 \left\| b_{t+1} \right\|_2 + \frac{\eta_t}{2} \left\| b_{t+1} \right\|_2^2 ,$$

(14)
where the second inequality is because of (13) and the last inequality holds because we assume that the loss function satisfies the PL condition with parameter $2\mu$, and $G' = \max\{2HM_{\zeta}, H\}$, 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 (14), we have

$$R_n(\hat{\theta}_{t+1}) - R_n(\theta_n^*) + \frac{\eta_t}{4} \left\| \nabla_{\theta} R_n(\hat{\theta}_t) \right\|_2^2 \leq \frac{t + \kappa - 1}{2\mu} \left\| \nabla_{\theta} R_n(\hat{\theta}_t) \right\|_2^2 + (t + \kappa - 1) H(1 - \alpha) \frac{2}{2(\alpha + 1)} \frac{2}{\mu} \left( \frac{2}{\mu} \right)^{1+\alpha}$$

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

$$R_n(\hat{\theta}_{t+1}) - R_n(\theta_n^*) + \frac{1}{2\mu(t + \kappa)} \left\| \nabla_{\theta} R_n(\hat{\theta}_t) \right\|_2^2 \leq \frac{t + \kappa - 1}{2\mu} \left\| \nabla_{\theta} R_n(\hat{\theta}_t) \right\|_2^2 + \frac{2}{\mu} \left( \frac{2}{\mu} \right)^{1+\alpha}$$

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

$$(t + \kappa)(t + \kappa - 1) \left( R_n(\hat{\theta}_{t+1}) - R_n(\theta_n^*) \right) + \frac{t + \kappa - 1}{2\mu} \left\| \nabla_{\theta} R_n(\hat{\theta}_t) \right\|_2^2 \leq (t + \kappa - 1)(t + \kappa - 2) \left( R_n(\hat{\theta}_t) - R_n(\theta_n^*) \right) + (t + \kappa - 1) H(1 - \alpha) \frac{2}{2(\alpha + 1)} \frac{2}{\mu} \left( \frac{2}{\mu} \right)^{1+\alpha}$$

With Lemma 5 for each $t = 1, \cdots, T$, with probability at least $1 - \xi$, we have:

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

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

$$(t + \kappa)(t + \kappa - 1) \left( R_n(\hat{\theta}_{t+1}) - R_n(\theta_n^*) \right) + \frac{t + \kappa - 1}{2\mu} \left\| \nabla_{\theta} R_n(\hat{\theta}_t) \right\|_2^2 \leq (t + \kappa - 1)(t + \kappa - 2) \left( R_n(\hat{\theta}_t) - R_n(\theta_n^*) \right) + (t + \kappa - 1) H(1 - \alpha) \frac{2}{2(\alpha + 1)} \frac{2}{\mu} \left( \frac{2}{\mu} \right)^{1+\alpha}$$

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

$$(T + \kappa)(T + \kappa - 1) \left( R_n(\hat{\theta}_{T+1}) - R_n(\theta_n^*) \right) + \sum_{t=1}^{T} \frac{t + \kappa - 1}{2\mu} \left\| \nabla_{\theta} R_n(\hat{\theta}_t) \right\|_2^2 \leq \kappa(T + \kappa - 1) \left( R_n(\hat{\theta}) - R_n(\theta_n^*) \right) + \sum_{t=1}^{T} (t + \kappa - 1) H(1 - \alpha) \frac{2}{2(\alpha + 1)} \frac{2}{\mu} \left( \frac{2}{\mu} \right)^{1+\alpha}$$

Here, for simplicity, we represent $t = 0, \cdots, T - 1$ by $t = 1, \cdots, T$. 
Then we bound term \( \sum_{t=1}^{T} (t + \kappa)^{-\alpha} (t + \kappa - 1)^{H(1-\alpha)/2(\alpha+1)} \left( \frac{2}{\mu} \right)^{1+\alpha} \).

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 (10), and note that \( 0 \leq \ell(\cdot, \cdot) \leq M_t \), we have:
\[
(T + \kappa)(T + \kappa - 1) \left( R_n(\hat{\theta}_{T+1}) - R_n(\theta^*_n) \right) \leq \kappa(\kappa - 1)M_t + \frac{(T + \kappa)^{2-\alpha} H(1-\alpha)}{2(\alpha+1)} \left( \frac{2}{\mu} \right)^{1+\alpha} + \left( T\kappa + \frac{T(T - 1)}{2} \right) \frac{2G'\sqrt{T}}{\sigma} \left( 1 + \left( \frac{8\log(T/\zeta)}{p} \right)^{1/4} \right) + \left( T\kappa + \frac{T(T - 1)}{2} \right) \frac{\sigma^2p}{\mu} \left( 1 + \left( \frac{8\log(T/\zeta)}{p} \right)^{1/4} \right)^2.
\]

As a result, taking \( \sigma \) 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 \( T = \mathcal{O}\left( n^{2/(2\alpha)} \right) \), with probability at least \( 1 - \zeta \), we have:
\[
R_n(\hat{\theta}_{T+1}) - R_n(\theta^*_n) \lesssim \frac{G^2\sqrt{p\log(1/\delta)}}{n^{1/(2\alpha)}} \epsilon \left( 1 + \left( \frac{8\log(T/\zeta)}{p} \right)^{1/4} \right).
\]

The result follows.

A.3. Proof of Theorem 1

Before the detailed proof, we first prove the following lemma 10. To get Lemma 10, we need the following lemmas given in (Bousquet, Klochkov, and Zhivotovskiy 2020).

**Lemma 8 ([Bousquet, Klochkov, and Zhivotovskiy 2020]).** Assume that \( z_1, \cdots, z_n \) are independent variables and the function \( g_i : \mathbb{Z}^n \rightarrow \mathbb{R} \) satisfy the following properties for \( i = 1, \cdots, n \):

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

Then the following inequality holds for all \( q \geq 2 \).
\[
\left\| \sum_{i=1}^{n} g_i \right\|_q \leq 12\sqrt{2}K\epsilon n \log(n) + 4K\sqrt{qn}.
\]

**Lemma 9 ([Bousquet, Klochkov, and Zhivotovskiy 2020]).** Under the uniform stability condition with parameter \( \gamma \) and uniformly bounded loss function \( \ell(\cdot, \cdot) \leq M_t \), we have for \( g_i = \mathbb{E}_z \ell(z, \theta^{(i)}_{t_n}) - \mathbb{E}_z \ell(z, \theta^{(i)}_{t_n}) \),
\[
\left| \mathbb{E} \left[ n (R_n(\theta_n) - R(\theta_n)) - \sum_{i=1}^{n} g_i \right] \right| \leq 2\gamma n.
\]

**Lemma 10.** 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 \( \tilde{\theta}_n = \tilde{\theta}(z_1, \cdots, z_n) \). Then for all \( q \geq 2 \),
\[
\left\| R_n(\hat{\theta}_n) - R(\hat{\theta}_n) - \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[\ell(z_i, \hat{\theta}^{(i)}_n)|z_i] + \mathbb{E}R(\hat{\theta}_n) \right\|_{q} \lesssim Gq\log(n) \left( \sqrt{\frac{R_n(\hat{\theta}_n) - R_n(\theta^*_n)}{\mu}} + \frac{G}{\mu} \right).
\]
Proof. Via Lemma 4

\[ \| \hat{\theta}_n - \hat{\theta}'_n \|_2 \leq 2\sqrt{2} \sqrt{\frac{R_n(\hat{\theta}_n) - R_n(\hat{\theta}'_n)}{\mu}} + \frac{4G}{\mu} \]

Recalling the definition of \( \gamma \)-uniformly stability: If for any \( z, z', z_1, \cdots, z_n \in Z \) and \( i = 1, \cdots, n \), it holds that

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

Then with \( G \)-Lipschitzness, we have:

\[ \| \ell(z, \hat{\theta}_n) - \ell(z, \hat{\theta}'_n) \| \leq 2\sqrt{2}G \sqrt{\frac{R_n(\hat{\theta}_n) - R_n(\hat{\theta}'_n)}{\mu}} + \frac{4G^2}{\mu}. \]

where \( \hat{\theta}_n \) and \( \hat{\theta}'_n \) are private models derived from any adjacent datasets. In the following, we use \( \theta^{(i)} \) to represent \( \hat{\theta}'_n \), which means that the single different data instance is the \( i^{th} \) one.

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

\[ E_z, g_i(z_1, \cdots, z_n) = 0 \]

almost surely, and

\[ h_i(z_1, \cdots, z_n) = g_i(z_1, \cdots, z_n) - E[g_i(z_1, \cdots, z_n)|z_i], \]

we have:

\[ E_z, h_i(z_1, \cdots, z_n) = 0 \]

Via the definition of \( h_i \), we observe that \( E[h_i|z_i] = 0 \) almost surely, which further implies \( K = 0 \) in Lemma 8 so we have for \( q \geq 2 \):

\[ \| \sum_{i=1}^n h_i \|_q \leq 192Gq n \log(n) \left( \frac{R_n(\hat{\theta}_n) - R_n(\hat{\theta}'_n)}{2\mu} + \frac{G}{\mu} \right). \]

Via Lemma 9 we have:

\[ \| n \left( R_n(\hat{\theta}_n) - R(\hat{\theta}_n) \right) - \sum_{i=1}^n g_i \| \leq 2n \left( 2\sqrt{2G} \sqrt{\frac{R_n(\hat{\theta}_n) - R_n(\hat{\theta}'_n)}{\mu}} + \frac{4G^2}{\mu} \right). \]

Noting that

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

we have:

\[ \| R_n(\hat{\theta}_n) - R(\hat{\theta}_n) - \frac{1}{n} \sum_{i=1}^n E \left[ \ell(z_i, \hat{\theta}_n^{(i)}|z_i) \right] + ER(\hat{\theta}_n') \| \leq Gq n \log(n) \left( \frac{R_n(\hat{\theta}_n) - R_n(\hat{\theta}'_n)}{\mu} + \frac{G}{\mu} \right). \]

The result follows. \( \square \)

To get Theorem 1, we further need the following lemma given in (Boucheron, Lugosi, and Massart 2013).

**Lemma 11** (Boucheron, Lugosi, and Massart 2013). If \( X_1, \cdots, X_n \) are zero mean, independent and bounded \( |X_i| \leq M \) almost surely, then for \( q \geq 2 \),

\[ \| X_1 + \cdots + X_n \|_q \leq 6 \left( \sum_{i=1}^n E[X_i^q] \right)^{1/q} + 4qM. \]
Then, we can start our proof.

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

\[
R(\hat{\theta}_n) - R(\theta^*) \leq c_1 \frac{G^2 p \log(n) \log(1/\delta)}{n^2 \epsilon^2} \left(1 + \left(\frac{8 \log(T/\zeta)}{p}\right)^{1/4}\right)^2
+ c_2 G \log^{1.5}(n) \sqrt{p \log(1/\delta)} \left(1 + \left(\frac{8 \log(T/\zeta)}{p}\right)^{1/4}\right) + c_3 \log(n) / n.
\]

for some constants \(c_1, c_2, c_3 > 0\).

**Proof.** Via Lemma 1, we have:

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

where \(||\rho||_q \lesssim G_q \log(n) \left(\sqrt{\frac{R_n(\hat{\theta}_n) - R_n(\theta^*_n)}{\mu}} + \frac{G}{\mu n}\right)\) for \(q \geq 2\) and \(\mathbb{E}'\) denotes the expectation taken over the independent copy.

Plugging this back to (9), we have:

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

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

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

Based on the definition of \(R(\theta)\), Assumption 5 is equivalent to:

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

(18)

With \(G\)-Lipschitz and PL inequality with parameter \(\mu\), we have \(B = 2G^2 / \mu\).

So, via (18),

\[
\mathbb{E} \left[ \left( \mathbb{E}' \ell(z_i, \hat{\theta}'_n) - \ell(z_i, \theta^*_n) \right)^2 \right] \leq \frac{2G^2}{\mu} \mathbb{E} \left( \mathbb{E}' \ell(z_i, \hat{\theta}'_n) - \ell(z_i, \theta^*_n) \right)
= \frac{2G^2}{\mu} \mathbb{E} \left[ \mathbb{E}[R(\hat{\theta}_n)] - R(\theta^*) \right],
\]

(19)

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

\[
\mathbb{E}' \ell(z_i, \hat{\theta}'_n) - \ell(z_i, \theta^*_n) = \mathbb{E}' \ell(z_i, \hat{\theta}'_n) - \mathbb{E}' \ell(z_i, \theta'^*_n) + \mathbb{E}' \ell(z_i, \theta'^*_n) - \ell(z_i, \theta^*_n).
\]

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:

\[
\left| \frac{1}{n} \sum_{i=1}^{n} X_i \right|_q = R_n(\hat{\theta}_n) - R_n(\theta^*_n).
\]

(20)

Via Lemma 11 since \(\mathbb{E}[R(\theta'_n)] - R(\theta^*)\) is exactly the expectation of each \(X''_i\), we have for \(q \geq 2\),

\[
\left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}' [\ell(z_i, \theta'^*_n)] - \ell(z_i, \theta^*_n) - \mathbb{E}[R(\theta'_n)] + R(\theta^*) \right|_q \leq \sqrt{\mathbb{E} \left[ \left( \mathbb{E}' \ell(z_i, \hat{\theta}'_n) - \ell(z_i, \theta^*_n) \right)^2 \right]} \frac{q}{n}
\leq \sqrt{\frac{2G^2}{\mu} \mathbb{E}[R(\theta'_n)] - R(\theta^*)} \frac{q}{n} + qM_{\ell} \frac{q}{n},
\]

(21)
where the last inequality holds because of (19) and \(E[R(\theta_n)] = E[R(\hat{\theta}_n)]\).

Plugging (20) and (21) back into (17), we obtain for each \(q \geq 2\) and some constant \(C > 0\),

\[
\begin{align*}
\left\| R(\hat{\theta}_n) - R(\theta^*) - \left( R_n(\hat{\theta}_n) - R_n(\theta^*_n) \right) \right\|_q \\
\leq C \left( Gq\log(n) \left( \frac{R_n(\hat{\theta}_n) - R_n(\theta^*_n)}{\mu} + \frac{G}{\mu n} \right) + \left( R_n(\hat{\theta}_n) - R_n(\theta^*_n) \right) \right) \\
\leq \varphi C \left( E[R(\theta_n)] - R(\theta^*) \right) + C \left( 2G\log(n) \sqrt{\frac{R_n(\hat{\theta}_n) - R_n(\theta^*_n)}{\mu}} + \left( R_n(\hat{\theta}_n) - R_n(\theta^*_n) \right) \right) \\
+ C \left( \frac{G^2 \log(n)}{\mu} + \frac{2G^2}{\mu \varphi} + M_\epsilon \right) \left( \frac{q}{n} \right),
\end{align*}
\]

(22)

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{align*}
E[R(\hat{\theta}_n)] - R(\theta^*) &

\leq \frac{1}{1 - \varphi C} 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{2CG\log(n)}{1 - \varphi C} \left( \frac{R_n(\hat{\theta}_n) - R_n(\theta^*_n)}{\mu} \right) \\
&

+ \frac{2C}{(1 - \varphi C)n} \left( \frac{G^2 \log(n)}{\mu} + \frac{2G^2}{\mu \varphi} + M_\epsilon \right).
\end{align*}
\]

Taking this back to (22), we have:

\[
R(\hat{\theta}_n) - R(\theta^*) \leq \frac{c_1}{n} \left( R_n(\hat{\theta}_n) - R_n(\theta^*_n) \right) + c_2 \log(n) \sqrt{R_n(\hat{\theta}_n) - R_n(\theta^*_n)} + c_3 \frac{\log(n)}{n},
\]

(23)

for some constants \(c_1, c_2\) and \(c_3\), where we combine \(R_n(\hat{\theta}_n) - R_n(\theta^*_n)\) and its expectation together.

Then via Lemma [3] with probability at least \(1 - \zeta\), we have:

\[
R(\hat{\theta}_n) - R(\theta^*) \leq c_1 \frac{G^2 p \log(n) \log(1/\delta)}{n^2 \epsilon^2} \left( 1 + \left( \frac{8 \log(T/\zeta)}{p} \right)^{1/4} \right)^2 \\
+ c_2 \frac{G \log^{1.5}(n) \sqrt{p \log(1/\delta)}}{n \epsilon} \left( 1 + \left( \frac{8 \log(T/\zeta)}{p} \right)^{1/4} \right) + c_3 \frac{\log(n)}{n},
\]

for some constants \(c_1, c_2, c_3 > 0\).

The result follows.

\(\Box\)

A.4. Proof of Theorem 5

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

$$R(\hat{\theta}_n) - R(\theta^*) \leq c_1 \frac{G'^2 \sqrt{p \log(1/\delta)}}{n \frac{\log(n)}{1/2}} \left(1 + \left(\frac{8 \log(T/\zeta)}{p}\right)^{1/4}\right)$$

$$+ c_2 \frac{G' \log(n) \sqrt{p \log(1/\delta)}}{n \frac{\log(n)}{1/2}} \left(1 + \left(\frac{8 \log(T/\zeta)}{p}\right)^{1/4}\right)^{1/2} + c_3 \frac{\log(n)}{n},$$

for some constants $c_1, c_2$ and $c_3$, where $G' = \max\{2HC, H\}$.

**Proof.** Like inequality (23) in the proof of Theorem 1 (Appendix A.3), we have:

$$R(\hat{\theta}_n) - R(\theta^*) \leq c_1 \left(R_n(\hat{\theta}_n) - R_n(\theta^*_n)\right) + c_2 \log(n) \sqrt{R_n(\hat{\theta}_n) - R_n(\theta^*_n)} + c_3 \frac{\log(n)}{n},$$

(24)

for some constants $c_1, c_2$ and $c_3$.

The differences between inequalities (23) and (24) are in constants, for example, Lipschitz constant $G$ discussed in Appendix A.3 comes to $G' = \max\{2HC, H\}$ here, as discussed before; and in Appendix A.3, the PL condition is with parameter $\mu$, rather than $2\mu$ here.

Combining the result obtained by Lemma 7, taking $T = O\left(n^{2/3}\right)$ and $\eta_b = \frac{2}{\mu(1+\kappa)}$ with $\kappa = \frac{2H^1/\mu}{\mu}$, then with probability at least $1 - \zeta$,

$$R(\hat{\theta}_n) - R(\theta^*) \leq c_1 \frac{G'^2 \sqrt{p \log(1/\delta)}}{n \frac{\log(n)}{1/2}} \left(1 + \left(\frac{8 \log(T/\zeta)}{p}\right)^{1/4}\right)$$

$$+ c_2 \frac{G' \log(n) \sqrt{p \log(1/\delta)}}{n \frac{\log(n)}{1/2}} \left(1 + \left(\frac{8 \log(T/\zeta)}{p}\right)^{1/4}\right)^{1/2} + c_3 \frac{\log(n)}{n},$$

for some constants $c_1, c_2$ and $c_3$, where $G' = \max\{2HC, H\}$.

The result holds.

---

**A.5. Proof of Theorem 3**

**Theorem 6.** If Assumptions 3 hold, the loss function and the parameter space are bounded, i.e. $0 \leq \ell(\cdot, \cdot) \leq M_\ell$, $\|C\|_2 \leq M_C$. Taking $\sigma$ given by Lemma 2, $T = O\left(\log(n)\right)$, and $\eta_1 = \cdots = \eta_T = \eta$, where $\eta = \left(\frac{1}{T}\right)^{1/\alpha}$, if $\zeta \in (\exp(-p/8), 1)$, then with probability at least $1 - \zeta$,

$$R(\hat{\theta}_n) - R(\theta^*) \leq c_1 \frac{G'^2 \log(n)p \log(1/\delta)}{n^{2\alpha^2}} \left(1 + \left(\frac{8 \log(T/\zeta)}{p}\right)^{1/4}\right)^{1/2}$$

$$+ c_2 \frac{G' \log(n) \sqrt{p \log(1/\delta)}}{n \frac{\log(n)}{1/2}} \left(1 + \left(\frac{8 \log(T/\zeta)}{p}\right)^{1/4}\right)^{1/2} + c_3 \frac{\log(n)}{n},$$

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

**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, if taking $\eta_t = H^{-1/\alpha}$ we have:

$$R_n(\hat{\theta}_{t+1}) - R_n(\hat{\theta}_t) \overset{(\alpha)}{=} \left\langle \nabla \theta R_n(\hat{\theta}_t), \hat{\theta}_{t+1} - \hat{\theta}_t \right\rangle + \frac{H}{2} \left\| \hat{\theta}_{t+1} - \hat{\theta}_t \right\|^{\alpha+1}_2$$

$$= -\eta_t \left\langle \nabla \theta R_n(\hat{\theta}_t), \nabla \theta R_n(\hat{\theta}_t) + b_t \right\rangle + \frac{H\eta_t^{\alpha+1}}{2} \left(\left\| \nabla \theta R_n(\hat{\theta}_t) + b_t \right\|^{\alpha+1}_2 \right)$$

$$\leq -\eta_t \left\| \nabla \theta R_n(\hat{\theta}_t) \right\|_2^2 + \frac{H\eta_t^{\alpha+1}}{2} \left(\left\| \nabla \theta R_n(\hat{\theta}_t) \right\|_2 + \|b_t\|_2 \right)^2 + (H\eta_t^{\alpha+1} - \eta_t) \left\langle \nabla \theta R_n(\hat{\theta}_t), b_t \right\rangle$$

$$\leq -\frac{\eta_t}{2} \left\| \nabla \theta R_n(\hat{\theta}_t) \right\|_2^2 + \frac{H\eta_t^{\alpha+1}}{2} \|b_t\|_2^2$$

$$\overset{(PL)}{\leq} -\mu \eta_t \left( R_n(\hat{\theta}_t) - R_n(\theta^*_n) \right) + \frac{\eta_t}{2} \|b_t\|_2^2, $$

where $\mu = \frac{1}{2}$. \hfill \Box

---
where the second inequality holds because by normalization, and the third inequality holds because \( \eta_t = \left( \frac{1}{p} \right)^{1/\alpha} \).

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

\[
R_n(\theta_{t+1}) - R_n(\theta_n^*) \leq (1 - \mu \eta_t) \left(R_n(\theta_t) - R_n(\theta_n^*)\right) + \frac{\eta_t}{2} \|b\|^2.
\]

Summing over \( T \) iterations,

\[
R_n(\hat{\theta}_T) - R_n(\theta_n^*) \leq (1 - \mu \eta)^T \left(R_n(\hat{\theta}_0) - R_n(\theta_n^*)\right) + \frac{T - 1}{2} \sum_{t=0}^{T-1} (1 - \mu \eta)^t \|b_t\|^2,
\]

where \( \eta = \left( \frac{1}{p} \right)^{1/\alpha} \).

With Lemma 5, with probability at least \( 1 - \xi \),

\[
R_n(\hat{\theta}_n) - R_n(\theta_n^*) \leq (1 - \mu \eta)^T M_T + \sigma^2 p \frac{2}{2 \mu} \left(1 + \left(\frac{8 \log(T/\zeta)}{p}\right)^{1/4}\right)^2.
\]

Noting that by definition, \( \mu \eta \leq 1 \), so if taking \( T = O(\log(n)) \), with probability at least \( 1 - \xi \), we have:

\[
R_n(\hat{\theta}_n) - R_n(\theta_n^*) \leq c G^2 \log(n) p \log(1/\delta) \left(1 + \left(\frac{8 \log(T/\zeta)}{p}\right)^{1/4}\right)^2.
\]

Then, like in (24), we have:

\[
R(\hat{\theta}_n) - R(\theta^*) \leq c_1 \left(R_n(\hat{\theta}_n) - R_n(\theta_n^*)\right) + c_2 \log(n) \sqrt{R_n(\hat{\theta}_n) - R_n(\theta_n^*)} + c_3 \frac{\log(n)}{n}
\]

\[
\leq c_1 \frac{G^2 \log(n) p \log(1/\delta)}{n^2 \epsilon^2} \left(1 + \left(\frac{8 \log(T/\zeta)}{p}\right)^{1/4}\right)^2
\]

\[
+ c_2 \frac{G^2 \log(1.5(n) \sqrt{p \log(1/\delta)}}{n \epsilon} \left(1 + \left(\frac{8 \log(T/\zeta)}{p}\right)^{1/4}\right) + c_3 \frac{\log(n)}{n},
\]

for some constants \( c_1, c_2, c_3 > 0 \).

The result follows.

---

**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.

**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 \( \epsilon \) 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 multiple 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.

**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 \( \epsilon \) 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.

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

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

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