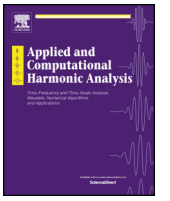
Appl. Comput. Harmon. Anal. 56 (2022) 306–336

Contents lists available at ScienceDirect

Applied and Computational Harmonic Analysis

www.elsevier.com/locate/acha



# Differentially private SGD with non-smooth losses

Puyu Wang [a], Yunwen Lei [b], Yiming Ying [c,*], Hai Zhang [a]

[a] *School of Mathematics, Northwest University, Xi'an, 710127, China*
[b] *School of Computer Science, University of Birmingham, Birmingham B15 2TT, UK*
[c] *Department of Mathematics and Statistics, State University of New York at Albany, Albany, NY, 12222, USA*

ARTICLE INFO

*Article history:*
Received 21 January 2021
Received in revised form 25 June 2021
Accepted 2 September 2021
Available online 15 September 2021
Communicated by Sergei Pereverzyev

*Keywords:*
Stochastic gradient descent
Algorithmic stability
Differential privacy
Generalization

ABSTRACT

In this paper, we are concerned with differentially private stochastic gradient descent (SGD) algorithms in the setting of stochastic convex optimization (SCO). Most of the existing work requires the loss to be Lipschitz continuous and strongly smooth, and the model parameter to be uniformly bounded. However, these assumptions are restrictive as many popular losses violate these conditions including the hinge loss for SVM, the absolute loss in robust regression, and even the least square loss in an unbounded domain. We significantly relax these restrictive assumptions and establish privacy and generalization (utility) guarantees for private SGD algorithms using output and gradient perturbations associated with non-smooth convex losses. Specifically, the loss function is relaxed to have an $\alpha$-Hölder continuous gradient (referred to as $\alpha$-*Hölder smoothness*) which instantiates the Lipschitz continuity ($\alpha = 0$) and the strong smoothness ($\alpha = 1$). We prove that noisy SGD with $\alpha$-Hölder smooth losses using gradient perturbation can guarantee $(\epsilon, \delta)$-differential privacy (DP) and attain optimal excess population risk $O\Big(\frac{\sqrt{d\log(1/\delta)}}{n\epsilon} + \frac{1}{\sqrt{n}}\Big)$, up to logarithmic terms, with the gradient complexity $O(n^{\frac{2-\alpha}{1+\alpha}} + n)$. This shows an important trade-off between $\alpha$-Hölder smoothness of the loss and the computational complexity for private SGD with statistically optimal performance. In particular, our results indicate that $\alpha$-Hölder smoothness with $\alpha \geq 1/2$ is sufficient to guarantee $(\epsilon, \delta)$-DP of noisy SGD algorithms while achieving optimal excess risk with a linear gradient complexity $O(n)$.

© 2021 Elsevier Inc. All rights reserved.

## 1. Introduction

Stochastic gradient descent (SGD) algorithms are widely employed to train a wide range of machine learning (ML) models such as SVM, logistic regression, and deep neural networks. It is an iterative algorithm which replaces the true gradient on the entire training data by a randomized gradient estimated from a random subset (mini-batch) of the available data. As opposed to gradient descent algorithms, this reduces

* Corresponding author.
*E-mail address:* yying@albany.edu (Y. Ying).

https://doi.org/10.1016/j.acha.2021.09.001
1063-5203/© 2021 Elsevier Inc. All rights reserved.

the computational burden at each iteration trading for a lower convergence rate [5]. There is a large amount of work considering the optimization error (convergence analysis) of SGD and its variants in the linear case [2,19,20,29,30] as well as the general setting of reproducing kernel Hilbert spaces [10,23,28,36,37,32].

At the same time, data collected often contain sensitive information such as individual records from schools and hospitals, financial records for fraud detection, online behavior from social media and genomic data from cancer diagnosis. Modern ML algorithms can explore the fine-grained information about data in order to make a perfect prediction which, however, can lead to privacy leakage [8,31]. To a large extent, SGD algorithms have become the workhorse behind the remarkable progress of ML and AI. Therefore, it is of pivotal importance for developing privacy-preserving SGD algorithms to protect the privacy of the data. Differential privacy (DP) [12,14] has emerged as a well-accepted mathematical definition of privacy which ensures that an attacker gets roughly the same information from the dataset regardless of whether an individual is present or not. Its related technologies have been adopted by Google [15], Apple [25], Microsoft [11] and the US Census Bureau [1].

In this paper, we are concerned with differentially private SGD algorithms in the setting of stochastic convex optimization (SCO). Specifically, let the input space $\mathcal{X}$ be a domain in some Euclidean space, the output space $\mathcal{Y} \subseteq \mathbb{R}$, and $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Denote the loss function by $\ell : \mathbb{R}^d \times \mathcal{Z} \mapsto [0, \infty)$ and assume, for any $z \in \mathcal{Z}$, that $\ell(\cdot, z)$ is convex with respect to (w.r.t.) the first argument. SCO aims to minimize the expected (population) risk, i.e. $\mathcal{R}(\mathbf{w}) := \mathbb{E}_z[\ell(\mathbf{w}, z)]$, where the model parameter $\mathbf{w}$ belongs to a (not necessarily bounded) domain $\mathcal{W} \subseteq \mathbb{R}^d$, and the expectation is taken w.r.t. $z$ according to a population distribution $\mathcal{D}$. While the population distribution is usually unknown, we have access to a finite set of $n$ training data points denoted by $S = \{z_i \in \mathcal{Z} : i = 1, 2, \ldots, n\}$. It is assumed to be independently and identically distributed (i.i.d.) according to the distribution $\mathcal{D}$ on $\mathcal{Z}$. In this context, one often considers SGD algorithms to solve the Empirical Risk Minimization (ERM) problem defined by

$$\min_{\mathbf{w} \in \mathcal{W}} \Big\{ \mathcal{R}_S(\mathbf{w}) := \frac{1}{n} \sum_{i=1}^{n} \ell(\mathbf{w}, z_i) \Big\}.$$

For a randomized algorithm (e.g., SGD) $\mathcal{A}$ to solve the above ERM problem, let $\mathcal{A}(S)$ be the output of algorithm $\mathcal{A}$ based on the dataset $S$. Then, its statistical generalization performance is measured by the excess (population) risk, i.e., the discrepancy between the expected risk $\mathcal{R}(\mathcal{A}(S))$ and the least possible one in $\mathcal{W}$, which is defined by

$$\epsilon_{\text{risk}}(\mathcal{A}(S)) = \mathcal{R}(\mathcal{A}(S)) - \min_{\mathbf{w} \in \mathcal{W}} \mathcal{R}(\mathbf{w}).$$

Along this line, there are a considerable amount of work [35,4,16] on analyzing the excess risk of private SGD algorithms in the setting of SCO. However, most of such approaches often require two assumptions: 1) the loss $\ell$ is $L$-Lipschitz and $\beta$-smooth; 2) the domain $\mathcal{W}$ is uniformly bounded. These assumptions are very restrictive as many popular losses violate these conditions including the hinge loss $(1 - y\mathbf{w}^T x)_+^q$ for $q$-norm soft margin SVM and the $q$-norm loss $|y - \mathbf{w}^T x|^q$ in regression with $1 \leq q \leq 2$. More specifically, the work [35] assumed the loss to be Lipschitz continuous and strongly smooth and showed that the private SGD algorithm with output perturbation can achieve $(\epsilon, \delta)$-DP and an excess risk rate $\mathcal{O}(\frac{(d \log(1/\delta))^{1/4}}{\sqrt{n\epsilon}})$ when the gradient complexity (i.e. the number of computing gradients) $T = n$. The study [4] proved, under the same assumptions, that the private SGD algorithm with gradient perturbation can achieve an optimal excess risk rate $\mathcal{O}\Big(\frac{\sqrt{d \log(1/\delta)}}{n\epsilon} + \frac{1}{\sqrt{n}}\Big)$ while guaranteeing its $(\epsilon, \delta)$-DP. To deal with the non-smoothness, it used the Moreau envelope technique to smooth the loss function and got the optimal rate. However, the algorithm is computationally inefficient with a gradient complexity $\mathcal{O}\Big(n^{4.5}\sqrt{\epsilon} + \frac{n^{6.5}\epsilon^{4.5}}{(d \log(1/\delta))^2}\Big)$. The work [16] improved the gradient complexity of the algorithm to $\mathcal{O}(n^2 \log(\frac{1}{\delta}))$ by localizing the approximate minimizer of the population loss on each phase. Recently, [3] showed that a simple variant of noisy projected SGD yields the

optimal rate with gradient complexity $O(n^2)$. However, it only focused on the Lipschitz continuous losses and assumed that the parameter domain $\mathcal{W}$ is bounded.

Our main contribution is to significantly relax these restrictive assumptions and to prove both privacy and generalization (utility) guarantees for private SGD algorithms with non-smooth convex losses in both bounded and unbounded domains. Specifically, the loss function $\ell(\mathbf{w}, z)$ is relaxed to have an $\alpha$-Hölder continuous gradient w.r.t. the first argument, i.e., there exists $L > 0$ such that, for any $\mathbf{w}, \mathbf{w}' \in \mathcal{W}$ and any $z \in \mathcal{Z}$,

$$\|\partial\ell(\mathbf{w}, z) - \partial\ell(\mathbf{w}', z)\|_2 \leq L\|\mathbf{w} - \mathbf{w}'\|_2^{\alpha},$$

where $\|\cdot\|_2$ denotes the Euclidean norm, $\partial\ell(\mathbf{w}, z)$ denotes a subgradient of $\ell$ w.r.t. the first argument. For the sake of notional simplicity, we refer to this condition as *$\alpha$-Hölder smoothness* with parameter $L$. The smoothness parameter $\alpha \in [0, 1]$ characterizes the smoothness of the loss function $\ell(\cdot, z)$. The case of $\alpha = 0$ corresponds to the Lipschitz continuity of the loss $\ell$ while $\alpha = 1$ means its strong smoothness. This definition instantiates many non-smooth loss functions mentioned above. For instance, the hinge loss for $q$-norm soft-margin SVM and $q$-norm loss for regression mentioned above with $q \in [1, 2]$ are $(q-1)$-Hölder smooth. In particular, we prove that noisy SGD with $\alpha$-Hölder smooth losses using gradient perturbation can guarantee $(\epsilon, \delta)$-DP and attain the optimal excess population risk $O\Big(\frac{\sqrt{d\log(1/\delta)}}{n\epsilon} + \frac{1}{\sqrt{n}}\Big)$, up to logarithmic terms, with gradient complexity $O(n^{\frac{2-\alpha}{1+\alpha}} + n)$. This shows an important trade-off between $\alpha$-Hölder smoothness of the loss and the computational complexity for private SGD in order to achieve statistically optimal performance. In particular, our results indicate that $\alpha$-Hölder smoothness with $\alpha \geq 1/2$ is sufficient to guarantee $(\epsilon, \delta)$-DP of noisy SGD algorithms while achieving the optimal excess risk with linear gradient complexity $O(n)$. Table 1 summarizes the upper bound of the excess population risk, gradient complexity of the aforementioned algorithms in comparison to our methods.

Our key idea to handle general Hölder smooth losses is to establish the approximate non-expansiveness of the gradient mapping, and the refined boundedness of the iterates of SGD algorithms when domain $\mathcal{W}$ is unbounded. This allows us to show the uniform argument stability [24] of the iterates of SGD algorithms with high probability w.r.t. the internal randomness of the algorithm (not w.r.t. the data $S$), and consequently estimate the generalization error of differentially private SGD with non-smooth losses.

**Organization of the Paper.** The rest of the paper is organized as follows. The formulation of SGD algorithms and the main results are given in Section 2. We provide the proofs in Section 3 and conclude the paper in Section 4.

## 2. Problem formulation and main results

### 2.1. Preliminaries

Throughout the paper, we assume that the loss function $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}$ is convex w.r.t. the first argument, i.e., for any $z \in \mathcal{Z}$ and $\mathbf{w}, \mathbf{w}' \in \mathcal{W}$, there holds $\ell(\mathbf{w}, z) \geq \ell(\mathbf{w}', z) + \langle \partial\ell(\mathbf{w}', z), \mathbf{w} - \mathbf{w}' \rangle$ where $\partial\ell(\mathbf{w}', z)$ denotes a subgradient of $\ell(\cdot, z)$ in the first argument. We restrict our attention to the (projected) stochastic gradient descent algorithm which is defined as below.

**Definition 1** *(Stochastic Gradient Descent).* Let $\mathcal{W} \subseteq \mathbb{R}^d$ be convex, $T$ denote the number of iterations, and $\mathrm{Proj}_{\mathcal{W}}$ denote the projection to $\mathcal{W}$. Let $\mathbf{w}_1 = \mathbf{0} \in \mathbb{R}^d$ be an initial point, and $\{\eta_t\}_{t=1}^T$ be a sequence of positive step sizes. At step $t \in \{1, \ldots, T\}$, the update rule of (projected) stochastic gradient decent is given by

$$\mathbf{w}_{t+1} = \mathrm{Proj}_{\mathcal{W}}\big(\mathbf{w}_t - \eta_t \partial\ell(\mathbf{w}_t, z_{i_t})\big), \tag{1}$$

**Table 1**
Comparison of different $(\epsilon, \delta)$-DP algorithms. We report the method, utility (generalization) bound, gradient complexity and parameter domain for three types of convex losses, i.e. Lipschitz, Lipschitz and smooth, and $\alpha$-Hölder smooth. Here *Output*, *Gradient*, *Phased Output* and *Phased ERM* denote output perturbation which adds Gaussian noise to the output of non-private SGD, gradient perturbation which adds Gaussian noise at each SGD update, phased output perturbation and phased ERM output perturbation [16], respectively. The gradient complexity is the total number of computing the gradient on one datum in the algorithm.

| Reference | Loss | Method | Utility bounds | Gradient Complexity | Domain |
|---|---|---|---|---|---|
| [35] | Lipschitz & smooth | *Output* | $\mathcal{O}\Big(\frac{(d\log(\frac{1}{\delta}))^{\frac{1}{4}}}{\sqrt{n\epsilon}}\Big)$ | $\mathcal{O}(n)$ | bounded |
| [4] | Lipschitz & smooth | *Gradient* | $\mathcal{O}\Big(\frac{\sqrt{d\log(\frac{1}{\delta})}}{n\epsilon}+\frac{1}{\sqrt{n}}\Big)$ | $\mathcal{O}\Big(n^{1.5}\sqrt{\epsilon}+\frac{(n\epsilon)^{2.5}}{d\log(\frac{1}{\delta})}\Big)$ | bounded |
| | Lipschitz | *Gradient* | $\mathcal{O}\Big(\frac{\sqrt{d\log(\frac{1}{\delta})}}{n\epsilon}+\frac{1}{\sqrt{n}}\Big)$ | $\mathcal{O}\Big(n^{4.5}\sqrt{\epsilon}+\frac{n^{6.5}\epsilon^{4.5}}{(d\log(\frac{1}{\delta}))^2}\Big)$ | bounded |
| [16] | Lipschitz & smooth | *Phased Output* | $\mathcal{O}\Big(\frac{\sqrt{d\log(\frac{1}{\delta})}}{n\epsilon}+\frac{1}{\sqrt{n}}\Big)$ | $\mathcal{O}(n)$ | bounded |
| | Lipschitz | *Phased ERM* | $\mathcal{O}\Big(\frac{\sqrt{d\log(\frac{1}{\delta})}}{n\epsilon}+\frac{1}{\sqrt{n}}\Big)$ | $\mathcal{O}(n^2\log(\frac{1}{\delta}))$ | bounded |
| [3] | Lipschitz | *Gradient* | $\mathcal{O}\Big(\frac{\sqrt{d\log(\frac{1}{\delta})}}{n\epsilon}+\frac{1}{\sqrt{n}}\Big)$ | $\mathcal{O}(n^2)$ | bounded |
| Ours | $\alpha$-Hölder smooth | *Output* | $\mathcal{O}\Big(\frac{(d\log(\frac{1}{\delta}))^{\frac{1}{4}}\sqrt{\log(\frac{n}{\delta})}}{\sqrt{n\epsilon}}\Big)$ | $\mathcal{O}(n^{\frac{2-\alpha}{1+\alpha}}+n)$ | bounded |
| | $\alpha$-Hölder smooth | *Output* | $\mathcal{O}\Big(\frac{\sqrt{d\log(\frac{1}{\delta})\log(\frac{n}{\delta})}}{n^{\frac{2}{3+\alpha}}\epsilon}+\frac{\log(\frac{n}{\delta})}{n^{\frac{1}{3+\alpha}}}\Big)$ | $\mathcal{O}(n^{\frac{-\alpha^2-3\alpha+6}{(1+\alpha)(3+\alpha)}}+n)$ | unbounded |
| | $\alpha$-Hölder smooth | *Gradient* | $\mathcal{O}\Big(\frac{\sqrt{d\log(\frac{1}{\delta})}}{n\epsilon}+\frac{1}{\sqrt{n}}\Big)$ | $\mathcal{O}(n^{\frac{2-\alpha}{1+\alpha}}+n)$ | bounded |

where $\{i_t\}$ is uniformly drawn from $[n] := \{1, 2, \ldots, n\}$. When $\mathcal{W} = \mathbb{R}^d$, then (1) is reduced to $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \partial\ell(\mathbf{w}_t, z_{i_t})$.

For a randomized learning algorithm $\mathcal{A} : \mathcal{Z}^n \to \mathcal{W}$, let $\mathcal{A}(S)$ denote the model produced by running $\mathcal{A}$ over the training dataset $S$. We say two datasets $S$ and $S'$ are *neighboring* datasets, denoted by $S \simeq S'$, if they differ by a single datum. We consider the following high-probabilistic version of the uniform argument stability (UAS), which is an extension of the UAS in expectation [24].

**Definition 2** *(Uniform argument stability).* We say an algorithm $\mathcal{A}$ has $\Delta_{\mathcal{A}}$-UAS with probability at least $1-\gamma$ $(\gamma \in (0,1))$ if

$$\mathbb{P}_{\mathcal{A}}\big(\sup_{S\simeq S'} \delta_{\mathcal{A}}(S, S') \geq \Delta_{\mathcal{A}}\big) \leq \gamma,$$

where $\delta_{\mathcal{A}}(S, S') := \|\mathcal{A}(S) - \mathcal{A}(S')\|_2$.

We will use UAS to study generalization bounds with high probability. In particular, the following lemma as a straightforward extension of Corollary 8 in [7] establishes the relationship between UAS and generalization errors. The proof is given in the Appendix for completeness.

**Lemma 1.** *Suppose $\ell$ is nonnegative, convex and $\alpha$-Hölder smooth with parameter $L$. Let $M_0 = \sup_{z\in\mathcal{Z}} \ell(0, z)$ and $M = \sup_{z\in\mathcal{Z}} \|\partial\ell(0, z)\|_2$. Let $\mathcal{A}$ be a randomized algorithm with the output of $\mathcal{A}$ bounded by $G$ and*

$$\mathbb{P}_{\mathcal{A}}\big(\sup_{S\simeq S'} \delta_{\mathcal{A}}(S, S') \geq \Delta_{\mathcal{A}}\big) \leq \gamma_0.$$

*Then there exists a constant $c > 0$ such that for any distribution $\mathcal{D}$ over $\mathcal{Z}$ and any $\gamma \in (0, 1)$, there holds*

$$\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^n,\mathcal{A}}\Bigg[|\mathcal{R}(\mathcal{A}(\mathcal{S})) - \mathcal{R}_{\mathcal{S}}(\mathcal{A}(\mathcal{S}))| \\ \geq c\bigg((M+LG^{\alpha})\Delta_{\mathcal{A}}\log(n)\log(1/\gamma) + \big(M_0 + (M+LG^{\alpha})G\big)\sqrt{n^{-1}\log(1/\gamma)}\bigg)\Bigg] \\ \leq \gamma_0 + \gamma.$$

Differential privacy [13] is a *de facto* standard privacy measure for a randomized algorithm $\mathcal{A}$.

**Definition 3** *(Differential Privacy).* We say a randomized algorithm $\mathcal{A}$ satisfies $(\epsilon,\delta)$-DP if, for any two neighboring datasets $S$ and $S'$ and any event $E$ in the output space of $\mathcal{A}$, there holds

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

In particular, we call it satisfies $\epsilon$-DP if $\delta = 0$.

We also need the following concept called $\ell_2$-sensitivity.

**Definition 4** *($\ell_2$-sensitivity).* The $\ell_2$-sensitivity of a function (mechanism) $\mathcal{M}: \mathcal{Z}^n \to \mathcal{W}$ is defined as $\Delta = \sup_{S\simeq S'}\|\mathcal{M}(S) - \mathcal{M}(S')\|_2$, where $S$ and $S'$ are neighboring datasets.

A basic mechanism to obtain $(\epsilon,\delta)$-DP from a given function $\mathcal{M}: \mathcal{Z}^n \to \mathcal{W}$ is to add a random noise from a Gaussian distribution $\mathcal{N}(0,\sigma^2\mathbf{I}_d)$ where $\sigma$ is proportional to its $\ell_2$-sensitivity. This mechanism is often referred to as Gaussian mechanism as stated in the following lemma.

**Lemma 2** *([14]). Given a function $\mathcal{M}: \mathcal{Z}^n \to \mathcal{W}$ with the $\ell_2$-sensitivity $\Delta$ and a dataset $S\subset\mathcal{Z}^n$, and assume that $\sigma \geq \frac{\sqrt{2\log(1.25/\delta)}\Delta}{\epsilon}$. The following Gaussian mechanism yields $(\epsilon,\delta)$-DP:*

$$\mathcal{G}(S,\sigma) := \mathcal{M}(S) + \mathbf{b}, \quad \mathbf{b}\sim\mathcal{N}(0,\sigma^2\mathbf{I}_d),$$

*where $\mathbf{I}_d$ is the identity matrix in $\mathbb{R}^{d\times d}$.*

Although the concept of $(\epsilon,\delta)$-DP is widely used in privacy-preserving methods, its composition and subsampling amplification results are relatively loose, which are not suitable for iterative SGD algorithms. Based on the Rényi divergence, the work [26] proposed Rényi differential privacy (RDP) as a relaxation of DP to achieve tighter analysis of composition and amplification mechanisms.

**Definition 5** *(RDP [26]).* For $\lambda > 1$, $\rho > 0$, a randomized mechanism $\mathcal{A}$ satisfies $(\lambda,\rho)$-RDP, if, for all neighboring datasets $S$ and $S'$, we have

$$D_{\lambda}\big(\mathcal{A}(S)\parallel\mathcal{A}(S')\big) := \frac{1}{\lambda-1}\log\int\Big(\frac{P_{\mathcal{A}(S)}(\theta)}{P_{\mathcal{A}(S')}(\theta)}\Big)^{\lambda}dP_{\mathcal{A}(S')}(\theta) \leq \rho,$$

where $P_{\mathcal{A}(S)}(\theta)$ and $P_{\mathcal{A}(S')}(\theta)$ are the density of $\mathcal{A}(S)$ and $\mathcal{A}(S')$, respectively.

As $\lambda\to\infty$, RDP reduces to $\epsilon$-DP, i.e., $\mathcal{A}$ satisfies $\epsilon$-DP if and only if $D_{\infty}\big(\mathcal{A}(S)||\mathcal{A}(S')\big) \leq \epsilon$ for any neighboring datasets $S$ and $S'$. Our analysis requires the introduction of several lemmas on useful properties of RDP listed below.

First, we introduce the privacy amplification of RDP by uniform subsampling, which is fundamental to establish privacy guarantees of noisy SGD algorithms. In general, a uniform subsampling scheme first draws a subset with size $pn$ uniformly at random with a subsampling rate $p\leq 1$, and then applies a known randomized mechanism to the subset.

**Lemma 3** *([22]). Consider a function $\mathcal{M}: \mathcal{Z}^n \to \mathcal{W}$ with the $\ell_2$-sensitivity $\Delta$, and a dataset $S \subset \mathcal{Z}^n$. The Gaussian mechanism $\mathcal{G}(S,\sigma) = \mathcal{M}(S) + \mathbf{b}$, where $\mathbf{b} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_d)$, applied to a subset of samples that are drawn uniformly without replacement with subsampling rate $p$ satisfies $(\lambda, 3.5p^2\lambda\Delta^2/\sigma^2)$-RDP if $\sigma^2 \geq 0.67\Delta^2$ and $\lambda - 1 \leq \frac{2\sigma^2}{3\Delta^2} \log\left(\frac{1}{\lambda p(1+\sigma^2/\Delta^2)}\right)$.*

The following adaptive composition theorem of RDP establishes the privacy of a composition of several adaptive mechanisms in terms of that of individual mechanisms. We say a sequence of mechanisms $(\mathcal{A}_1, \ldots, \mathcal{A}_k)$ are chosen adaptively if $\mathcal{A}_i$ can be chosen based on the outputs of the previous mechanisms $\mathcal{A}_1(S), \ldots, \mathcal{A}_{i-1}(S)$ for any $i \in [k]$.

**Lemma 4** *(Adaptive Composition of RDP [26]). If a mechanism $\mathcal{A}$ consists of a sequence of adaptive mechanisms $(\mathcal{A}_1, \ldots, \mathcal{A}_k)$ with $\mathcal{A}_i$ satisfying $(\lambda, \rho_i)$-RDP, $i \in [k]$, then $\mathcal{A}$ satisfies $(\lambda, \sum_{i=1}^k \rho_i)$-RDP.*

Lemma 4 tells us that the derivation of the privacy guarantee for a composition mechanism is simple and direct. This is the underlying reason that we adopt RDP in our subsequent privacy analysis. The following lemma allows us to further convert RDP back to $(\epsilon, \delta)$-DP.

**Lemma 5** *(From RDP to $(\epsilon,\delta)$-DP [26]). If a randomized mechanism $\mathcal{A}$ satisfies $(\lambda, \rho)$-RDP, then $\mathcal{A}$ satisfies $(\rho + \log(1/\delta)/(\lambda - 1), \delta)$-DP for all $\delta \in (0, 1)$.*

The following lemma shows that a post-processing procedure always preserves privacy.

**Lemma 6** *(Post-processing [26]). Let $\mathcal{A}: \mathcal{Z}^n \to \mathcal{W}_1$ satisfy $(\lambda, \rho)$-RDP and $f: \mathcal{W}_1 \to \mathcal{W}_2$ be an arbitrary function. Then $f \circ \mathcal{A}: \mathcal{Z}^n \to \mathcal{W}_2$ satisfies $(\lambda, \rho)$-RDP.*

### *2.2. Main results*

We present our main results here. First, we state a key bound of UAS for SGD when $\mathcal{W} \subseteq \mathbb{R}^d$ and the loss function is $\alpha$-Hölder smooth. Then, we propose two privacy-preserving SGD-type algorithms using output and gradient perturbations, and present the corresponding privacy and generalization (utility) guarantees. The utility guarantees in terms of the excess risk typically rely on two main errors: optimization errors and generalization errors, as shown soon in (3) and (4) for the algorithms with output and gradient perturbations, respectively. We will apply techniques in optimization theory to handle the optimization errors [27], and the concept of UAS [6,17,24], which was given in Definition 2 in Subsection 2.1, to estimate the generalization errors.

#### *2.2.1. UAS bound of SGD with non-smooth losses*

We begin by stating the key result on the distance between two iterate trajectories produced by SGD on neighboring datasets. Let

$$c_{\alpha,1} = \begin{cases} (1 + 1/\alpha)^{\frac{\alpha}{1+\alpha}} L^{\frac{1}{1+\alpha}}, & \text{if } \alpha \in (0, 1] \\ M + L, & \text{if } \alpha = 0, \end{cases} \tag{2}$$

and $c_{\alpha,2} = \sqrt{\frac{1-\alpha}{1+\alpha}}(2^{-\alpha}L)^{\frac{1}{1-\alpha}}$, where $M = \sup_{z\in Z} \|\partial\ell(0,z)\|_2$. In addition, define $C_\alpha = \frac{1-\alpha}{1+\alpha} c_{\alpha,1}^{\frac{2(1+\alpha)}{1-\alpha}} \left(\frac{\alpha}{1+\alpha}\right)^{\frac{2\alpha}{1-\alpha}} + 2\sup_{z\in Z} \ell(0;z)$. Furthermore, let $\mathcal{B}(0,r)$ denote the Euclidean ball of radius $r > 0$ centered at $0 \in \mathbb{R}^d$. Without loss of generality, we assume $\eta > 1/T$.

**Algorithm 1** Differentially Private SGD with Output perturbation (`DP-SGD-Output`)

**Inputs:** Data $S = \{z_i \in \mathcal{Z} : i = 1, \dots, n\}$, $\alpha$-Hölder smooth loss $\ell(\mathbf{w}, z)$ with parameter $L$, the convex set $\mathcal{W}$, step size $\eta$, number of iterations $T$, and privacy parameters $\epsilon$, $\delta$
**Set:** $\mathbf{w}_1 = \mathbf{0}$
**for** $t = 1$ to $T$ **do**
  Sample $i_t \sim \mathrm{Unif}([n])$
  $\mathbf{w}_{t+1} = \mathrm{Proj}_{\mathcal{W}}(\mathbf{w}_t - \eta \partial \ell(\mathbf{w}_t; z_{i_t}))$
**end for**
**if** $\mathcal{W} = \mathbb{R}^d$ **then**
  let $\Delta = \Delta_{\mathrm{SGD}}(\delta/2)$
**else if** $\mathcal{W} \subseteq \mathcal{B}(0, R)$ **then**
  let $\Delta = \tilde{\Delta}_{\mathrm{SGD}}(\delta/2)$
**end if**
**Compute:** $\sigma^2 = \frac{2\log(2.5/\delta)\Delta^2}{\epsilon^2}$
**return:** $\mathbf{w}_{\mathrm{priv}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{w}_t + \mathbf{b}$ where $\mathbf{b} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_d)$

**Theorem 7.** *Suppose that the loss function $\ell$ is convex and $\alpha$-Hölder smooth with parameter $L$. Let $\mathcal{A}$ be the SGD with $T$ iterations and $\eta_t = \eta < \min\{1, 1/L\}$, and $\bar{\mathbf{w}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{w}_t$ be the output produced by $\mathcal{A}$. Further, let $c_{\gamma,T} = \max\Big\{\big(3n\log(n/\gamma)/T\big)^{\frac{1}{2}}, 3n\log(n/\gamma)/T\Big\}$.*

*(a) If $\ell$ is nonnegative and $\mathcal{W} = \mathbb{R}^d$, then, for any $\gamma \in (0,1)$, there holds*

$$\mathbb{P}_{\mathcal{A}}\Big(\sup_{S \simeq S'} \delta_{\mathcal{A}}(S, S') \geq \Delta_{SGD}(\gamma)\Big) \leq \gamma,$$

*where* $\Delta_{SGD}(\gamma) = \Big(e\big(c_{\alpha,2}^2 T\eta^{\frac{2}{1-\alpha}} + 4\big(M + L(C_\alpha T\eta)^{\frac{\alpha}{2}}\big)^2\eta^2\Big(1 + \frac{T}{n}(1 + c_{\gamma,T})\Big)\frac{T}{n}(1 + c_{\gamma,T})\big)\Big)^{1/2}$.

*(b) If $\mathcal{W} \subseteq \mathcal{B}(0, R)$ with $R > 0$, then, for any $\gamma \in (0,1)$, there holds*

$$\mathbb{P}_{\mathcal{A}}\Big(\sup_{S \simeq S'} \delta_{\mathcal{A}}(S, S') \geq \tilde{\Delta}_{SGD}(\gamma)\Big) \leq \gamma,$$

*where* $\tilde{\Delta}_{SGD}(\gamma) = \Big(e\big(c_{\alpha,2}^2 T\eta^{\frac{2}{1-\alpha}} + 4\big(M + LR^\alpha\big)^2\eta^2\Big(1 + \frac{T}{n}(1 + c_{\gamma,T})\Big)\frac{T}{n}(1 + c_{\gamma,T})\big)\Big)^{1/2}$.

**Remark 1.** Under the reasonable assumption of $T \geq n$, we have $c_{\gamma,T} = \mathcal{O}(\log(n/\gamma))$. Then $\Delta_{\mathrm{SGD}}(\gamma) = \mathcal{O}\Big(\sqrt{T}\eta^{\frac{1}{1-\alpha}} + \frac{(T\eta)^{1+\alpha/2}\log(n/\gamma)}{n}\Big)$ and $\tilde{\Delta}_{\mathrm{SGD}}(\gamma) = \mathcal{O}\Big(\sqrt{T}\eta^{\frac{1}{1-\alpha}} + \frac{T\eta\log(n/\gamma)}{n}\Big)$. In addition, if $\ell$ is strongly smooth, i.e., $\alpha = 1$, the first term in the UAS bounds tends to 0 under the typical assumption of $\eta < 1$. In this case we have $\Delta_{\mathrm{SGD}}(\gamma) = \mathcal{O}\Big(\frac{(T\eta)^{3/2}\log(n/\gamma)}{n}\Big)$ and $\tilde{\Delta}_{\mathrm{SGD}}(\gamma) = \mathcal{O}\Big(\frac{T\eta\log(n/\gamma)}{n}\Big)$. The work [3] established the high probability upper bound of the random variable of the argument stability $\delta_{SGD}$ in the order of $\mathcal{O}(\sqrt{T}\eta + \frac{T\eta}{n})$ for Lipschitz continuous losses under an additional assumption $\gamma \geq \exp(-n/2)$. Our result gives the upper bound of $\sup_{S \simeq S'} \delta_{SGD}(S, S')$ in the order of $\mathcal{O}(\sqrt{T}\eta + \frac{T\eta\log(n/\gamma)}{n})$ for any $\gamma \in (0,1)$ for the case of $\alpha = 0$. The work [17] gave the bound of $\mathcal{O}(T\eta/n)$ in expectation for Lipschitz continuous and smooth loss functions. As a comparison, our stability bounds are stated with high probability and do not require the Lipschitz condition. Under a further Lipschitz condition, our stability bounds actually recover the bound $\mathcal{O}(T\eta/n)$ in [17] in the smooth case. Indeed, both the term $\big(M + (C_\alpha T\eta)^{\frac{\alpha}{2}}\big)^2$ and the term $\big(M + LR^\alpha\big)^2$ are due to controlling the magnitude of gradients, and can be replaced by $L^2$ for $L$-Lipschitz losses.

### *2.2.2. Differentially private SGD with output perturbation*

Output perturbation [9,13] is a common approach to achieve $(\epsilon, \delta)$-DP. The main idea is to add a random noise $\mathbf{b}$ to the output of the SGD algorithm, where $\mathbf{b}$ is randomly sampled from the Gaussian distribution with mean 0 and variance proportional to the $\ell_2$-sensitivity of SGD. In Algorithm 1, we propose the private SGD algorithm with output perturbation for non-smooth losses in both bounded domain $\mathcal{W} \subseteq \mathcal{B}(0, R)$ and unbounded domain $\mathcal{W} = \mathbb{R}^d$. The difference in these two cases is that we add random noise with different

variances according to the sensitivity analysis of SGD stated in Theorem 7. In the sequel, we present the privacy and utility guarantees for Algorithm 1.

**Theorem 8** *(Privacy guarantee). Suppose that the loss function $\ell$ is convex, nonnegative and $\alpha$-Hölder smooth with parameter $L$. Then Algorithm 1 (`DP-SGD-Output`) satisfies $(\epsilon, \delta)$-DP.*

According to the definitions, the $\ell_2$-sensitivity of SGD is identical to the UAS of SGD: $\sup_{S\simeq S'} \delta_{SGD}(S, S')$. In this sense, the proof of Theorem 8 directly follows from Theorem 7 and Lemma 2. For completeness, we include the detailed proof in Subsection 3.2.

Recall that the empirical risk is defined by $\mathcal{R}_S(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^n \ell(\mathbf{w}, z_i)$, and the population risk is $\mathcal{R}(\mathbf{w}) = \mathbb{E}_z[\ell(\mathbf{w}, z)]$. Let $\mathbf{w}^* \in \arg\min_{\mathbf{w}\in\mathcal{W}} \mathcal{R}(\mathbf{w})$ be the one with the best prediction performance over $\mathcal{W}$. We use the notation $B \asymp \tilde{B}$ if there exist constants $c_1, c_2 > 0$ such that $c_1\tilde{B} < B \leq c_2\tilde{B}$. Without loss of generality, we always assume $\|\mathbf{w}^*\|_2 \geq 1$.

**Theorem 9** *(Utility guarantee for unbounded domain). Suppose the loss function $\ell$ is nonnegative, convex and $\alpha$-Hölder smooth with parameter $L$. Let $\mathbf{w}_{priv}$ be the output produced by Algorithm 1 with $\mathcal{W} = \mathbb{R}^d$ and $\eta = n^{\frac{1}{3+\alpha}}/\big(T(\log(\frac{1}{\gamma}))^{\frac{1}{3+\alpha}}\big)$. Let $T \asymp n^{\frac{-\alpha^2-3\alpha+6}{(1+\alpha)(3+\alpha)}}$ if $0 \leq \alpha < \frac{\sqrt{73}-7}{4}$, and $T \asymp n$ else. Then, for any $\gamma \in (4\max\{\exp(-d/8), \delta\}, 1)$, with probability at least $1-\gamma$ over the randomness in both the sample and the algorithm, there holds*

$$\mathcal{R}(\mathbf{w}_{priv}) - \mathcal{R}(\mathbf{w}^*) = \|\mathbf{w}^*\|_2^2 \cdot \mathcal{O}\bigg(\frac{\sqrt{d\log(1/\delta)}\log(n/\delta)}{(\log(1/\gamma))^{\frac{1+\alpha}{4(3+\alpha)}} n^{\frac{2}{3+\alpha}}\epsilon} + \frac{\log(n)\big(\log(1/\gamma)\big)^{\frac{2}{3+\alpha}}\log(n/\delta)}{n^{\frac{1}{3+\alpha}}}\bigg).$$

To examine the excess population risk $\mathcal{R}(\mathbf{w}_{\rm priv}) - \mathcal{R}(\mathbf{w}^*)$, we use the following error decomposition:

$$\mathcal{R}(\mathbf{w}_{\rm priv}) - \mathcal{R}(\mathbf{w}^*) = [\mathcal{R}(\mathbf{w}_{\rm priv}) - \mathcal{R}(\bar{\mathbf{w}})] + [\mathcal{R}(\bar{\mathbf{w}}) - \mathcal{R}_S(\bar{\mathbf{w}})] + [\mathcal{R}_S(\bar{\mathbf{w}}) - \mathcal{R}_S(\mathbf{w}^*)] + [\mathcal{R}_S(\mathbf{w}^*) - \mathcal{R}(\mathbf{w}^*)], \quad (3)$$

where $\bar{\mathbf{w}} = \frac{1}{T}\sum_{t=1}^T \mathbf{w}_t$ is the output of non-private SGD. The first term is due to the added noise $\mathbf{b}$, which can be estimated by the Chernoff bound for Gaussian random vectors. The second term is the generalization error of SGD, which can be handled by the stability analysis. The third term is an optimization error and can be controlled by standard techniques in optimization theory. Finally, the last term can be bounded by $\mathcal{O}(1/\sqrt{n})$ by Hoeffding inequality. The proof of Theorem 9 is given in Subsection 3.2.

Now, we turn our attention to the utility guarantee for the case with a bounded domain.

**Theorem 10** *(Utility guarantees for bounded domain). If the loss function $\ell$ is nonnegative, convex and $\alpha$-Hölder smooth with parameter $L$. Let $\mathbf{w}_{priv}$ be the output produced by Algorithm 1 with $\mathcal{W} \subseteq \mathcal{B}(0, R)$. Let $T \asymp n^{\frac{2-\alpha}{1+\alpha}}$ if $\alpha < \frac{1}{2}$, $T \asymp n$ else, and choose $\eta = 1/\Big(T\max\Big\{\frac{\sqrt{\log(n/\delta)\log(n)\log(1/\gamma)}}{\sqrt{n}}, \frac{(d\log(1/\delta))^{1/4}\sqrt{\log(n/\delta)}(\log(1/\gamma))^{1/8}}{\sqrt{n\epsilon}}\Big\}\Big)$. Then for any $\gamma \in (4\max\{\exp(-d/8), \delta\}, 1)$, with probability at least $1-\gamma$ over the randomness in both the sample and the algorithm, there holds*

$$\mathcal{R}(\mathbf{w}_{priv}) - \mathcal{R}(\mathbf{w}^*) = \|\mathbf{w}^*\|_2^2 \cdot \mathcal{O}\bigg(\frac{\big(d\log(1/\delta)\big)^{\frac{1}{4}}(\log(1/\gamma))^{\frac{1}{8}}\sqrt{\log(n/\delta)}}{\sqrt{n\epsilon}} + \frac{\sqrt{\log(n)\log(1/\gamma)\log(n/\delta)}}{\sqrt{n}}\bigg).$$

The definition of $\alpha$-Hölder smoothness and the convexity of $\ell$ imply the following inequalities

$$\|\partial\ell(\mathbf{w}, z)\|_2 \leq M + LR^\alpha \text{ and } \ell(\mathbf{w}, z) \leq \ell(0, z) + MR + LR^{1+\alpha}, \quad \forall z \in \mathcal{Z}, \mathbf{w} \in \mathcal{W}.$$

**Algorithm 2** Differentially Private SGD with Gradient perturbation (`DP-SGD-Gradient`)

**Inputs:** Data $S = \{z_i \in \mathcal{Z} : i = 1, \dots, n\}$, loss function $\ell(\mathbf{w}, z)$ with Hölder parameters $\alpha$ and $L$, the convex set $\mathcal{W} \subseteq \mathcal{B}(0, R)$, step size $\eta$, number of iterations $T$, privacy parameters $\epsilon$, $\delta$, and constant $\beta$.
**Set:** $\mathbf{w}_1 = \mathbf{0}$
Compute $\sigma^2 = \frac{14(M+LR^\alpha)^2 T}{\beta n^2 \epsilon}\left(\frac{\log(1/\delta)}{(1-\beta)\epsilon} + 1\right)$
**for** $t = 1$ to $T$ **do**
  Sample $i_t \sim \mathrm{Unif}([n])$
  $\mathbf{w}_{t+1} = \mathrm{Proj}_{\mathcal{W}}(\mathbf{w}_t - \eta(\partial\ell(\mathbf{w}_t; z_{i_t}) + \mathbf{b}_t))$, where $\mathbf{b}_t \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_d)$
**end for**
**return:** $\mathbf{w}_{\mathrm{priv}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{w}_t$

These together with Theorem 9 imply the utility guarantee in the above theorem. The detailed proof is given in Subsection 3.2.

**Remark 2.** The private SGD algorithm with output perturbation was studied in [35] under both the Lipschitz continuity and the strong smoothness assumption, where the excess population risk for one-pass private SGD (i.e. the total iteration number $T = n$) with a bounded parameter domain was bounded by $\mathcal{O}\big((n\epsilon)^{-\frac{1}{2}}(d\log(1/\delta)^{\frac{1}{4}}\big)$. As a comparison, we show that the same rate (up to a logarithmic factor) $\mathcal{O}\big((n\epsilon)^{-\frac{1}{2}}(d\log(1/\delta))^{\frac{1}{4}}\log^{\frac{1}{2}}(n/\delta)\big)$ can be achieved for general $\alpha$-Hölder smooth losses by taking $T = \mathcal{O}(n^{\frac{2-\alpha}{1+\alpha}} + n)$. Our results extend the output perturbation for private SGD algorithms to a more general class of non-smooth losses.

*2.2.3. Differentially private SGD with gradient perturbation*

An alternative approach to achieve $(\epsilon, \delta)$-DP is gradient perturbation, i.e., adding Gaussian noise to the stochastic gradient at each update. The detailed algorithm is described in Algorithm 2, whose privacy guarantee is established in the following theorem.

**Theorem 11** *(Privacy guarantee). Suppose the loss function $\ell$ is nonnegative, convex and $\alpha$-Hölder smooth with parameter $L$. Then Algorithm 2 (`DP-SGD-Gradient`) satisfies $(\epsilon, \delta)$-DP if there exists $\beta \in (0, 1)$ such that $\frac{\sigma^2}{4(M+LR^\alpha)^2} \geq 0.67$ and $\lambda - 1 \leq \frac{\sigma^2}{6(M+LR^\alpha)^2}\log\left(\frac{n}{\lambda(1+\frac{\sigma^2}{4(M+LR^\alpha)^2})}\right)$ hold with $\lambda = \frac{\log(1/\delta)}{(1-\beta)\epsilon} + 1$.*

Since $\mathcal{W} \subseteq \mathcal{B}(0, R)$, the Hölder smoothness of $\ell$ implies that $\|\partial\ell(\mathbf{w}_t, z)\|_2 \leq M + LR^\alpha$ for any $t \in [T]$ and any $z \in \mathcal{Z}$, from which we know that the $\ell_2$-sensitivity of the function $\mathcal{M}_t = \partial\ell(\mathbf{w}_t, z)$ can be bounded by $2(M + LR^\alpha)$. By Lemma 3 and the post-processing property of DP, it is easy to show that the update of $\mathbf{w}_t$ satisfies $(\frac{\log(1/\delta)}{(1-\beta)\epsilon} + 1, \frac{\beta\epsilon}{T})$-RDP for any $t \in [T]$. Furthermore, by the composition theorem of RDP and the relationship between $(\epsilon, \delta)$-DP and RDP, we can show that the proposed algorithm satisfies $(\epsilon, \delta)$-DP. The detailed proof can be found in Subsection 3.3.

Other than the privacy guarantees, the `DP-SGD-Gradient` algorithm also enjoys utility guarantees as stated in the following theorem.

**Theorem 12** *(Utility guarantee). Suppose the loss function $\ell$ is nonnegative, convex and $\alpha$-Hölder smooth with parameter $L$. Let $\mathbf{w}_{priv}$ be the output produced by Algorithm 2 with $\eta = \frac{1}{T}\max\big\{\frac{\sqrt{\log(n)\log(n/\gamma)\log(1/\gamma)}}{\sqrt{n}}, \frac{\sqrt{d\log(1/\delta)}(\log(1/\gamma))^{\frac{1}{4}}}{n\epsilon}\big\}$. Furthermore, let $T \asymp n^{\frac{2-\alpha}{1+\alpha}}$ if $\alpha < \frac{1}{2}$, and $T \asymp n$ else. Then, for any $\gamma \in (18\exp(-Td/8), 1)$, with probability at least $1-\gamma$ over the randomness in both the sample and the algorithm, there holds*

$$\mathcal{R}(\mathbf{w}_{priv}) - \mathcal{R}(\mathbf{w}^*) = \|\mathbf{w}^*\|_2^2 \cdot \mathcal{O}\left(\frac{\sqrt{d\log(1/\delta)\log(1/\gamma)}}{n\epsilon} + \frac{\sqrt{\log(n)\log(n/\gamma)\log(1/\gamma)}}{\sqrt{n}}\right).$$

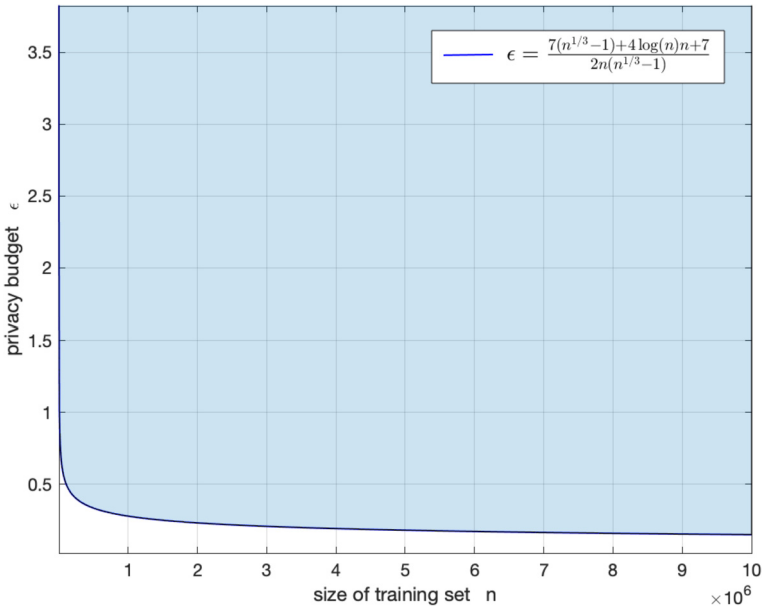



**Fig. 1.** The sufficient condition for the existence of $\beta$ in Lemma 13. The shaded area is the area where the sufficient condition in Lemma 13 holds true, i.e., $\epsilon \geq (7(n^{\frac{1}{3}}-1)+4\log(n)n+7)/(2n(n^{\frac{1}{3}}-1))$.

Our basic idea to prove Theorem 12 is to use the following error decomposition:

$$\mathcal{R}(\mathbf{w}_{\text{priv}}) - \mathcal{R}(\mathbf{w}^*) = [\mathcal{R}(\mathbf{w}_{\text{priv}}) - \mathcal{R}_S(\mathbf{w}_{\text{priv}})] + [\mathcal{R}_S(\mathbf{w}_{\text{priv}}) - \mathcal{R}_S(\mathbf{w}^*)] + [\mathcal{R}_S(\mathbf{w}^*) - \mathcal{R}(\mathbf{w}^*)]. \tag{4}$$

Similar to the proof of Theorem 9, the generalization error $\mathcal{R}(\mathbf{w}_{\text{priv}}) - \mathcal{R}_S(\mathbf{w}_{\text{priv}})$ can be handled by the UAS bound, the optimization error $\mathcal{R}_S(\mathbf{w}_{\text{priv}}) - \mathcal{R}_S(\mathbf{w}^*)$ can be estimated by standard techniques in optimization [e.g. 27], and the last term $\mathcal{R}_S(\mathbf{w}^*) - \mathcal{R}(\mathbf{w}^*)$ can be bounded by the Hoeffding inequality. The detailed proof can be found in Subsection 3.3.

**Remark 3.** We now compare our results with the related work under a bounded domain assumption. The work [4] established the optimal rate $\mathcal{O}(\frac{1}{n\epsilon}\sqrt{d\log(1/\delta)} + \frac{1}{\sqrt{n}})$ for the excess population risk of private SCO algorithm in either smooth case ($\alpha = 1$) or non-smooth case ($\alpha = 0$). However, their algorithm has a large gradient complexity $\mathcal{O}\Big(n^{4.5}\sqrt{\epsilon} + \frac{n^{6.5}\epsilon^{4.5}}{(d\log(\frac{1}{\delta}))^2}\Big)$. The work [16] proposed a private phased ERM algorithm for SCO, which can achieve the optimal excess population risk for non-smooth losses with a better gradient complexity of the order $\mathcal{O}(n^2\log(1/\delta))$. The very recent work [3] improved the gradient complexity to $\mathcal{O}(n^2)$. As a comparison, we show that SGD with gradient complexity $\mathcal{O}(n^{\frac{2-\alpha}{1+\alpha}}+n)$ is able to achieve the optimal (up to logarithmic terms) excess population risk $\mathcal{O}(\frac{1}{n\epsilon}\sqrt{d\log(1/\delta)}+\frac{1}{\sqrt{n}})$ for general $\alpha$-Hölder smooth losses. Our results match the existing gradient complexity for both the smooth case in [4] and the Lipschitz continuity case [3]. An interesting observation is that our algorithm can achieve the optimal utility guarantee with the linear gradient complexity $\mathcal{O}(n)$ for $\alpha \geq 1/2$, which shows that a relaxation of the strong smoothness from $\alpha = 1$ to $\alpha \geq 1/2$ does not bring any harm in both the generalization and computation complexity.

Now, we give a sufficient condition for the existence of $\beta$ in Theorem 11 under a specific parameter setting.

**Lemma 13.** *Let $n \geq 18$, $T = n$ and $\delta = 1/n^2$. If $\epsilon \geq \frac{7(n^{\frac{1}{3}}-1)+4\log(n)n+7}{2n(n^{\frac{1}{3}}-1)}$, then there exists $\beta \in (0,1)$ such that Algorithm 2 satisfies $(\epsilon,\delta)$-DP.*

**Remark 4.** Privacy parameters $\epsilon$ and $\delta$ together quantify the privacy risk. $\epsilon$ is often called the privacy budget controlling the degree of privacy leakage. A larger value of $\epsilon$ implies higher privacy risk. Therefore, the value of $\epsilon$ depends on how much privacy the user needs to protect. Theoretically, the value of $\epsilon$ is less than 1. However, in practice, to obtain the desired utility, a larger privacy budget, i.e., $\epsilon \geq 1$, is always acceptable [35,33]. For instance, Apple uses a privacy budget $\epsilon = 8$ for Safari Auto-play intent detection, and $\epsilon = 2$ for Health types[1]. Parameter $\delta$ is the probability with which $e^{\epsilon}$ fails to bound the ratio between the two probabilities in the definition of differential privacy, i.e., the probability of privacy protection failure. For meaningful privacy guarantees, according to [14] the value of $\delta$ should be much smaller than $1/n$. In particular, we always choose $\delta = 1/n^2$. For `DP-SGD-Gradient` algorithm, another constant we should discuss is $\beta$ which depends on the choice of the number of iterations $T$, size of training data $n$, privacy parameters $\epsilon$ and $\delta$. The appearance of this parameter is due to the use of subsampling result for RDP (see Lemma 3). The condition in Lemma 13 ensures the existence of $\beta \in (0,1)$ such that Algorithm 2 satisfies DP. Fig. 1 shows how the range of $\epsilon$ changes as we increase the size of training dataset $n$. In practical applications, we search in $(0,1)$ for all $\beta$ that satisfy the RDP conditions in Theorem 11. Note that the closer the $\beta$ is to $1/2$, the smaller the variance of the noise added to the algorithm in each iteration. Therefore, we choose the value that is closest to $1/2$ of all $\beta$ that meets the RDP conditions as the value of $\beta$.

We end this section with a final remark on the challenges of proving DP for Algorithm 2 when $\mathcal{W}$ is unbounded.

**Remark 5.** To make Algorithm 2 satisfy DP when $\mathcal{W} = \mathbb{R}^d$, the variance $\sigma_t$ of the noise $\mathbf{b}_t$ added in the $t$-th iteration should be proportional to the $\ell_2$-sensitivity $\Delta_t = \|\partial\ell(\mathbf{w}_t, z_{i_t}) - \partial\ell(\mathbf{w}_t, z'_{i_t})\|_2$. The definition of Hölder smoothness implies that $\Delta_t \leq 2(M + L\|\mathbf{w}_t\|_2^{\alpha})$. When $\alpha = 0$, we have $\Delta_t \leq 2(M + L)$ and the privacy guarantee can be established in a way similar to Theorem 11. When $\alpha \in (0,1]$, we have to establish an upper bound of $\|\mathbf{w}_t\|_2$. Since $\mathbf{w}_t = \mathbf{w}_{t-1} - \eta(\partial\ell(\mathbf{w}_{t-1}, z_{i_{t-1}}) + \mathbf{b}_{t-1})$ ($\mathbf{b}_{t-1} \sim \mathcal{N}(0, \sigma_{t-1}^2\mathbf{I}_d)$), we can only give a bound of $\|\mathbf{w}_t\|_2$ with high probability. Thus, the sensitivity $\Delta_t$ can not be uniformly bounded in this case. Therefore, the first challenge is how to analyze the privacy guarantee when the sensitivity changes at each iteration and all of them can not be uniformly bounded. Furthermore, by using the property of the Gaussian vector, we can prove that $\|\mathbf{w}_t\|_2 = \mathcal{O}(\sqrt{t\eta} + \eta\sum_{j=1}^{t-1}\sigma_j + \eta\sqrt{d\sum_{j=1}^{t-1}\sigma_j^2})$ with high probability. However, as mentioned above, the variance $\sigma_t$ should be proportional to $\Delta_t$ whose upper bound involves $\|\mathbf{w}_t\|_2^{\alpha}$. Thus, $\sigma_t$ is proportional to $(t\eta)^{\alpha/2} + \eta^{\alpha}(\sum_{j=1}^{t-1}\sigma_j)^{\alpha} + \eta^{\alpha}(d\sum_{j=1}^{t-1}\sigma_j^2)^{\alpha/2}$. For this reason, it seems difficult to give a clear expression for an upper bound of $\|\mathbf{w}_t\|_2$.

## 3. Proofs of main results

Before presenting the detailed proof, we first introduce some useful lemmas on the concentration behavior of random variables.

**Lemma 14** *(Chernoff bound for Bernoulli variable [34]). Let $X_1, \ldots, X_k$ be independent random variables taking values in $\{0,1\}$. Let $X = \sum_{i=1}^{k} X_i$ and $\mu = \mathbb{E}[X]$. The following statements hold.*

*(a) For any $\tilde{\gamma} \in (0,1)$, with probability at least $1 - \exp\big(-\mu\tilde{\gamma}^2/3\big)$, there holds $X \leq (1+\tilde{\gamma})\mu$.*
*(b) For any $\tilde{\gamma} \geq 1$, with probability at least $1 - \exp\big(-\mu\tilde{\gamma}/3\big)$, there holds $X \leq (1+\tilde{\gamma})\mu$.*

[1] https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf

**Lemma 15** *(Chernoff bound for the $\ell_2$-norm of Gaussian vector [34]). Let $X_1, \ldots, X_k$ be i.i.d. standard Gaussian random variables, and $\mathbf{X} = [X_1, \ldots, X_k] \in \mathbb{R}^k$. Then for any $t \in (0, 1)$, with probability at least $1 - \exp(-kt^2/8)$, there holds $\|\mathbf{X}\|_2^2 \leq k(1 + t)$.*

**Lemma 16** *(Hoeffding inequality [18]). Let $X_1, \ldots, X_k$ be independent random variables such that $a_i \leq X_i \leq b_i$ with probability 1 for all $i \in [k]$. Let $X = \frac{1}{k}\sum_{i=1}^k X_i$. Then for any $t > 0$, with probability at least $1 - \exp(-2t^2/\sum_i (b_i - a_i)^2)$, there holds $X - \mathbb{E}[X] \leq t$.*

**Lemma 17** *(Azuma-Hoeffding inequality [18]). Let $X_1, \ldots, X_k$ be a sequence of random variables where $X_i$ may depend on the previous random variables $X_1, \ldots, X_{i-1}$ for all $i = 1, \ldots, k$. Consider a sequence of functionals $\xi_i(X_1, \ldots, X_i)$, $i \in [k]$. If $|\xi_i - \mathbb{E}_{X_i}[\xi_i]| \leq b_i$ for each $i$. Then for all $t > 0$, with probability at least $1 - \exp(-t^2/(2\sum_i b_i^2))$, there holds $\sum_{i=1}^k \xi_i - \sum_{i=1}^k \mathbb{E}_{X_i}[\xi_i] \leq t$.*

**Lemma 18** *(Tail bound of sub-Gaussian variable [34]). Let $X$ be a sub-Gaussian random variable with mean $\mu$ and sub-Gaussian parameter $v^2$. Then, for any $t \geq 0$, we have, with probability at least $1 - \exp\left(-t^2/(2v^2)\right)$, that $X - \mu \leq t$.*

### *3.1. Proofs on UAS bound of SGD on non-smooth losses*

Our stability analysis for unbounded domain requires the following lemma on the self-bounding property for Hölder smooth losses.

**Lemma 19.** *([21,37]) Suppose the loss function $\ell$ is nonnegative, convex and $\alpha$-Hölder smooth with parameter $L$. Then for $c_{\alpha,1}$ defined as (2) we have*

$$\|\partial\ell(\mathbf{w}, z)\|_2 \leq c_{\alpha,1}\ell^{\frac{\alpha}{1+\alpha}}(\mathbf{w}, z), \quad \forall \mathbf{w} \in \mathbb{R}^d, z \in \mathcal{Z}.$$

Based on Lemma 19, we develop the following bound on the iterates produced by the SGD update (1) which is critical to analyze the privacy and utility guarantees in the case of unbounded domain. Recall that $M = \sup_{z\in\mathcal{Z}} \|\partial\ell(0, z)\|_2$.

**Lemma 20.** *Suppose the loss function $\ell$ is nonnegative, convex and $\alpha$-Hölder smooth with parameter $L$. Let $\{\mathbf{w}_t\}_{t=1}^T$ be the sequence produced by SGD with $T$ iterations when $\mathcal{W} = \mathbb{R}^d$ and $\eta_t < \min\{1, 1/L\}$. Then, for any $t \in [T]$, there holds*

$$\|\mathbf{w}_{t+1}\|_2^2 \leq C_\alpha \sum_{j=1}^t \eta_j,$$

*where $C_\alpha = \frac{1-\alpha}{1+\alpha} c_{\alpha,1}^{\frac{2(1+\alpha)}{1-\alpha}} \left(\frac{\alpha}{1+\alpha}\right)^{\frac{2\alpha}{1-\alpha}} + 2\sup_{z\in\mathcal{Z}} \ell(0; z)$.*

**Proof.** The update rule $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t\partial\ell(\mathbf{w}_t, z_{i_t})$ implies that

$$\|\mathbf{w}_{t+1}\|_2^2 = \|\mathbf{w}_t - \eta_t\partial\ell(\mathbf{w}_t, z_{i_t})\|_2^2 = \|\mathbf{w}_t\|_2^2 + \eta_t^2\|\partial\ell(\mathbf{w}_t, z_{i_t})\|_2^2 - 2\eta_t\langle\mathbf{w}_t, \partial\ell(\mathbf{w}_t, z_{i_t})\rangle. \tag{5}$$

First, we consider the case $\alpha = 0$. By the definition of Hölder smoothness, we know $\ell$ is $(M + L)$-Lipschitz continuous. Furthermore, by the convexity of $\ell$, we have

$$\begin{aligned}\eta_t\|\partial\ell(\mathbf{w}_t, z_{i_t})\|_2^2 - 2\langle\mathbf{w}_t, \partial\ell(\mathbf{w}_t, z_{i_t}) &\leq \eta_t\|\partial\ell(\mathbf{w}_t, z_{i_t})\|_2^2 + 2\big(\ell(0, z_{i_t}) - \ell(\mathbf{w}_t, z_{i_t})\big)\\ &\leq (M + L)^2 + 2\sup_{z\in\mathcal{Z}} \ell(0, z),\end{aligned}$$

where in the last inequality we have used $\eta_t < 1$ and the nonnegativity of $\ell$. Now, putting the above inequality back into (5) and taking the summation gives

$$\|\mathbf{w}_{t+1}\|_2^2 \leq \big((M+L)^2 + 2\sup_{z\in\mathcal{Z}} \ell(0,z)\big) \sum_{j=1}^{t} \eta_j. \tag{6}$$

Then, we consider the case $\alpha = 1$. In this case, Lemma 19 implies $\|\partial\ell(\mathbf{w},z)\|_2^2 \leq 2L\ell(\mathbf{w},z)$. Therefore,

$$\eta_t \|\partial\ell(\mathbf{w}_t, z_{i_t})\|_2^2 - 2\langle \mathbf{w}_t, \partial\ell(\mathbf{w}_t, z_{i_t})\rangle \leq 2\eta_t L\ell(\mathbf{w}_t, z_{i_t}) + 2\ell(0, z_{i_t}) - 2\ell(\mathbf{w}_t, z_{i_t}) \leq 2\ell(0, z_{i_t}),$$

where we have used the convexity of $\ell$ and $\eta_t < 1/L$. Plugging the above inequality back into (5) and taking the summation yield that

$$\|\mathbf{w}_{t+1}\|_2^2 \leq 2\sup_{z\in\mathcal{Z}} \ell(0,z) \sum_{j=1}^{t} \eta_j. \tag{7}$$

Finally, we consider the case $\alpha \in (0,1)$. According to the self-bounding property and the convexity, we know

$$\|\partial\ell(\mathbf{w}_t, z_{i_t})\|_2 \leq c_{\alpha,1} \ell^{\frac{\alpha}{1+\alpha}}(\mathbf{w}_t, z_{i_t}) \leq c_{\alpha,1} \big(\langle \mathbf{w}_t, \partial\ell(\mathbf{w}_t, z_{i_t})\rangle + \ell(0, z_{i_t})\big)^{\frac{\alpha}{1+\alpha}}.$$

Therefore, for $\alpha \in (0,1)$ there holds

$$\begin{aligned}
\|\partial\ell(\mathbf{w}_t, z_{i_t})\|_2^2 &\leq c_{\alpha,1}^2 \big(\langle \mathbf{w}_t, \partial\ell(\mathbf{w}_t, z_{i_t})\rangle + \ell(0, z_{i_t})\big)^{\frac{2\alpha}{1+\alpha}} \\
&= \Big(\frac{1+\alpha}{\alpha\eta_t}\big(\langle \mathbf{w}_t, \partial\ell(\mathbf{w}_t, z_{i_t})\rangle + \ell(0, z_{i_t})\big)\Big)^{\frac{2\alpha}{1+\alpha}} \cdot \Big(c_{\alpha,1}^2 \big(\frac{1+\alpha}{\alpha\eta_t}\big)^{-\frac{2\alpha}{1+\alpha}}\Big) \\
&\leq \frac{2\alpha}{1+\alpha}\Big(\frac{1+\alpha}{\alpha\eta_t}\big(\langle \mathbf{w}_t, \partial\ell(\mathbf{w}_t, z_{i_t})\rangle + \ell(0, z_{i_t})\big)\Big) + \frac{1-\alpha}{1+\alpha}\Big(c_{\alpha,1}^2 \big(\frac{1+\alpha}{\alpha\eta_t}\big)^{-\frac{2\alpha}{1+\alpha}}\Big)^{\frac{1+\alpha}{1-\alpha}} \\
&= 2\eta_t^{-1}\big(\langle \mathbf{w}_t, \partial\ell(\mathbf{w}_t, z_{i_t})\rangle + \ell(0, z_{i_t})\big) + \frac{1-\alpha}{1+\alpha} c_{\alpha,1}^{\frac{2(1+\alpha)}{1-\alpha}} \big(\frac{\alpha}{1+\alpha}\big)^{\frac{2\alpha}{1-\alpha}} \eta_t^{\frac{2\alpha}{1-\alpha}},
\end{aligned}$$

where the last inequality used Young's inequality $ab \leq \frac{1}{p}a^p + \frac{1}{q}b^q$ with $\frac{1}{p} + \frac{1}{q} = 1$. Putting the above inequality into (5), we have

$$\|\mathbf{w}_{t+1}\|_2^2 \leq \|\mathbf{w}_t\|_2^2 + \frac{1-\alpha}{1+\alpha} c_{\alpha,1}^{\frac{2(1+\alpha)}{1-\alpha}} \big(\frac{\alpha}{1+\alpha}\big)^{\frac{2\alpha}{1-\alpha}} \eta_t^{\frac{2}{1-\alpha}} + 2\ell(0, z_{i_t})\eta_t.$$

If the step size $\eta_t < 1$, then

$$\|\mathbf{w}_{t+1}\|_2^2 \leq \|\mathbf{w}_t\|_2^2 + \left(\frac{1-\alpha}{1+\alpha} c_{\alpha,1}^{\frac{2(1+\alpha)}{1-\alpha}} \big(\frac{\alpha}{1+\alpha}\big)^{\frac{2\alpha}{1-\alpha}} + 2\sup_{z\in\mathcal{Z}} \ell(0;z)\right)\eta_t.$$

Taking a summation of the above inequality, we get

$$\|\mathbf{w}_{t+1}\|_2^2 \leq \Big(\frac{1-\alpha}{1+\alpha} c_{\alpha,1}^{\frac{2(1+\alpha)}{1-\alpha}} \big(\frac{\alpha}{1+\alpha}\big)^{\frac{2\alpha}{1-\alpha}} + 2\sup_{z\in\mathcal{Z}} \ell(0;z)\Big) \sum_{j=1}^{t} \eta_j. \tag{8}$$

The desired result follows directly from (6), (7) and (8) for different values of $\alpha$. □

The following lemma shows the approximately non-expensive behavior of the gradient mapping $\mathbf{w} \mapsto \mathbf{w} - \eta\partial\ell(\mathbf{w}, z)$. The case $\alpha \in [0, 1)$ can be found in Lei and Ying [21], and the case $\alpha = 1$ can be found in Hardt [17].

**Lemma 21.** *Suppose the loss function $\ell$ is convex and $\alpha$-Hölder smooth with parameter $L$. Then for all $\mathbf{w}, \mathbf{w}' \in \mathbb{R}^d$ and $\eta \leq 2/L$ there holds*

$$\|\mathbf{w} - \eta\partial\ell(\mathbf{w}, z) - \mathbf{w}' + \eta\partial\ell(\mathbf{w}', z)\|_2^2 \leq \|\mathbf{w} - \mathbf{w}'\|_2^2 + \frac{1-\alpha}{1+\alpha}(2^{-\alpha}L)^{\frac{2}{1-\alpha}}\eta^{\frac{2}{1-\alpha}}.$$

With the above preparation, we are now ready to prove Theorem 7.

**Proof of Theorem 7.** (a) Assume that $S$ and $S'$ differ by the $i$-th datum, i.e., $z_i \neq z_i'$. Let $\{\mathbf{w}_t\}_{t=1}^T$ and $\{\mathbf{w}_t'\}_{t=1}^T$ be the sequence produced by SGD update (1) based on $S$ and $S'$, respectively. For simplicity, let $c_{\alpha,2}^2 = \frac{1-\alpha}{1+\alpha}(2^{-\alpha}L)^{\frac{2}{1-\alpha}}$. Note that when $\mathcal{W} = \mathbb{R}^d$, Eq. (1) reduces to $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta\partial\ell(\mathbf{w}_t, z_{i_t})$. For any $t \in [T]$, we consider the following two cases.
*Case 1:* If $i_t \neq i$, Lemma 21 implies that

$$\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}'\|_2^2 = \|\mathbf{w}_t - \eta_t\partial\ell(\mathbf{w}_t, z_{i_t}) - \mathbf{w}_t' + \eta_t\partial\ell(\mathbf{w}_t', z_{i_t})\|_2^2 \leq \|\mathbf{w}_t - \mathbf{w}_t'\|_2^2 + c_{\alpha,2}^2\eta_t^{\frac{2}{1-\alpha}}.$$

*Case 2:* If $i_t = i$, it follows from the elementary inequality $(a+b)^2 \leq (1+p)a^2 + (1+1/p)b^2$ that

$$\begin{aligned}\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}'\|_2^2 &= \|\mathbf{w}_t - \eta_t\partial\ell(\mathbf{w}_t, z_i) - \mathbf{w}_t' + \eta_t\partial\ell(\mathbf{w}_t', z_i')\|_2^2 \\ &\leq (1+p)\|\mathbf{w}_t - \mathbf{w}_t'\|_2^2 + (1+1/p)\eta_t^2\|\partial\ell(\mathbf{w}_t', z_i') - \partial\ell(\mathbf{w}_t, z_i)\|_2^2.\end{aligned}$$

According to the definition of Hölder smoothness and Lemma 20, we know

$$\|\partial\ell(\mathbf{w}_t, z)\|_2 \leq M + L\Big(C_\alpha\sum_{j=1}^{t-1}\eta_j\Big)^{\frac{\alpha}{2}} := c_{\alpha,t}. \tag{9}$$

Combining the above two cases and (9) together, we have

$$\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}'\|_2^2 \leq (1+p)^{\mathbb{I}_{[i_t=i]}}\|\mathbf{w}_t - \mathbf{w}_t'\|_2^2 + c_{\alpha,2}^2\eta_t^{\frac{2}{1-\alpha}} + 4(1+1/p)\mathbb{I}_{[i_t=i]}c_{\alpha,t}^2\eta_t^2,$$

where $\mathbb{I}_{[i_t=i]}$ is the indicator function, i.e., $\mathbb{I}_{[i_t=i]} = 1$ if $i_t = i$ and 0 otherwise. Applying the above inequality recursively, we get

$$\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}'\|_2^2 \leq \prod_{k=1}^t(1+p)^{\mathbb{I}_{[i_k=i]}}\|\mathbf{w}_1 - \mathbf{w}_1'\|_2^2 + \Big(c_{\alpha,2}^2\sum_{k=1}^t\eta_k^{\frac{2}{1-\alpha}} + 4\sum_{k=1}^t c_{\alpha,k}^2\eta_k^2(1+1/p)\mathbb{I}_{[i_k=i]}\Big)\prod_{j=k+1}^t(1+p)^{\mathbb{I}_{[i_j=i]}}.$$

Since $\mathbf{w}_1 = \mathbf{w}_1'$ and $\eta_t = \eta$, we further get

$$\begin{aligned}\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}'\|_2^2 &\leq \prod_{j=2}^t(1+p)^{\mathbb{I}_{[i_j=i]}}\Big(c_{\alpha,2}^2 t\eta^{\frac{2}{1-\alpha}} + 4\eta^2\sum_{k=1}^t c_{\alpha,k}^2(1+1/p)\mathbb{I}_{[i_k=i]}\Big) \\ &\leq (1+p)^{\sum_{j=2}^t\mathbb{I}_{[i_j=i]}}\Big(c_{\alpha,2}^2 t\eta^{\frac{2}{1-\alpha}} + 4c_{\alpha,t}^2\eta^2(1+1/p)\sum_{k=1}^t\mathbb{I}_{[i_k=i]}\Big).\end{aligned} \tag{10}$$

Applying Lemma 14 with $X_j = \mathbb{I}_{[i_j=i]}$ and $X = \sum_{j=1}^t X_j$, for any $\exp(-t/3n) \le \gamma \le 1$, with probability at least $1 - \frac{\gamma}{n}$, there holds

$$\sum_{j=1}^t \mathbb{I}_{[i_j=i]} \le \frac{t}{n}\Big(1 + \frac{\sqrt{3\log(n/\gamma)}}{\sqrt{t/n}}\Big).$$

For any $0 < \gamma < \exp(-t/3n)$, with probability at least $1 - \frac{\gamma}{n}$, there holds

$$\sum_{j=1}^t \mathbb{I}_{[i_j=i]} \le \frac{t}{n}\Big(1 + \frac{3\log(n/\gamma)}{t/n}\Big).$$

Plug the above two inequalities back into (10), and let $c_{\gamma,t} = \max\Big\{\sqrt{\frac{3\log(n/\gamma)}{t/n}}, \frac{3\log(n/\gamma)}{t/n}\Big\}$. Then, for any $\gamma \in (0,1)$, with probability at least $1 - \frac{\gamma}{n}$, we have

$$\|\mathbf{w}_{t+1} - \mathbf{w}'_{t+1}\|_2^2 \le (1+p)^{\frac{t}{n}(1+c_{\gamma,t})}\Big(c_{\alpha,2}^2 t\eta^{\frac{2}{1-\alpha}} + 4c_{\alpha,t}^2\eta^2(1+1/p)\frac{t}{n}(1+c_{\gamma,t})\Big).$$

Let $p = \frac{1}{\frac{t}{n}(1+c_{\gamma,t})}$. Then we know $(1+p)^{\frac{t}{n}(1+c_{\gamma,t})} \le e$ and therefore

$$\|\mathbf{w}_{t+1} - \mathbf{w}'_{t+1}\|_2^2 \le e\Big(c_{\alpha,2}^2 t\eta^{\frac{2}{1-\alpha}} + 4c_{\alpha,t}^2\eta^2\Big(1 + \frac{t}{n}(1+c_{\gamma,t})\Big)\frac{t}{n}(1+c_{\gamma,t})\Big). \tag{11}$$

This together with the inequality $c_{\alpha,t}^2 \le \big(M + L(C_\alpha t\eta)^{\frac{\alpha}{2}}\big)^2$ due to Lemma 20, we have, with probability at least $1 - \frac{\gamma}{n}$, that

$$\|\mathbf{w}_{t+1} - \mathbf{w}'_{t+1}\|_2^2 \le e\Big(c_{\alpha,2}^2 t\eta^{\frac{2}{1-\alpha}} + 4\big(M + L(C_\alpha t\eta)^{\frac{\alpha}{2}}\big)^2\eta^2\Big(1 + \frac{t}{n}(1+c_{\gamma,t})\Big)\frac{t}{n}(1+c_{\gamma,t})\Big).$$

By taking a union bound of probabilities over $i = 1, \dots, n$, with probability at least $1 - \gamma$, there holds

$$\sup_{S\simeq S'} \|\mathbf{w}_{t+1} - \mathbf{w}'_{t+1}\|_2^2 \le e\Big(c_{\alpha,2}^2 t\eta^{\frac{2}{1-\alpha}} + 4\big(M + L(C_\alpha t\eta)^{\frac{\alpha}{2}}\big)^2\eta^2\Big(1 + \frac{t}{n}(1+c_{\gamma,t})\Big)\frac{t}{n}(1+c_{\gamma,t})\Big).$$

Let $\Delta_{SGD}(\gamma) = \Big(e\big(c_{\alpha,2}^2 T\eta^{\frac{2}{1-\alpha}} + 4\big(M + L(C_\alpha T\eta)^{\frac{\alpha}{2}}\big)^2\eta^2\Big(1 + \frac{T}{n}(1+c_{\gamma,T})\Big)\frac{T}{n}(1+c_{\gamma,T})\big)\Big)^{1/2}$. Recall that $\mathcal{A}$ is the SGD with $T$ iterations, and $\bar{\mathbf{w}} = \frac{1}{T}\sum_{t=1}^T \mathbf{w}_t$ is the output produced by $\mathcal{A}$. Hence, $\sup_{S\simeq S'} \delta_{\mathcal{A}}(S,S') = \sup_{S\simeq S'} \|\bar{\mathbf{w}} - \bar{\mathbf{w}}'\|_2$. By the convexity of the $\ell_2$-norm, with probability at least $1-\gamma$, we have

$$\sup_{S\simeq S'} \delta_{\mathcal{A}}(S,S') \le \frac{1}{T}\sum_{t=1}^T \sup_{S\simeq S'} \|\mathbf{w}_t - \mathbf{w}'_t\|_2 \le \Delta_{SGD}(\gamma).$$

This completes the proof of part (a).

(b) For the case $\mathcal{W} \subseteq \mathcal{B}(0,R)$, the analysis is similar to the case $\mathcal{W} = \mathbb{R}^d$ except using a different estimate for the term $\|\partial\ell(\mathbf{w}_t, z)\|_2$. Indeed, in this case we have $\|\mathbf{w}_t\|_2 \le R$, which together with the Hölder smoothness, implies $\|\partial\ell(\mathbf{w}_t, z)\|_2 \le M + LR^\alpha$ for any $t \in [T]$ and $z \in \mathcal{Z}$. Now, replacing $c_{\alpha,t} = M + LR^\alpha$ in (9) and putting $c_{\alpha,t}$ back into (11), with probability at least $1 - \frac{\gamma}{n}$, we obtain

$$\sup_{S\simeq S'} \|\mathbf{w}_{t+1} - \mathbf{w}'_{t+1}\|_2^2 \le e\Big(c_{\alpha,2}^2 t\eta^{\frac{2}{1-\alpha}} + 4\big(M + LR^\alpha\big)^2\eta^2\Big(1 + \frac{t}{n}(1+c_{\gamma,t})\Big)\frac{t}{n}(1+c_{\gamma,t})\Big).$$

Now, let $\tilde{\Delta}_{SGD}(\gamma) = \Big(e\big(c_{\alpha,2}^2 T\eta^{\frac{2}{1-\alpha}} + 4\big(M + LR^\alpha\big)^2\eta^2\Big(1 + \frac{T}{n}(1+c_{\gamma,T})\Big)\frac{T}{n}(1+c_{\gamma,T})\big)\Big)^{1/2}$. The convexity of a norm implies, with probability at least $1-\gamma$, that

$$\sup_{S\simeq S'} \delta_{\mathcal{A}}(S,S') \le \frac{1}{T}\sum_{t=1}^{T}\sup_{S\simeq S'}\|\mathbf{w}_t - \mathbf{w}_t'\|_2 \le \tilde{\Delta}_{SGD}(\gamma).$$

The proof of the theorem is completed. □

### 3.2. Proofs on differentially private SGD with output perturbation

In this subsection, we prove the privacy and utility guarantees for output perturbation (i.e. Algorithm 1). We consider both the unbounded domain $\mathcal{W} = \mathbb{R}^d$ and bounded domain $\mathcal{W} \subseteq \mathcal{B}(0,R)$.

We first prove Theorem 8 on the privacy guarantee of Algorithm 1.

**Proof of Theorem 8.** Let $\mathcal{A}$ be the SGD with $T$ iterations, $\bar{\mathbf{w}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{w}_t$ be the output of $\mathcal{A}$. First, consider the unbounded domain case, i.e., $\mathcal{W} = \mathbb{R}^d$. Let $I = \{i_1,\dots,i_T\}$ be the sequence of sampling after $T$ iterations in $\mathcal{A}$. Define

$$\mathcal{B} = \big\{I : \sup_{S\simeq S'}\delta_{\mathcal{A}}(S,S') \le \Delta_{\mathrm{SGD}}(\delta/2)\big\}.$$

Part (a) in Theorem 7 implies that $\mathbb{P}(I\in\mathcal{B}) \ge 1-\delta/2$. Further, according to the definitions, we know the $\ell_2$-sensitivity of $\mathcal{A}$ is identical to the UAS of $\mathcal{A}$. Thus, if $I\in\mathcal{B}$, then Lemma 2 with $\delta' = \delta/2$ implies Algorithm 1 satisfies $(\epsilon,\delta/2)$-DP. For any neighboring datasets $S$ and $S'$, let $\mathbf{w}_{\mathrm{priv}}$ and $\mathbf{w}'_{\mathrm{priv}}$ be the output produced by Algorithm 1 based on $S$ and $S'$, respectively. Hence, for any $E\subseteq\mathbb{R}^d$ we have

$$\begin{aligned}\mathbb{P}(\mathbf{w}_{\mathrm{priv}}\in E) &= \mathbb{P}(\mathbf{w}_{\mathrm{priv}}\in E\cap I\in\mathcal{B}) + \mathbb{P}(\mathbf{w}_{\mathrm{priv}}\in E\cap I\in\mathcal{B}^c)\\ &\le \mathbb{P}(\mathbf{w}_{\mathrm{priv}}\in E|I\in\mathcal{B})\mathbb{P}(I\in\mathcal{B}) + \frac{\delta}{2} \le \Big(e^\epsilon\mathbb{P}(\mathbf{w}'_{\mathrm{priv}}\in E|I\in\mathcal{B}) + \frac{\delta}{2}\Big)\mathbb{P}(I\in\mathcal{B}) + \frac{\delta}{2}\\ &\le e^\epsilon\mathbb{P}(\mathbf{w}'_{\mathrm{priv}}\in E\cap I\in\mathcal{B}) + \delta \le e^\epsilon\mathbb{P}(\mathbf{w}'_{\mathrm{priv}}\in E) + \delta,\end{aligned}$$

where in the second inequality we have used the definition of DP. Therefore, Algorithm 1 satisfies $(\epsilon,\delta)$-DP when $\mathcal{W} = \mathbb{R}^d$. The bounded domain case can be proved in a similar way by using part (b) of Theorem 7. The proof is completed. □

Now, we turn to the utility guarantees of Algorithm 1. Recall that the excess population risk $\mathcal{R}(\mathbf{w}_{\mathrm{priv}}) - \mathcal{R}(\mathbf{w}^*)$ can be decomposed as follows ($\bar{\mathbf{w}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{w}_t$)

$$\mathcal{R}(\mathbf{w}_{\mathrm{priv}}) - \mathcal{R}(\mathbf{w}^*) = [\mathcal{R}(\mathbf{w}_{\mathrm{priv}}) - \mathcal{R}(\bar{\mathbf{w}})] + [\mathcal{R}(\bar{\mathbf{w}}) - \mathcal{R}_{\mathcal{S}}(\bar{\mathbf{w}})] + [\mathcal{R}_{\mathcal{S}}(\bar{\mathbf{w}}) - \mathcal{R}_{\mathcal{S}}(\mathbf{w}^*)] + [\mathcal{R}_{\mathcal{S}}(\mathbf{w}^*) - \mathcal{R}(\mathbf{w}^*)]. \tag{12}$$

We now introduce three lemmas to control the first three terms on the right hand side of (12). The following lemma controls the error resulting from the added noise.

**Lemma 22.** *Suppose the loss function $\ell$ is nonnegative, convex and $\alpha$-Hölder smooth with parameter $L$. Let $\mathbf{w}_{priv}$ be the output produced by Algorithm 1 based on the dataset $S = \{z_1,\cdots,z_n\}$ with $\eta_t = \eta < \min\{1, 1/L\}$. Then for any $\gamma\in(4\exp(-d/8),1)$, the following statements hold true.*

*(a) If $\mathcal{W} = \mathbb{R}^d$, then, with probability at least $1 - \frac{\gamma}{4}$, there holds*

$$\mathcal{R}(\mathbf{w}_{priv}) - \mathcal{R}(\bar{\mathbf{w}}) = \mathcal{O}\Big((T\eta)^{\frac{\alpha}{2}}\sigma\sqrt{d}(\log(1/\gamma))^{\frac{1}{4}} + \sigma^{1+\alpha}d^{\frac{1+\alpha}{2}}(\log(1/\gamma))^{\frac{1+\alpha}{4}}\Big).$$

*(b) If $\mathcal{W} \subseteq \mathcal{B}(0, R)$ with $R > 0$, then, with probability at least $1 - \frac{\gamma}{4}$, we have*

$$\mathcal{R}(\mathbf{w}_{priv}) - \mathcal{R}(\bar{\mathbf{w}}) = \mathcal{O}\Big(\sigma\sqrt{d}(\log(1/\gamma))^{\frac{1}{4}} + \sigma^{1+\alpha}d^{\frac{1+\alpha}{2}}(\log(1/\gamma))^{\frac{1+\alpha}{4}}\Big).$$

**Proof.** (a) First, we consider the case $\mathcal{W} = \mathbb{R}^d$. Note that

$$\begin{aligned}\mathcal{R}(\mathbf{w}_{\rm priv}) - \mathcal{R}(\bar{\mathbf{w}}) &= \mathbb{E}_z[\ell(\mathbf{w}_{\rm priv}, z) - \ell(\bar{\mathbf{w}}, z)] \le \mathbb{E}_z[\langle\partial\ell(\mathbf{w}_{\rm priv}, z), \mathbf{w}_{\rm priv} - \bar{\mathbf{w}}\rangle] \\ &\le \mathbb{E}_z[\|\partial\ell(\mathbf{w}_{\rm priv}, z)\|_2\|\mathbf{b}\|_2] \le (M + L\|\mathbf{w}_{\rm priv}\|_2^{\alpha})\|\mathbf{b}\|_2 \\ &\le (M + L\|\bar{\mathbf{w}}\|_2^{\alpha})\|\mathbf{b}\|_2 + L\|\mathbf{b}\|_2^{1+\alpha}, \end{aligned} \tag{13}$$

where the first inequality is due to the convexity of $\ell$, the second inequality follows from the Cauchy-Schwartz inequality, the third inequality is due to the definition of Hölder smoothness, and the last inequality uses $\mathbf{w}_{\rm priv} = \bar{\mathbf{w}} + \mathbf{b}$. Hence, to estimate $\mathcal{R}(\mathbf{w}_{\rm priv}) - \mathcal{R}(\bar{\mathbf{w}})$, it suffices to bound $\|\mathbf{b}\|_2$ and $\|\bar{\mathbf{w}}\|_2$. Since $\mathbf{b} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$, then for any $\gamma \in (4\exp(-d/8), 1)$, Lemma 15 implies, with probability at least $1 - \frac{\gamma}{4}$, that

$$\|\mathbf{b}\|_2 \le \sigma\sqrt{d}\Big(1 + \Big(\frac{8}{d}\log\big(4/\gamma\big)\Big)^{\frac{1}{4}}\Big). \tag{14}$$

Further, by the convexity of a norm and Lemma 20, we know

$$\|\bar{\mathbf{w}}\|_2 \le \frac{1}{T}\sum_{t=1}^{T}\|\mathbf{w}_t\|_2 \le \big(C_\alpha T\eta\big)^{\frac{1}{2}}. \tag{15}$$

Putting the above inequality and (14) back into (13) yields

$$\begin{aligned}\mathcal{R}(\mathbf{w}_{\rm priv}) - \mathcal{R}(\bar{\mathbf{w}}) &\le \big(M + L(C_\alpha T\eta)^{\frac{\alpha}{2}}\big)\sigma\sqrt{d}\Big(1 + \Big(\frac{8}{d}\log\big(4/\gamma\big)\Big)^{\frac{1}{4}}\Big) + L\sigma^{1+\alpha}d^{\frac{1+\alpha}{2}}\Big(1 + \Big(\frac{8}{d}\log\big(4/\gamma\big)\Big)^{\frac{1}{4}}\Big)^{1+\alpha} \\ &= \mathcal{O}\Big((T\eta)^{\frac{\alpha}{2}}\sigma\sqrt{d}\big(\log(1/\gamma)\big)^{\frac{1}{4}} + \sigma^{1+\alpha}d^{\frac{1+\alpha}{2}}\big(\log(1/\gamma)\big)^{\frac{1+\alpha}{4}}\Big).\end{aligned}$$

This completes the proof of part (a).

(b) The proof for the unbounded domain case is similar to that of the bounded domain. Since $\|\mathbf{w}_t\|_2 \le R$ for $t \in [T]$ in this case, then

$$\|\bar{\mathbf{w}}\|_2 \le \frac{1}{T}\sum_{t=1}^{T}\|\mathbf{w}_t\|_2 \le R. \tag{16}$$

Plugging (16) and (14) back into (13) yield the result in part (b). □

In the following lemma, we use the stability of SGD to control the generalization error $\mathcal{R}(\bar{\mathbf{w}}) - \mathcal{R}_S(\bar{\mathbf{w}})$.

**Lemma 23.** *Suppose the loss function $\ell$ is nonnegative, convex, and $\alpha$-Hölder smooth with parameter $L$. Let $\mathcal{A}$ be the SGD with $T$ iterations and $\eta_t = \eta < \min\{1, 1/L\}$ based on the dataset $S = \{z_1, \cdots, z_n\}$, and $\bar{\mathbf{w}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{w}_t$ be the output produced by $\mathcal{A}$. Then for any $\gamma \in (4\delta, 1)$, the following statements hold true.*

*(a) If $\mathcal{W} = \mathbb{R}^d$, then, with probability at least $1 - \frac{\gamma}{4}$, there holds*

$$\mathcal{R}(\bar{\mathbf{w}}) - \mathcal{R}_S(\bar{\mathbf{w}}) = \mathcal{O}\Big((T\eta)^{\frac{\alpha}{2}}\Delta_{SGD}(\delta/2)\log(n)\log(1/\gamma) + (T\eta)^{\frac{1+\alpha}{2}}\sqrt{n^{-\frac{1}{2}}\log(1/\gamma)}\Big).$$

*(b) If $\mathcal{W} \subseteq \mathcal{B}(0,R)$ with $R > 0$, then, with probability at least $1 - \frac{\gamma}{4}$, we have*

$$\mathcal{R}(\bar{\mathbf{w}}) - \mathcal{R}_S(\bar{\mathbf{w}}) = \mathcal{O}\Big(\tilde{\Delta}_{SGD}(\delta/2)\log(n)\log(1/\gamma) + \sqrt{n^{-\frac{1}{2}}\log(1/\gamma)}\Big).$$

**Proof.** (a) Consider the unbounded domain case. Part (a) in Theorem 7 implies, with probability at least $1 - \frac{\delta}{2}$, that

$$\sup_{S \simeq S'} \delta_{\mathcal{A}}(S, S') \leq \Delta_{\mathrm{SGD}}(\delta/2). \tag{17}$$

Since $\gamma \geq 4\delta$, then we know (17) holds with probability at least $1 - \frac{\gamma}{8}$. According to the result $\|\bar{\mathbf{w}}\|_2 \leq \sqrt{C_\alpha T\eta}$ by (15) and Lemma 1 with $G = \sqrt{C_\alpha T\eta}$ together, we derive the following inequality with probability at least $1 - \frac{\gamma}{8} - \frac{\gamma}{8} = 1 - \frac{\gamma}{4}$

$$\begin{aligned}
\mathcal{R}(\bar{\mathbf{w}}) - \mathcal{R}_S(\bar{\mathbf{w}}) &\leq c\Bigg((M + L(C_\alpha T\eta)^{\frac{\alpha}{2}})\Delta_{\mathrm{SGD}}(\delta/2)\log(n)\log(8/\gamma) \\
&\quad + \Big(\sup_{z\in Z}\ell(0,z) + (M + L(T\eta)^{\frac{\alpha}{2}})\sqrt{T\eta}\Big)\sqrt{\frac{\log(8/\gamma)}{n}}\Bigg) \\
&= \mathcal{O}\Bigg((T\eta)^{\frac{\alpha}{2}}\Delta_{\mathrm{SGD}}(\delta/2)\log(n)\log(1/\gamma) + (T\eta)^{\frac{1+\alpha}{2}}\sqrt{\frac{\log(1/\gamma)}{n}}\Bigg),
\end{aligned}$$

where $c > 0$ is a constant. The proof of part (a) is completed.

(b) For the case $\mathcal{W} \subseteq \mathcal{B}(0,R)$, the proof follows a similar argument as part (a). Indeed, part (b) in Theorem 7 implies, with probability at least $1 - \frac{\gamma}{8}$, that

$$\sup_{S \simeq S'} \delta_{\mathcal{A}}(S, S') \leq \tilde{\Delta}_{\mathrm{SGD}}(\delta/2). \tag{18}$$

Note that $\|\bar{\mathbf{w}}\|_2 \leq R$ in this case, then combining (18) and Lemma 1 with $G = R$ together, with probability at least $1 - \frac{\gamma}{4}$, we have

$$\begin{aligned}
\mathcal{R}(\bar{\mathbf{w}}) - \mathcal{R}_S(\bar{\mathbf{w}}) &\leq c\Bigg((M + LR^\alpha)\tilde{\Delta}_{\mathrm{SGD}}(\delta/2)\log(n)\log(8/\gamma) + \Big(\sup_{z\in Z}\ell(0,z) + (M + LR^\alpha)R\Big)\sqrt{\frac{\log(8/\gamma)}{n}}\Bigg) \\
&= \mathcal{O}\Bigg(\tilde{\Delta}_{\mathrm{SGD}}(\delta/2)\log(n)\log(1/\gamma) + \sqrt{\frac{\log(1/\gamma)}{n}}\Bigg),
\end{aligned}$$

where $c > 0$ is a constant. This completes the proof of part (b). □

In the following lemma, we use techniques in optimization theory to control the optimization error $\mathcal{R}_S(\bar{\mathbf{w}}) - \mathcal{R}_S(\mathbf{w}^*)$.

**Lemma 24.** *Suppose the loss function $\ell$ is nonnegative, convex and $\alpha$-Hölder smooth with parameter $L$. Let $\mathcal{A}$ be the SGD with $T$ iterations and $\eta_t = \eta < \min\{1, 1/L\}$ based on the dataset $S = \{z_1, \cdots, z_n\}$, and $\bar{\mathbf{w}} = \frac{1}{T}\sum_{t=1}^T \mathbf{w}_t$ be the output produced by $\mathcal{A}$. Then, for any $\gamma \in (0,1)$, the following statements hold true.*

*(a) If $\mathcal{W} = \mathbb{R}^d$, then, with probability at least $1 - \frac{\gamma}{4}$, there holds*

$$\mathcal{R}_S(\bar{\mathbf{w}}) - \mathcal{R}_S(\mathbf{w}^*) = \mathcal{O}\bigg(\eta^{\frac{1+\alpha}{2}} T^{\frac{\alpha}{2}} \sqrt{\log(1/\gamma)} + \|\mathbf{w}^*\|_2^{1+\alpha} \sqrt{\frac{\log(1/\gamma)}{T}} + \frac{\|\mathbf{w}^*\|_2^2}{\eta T} + \|\mathbf{w}^*\|_2^{1+\alpha} \eta\bigg).$$

*(b) If $\mathcal{W} \subseteq \mathcal{B}(0, R)$ with $R > 0$, then, with probability at least $1 - \frac{\gamma}{4}$, we have*

$$\mathcal{R}_S(\bar{\mathbf{w}}) - \mathcal{R}_S(\mathbf{w}^*) = \mathcal{O}\bigg(\|\mathbf{w}^*\|_2^{1+\alpha} \sqrt{\frac{\log(1/\gamma)}{T}} + \frac{\|\mathbf{w}^*\|_2^2}{\eta T} + \|\mathbf{w}^*\|_2^{1+\alpha} \eta\bigg).$$

**Proof.** (a) We first consider the case $\mathcal{W} = \mathbb{R}^d$. From the convexity of $\ell$, we have

$$\begin{aligned}
\mathcal{R}_S(\bar{\mathbf{w}}) - \mathcal{R}_S(\mathbf{w}^*) &\leq \frac{1}{T} \sum_{t=1}^{T} \mathcal{R}_S(\mathbf{w}_t) - \mathcal{R}_S(\mathbf{w}^*) \\
&= \frac{1}{T} \sum_{t=1}^{T} [\mathcal{R}_S(\mathbf{w}_t) - \ell(\mathbf{w}_t, z_{i_t})] + \frac{1}{T} \sum_{t=1}^{T} [\ell(\mathbf{w}^*, z_{i_t}) - \mathcal{R}_S(\mathbf{w}^*)] + \frac{1}{T} \sum_{t=1}^{T} [\ell(\mathbf{w}_t, z_{i_t}) - \ell(\mathbf{w}^*, z_{i_t})].
\end{aligned} \tag{19}$$

First, we consider the upper bound of $\frac{1}{T} \sum_{t=1}^{T} [\mathcal{R}_S(\mathbf{w}_t) - \ell(\mathbf{w}_t; z_{i_t})]$. Since $\{z_{i_t}\}$ is uniformly sampled from the dataset $S$, then for all $t = 1, \ldots, T$ we obtain

$$\mathbb{E}_{z_{i_t}} [\ell(\mathbf{w}_t, z_{i_t}) | \mathbf{w}_1, ..., \mathbf{w}_{t-1}] = \mathcal{R}_S(\mathbf{w}_t).$$

By the convexity of $\ell$, the definition of Hölder smoothness and Lemma 20, for any $z \in \mathcal{Z}$ and all $t \in [T]$, there holds

$$\begin{aligned}
\ell(\mathbf{w}_t, z) &\leq \sup_z \ell(0, z) + \langle \partial \ell(\mathbf{w}_t, z), \mathbf{w}_t \rangle \leq \sup_z \ell(0, z) + \|\partial \ell(\mathbf{w}_t, z)\|_2 \|\mathbf{w}_t\|_2 \\
&\leq \sup_z \ell(0, z) + (M + L\|\mathbf{w}_t\|_2^{\alpha}) \|\mathbf{w}_t\|_2 \leq \sup_z \ell(0, z) + M(C_\alpha T \eta)^{\frac{1}{2}} + L(C_\alpha T \eta)^{\frac{1+\alpha}{2}}.
\end{aligned} \tag{20}$$

Similarly, for any $z \in \mathcal{Z}$, we have

$$\ell(\mathbf{w}^*, z) \leq \sup_z \ell(0; z) + M\|\mathbf{w}^*\|_2 + L\|\mathbf{w}^*\|_2^{1+\alpha}. \tag{21}$$

Now, combining Lemma 17 with (20) and noting $\eta > 1/T$, we get the following inequality with probability at least $1 - \frac{\gamma}{8}$

$$\begin{aligned}
\frac{1}{T} \sum_{t=1}^{T} [\mathcal{R}_S(\mathbf{w}_t) - \ell(\mathbf{w}_t, z_{i_t})] &\leq \Big(\sup_z \ell(0, z) + M(C_\alpha T \eta)^{\frac{1}{2}} + L(C_\alpha T \eta)^{\frac{1+\alpha}{2}}\Big) \sqrt{\frac{2 \log(\frac{8}{\gamma})}{T}} \\
&= \mathcal{O}\Big(\eta^{\frac{1+\alpha}{2}} T^{\frac{\alpha}{2}} \sqrt{\log(1/\gamma)}\Big).
\end{aligned} \tag{22}$$

According to Lemma 16, with probability at least $1 - \frac{\gamma}{8}$, there holds

$$\frac{1}{T} \sum_{t=1}^{T} [\ell(\mathbf{w}^*; z_{i_t}) - \mathcal{R}_S(\mathbf{w}^*)] \leq \Big(\sup_z \ell(0, z) + M\|\mathbf{w}^*\|_2 + L\|\mathbf{w}^*\|_2^{1+\alpha}\Big) \sqrt{\frac{\log(8/\gamma)}{2T}} = \mathcal{O}\Big(\|\mathbf{w}^*\|_2^{1+\alpha} \sqrt{\frac{\log(1/\gamma)}{T}}\Big). \tag{23}$$

Finally, we consider the term $\frac{1}{T}\sum_{t=1}^{T}[\ell(\mathbf{w}_t, z_{i_t}) - \ell(\mathbf{w}^*, z_t)]$. The update rule implies $\mathbf{w}_{t+1} - \mathbf{w}^* = \big(\mathbf{w}_t - \mathbf{w}^*\big) - \eta\partial\ell(\mathbf{w}_t, z_{i_t})$, from which we know

$$\begin{aligned}\|\mathbf{w}_{t+1} - \mathbf{w}^*\|_2^2 &= \|\big(\mathbf{w}_t - \mathbf{w}^*\big) - \eta\partial\ell(\mathbf{w}_t, z_{i_t})\|_2^2 \\ &= \|\mathbf{w}_t - \mathbf{w}^*\|_2^2 + \eta^2\|\partial\ell(\mathbf{w}_t, z_{i_t})\|_2^2 - 2\eta\langle\partial\ell(\mathbf{w}_t, z_{i_t}), \mathbf{w}_t - \mathbf{w}^*\rangle.\end{aligned}$$

It then follows that

$$\langle\partial\ell(\mathbf{w}_t, z_{i_t}), \mathbf{w}_t - \mathbf{w}^*\rangle = \frac{1}{2\eta}\big(\|\mathbf{w}_t - \mathbf{w}^*\|_2^2 - \|\mathbf{w}_{t+1} - \mathbf{w}^*\|_2^2\big) + \frac{\eta}{2}\|\partial\ell(\mathbf{w}_t, z_{i_t})\|_2^2.$$

Combining the above inequality and the convexity of $\ell$ together, we derive

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}[\ell(\mathbf{w}_t, z_{i_t}) - \ell(\mathbf{w}^*, z_{i_t})] &\leq \frac{1}{T}\sum_{t=1}^{T}\Big[\frac{1}{2\eta}\big(\|\mathbf{w}_t - \mathbf{w}^*\|_2^2 - \|\mathbf{w}_{t+1} - \mathbf{w}^*\|_2^2\big) + \frac{\eta}{2}\|\partial\ell(\mathbf{w}_t, z_{i_t})\|_2^2\Big] \\ &\leq \frac{1}{2T\eta}\|\mathbf{w}_1 - \mathbf{w}^*\|_2^2 + \frac{\eta}{2T}\sum_{t=1}^{T}\|\partial\ell(\mathbf{w}_t, z_{i_t})\|_2^2. \end{aligned} \tag{24}$$

Since $0 \leq \frac{2\alpha}{1+\alpha} \leq 1$, Lemma 19 implies the following inequality for any $t = 1, \ldots, T$

$$\|\partial\ell(\mathbf{w}_t; z_{i_t})\|_2^2 \leq c_{\alpha,1}\ell^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_t; z_{i_t}) \leq c_{\alpha,1}\max\{\ell(\mathbf{w}_t; z_{i_t}), 1\} \leq c_{\alpha,1}\ell(\mathbf{w}_t; z_{i_t}) + c_{\alpha,1}.$$

Putting $\|\partial\ell(\mathbf{w}_t; z_{i_t})\|_2^2 \leq c_{\alpha,1}\ell(\mathbf{w}_t; z_{i_t}) + c_{\alpha,1}$ back into (24) and noting $\|\mathbf{w}_1\|_2 = 0$, we have

$$\frac{1}{T}\sum_{t=1}^{T}[\ell(\mathbf{w}_t, z_{i_t}) - \ell(\mathbf{w}^*, z_{i_t})] \leq \frac{\|\mathbf{w}^*\|_2^2}{2\eta T} + \frac{c_{\alpha,1}\eta}{2T}\sum_{t=1}^{T}\ell(\mathbf{w}_t, z_{i_t}) + \frac{c_{\alpha,1}\eta}{2}.$$

Rearranging the above inequality and using (21), we derive

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}[\ell(\mathbf{w}_t, z_{i_t}) - \ell(\mathbf{w}^*, z_{i_t})] &\leq \frac{1}{1 - \frac{c_{\alpha,1}\eta}{2}}\Big(\frac{\|\mathbf{w}^*\|_2^2}{2\eta T} + \frac{c_{\alpha,1}\eta}{2T}\sum_{t=1}^{T}\ell(\mathbf{w}^*, z_{i_t}) + \frac{c_{\alpha,1}\eta}{2}\Big) \\ &= \mathcal{O}\Big(\frac{\|\mathbf{w}^*\|_2^2}{\eta T} + \|\mathbf{w}^*\|_2^{1+\alpha}\eta\Big). \end{aligned} \tag{25}$$

Now, plugging (22), (23) and (25) back into (19), we derive

$$\mathcal{R}_{\mathcal{S}}(\bar{\mathbf{w}}) - \mathcal{R}_{\mathcal{S}}(\mathbf{w}^*) = \mathcal{O}\Big(\eta^{\frac{1+\alpha}{2}}T^{\frac{\alpha}{2}}\sqrt{\log(1/\gamma)} + \|\mathbf{w}^*\|_2^{1+\alpha}\sqrt{\frac{\log(1/\gamma)}{T}} + \frac{\|\mathbf{w}^*\|_2^2}{\eta T} + \|\mathbf{w}^*\|_2^{1+\alpha}\eta\Big)$$

with probability at least $1 - \frac{\gamma}{4}$, which completes the proof of part (a).

(b) Consider the bounded domain case. Since $\|\mathbf{w}_t\|_2 \leq R$ for any $t \in [T]$, then by the convexity of $\ell$ and the definition of Hölder smoothness, for any $z \in \mathcal{Z}$, there holds $\ell(\mathbf{w}_t, z) \leq \sup_z \ell(0, z) + (M + LR^\alpha)R$. Combining the above inequality and Lemma 17 together, with probability at least $1 - \frac{\gamma}{8}$, we obtain

$$\frac{1}{T}\sum_{t=1}^{T}[\mathcal{R}_{\mathcal{S}}(\mathbf{w}_t) - \ell(\mathbf{w}_t, z_{i_t})] \leq \big(\sup_z \ell(0, z) + (M + LR^\alpha)R\big)\sqrt{\frac{2\log(\frac{8}{\gamma})}{T}} = \mathcal{O}\Big(\sqrt{\frac{\log(1/\gamma)}{T}}\Big). \tag{26}$$

Since $\|\mathbf{w}_{t+1}-\mathbf{w}^*\|_2^2 = \|\mathrm{Proj}_{\mathcal{W}}\big(\mathbf{w}_t-\eta\partial\ell(\mathbf{w}_t,z_{i_t})\big)-\mathbf{w}^*\|_2^2 \le \|(\mathbf{w}_t-\mathbf{w}^*)-\eta\partial\ell(\mathbf{w}_t,z_{i_t})\|_2^2$, then (25) also holds true in this case. Putting (26), (23) and (25) back into (19), with probability at least $1-\frac{\gamma}{4}$, we have

$$\mathcal{R}_S(\bar{\mathbf{w}})-\mathcal{R}_S(\mathbf{w}^*)=\mathcal{O}\bigg(\|\mathbf{w}^*\|_2^{1+\alpha}\sqrt{\frac{\log(1/\gamma)}{T}}+\frac{\|\mathbf{w}^*\|_2^2}{\eta T}+\|\mathbf{w}^*\|_2^{1+\alpha}\eta\bigg).$$

The proof is completed. □

Now, we are in a position to prove the utility guarantee for `DP-SGD-Output` algorithm. First, we give the proof for the unbounded domain case (i.e. Theorem 9).

**Proof of Theorem 9.** Note that $\mathcal{R}_S(\mathbf{w}^*)-\mathcal{R}(\mathbf{w}^*)=\mathcal{R}_S(\mathbf{w}^*)-\mathbb{E}_S[\mathcal{R}_S(\mathbf{w}^*)]$. By Hoeffding inequality and (21), with probability at least $1-\frac{\gamma}{4}$, there holds

$$\mathcal{R}_S(\mathbf{w}^*)-\mathcal{R}(\mathbf{w}^*)\le\Big(\sup_{z\in\mathcal{Z}}\ell(0,z)+M\|\mathbf{w}^*\|_2+L\|\mathbf{w}^*\|_2^{1+\alpha}\Big)\sqrt{\frac{\log(4/\gamma)}{2n}}=\mathcal{O}\bigg(\|\mathbf{w}^*\|_2^{1+\alpha}\sqrt{\frac{\log(1/\gamma)}{n}}\bigg). \tag{27}$$

Combining part (a) in Lemmas 22, 23, 24 and (27) together, with probability at least $1-\gamma$, the population excess risk can be bounded as follows

$$\begin{aligned}
&\mathcal{R}(\mathbf{w}_{\mathrm{priv}})-\mathcal{R}(\mathbf{w}^*)\\
&=[\mathcal{R}(\mathbf{w}_{\mathrm{priv}})-\mathcal{R}(\bar{\mathbf{w}})]+[\mathcal{R}(\bar{\mathbf{w}})-\mathcal{R}_S(\bar{\mathbf{w}})]+[\mathcal{R}_S(\bar{\mathbf{w}})-\mathcal{R}_S(\mathbf{w}^*)]+[\mathcal{R}_S(\mathbf{w}^*)-\mathcal{R}(\mathbf{w}^*)]\\
&=\mathcal{O}\bigg((T\eta)^{\frac{\alpha}{2}}\sigma\sqrt{d}\big(\log(1/\gamma)\big)^{\frac{1}{4}}+\sigma^{1+\alpha}d^{\frac{1+\alpha}{2}}(\log(1/\gamma))^{\frac{1+\alpha}{4}}+(T\eta)^{\frac{\alpha}{2}}\Delta_{\mathrm{SGD}}(\delta/2)\log(n)\log(1/\gamma)\\
&\quad+\eta^{\frac{1+\alpha}{2}}\Big(T^{\frac{1+\alpha}{2}}\sqrt{\frac{\log(1/\gamma)}{n}}+T^{\frac{\alpha}{2}}\sqrt{\log(1/\gamma)}\Big)+\frac{\|\mathbf{w}^*\|_2^2}{\eta T}+\|\mathbf{w}^*\|_2^{1+\alpha}\eta+\|\mathbf{w}^*\|_2^{1+\alpha}\sqrt{\frac{\log(1/\gamma)}{n}}\bigg).
\end{aligned}\tag{28}$$

Plugging $\Delta_{\mathrm{SGD}}(\delta/2)=\mathcal{O}\Big(\sqrt{T}\eta^{\frac{1}{1-\alpha}}+\frac{(T\eta)^{1+\frac{\alpha}{2}}\log(n/\delta)}{n}\Big)$ and $\sigma=\mathcal{O}(\frac{\sqrt{\log(1/\delta)}\Delta_{\mathrm{SGD}}(\delta/2)}{\epsilon})$ back into (28), we have

$$\begin{aligned}
&\mathcal{R}(\mathbf{w}_{\mathrm{priv}})-\mathcal{R}(\mathbf{w}^*)\\
&=\mathcal{O}\bigg(T^{\frac{1+\alpha}{2}}\sqrt{\frac{\log(1/\gamma)}{n}}\eta^{\frac{1+\alpha}{2}}+\frac{T^{1+\alpha}\sqrt{d\log(1/\delta)}\big(\log(1/\gamma)\big)^{\frac{1}{4}}\log(n/\delta)}{n\epsilon}\eta^{1+\alpha}\\
&\quad+\frac{d^{\frac{1+\alpha}{2}}(\log(1/\gamma))^{\frac{1+\alpha}{4}}\big(T\log(\frac{1}{\delta})\big)^{\frac{1+\alpha}{2}}}{\epsilon^{1+\alpha}}\eta^{\frac{1+\alpha}{1-\alpha}}\\
&\quad+\frac{\big(d\log(1/\delta)\big)^{\frac{1+\alpha}{2}}(\log(1/\gamma))^{\frac{1+\alpha}{4}}T^{(1+\frac{\alpha}{2})(1+\alpha)}\big(\log(n/\delta)\big)^{1+\alpha}}{(n\epsilon)^{1+\alpha}}\eta^{(1+\frac{\alpha}{2})(1+\alpha)}\\
&\quad+\frac{T^{\frac{1+\alpha}{2}}\sqrt{d\log(\frac{1}{\delta})}\big(\log(1/\gamma)\big)^{\frac{1}{4}}}{\epsilon}\eta^{\frac{2+\alpha-\alpha^2}{2(1-\alpha)}}+T^{\frac{1+\alpha}{2}}\log(n)\log(1/\gamma)\eta^{\frac{2+\alpha-\alpha^2}{2(1-\alpha)}}\\
&\quad+\frac{T^{1+\alpha}\log(n/\delta)\log(n)\log(1/\gamma)}{n}\eta^{1+\alpha}+\frac{1}{\eta T}+\eta+\sqrt{\frac{\log(1/\gamma)}{n}}\bigg)\cdot\|\mathbf{w}^*\|_2^2.
\end{aligned}\tag{29}$$

Taking the derivative of $\frac{1}{T\eta}+T^{\frac{1+\alpha}{2}}\sqrt{\frac{\log(1/\gamma)}{n}}\eta^{\frac{1+\alpha}{2}}$ w.r.t $\eta$ and setting it to 0, then we have $\eta=n^{\frac{1}{3+\alpha}}/\big(T(\log(1/\gamma))^{\frac{1}{3+\alpha}}\big)$. Putting this $\eta$ back into (29), we obtain

$$
\begin{aligned}
&\mathcal{R}(\mathbf{w}_{\text{priv}}) - \mathcal{R}(\mathbf{w}^*) \\
&= \mathcal{O}\Bigg( \frac{n^{\frac{(2-\alpha)(1+\alpha)}{2(1-\alpha)(3+\alpha)}} \sqrt{d\log(1/\delta)}}{T^{\frac{1+\alpha}{2(1-\alpha)}} \epsilon \big(\log(1/\gamma)\big)^{\frac{1+4\alpha-\alpha^2}{4(1-\alpha)(3+\alpha)}}} + \frac{\sqrt{d\log(1/\delta)}\log(n/\delta)}{n^{\frac{2}{3+\alpha}} \epsilon \big(\log(1/\gamma)\big)^{\frac{1+\alpha}{4(3+\alpha)}}} + \Big( \frac{\sqrt{d\log(1/\delta)}\log(n/\delta)}{n^{\frac{4+\alpha}{2(3+\alpha)}} \epsilon \big(\log(1/\gamma)\big)^{\frac{1+\alpha}{4(3+\alpha)}}} \Big)^{1+\alpha} \\
&\quad + \Big( \frac{n^{\frac{1}{(1-\alpha)(3+\alpha)}} \sqrt{d\log(1/\delta)}}{T^{\frac{1+\alpha}{2(1-\alpha)}} \epsilon \big(\log(1/\gamma)\big)^{\frac{(1+\alpha)^2}{4(1-\alpha)(3+\alpha)}}} \Big)^{1+\alpha} \\
&\quad + \log(n)\log(n/\delta)\big(\log(1/\gamma)\big)^{\frac{2}{3+\alpha}} \Big( \frac{n^{\frac{2+\alpha-\alpha^2}{2(3+\alpha)(1-\alpha)}}}{T^{\frac{1+\alpha}{2(1-\alpha)}}} + \frac{1}{n^{\frac{2}{3+\alpha}}} + \frac{1}{n^{\frac{1}{3+\alpha}}} + \frac{n^{\frac{1}{3+\alpha}}}{T} \Big) \Bigg) \cdot \|\mathbf{w}^*\|_2^2. \qquad (30)
\end{aligned}
$$

To achieve the best rate with a minimal computational cost, we choose the smallest $T$ such that $\frac{n^{\frac{(2-\alpha)(1+\alpha)}{2(1-\alpha)(3+\alpha)}}}{T^{\frac{1+\alpha}{2(1-\alpha)}}} = \mathcal{O}(\frac{1}{n^{\frac{2}{3+\alpha}}})$, $\frac{n^{\frac{1+\alpha}{(1-\alpha)(3+\alpha)}}}{T^{\frac{(1+\alpha)^2}{2(1-\alpha)}}} = \mathcal{O}(\frac{1}{n^{\frac{(4+\alpha)(1+\alpha)}{2(3+\alpha)}}})$ and $\frac{n^{\frac{2+\alpha-\alpha^2}{2(3+\alpha)(1-\alpha)}}}{T^{\frac{1+\alpha}{2(1-\alpha)}}} + \frac{1}{n^{\frac{2}{3+\alpha}}} + \frac{n^{\frac{1}{3+\alpha}}}{T} = \mathcal{O}(\frac{1}{n^{\frac{1}{3+\alpha}}})$. Hence, we set $T \asymp n^{\frac{-\alpha^2-3\alpha+6}{(1+\alpha)(3+\alpha)}}$ if $0 \le \alpha \le \frac{\sqrt{73}-7}{4}$, and $T \asymp n$ else. Now, putting the choice of $T$ back into (30), we derive

$$
\begin{aligned}
\mathcal{R}(\mathbf{w}_{\text{priv}}) - \mathcal{R}(\mathbf{w}^*) =& \mathcal{O}\Bigg( \frac{\sqrt{d\log(1/\delta)}\log(n/\delta)}{(\log(1/\gamma))^{\frac{1+\alpha}{4(3+\alpha)}} n^{\frac{2}{3+\alpha}} \epsilon} + \Big( \frac{\sqrt{d\log(1/\delta)}\log(n/\delta)}{(\log(1/\gamma))^{\frac{1+\alpha}{4(3+\alpha)}} n^{\frac{4+\alpha}{2(3+\alpha)}} \epsilon} \Big)^{1+\alpha} \\
&+ \frac{\log(n)\big(\log(1/\gamma)\big)^{\frac{2}{3+\alpha}} \log(n/\delta)}{n^{\frac{1}{3+\alpha}}} \Bigg) \cdot \|\mathbf{w}^*\|_2^2.
\end{aligned}
$$

Without loss of generality, we assume the first term of the above utility bound is less than 1. Therefore, with probability at least $1-\gamma$, there holds

$$
\mathcal{R}(\mathbf{w}_{\text{priv}}) - \mathcal{R}(\mathbf{w}^*) = \|\mathbf{w}^*\|_2^2 \cdot \mathcal{O}\Bigg( \frac{\sqrt{d\log(1/\delta)}\log(n/\delta)}{(\log(1/\gamma))^{\frac{1+\alpha}{4(3+\alpha)}} n^{\frac{2}{3+\alpha}} \epsilon} + \frac{\log(n)\big(\log(1/\gamma)\big)^{\frac{2}{3+\alpha}} \log(n/\delta)}{n^{\frac{1}{3+\alpha}}} \Bigg).
$$

The proof is completed. □

Finally, we provide the proof of utility guarantee for the `DP-SGD-Output` algorithm when $\mathcal{W} \subseteq \mathcal{B}(0, R)$ (i.e. Theorem 10).

**Proof of Theorem 10.** The proof is similar to that of Theorem 9. Indeed, plugging part (b) in Lemmas 22, 23, 24 and (27) back into (12), with probability at least $1-\gamma$, the population excess risk can be bounded as follows

$$
\begin{aligned}
\mathcal{R}(\mathbf{w}_{\text{priv}}) - \mathcal{R}(\mathbf{w}^*) = \mathcal{O}\Big( & \sigma\sqrt{d}(\log(1/\gamma))^{\frac{1}{4}} + \sigma^{1+\alpha} d^{\frac{1+\alpha}{2}} (\log(1/\gamma))^{\frac{1+\alpha}{4}} + \tilde{\Delta}_{\text{SGD}}(\delta/2)\log(n)\log(1/\gamma) \\
& + \sqrt{\frac{\log(1/\gamma)}{n}} + \frac{\|\mathbf{w}^*\|_2^2}{T\eta} + \|\mathbf{w}^*\|_2^{1+\alpha}\eta + \|\mathbf{w}^*\|_2^{1+\alpha}\sqrt{\frac{\log(1/\gamma)}{n}} \Big).
\end{aligned}
$$

Note that $\tilde{\Delta}_{\text{SGD}}(\delta/2) = \mathcal{O}(\sqrt{T}\eta^{\frac{1}{1-\alpha}} + \frac{T\eta\log(n/\delta)}{n})$ and $\sigma = \frac{\sqrt{2\log(2.5/\delta)}\tilde{\Delta}_{\text{SGD}}(\delta/2)}{\epsilon}$. Then we have

$$
\begin{aligned}
\mathcal{R}(\mathbf{w}_{\text{priv}}) - \mathcal{R}(\mathbf{w}^*) = \mathcal{O}\Bigg( & \Big(\frac{T\log(n/\delta))\log(n)\log(1/\gamma)}{n} + \frac{T\sqrt{d\log(1/\delta)}\log(n/\delta)(\log(1/\gamma))^{\frac{1}{4}}}{n\epsilon}\Big)\eta \\
& + \Big(\frac{\sqrt{\log(1/\delta)Td}(\log(1/\gamma))^{\frac{1}{4}}}{\epsilon} + \sqrt{T}\log(n)\log(1/\gamma)\Big)\eta^{\frac{1}{1-\alpha}}
\end{aligned}
$$

$$+\frac{(Td\log(1/\delta))^{\frac{1+\alpha}{2}}(\log(1/\gamma))^{\frac{1+\alpha}{4}}}{\epsilon^{1+\alpha}}\eta^{\frac{1+\alpha}{1-\alpha}}$$
$$+\Big(\frac{T\sqrt{d\log(1/\delta)}\log(n/\delta)(\log(1/\gamma))^{\frac{1}{4}}}{n\epsilon}\Big)^{1+\alpha}\eta^{1+\alpha}+\frac{1}{T\eta}+\sqrt{\frac{\log(1/\gamma)}{n}}\Big)\cdot\|\mathbf{w}^*\|_2^2. \tag{31}$$

We consider the tradeoff between $1/\eta$ and $\eta$. Taking the derivative of $\frac{1}{T\eta}+\big(\frac{T\log(n/\delta)\log(n)\log(1/\gamma)}{n}+\frac{T\sqrt{d\log(1/\delta)}\log(n/\delta)(\log(1/\gamma))^{1/4}}{n\epsilon}\big)\eta$ w.r.t $\eta$ and setting it to 0, we have $\eta=1/\Big(T\max\Big\{\frac{\sqrt{\log(n/\delta)\log(n)\log(1/\gamma)}}{\sqrt{n}},\frac{(d\log(1/\delta))^{1/4}\sqrt{\log(n/\delta)}(\log(1/\gamma))^{1/8}}{\sqrt{n\epsilon}}\Big\}\Big)$. Then putting the value of $\eta$ back into (31), we obtain

$$\begin{aligned}\mathcal{R}(\mathbf{w}_{\mathrm{priv}})-\mathcal{R}(\mathbf{w}^*)=&\mathcal{O}\Big(\frac{\big(d\log(1/\delta)\big)^{\frac{1}{4}}(\log(1/\gamma))^{\frac{1}{8}}\sqrt{\log(n/\delta)}}{\sqrt{n\epsilon}}+\Big(\frac{\big(d\log(1/\delta)\big)^{\frac{1}{4}}(\log(1/\gamma))^{\frac{1}{8}}\sqrt{\log(n/\delta)}}{\sqrt{n\epsilon}}\Big)^{1+\alpha}\\&+\frac{\big(d\log(1/\delta)\big)^{\frac{1-2\alpha}{4(1-\alpha)}}(\log(1/\gamma))^{\frac{1-2\alpha}{8(1-\alpha)}}n^{\frac{1}{2(1-\alpha)}}\epsilon^{\frac{2\alpha-1}{2(1-\alpha)}}}{T^{\frac{1+\alpha}{2(1-\alpha)}}(\log(n/\delta))^{\frac{1}{2(1-\alpha)}}}\\&+\Big(\frac{\big(d\log(1/\delta)\big)^{\frac{1-2\alpha}{4(1-\alpha)}}(\log(1/\gamma))^{\frac{1-2\alpha}{8(1-\alpha)}}n^{\frac{1}{2(1-\alpha)}}\epsilon^{\frac{2\alpha-1}{2(1-\alpha)}}}{T^{\frac{1+\alpha}{2(1-\alpha)}}(\log(n/\delta))^{\frac{1}{2(1-\alpha)}}}\Big)^{1+\alpha}\\&+\sqrt{\log(n)\log(1/\gamma)\log(1/\delta)}\Big(\frac{1}{\sqrt{n}}+\frac{n^{\frac{1}{2(1-\alpha)}}}{T^{\frac{1+\alpha}{2(1-\alpha)}}}\Big)\Big)\cdot\|\mathbf{w}^*\|_2^2.\end{aligned}$$

Similarly, we choose the smallest $T$ such that $\frac{n^{\frac{1}{2(1-\alpha)}}}{T^{\frac{1+\alpha}{2(1-\alpha)}}}=\mathcal{O}(\frac{1}{\sqrt{n}})$. Hence, we set $T\asymp n^{\frac{2-\alpha}{1+\alpha}}$ if $\alpha<\frac{1}{2}$, and $T\asymp n$ else. Since $\frac{1}{4}\geq\frac{1-2\alpha}{2(1-\alpha)}$, we have

$$\begin{aligned}\mathcal{R}(\mathbf{w}_{\mathrm{priv}})-\mathcal{R}(\mathbf{w}^*)=\mathcal{O}\Big(&\frac{\big(d\log(1/\delta)\big)^{\frac{1}{4}}(\log(1/\gamma))^{\frac{1}{8}}\sqrt{\log(n/\delta)}}{\sqrt{n\epsilon}}+\Big(\frac{\big(d\log(1/\delta)\big)^{\frac{1}{4}}(\log(1/\gamma))^{\frac{1}{8}}\sqrt{\log(n/\delta)}}{\sqrt{n\epsilon}}\Big)^{1+\alpha}\\&+\frac{\sqrt{\log(n)\log(1/\gamma)\log(n/\delta)}}{\sqrt{n}}\Big)\cdot\|\mathbf{w}^*\|_2^2.\end{aligned}$$

It is reasonable to assume the first term is less than 1 here. Therefore, with probability at least $1-\gamma$, there holds

$$\mathcal{R}(\mathbf{w}_{\mathrm{priv}})-\mathcal{R}(\mathbf{w}^*)=\|\mathbf{w}^*\|_2^2\cdot\mathcal{O}\Big(\frac{\big(d\log(1/\delta)\big)^{\frac{1}{4}}(\log(1/\gamma))^{\frac{1}{8}}\sqrt{\log(n/\delta)}}{\sqrt{n\epsilon}}+\frac{\sqrt{\log(n)\log(1/\gamma)\log(n/\delta)}}{\sqrt{n}}\Big).$$

The proof is completed. □

### *3.3. Proofs on differential privacy of SGD with gradient perturbation*

We now turn to the analysis for `DP-SGD-Gradient` algorithm (i.e. Algorithm 2) and provide the proofs for Theorems 11 and 12. We start with the proof of Theorem 11 on the privacy guarantee for Algorithm 2.

**Proof of Theorem 11.** Consider the mechanism $\mathcal{G}_t=\mathcal{M}_t+\mathbf{b}_t$, where $\mathcal{M}_t=\partial\ell(\mathbf{w}_t,z_{i_t})$. For any $\mathbf{w}_t\in\mathcal{W}$ and any $z_{i_t},z'_{i_t}\in\mathcal{Z}$, the definition of $\alpha$-Hölder smoothness implies that

$$\|\partial\ell(\mathbf{w}_t,z_{i_t})-\partial\ell(\mathbf{w}_t,z'_{i_t})\|_2\leq 2\big(M+L\|\mathbf{w}_t\|_2^{\alpha}\big)\leq 2(M+LR^{\alpha}).$$

Therefore, the $\ell_2$-sensitivity of $\mathcal{M}_t$ is $2(M+LR^\alpha)$. Let

$$\sigma^2 = \frac{14(M+LR^\alpha)^2 T}{\beta n^2 \epsilon}\Big(\frac{\log(1/\delta)}{(1-\beta)\epsilon}+1\Big).$$

Lemma 3 with $p=\frac{1}{n}$ implies that $\mathcal{G}_t$ satisfies $\Big(\lambda, \frac{\lambda\beta\epsilon}{T(\frac{\log(1/\delta)}{(1-\beta)\epsilon}+1)}\Big)$-RDP if the following conditions hold

$$\frac{\sigma^2}{4(M+LR^\alpha)^2} \geq 0.67 \tag{32}$$

and

$$\lambda - 1 \leq \frac{\sigma^2}{6(M+LR^\alpha)^2}\log\Big(\frac{n}{\lambda(1+\frac{\sigma^2}{4(M+LR^\alpha)^2})}\Big). \tag{33}$$

Let $\lambda = \frac{\log(1/\delta)}{(1-\beta)\epsilon}+1$. We obtain that $\mathcal{G}_t$ satisfies $(\frac{\log(1/\delta)}{(1-\beta)\epsilon}+1, \frac{\beta\epsilon}{T})$-RDP. Then by the post-processing property of RDP (see Lemma 6), we know $\mathbf{w}_{t+1}$ also satisfies $(\frac{\log(1/\delta)}{(1-\beta)\epsilon}+1, \frac{\beta\epsilon}{T})$-RDP for any $t=0,...,T-1$. Furthermore, according to the adaptive composition theorem of RDP (see Lemma 4), Algorithm 2 satisfies $(\frac{\log(1/\delta)}{(1-\beta)\epsilon}+1, \beta\epsilon)$-RDP. Finally, by Lemma 5, the output of Algorithm 2 satisfies $(\epsilon,\delta)$-DP as long as (32) and (33) hold. □

Now, we turn to the generalization analysis of Algorithm 2. First, we estimate the generalization error $\mathcal{R}(\mathbf{w}_{\rm priv}) - \mathcal{R}_S(\mathbf{w}_{\rm priv})$ in (4).

**Lemma 25.** *Suppose the loss function $\ell$ is nonnegative, convex and $\alpha$-Hölder smooth with parameter $L$. Let $\mathbf{w}_{priv}$ be the output produced by Algorithm 2 based on $S=\{z_1,\cdots,z_n\}$ with $\eta_t=\eta<\min\{1,1/L\}$. Then for any $\gamma\in(0,1)$, with probability at least $1-\frac{\gamma}{3}$, there holds*

$$\mathcal{R}(\mathbf{w}_{priv}) - \mathcal{R}_S(\mathbf{w}_{priv}) = \mathcal{O}\Big(\tilde{\Delta}_{SGD}(\gamma/6)\log(n)\log(1/\gamma) + \sqrt{\frac{\log(1/\gamma)}{n}}\Big).$$

**Proof.** Part (b) in Theorem 7 implies that $\tilde{\Delta}_{\rm SGD}(\gamma/6) = \mathcal{O}\Big(\sqrt{T}\eta^{\frac{1}{1-\alpha}} + \frac{T\eta\log(n/\gamma)}{n}\Big)$ with probability at least $1-\frac{\gamma}{6}$. Since the noise added to the gradient in each iteration is the same for the neighboring datasets $S$ and $S'$, the noise addition does not impact the stability analysis. Therefore, the UAS bound of the noisy SGD is equivalent to the SGD. According to Lemma 1 and $\|\mathbf{w}_{\rm priv}\|_2\leq R$, we derive the following inequality with probability at least $1-(\frac{\gamma}{6}+\frac{\gamma}{6})$

$$\begin{aligned}\mathcal{R}(\mathbf{w}_{\rm priv}) - \mathcal{R}_S(\mathbf{w}_{\rm priv}) &\leq c\Big((M+LR^\alpha)\tilde{\Delta}_{\rm SGD}(\gamma/6)\log(n)\log(6/\gamma) + \big(M_0+(M+LR^\alpha)R\big)\sqrt{\frac{\log(6/\gamma)}{n}}\Big)\\ &= \mathcal{O}\Big(\tilde{\Delta}_{\rm SGD}(\gamma/6)\log(n)\log(1/\gamma) + \sqrt{\frac{\log(1/\gamma)}{n}}\Big),\end{aligned}$$

where $c>0$ is a constant. The proof is completed. □

The following lemma gives an upper bound for the second term $\mathcal{R}_S(\mathbf{w}_{\rm priv}) - \mathcal{R}_S(\mathbf{w}^*)$ in (4).

**Lemma 26.** *Suppose the loss function $\ell$ is nonnegative, convex and $\alpha$-Hölder smooth with parameter $L$. Let $\mathbf{w}_{priv}$ be the output produced by Algorithm 2 based on $S=\{z_1,\cdots,z_n\}$ with $\eta_t=\eta<\min\{1,1/L\}$. Then, for any $\gamma\in(18\exp(-dT/8),1)$, with probability at least $1-\frac{\gamma}{3}$, there holds*

$$\mathcal{R}_S(\mathbf{w}_{priv}) - \mathcal{R}_S(\mathbf{w}^*) = \mathcal{O}\Big(\|\mathbf{w}^*\|_2^{1+\alpha}\sqrt{\frac{\log(1/\gamma)}{T}} + \frac{\|\mathbf{w}^*\|_2^2}{T\eta} + \eta + \frac{\sqrt{\log(1/\delta)\log(1/\gamma)}(\|\mathbf{w}^*\|_2+\eta)}{n\epsilon} + \frac{\eta T d\log(\frac{1}{\delta})\sqrt{\log(\frac{1}{\gamma})}}{n^2\epsilon^2}\Big).$$

**Proof.** To estimate the term $\mathcal{R}_S(\mathbf{w}_{\rm priv}) - \mathcal{R}_S(\mathbf{w}^*)$, we decompose it as

$$\mathcal{R}_S(\mathbf{w}_{\rm priv}) - \mathcal{R}_S(\mathbf{w}^*) \le \frac{1}{T}\sum_{t=1}^T[\mathcal{R}_S(\mathbf{w}_t) - \ell(\mathbf{w}_t, z_{i_t})] + \frac{1}{T}\sum_{t=1}^T[\ell(\mathbf{w}^*, z_{i_t}) - \mathcal{R}_S(\mathbf{w}^*)] + \frac{1}{T}\sum_{t=1}^T[\ell(\mathbf{w}_t, z_{i_t}) - \ell(\mathbf{w}^*, z_{i_t})]. \tag{34}$$

Similar to the analysis in (20) and (21), we have $\ell(\mathbf{w}^*, z) = \mathcal{O}(\|\mathbf{w}^*\|_2^{1+\alpha})$ for all $z \in \mathcal{Z}$ and $\ell(\mathbf{w}_t, z) = \mathcal{O}(R + R^{1+\alpha})$ for all $t = 1, \dots, T$ and $z \in \mathcal{Z}$. Therefore, Azuma-Hoeffding inequality (see Lemma 17) yields, with probability at least $1 - \frac{\gamma}{9}$, that

$$\frac{1}{T}\sum_{t=1}^T[\mathcal{R}_S(\mathbf{w}_t) - \ell(\mathbf{w}_t, z_t)] \le \big(\sup_{z\in\mathcal{Z}}\ell(0,z) + \sup_{t=1,\dots,T;z\in\mathcal{Z}}\ell(\mathbf{w}_t,z)\big)\sqrt{\frac{\log(9/\gamma)}{2T}} = \mathcal{O}\Big((R+R^{1+\alpha})\sqrt{\frac{\log(1/\gamma)}{T}}\Big). \tag{35}$$

In addition, Hoeffding inequality (see Lemma 16) implies, with probability at least $1 - \frac{\gamma}{9}$, that

$$\frac{1}{T}\sum_{t=1}^T[\ell(\mathbf{w}^*, z_{i_t}) - \mathcal{R}_S(\mathbf{w}^*)] \le (\sup_{z\in\mathcal{Z}}\ell(0,z) + \sup_{z\in\mathcal{Z}}\ell(\mathbf{w}^*,z))\sqrt{\frac{\log(9/\gamma)}{2T}} = \mathcal{O}\Big(\|\mathbf{w}^*\|_2^{1+\alpha}\sqrt{\frac{\log(1/\gamma)}{T}}\Big). \tag{36}$$

Finally, we try to bound $\frac{1}{T}\sum_{t=1}^T[\ell(\mathbf{w}_t, z_{i_t}) - \ell(\mathbf{w}^*, z_{i_t})]$. The SGD update rule implies that $\|\mathbf{w}_{t+1} - \mathbf{w}^*\|_2^2 = \|\mathrm{Proj}_{\mathcal{W}}\big(\mathbf{w}_t - \eta(\partial\ell(\mathbf{w}_t, z_{i_t}) + \mathbf{b}_t)\big) - \mathbf{w}^*\|_2^2 \le \|(\mathbf{w}_t - \mathbf{w}^*) - \eta(\partial\ell(\mathbf{w}_t, z_{i_t}) + \mathbf{b}_t)\|_2^2$, then we have $\langle \mathbf{w}_t - \mathbf{w}^*, \partial\ell(\mathbf{w}_t, z_{i_t})\rangle \le \frac{1}{2\eta}\big(\|\mathbf{w}_t - \mathbf{w}^*\|_2^2 - \|\mathbf{w}_{t+1} - \mathbf{w}^*\|_2^2\big) + \frac{\eta}{2}\big(\|\partial\ell(\mathbf{w}_t, z_{i_t})\|_2^2 + \|\mathbf{b}_t\|_2^2\big) - \langle \mathbf{b}_t, \mathbf{w}_t - \mathbf{w}^* - \eta\partial\ell(\mathbf{w}_t, z_{i_t})\rangle$. Further, noting $\|\mathbf{w}_1\|_2 = 0$, then by the convexity of $\ell$ we have

$$\frac{1}{T}\sum_{t=1}^T[\ell(\mathbf{w}_t, z_{i_t}) - \ell(\mathbf{w}^*, z_{i_t})] \le \frac{\|\mathbf{w}^*\|_2^2}{2T\eta} + \frac{\eta}{2T}\sum_{t=1}^T\|\partial\ell(\mathbf{w}_t, z_{i_t})\|_2^2 - \frac{1}{T}\sum_{t=1}^T\langle \mathbf{b}_t, \mathbf{w}_t - \mathbf{w}^* - \eta\partial\ell(\mathbf{w}_t, z_{i_t})\rangle + \frac{\eta}{2T}\sum_{t=1}^T\|\mathbf{b}_t\|_2^2.$$

The definition of $\alpha$-Hölder smoothness implies that $\|\partial\ell(\mathbf{w}_t, z_{i_t})\|_2 \le M + L\|\mathbf{w}_t\|_2^\alpha \le M + LR^\alpha$ for any $t$. Then, there hold

$$\frac{\eta}{2T}\sum_{t=1}^T\|\partial\ell(\mathbf{w}_t, z_{i_t})\|_2^2 \le \frac{\eta}{2T}\sum_{t=1}^T(M + L\|\mathbf{w}_t\|_2^\alpha)^2 = \mathcal{O}(\eta),$$

and

$$\|\mathbf{w}_t - \mathbf{w}^* - \eta\partial\ell(\mathbf{w}_t, z_{i_t})\|_2 \le \|\mathbf{w}^*\|_2 + R + \eta(M + LR^\alpha).$$

Since $\mathbf{b}_t$ is an $\sigma^2$-sub-Gaussian random vector, $\frac{1}{T}\langle \mathbf{b}_t, \mathbf{w}_t - \mathbf{w}^* - \eta\partial\ell(\mathbf{w}_t, z_{i_t})\rangle$ is an $\frac{\sigma^2}{T^2}\big(\|\mathbf{w}^*\|_2 + R + \eta(M + LR^\alpha)\big)^2$-sub-Gaussian random vector. Note that the sub-Gaussian parameter $\frac{\sigma^2}{T^2}\big(\|\mathbf{w}^*\|_2 + R +$

$\eta(M + LR^\alpha)\big)^2$ is independent of $\mathbf{w}_{t-1}$ and $\mathbf{b}_{t-1}$. Hence, $\frac{1}{T}\sum_{t=1}^T\langle \mathbf{b}_t, \mathbf{w}_t - \mathbf{w}^* - \eta\partial\ell(\mathbf{w}_t, z_{i_t})\rangle$ is an $\frac{\sigma^2\sum_{t=1}^T(\|\mathbf{w}^*\|_2+R+\eta(M+LR^\alpha))^2}{T^2}$-sub-Gaussian random vector. Since $\sigma^2 = \mathcal{O}(\frac{T\log(1/\delta)}{n^2\epsilon^2})$, the tail bound of Sub-Gaussian variables (see Lemma 18) implies, with probability at least $1-\frac{\gamma}{18}$, that

$$\frac{1}{T}\sum_{t=1}^T\langle \mathbf{b}_t, \mathbf{w}_t - \mathbf{w}^* - \eta\partial\ell(\mathbf{w}_t, z_{i_t})\rangle \le \frac{\Big(\sigma^2\big(\|\mathbf{w}^*\|_2 + R + \eta(M+LR^\alpha)\big)^2\Big)^{\frac{1}{2}}}{\sqrt{T}}\sqrt{2\log(18/\gamma)}$$
$$= \mathcal{O}\Big(\sigma(\|\mathbf{w}^*\|_2+\eta)\sqrt{\frac{\log(1/\gamma)}{T}}\Big) = \mathcal{O}\Big(\frac{\sqrt{\log(1/\delta)\log(1/\gamma)}(\|\mathbf{w}^*\|_2+\eta)}{n\epsilon}\Big).$$

According to the Chernoff bound for the $\ell_2$-norm of Gaussian vector with $\mathbf{X} = [\mathbf{b}_{11}, ..., \mathbf{b}_{1d}, \mathbf{b}_{21}..., \mathbf{b}_{Td}] \in \mathbb{R}^{Td}$(see Lemma 15), for any $\gamma \in (18\exp(-dT/8), 1)$, with probability at least $1-\frac{\gamma}{18}$, there holds

$$\frac{\eta}{2T}\sum_{t=1}^T\|\mathbf{b}_t\|_2^2 \le \frac{\eta d}{2T}\Big(1+(\frac{1}{d}\log(18/\gamma))^{\frac{1}{2}}\Big)T\sigma^2 = \mathcal{O}\Big(\frac{\eta T d\log(\frac{1}{\delta})\sqrt{\log(\frac{1}{\gamma})}}{n^2\epsilon^2}\Big).$$

Therefore, with probability at least $1-\frac{\gamma}{9}$, there holds

$$\frac{1}{T}\sum_{t=1}^T[\ell(\mathbf{w}_t, z_{i_t}) - \ell(\mathbf{w}^*, z_{i_t})] \le \mathcal{O}\Big(\frac{\|\mathbf{w}^*\|_2^2}{T\eta} + \eta + \frac{\sqrt{\log(1/\delta)\log(1/\gamma)}(\|\mathbf{w}^*\|_2+\eta)}{n\epsilon}$$
$$+ \frac{\eta T d\log(1/\delta)\sqrt{\log(1/\gamma)}}{n^2\epsilon^2}\Big). \tag{37}$$

Putting (35), (36) and (37) back into (34), we obtain, with probability at least $1-\frac{\gamma}{3}$, that

$$\mathcal{R}_S(\mathbf{w}_{\text{priv}}) - \mathcal{R}_S(\mathbf{w}^*) = \mathcal{O}\Big(\|\mathbf{w}^*\|_2^{1+\alpha}\sqrt{\frac{\log(1/\gamma)}{T}} + \frac{\|\mathbf{w}^*\|_2^2}{T\eta} + \eta + \frac{\sqrt{\log(1/\delta)\log(1/\gamma)}(\|\mathbf{w}^*\|_2+\eta)}{n\epsilon}$$
$$+ \frac{\eta T d\log(1/\delta)\sqrt{\log(1/\gamma)}}{n^2\epsilon^2}\Big).$$

The proof is completed. □

Now, we are ready to prove the utility theorem for *DP-SGD-Gradient* algorithm.

**Proof of Theorem 12.** The Hoeffding inequality implies, with probability at least $1-\frac{\gamma}{3}$, that

$$\mathcal{R}_S(\mathbf{w}^*) - \mathcal{R}(\mathbf{w}^*) \le \big(\sup_{z\in\mathcal{Z}}\ell(0,z) + \sup_{z\in\mathcal{Z}}\ell(\mathbf{w}^*, z)\big)\sqrt{\frac{\log(3/\gamma)}{2n}} = \mathcal{O}\Big(\|\mathbf{w}^*\|_2^{1+\alpha}\sqrt{\frac{\log(1/\gamma)}{n}}\Big).$$

Combining Lemma 25, Lemma 26 and the above inequality together, with probability at least $1-\gamma$, we obtain

$$\mathcal{R}(\mathbf{w}_{\text{priv}}) - \mathcal{R}(\mathbf{w}^*) = \mathcal{O}\Big(\tilde{\Delta}_{\text{SGD}}(\gamma/6)\log(n)\log(1/\gamma) + \frac{\|\mathbf{w}^*\|_2^2}{T\eta} + \eta + \frac{\sqrt{\log(1/\delta)\log(1/\gamma)}(\|\mathbf{w}^*\|_2+\eta)}{n\epsilon}$$
$$+ \frac{\eta T d\log(\frac{1}{\delta})\sqrt{\log(\frac{1}{\gamma})}}{n^2\epsilon^2} + \|\mathbf{w}^*\|_2^{1+\alpha}\sqrt{\frac{\log(1/\gamma)}{n}}\Big).$$

Now, putting $\tilde{\Delta}_{\text{SGD}}(\gamma/6) = \mathcal{O}(\sqrt{T}\eta^{\frac{1}{1-\alpha}} + \frac{T\eta\log(n/\gamma)}{n})$ back into the above estimate, we have

$$\mathcal{R}(\mathbf{w}_{\text{priv}}) - \mathcal{R}(\mathbf{w}^*) = \mathcal{O}\Big(\sqrt{T}\log(n)\log(1/\gamma)\eta^{\frac{1}{1-\alpha}} + \frac{\|\mathbf{w}^*\|_2^2}{T\eta} + \eta\Big(\frac{Td\log(1/\delta)\sqrt{\log(1/\gamma)}}{n^2\epsilon^2} + \frac{T\log(n)\log(n/\gamma)\log(1/\gamma)}{n}\Big) + \|\mathbf{w}^*\|_2^{1+\alpha}\sqrt{\frac{\log(1/\gamma)}{n}} + \frac{\|\mathbf{w}^*\|_2\sqrt{\log(1/\delta)\log(1/\gamma)}}{n\epsilon}\Big). \tag{38}$$

To choose a suitable $\eta$ and $T$ such that the algorithm achieves the optimal rate, we consider the trade-off between $1/\eta$ and $\eta$. We take the derivative of $\frac{1}{T\eta} + \eta\big(\frac{Td\log(1/\delta)\sqrt{\log(1/\gamma)}}{n^2\epsilon^2} + \frac{T\log(n)\log(n/\gamma)\log(1/\gamma)}{n}\big)$ w.r.t $\eta$ and set it to 0, then we have $\eta = 1/T \cdot \max\big\{\frac{\sqrt{\log(n)\log(n/\gamma)\log(1/\gamma)}}{\sqrt{n}}, \frac{\sqrt{d\log(1/\delta)}(\log(1/\gamma))^{\frac{1}{4}}}{n\epsilon}\big\}$. Putting the value of $\eta$ back into (38), we obtain

$$\mathcal{R}(\mathbf{w}_{\text{priv}}) - \mathcal{R}(\mathbf{w}^*) = \mathcal{O}\Big(\frac{\big(\log(n)\log(1/\gamma)\big)^{\frac{1-2\alpha}{2(1-\alpha)}} n^{\frac{1}{2(1-\alpha)}}}{T^{\frac{1+\alpha}{2(1-\alpha)}}\big(\log(n/\gamma)\big)^{\frac{1}{2(1-\alpha)}}} + \frac{\sqrt{d\log(1/\delta)\log(1/\gamma)}}{n\epsilon} + \frac{\sqrt{\log(n)\log(n/\gamma)\log(1/\gamma)}}{\sqrt{n}}\Big) \cdot \|\mathbf{w}^*\|_2^2.$$

In addition, if $n = \mathcal{O}(T^{\frac{1+\alpha}{2-\alpha}})$, then there holds

$$\mathcal{R}(\mathbf{w}_{\text{priv}}) - \mathcal{R}(\mathbf{w}^*) = \|\mathbf{w}^*\|_2^2 \cdot \mathcal{O}\Big(\frac{\sqrt{d\log(1/\delta)\log(1/\gamma)}}{n\epsilon} + \frac{\sqrt{\log(n)\log(n/\gamma)\log(1/\gamma)}}{\sqrt{n}}\Big).$$

The above bound matches the optimal rate $\mathcal{O}\big(\frac{\sqrt{d\log(1/\delta)}}{n\epsilon} + \frac{1}{\sqrt{n}}\big)$. Furthermore, we want the algorithm to achieve the optimal rate with a low computational cost. Therefore, we set $T \asymp n^{\frac{2-\alpha}{1+\alpha}}$ if $\alpha < \frac{1}{2}$, and $T \asymp n$ else. The proof is completed. □

Finally, we give the proof of Lemma 13 on the existence of $\beta$ for Algorithm 2 to be $(\epsilon, \delta)$-DP.

**Proof of Lemma 13.** We give sufficient conditions for the existence of $\beta \in (0,1)$ such that RDP conditions (32) and (33) hold with $\sigma^2 = \frac{14(M+LR^\alpha)^2\lambda}{\beta n\epsilon}$ and $\lambda = \frac{2\log(n)}{(1-\beta)\epsilon} + 1$ in Theorem 11. Condition (32) with $T = n$ and $\delta = \frac{1}{n^2}$ is equivalent to

$$f(\beta) := \beta^2 - \Big(1 + \frac{7}{1.34n\epsilon}\Big)\beta + \frac{7(2\log(n)+\epsilon)}{1.34n\epsilon^2} \geq 0. \tag{39}$$

If $\big(1+\frac{7}{1.34n\epsilon}\big)^2 < \frac{28(2\log(n)+\epsilon)}{1.34n\epsilon^2}$, then $f(\beta) \geq 0$ for all $\beta$. Then (32) holds for any $\beta \in (0,1)$. If $\big(1+\frac{7}{1.34n\epsilon}\big)^2 \geq \frac{28(2\log(n)+\epsilon)}{1.34n\epsilon^2}$, then $\beta \in (0,\beta_1] \cup [\beta_2, +\infty)$ such that the above condition holds, where $\beta_{1,2} = \frac{1}{2}\Big(\big(1+\frac{7}{1.34n\epsilon}\big) \mp \sqrt{\big(1+\frac{7}{1.34n\epsilon}\big)^2 - \frac{28(2\log(n)+\epsilon)}{1.34n\epsilon^2}}\Big)$ are two roots of $f(\beta) = 0$.

Now, we consider the second RDP condition. Plugging $\sigma^2 = \frac{14(M+LR^\alpha)^2\lambda}{\beta n\epsilon}$ back into (33), we derive

$$\frac{3\beta n\epsilon(\lambda-1)}{7\lambda} + \log(\lambda) + \log(1 + \frac{7\lambda}{2\beta n\epsilon}) \leq \log(n). \tag{40}$$

To guarantee (40), it suffices that the following three inequalities hold

$$\frac{3\beta n\epsilon(\lambda-1)}{7\lambda} \leq \frac{\log(n)}{3}, \tag{41}$$

$$\log(\lambda) \leq \frac{\log(n)}{3}, \tag{42}$$

$$\log\big(1+\frac{7\lambda}{2\beta n\epsilon}\big) \leq \frac{\log(n)}{3}. \tag{43}$$

We set $\lambda = \frac{2\log(n)}{(1-\beta)\epsilon} + 1$ in the above three inequalities. Since $\lambda > 1$, then (41) holds if $\beta \leq 7\log(n)/9n\epsilon$. Eq. (42) reduces to $\beta \leq 1 - \frac{2\log(n)}{(n^{1/3}-1)\epsilon}$. Moreover, (43) is equivalent to the following inequality

$$g(\beta) := \beta^2 - (1 + \frac{7}{2n(n^{\frac{1}{3}}-1)\epsilon})\beta + \frac{7(2\log(n)+\epsilon)}{2n(n^{\frac{1}{3}}-1)\epsilon^2} \leq 0. \tag{44}$$

There exists at least one $\beta$ such that $g(\beta) \leq 0$ if $(1 + \frac{7}{2n(n^{1/3}-1)\epsilon})^2 - \frac{14(2\log(n)+\epsilon)}{n(n^{1/3}-1)\epsilon^2} \geq 0$, which can be ensured by the condition $\epsilon \geq \frac{7}{2n(n^{1/3}-1)} + 2\sqrt{\frac{7\log(n)}{n(n^{1/3}-1)}}$. Furthermore, $g(\beta) \leq 0$ for all $\beta \in [\beta_3, \beta_4]$, where $\beta_{3,4} = \frac{1}{2}\Big(\big(1 + \frac{7}{2n(n^{1/3}-1)\epsilon}\big) \mp \sqrt{\big(1 + \frac{7}{2n(n^{1/3}-1)\epsilon}\big)^2 - \frac{14(2\log(n)+\epsilon)}{n(n^{1/3}-1)\epsilon^2}}\Big)$ are two roots of $g(\beta) = 0$. Finally, note that

$$\max\left\{\frac{7}{2n(n^{\frac{1}{3}}-1)} + 2\sqrt{\frac{7\log(n)}{n(n^{\frac{1}{3}}-1)}}, \frac{\log(n)\big(14\log(n)(n^{\frac{1}{3}}-1)+162n-63\big)}{9n\big(2\log(n)(n^{\frac{1}{3}}-1)-9\big)}\right\} \leq \frac{7(n^{\frac{1}{3}}-1)+4\log(n)n+7}{2n(n^{\frac{1}{3}}-1)}.$$

Then if $n \geq 18$ and

$$\epsilon \geq \frac{7(n^{\frac{1}{3}}-1)+4\log(n)n+7}{2n(n^{\frac{1}{3}}-1)},$$

there hold

$$\beta_3 \leq \min\left\{\frac{7\log(n)}{9n\epsilon}, 1 - \frac{2\log(n)}{(n^{\frac{1}{3}}-1)\epsilon}\right\} \tag{45}$$

and

$$\beta_3 \leq \beta_1 \ \text{ if } \big(1 + \frac{7}{1.34n\epsilon^2}\big)^2 \geq \frac{28(2\log(n)+\epsilon)}{1.34n\epsilon^2}. \tag{46}$$

Conditions (45) and (46) ensure the existence of at least one consistent $\beta \in (0,1)$ such that (39), (41), (42), (43) and (44) hold, which imply that (32) and (33) hold. The proof is completed. □

## 4. Conclusion

In this paper, we are concerned with differentially private SGD algorithms with non-smooth losses in the setting of stochastic convex optimization. In particular, we assume that the loss function is $\alpha$-Hölder smooth (i.e., the gradient is $\alpha$-Hölder continuous). We systematically studied the output and gradient perturbations for SGD and established their privacy as well as utility guarantees. For the output perturbation, we proved that our private SGD with $\alpha$-Hölder smooth losses in a bounded $\mathcal{W}$ can achieve $(\epsilon, \delta)$-DP with the excess risk rate $\mathcal{O}\Big(\frac{(d\log(1/\delta))^{1/4}\sqrt{\log(n/\delta)}}{\sqrt{n\epsilon}}\Big)$, up to some logarithmic terms, and gradient complexity $T = \mathcal{O}(n^{\frac{2-\alpha}{1+\alpha}} + n)$, which extends the results of [35] in the strongly-smooth case. We also established similar results for SGD algorithms with output perturbation in an unbounded domain $\mathcal{W} = \mathbb{R}^d$ with excess risk $\mathcal{O}\Big(\frac{\sqrt{d\log(1/\delta)}\log(n/\delta)}{n^{\frac{2}{3+\alpha}}\epsilon} + \frac{\log(n/\delta)}{n^{\frac{1}{3+\alpha}}}\Big)$, up to some logarithmic terms, which are the first-ever known results

of this kind for unbounded domains. For the gradient perturbation, we show that private SGD with $\alpha$-Hölder smooth losses in a bounded domain $\mathcal{W}$ can achieve optimal excess risk $\mathcal{O}\Big(\frac{\sqrt{d\log(1/\delta)}}{n\epsilon}+\frac{1}{\sqrt{n}}\Big)$ with gradient complexity $T=\mathcal{O}(n^{\frac{2-\alpha}{1+\alpha}}+n)$. Whether one can derive privacy and utility guarantees for gradient perturbation in an unbounded domain still remains a challenging open question to us.

**Acknowledgment**

This work was done while Puyu Wang was a visiting student at SUNY Albany. The corresponding author is Yiming Ying, whose work is supported by National Science Foundation (NSF) under grants DMS-2110836, IIS-2110546, IIS-1816227, and IIS-2008532. The work of Hai Zhang is supported by National Science Foundation of China (NSFC) under grant U1811461.

## Appendix. Proof of Lemma 1

In the appendix, we present the proof of Lemma 1. To this aim, we introduce the following lemma.

**Lemma 27.** *Suppose $\ell$ is nonnegative, convex and $\alpha$-Hölder smooth. Let $\mathcal{A}$ be a randomized algorithm with $\sup_{S\simeq S'}\delta_{\mathcal{A}}(S,S')\le\Delta_{\mathcal{A}}$. Suppose the output of $\mathcal{A}$ is bounded by $G>0$ and let $M_0=\sup_{z\in\mathcal{Z}}\ell(0,z)$, $M=\sup_{z\in\mathcal{Z}}\|\partial\ell(0,z)\|_2$. Then for any $\gamma\in(0,1)$, there holds*

$$\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^n,\mathcal{A}}\Bigg[|\mathcal{R}(\mathcal{A}(\mathcal{S}))-\mathcal{R}_{\mathcal{S}}(\mathcal{A}(\mathcal{S}))|\ge c\Bigg((M+LG^\alpha)\Delta_{\mathcal{A}}\log(n)\log(1/\gamma)+\big(M_0+(M+LG^\alpha)G\big)\sqrt{n^{-1}\log(1/\gamma)}\Bigg)\Bigg]\le\gamma.$$

**Proof.** By the convexity of $\ell$ and the definition of $\alpha$-Hölder smoothness, we have for any $S$ and $S'$,

$$\begin{aligned}\ell(\mathcal{A}(S),z)&\le\sup_{z\in\mathcal{Z}}\ell(0,z)+\langle\partial\ell(\mathcal{A}(S),z),\mathcal{A}(S)\rangle\le M_0+\|\partial\ell(\mathcal{A}(S),z)\|_2\|\mathcal{A}(S)\|_2\\&\le M_0+(M+L\|\mathcal{A}(S)\|_2^\alpha)\|\mathcal{A}(S)\|_2\le M_0+(M+LG^\alpha)G\end{aligned}\tag{47}$$

and

$$\begin{aligned}\sup_{z\in\mathcal{Z}}|\ell(\mathcal{A}(S),z)-\ell(\mathcal{A}(S'),z)|&\le\max\big\{\|\partial\ell(\mathcal{A}(S),z)\|_2,\|\partial\ell(\mathcal{A}(S'),z)\|_2\big\}\|\mathcal{A}(S)-\mathcal{A}(S')\|_2\\&\le(M+LG^\alpha)\|\mathcal{A}(S)-\mathcal{A}(S')\|_2.\end{aligned}$$

Note $\sup_{S\simeq S'}\delta_{\mathcal{A}}(S,S')\le\Delta_{\mathcal{A}}$ and $\delta_{\mathcal{A}}(S,S')=\|\mathcal{A}(S)-\mathcal{A}(S')\|_2$. Then for any neighboring datasets $S\simeq S'$, we have

$$\sup_{z\in\mathcal{Z}}|\ell(\mathcal{A}(S),z)-\ell(\mathcal{A}(S'),z)|\le(M+LG^\alpha)\Delta_{\mathcal{A}}.\tag{48}$$

Combining Eq. (47), Eq. (48) and Corollary 8 in [7] together, we derive the following probabilistic inequality

$$\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^n,\mathcal{A}}\Bigg[|\mathcal{R}(\mathcal{A}(\mathcal{S}))-\mathcal{R}_{\mathcal{S}}(\mathcal{A}(\mathcal{S}))|\ge c\Bigg((M+LG^\alpha)\Delta_{\mathcal{A}}\log(n)\log(1/\gamma)+\big(M_0+(M+LG^\alpha)G\big)\sqrt{n^{-1}\log(1/\gamma)}\Bigg)\Bigg]\le\gamma.$$

The proof is completed. □

**Proof of Lemma 1.** Let $E_1 = \{\mathcal{A} : \sup_{S\simeq S'} \|\mathcal{A}(S) - \mathcal{A}(S')\|_2 \geq \Delta_{\mathcal{A}}\}$ and $E_2 = \Big\{(S,\mathcal{A}) : |\mathcal{R}(\mathcal{A}(\mathcal{S})) - \mathcal{R}_S(\mathcal{A}(\mathcal{S}))| \geq c\Big((M + LG^\alpha)\Delta_{\mathcal{A}}\log(n)\log(1/\gamma) + \big(M_0 + (M + LG^\alpha)G\big)\sqrt{n^{-1}\log(1/\gamma)}\Big)\Big\}$. Then by the assumption we have $\mathbb{P}_{\mathcal{A}}[\mathcal{A} \in E_1] \leq \gamma_0$. Further, according to Lemma 27, for any $\gamma \in (0,1)$, we have $\mathbb{P}_{S,\mathcal{A}}[(S,\mathcal{A}) \in E_2 \cap \mathcal{A} \notin E_1] \leq \gamma$. Therefore,

$$\begin{aligned}\mathbb{P}_{S,\mathcal{A}}[(S,\mathcal{A}) \in E_2] &= \mathbb{P}_{S,\mathcal{A}}[(S,\mathcal{A}) \in E_2 \cap \mathcal{A} \in E_1] + \mathbb{P}_{S,\mathcal{A}}[(S,\mathcal{A}) \in E_2 \cap \mathcal{A} \notin E_1] \\ &\leq \mathbb{P}[\mathcal{A} \in E_1] + \mathbb{P}_{S,\mathcal{A}}[(S,\mathcal{A}) \in E_2 \cap \mathcal{A} \notin E_1] \leq \gamma_0 + \gamma.\end{aligned}$$

The proof is completed. □

## References


[1] John M. Abowd, The challenge of scientific reproducibility and privacy protection for statistical agencies, in: Census Scientific Advisory Committee, 2016.
[2] Francis Bach, Eric Moulines, Non-strongly-convex smooth stochastic approximation with convergence rate o (1/n), in: Advances in Neural Information Processing Systems, 2013, pp. 773–781.
[3] Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, Kunal Talwar, Stability of stochastic gradient descent on nonsmooth convex losses, Adv. Neural Inf. Process. Syst. 33 (2020).
[4] Raef Bassily, Vitaly Feldman, Kunal Talwar, Abhradeep Guha Thakurta, Private stochastic convex optimization with optimal rates, in: Advances in Neural Information Processing Systems, 2019, pp. 11279–11288.
[5] Léon Bottou, Olivier Bousquet, The tradeoffs of large scale learning, in: Advances in Neural Information Processing Systems, 2008, pp. 161–168.
[6] Olivier Bousquet, André Elisseeff, Stability and generalization, J. Mach. Learn. Res. 2 (Mar 2002) 499–526.
[7] Olivier Bousquet, Yegor Klochkov, Nikita Zhivotovskiy, Sharper bounds for uniformly stable algorithms, in: Conference on Learning Theory, PMLR, 2020, pp. 610–626.
[8] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, Dawn Song, The secret sharer: evaluating and testing unintended memorization in neural networks, in: 28th {USENIX} Security Symposium ({USENIX} Security 19), 2019, pp. 267–284.
[9] Kamalika Chaudhuri, Claire Monteleoni, Anand D. Sarwate, Differentially private empirical risk minimization, J. Mach. Learn. Res. 12 (Mar 2011) 1069–1109.
[10] Aymeric Dieuleveut, Francis Bach, Nonparametric stochastic approximation with large step-sizes, Ann. Stat. 44 (4) (2016) 1363–1399.
[11] Bolin Ding, Janardhan Kulkarni, Sergey Yekhanin, Collecting telemetry data privately, in: Advances in Neural Information Processing Systems, 2017, pp. 3571–3580.
[12] Cynthia Dwork, Jing Lei, Differential privacy and robust statistics, in: Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing, 2009, pp. 371–380.
[13] Cynthia Dwork, Frank McSherry, Kobbi Nissim, Adam Smith, Calibrating noise to sensitivity in private data analysis, in: Theory of Cryptography Conference, Springer, 2006, pp. 265–284.
[14] Cynthia Dwork, Aaron Roth, et al., The algorithmic foundations of differential privacy, Found. Trends Theor. Comput. Sci. 9 (3–4) (2014) 211–407.
[15] Úlfar Erlingsson, Vasyl Pihur, Aleksandra Korolova Rappor, Randomized aggregatable privacy-preserving ordinal response, in: Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, 2014, pp. 1054–1067.
[16] Vitaly Feldman, Tomer Koren, Kunal Talwar, Private stochastic convex optimization: optimal rates in linear time, in: Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, 2020, pp. 439–449.
[17] Moritz Hardt, Ben Recht, Yoram Singer, Train faster, generalize better: stability of stochastic gradient descent, in: International Conference on Machine Learning, 2016, pp. 1225–1234.
[18] Wassily Hoeffding, Probability inequalities for sums of bounded random variables, in: The Collected Works of Wassily Hoeffding, Springer, 1994, pp. 409–426.
[19] Rie Johnson, Tong Zhang, Accelerating stochastic gradient descent using predictive variance reduction, in: Advances in Neural Information Processing Systems, 2013, pp. 315–323.
[20] Simon Lacoste-Julien, Mark Schmidt, Francis Bach, A simpler approach to obtaining an o (1/t) convergence rate for the projected stochastic subgradient method, arXiv preprint, arXiv:1212.2002, 2012.
[21] Yunwen Lei, Yiming Ying, Fine-grained analysis of stability and generalization for stochastic gradient descent, in: International Conference on Machine Learning, PMLR, 2020, pp. 5809–5819.
[22] Zhicong Liang, Bao Wang, Quanquan Gu, Stanley Osher, Yuan Yao, Exploring private federated learning with Laplacian smoothing, arXiv preprint, arXiv:2005.00218, 2020.
[23] Junhong Lin, Lorenzo Rosasco, Optimal learning for multi-pass stochastic gradient methods, in: Advances in Neural Information Processing Systems, 2016, pp. 4556–4564.
[24] Tongliang Liu, Gábor Lugosi, Gergely Neu, Dacheng Tao, Algorithmic stability and hypothesis complexity, in: International Conference on Machine Learning, PMLR, 2017, pp. 2159–2167.

[25] Robert McMillan, Apple tries to peek at user habits without violating privacy, Wall St. J. (2016).
[26] Ilya Mironov, Rényi differential privacy, in: 2017 IEEE 30th Computer Security Foundations Symposium (CSF), IEEE, 2017, pp. 263–275.
[27] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, Alexander Shapiro, Robust stochastic approximation approach to stochastic programming, SIAM J. Optim. 19 (4) (2009) 1574–1609.
[28] Francesco Orabona, Simultaneous model selection and optimization through parameter-free stochastic learning, in: Advances in Neural Information Processing Systems, 2014, pp. 1116–1124.
[29] Alexander Rakhlin, Ohad Shamir, Karthik Sridharan, Making gradient descent optimal for strongly convex stochastic optimization, in: Proceedings of the 29th International Conference on Machine Learning, 2012, pp. 449–456.
[30] Ohad Shamir, Tong Zhang, Stochastic gradient descent for non-smooth optimization: convergence results and optimal averaging schemes, in: International Conference on Machine Learning, 2013, pp. 71–79.
[31] Reza Shokri, Marco Stronati, Congzheng Song, Vitaly Shmatikov, Membership inference attacks against machine learning models, in: 2017 IEEE Symposium on Security and Privacy (SP), IEEE, 2017, pp. 3–18.
[32] Steve Smale, Yuan Yao, Online learning algorithms, Found. Comput. Math. 6 (2) (2006) 145–170.
[33] Shuang Song, Kamalika Chaudhuri, Anand Sarwate, Learning from data with heterogeneous noise using sgd, in: Artificial Intelligence and Statistics, 2015, pp. 894–902.
[34] Martin J. Wainwright, High-Dimensional Statistics: A Non-asymptotic Viewpoint, vol. 48, Cambridge University Press, 2019.
[35] Xi Wu, Fengan Li, Arun Kumar, Kamalika Chaudhuri, Somesh Jha, 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, 2017, pp. 1307–1322.
[36] Yiming Ying, Massimiliano Pontil, Online gradient descent learning algorithms, Found. Comput. Math. 8 (5) (2008) 561–596.
[37] Yiming Ying, Ding-Xuan Zhou, Unregularized online learning algorithms with general loss functions, Appl. Comput. Harmon. Anal. 42 (2) (2017) 224–244.