Neurocomputing 585 (2024) 127557

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Differentially private stochastic gradient descent with low-noise

Puyu Wang [a], Yunwen Lei [b], Yiming Ying [c,*,1], Ding-Xuan Zhou [d]

[a] Department of Mathematics, Hong Kong Baptist University, Hong Kong
[b] Department of Mathematics, The University of Hong Kong, Hong Kong
[c] Department of Mathematics and Statistics, State University of New York at Albany, USA
[d] School of Mathematics and Statistics, University of Sydney, Australia

ARTICLE INFO

Communicated by L. Oneto

Keywords:
Differential privacy
Stochastic gradient descent
Generalization
Low-noise

ABSTRACT

Modern machine learning algorithms aim to extract fine-grained information from data to provide accurate predictions, which often conflicts with the goal of privacy protection. This paper addresses the practical and theoretical importance of developing privacy-preserving machine learning algorithms that ensure good performance while preserving privacy. In this paper, we focus on the privacy and utility (measured by excess risk bounds) performances of differentially private stochastic gradient descent (SGD) algorithms in the setting of stochastic convex optimization. Specifically, we examine the pointwise problem in the low-noise setting for which we derive sharper excess risk bounds for the differentially private SGD algorithm. In the pairwise learning setting, we propose a simple differentially private SGD algorithm based on gradient perturbation. Furthermore, we develop novel utility bounds for the proposed algorithm, proving that it achieves optimal excess risk rates even for non-smooth losses. Notably, we establish fast learning rates for privacy-preserving pairwise learning under the low-noise condition, which is the first of its kind.

1. Introduction

Stochastic gradient descent (SGD) iteratively updates model parameters using the gradient information over a small batch of random examples, which reduces the computation cost and makes it amenable to solving large-scale problems. Due to its low computational overhead and easy implementation, it has become the workhorse algorithm for training many machine learning models [1–9].

On the other important front, we have witnessed a significant risk of privacy leakage by sharing gradient information of machine learning models because the gradient often embeds knowledge about the training data. For instance, [10] provides paradigms for breaching privacy and reconstructing training examples from publicly shared gradients and [11] shows that the membership of a data record can be inferred from a binary classifier trained on gradients. As SGD is widely deployed in machine learning models, it is crucial to develop private SGD algorithms to mitigate the privacy leakage posted by gradients.

In this paper, we are interested in differentially private SGD (DP-SGD) for both pointwise and pairwise learning problems. Differential privacy (DP) [12] is a de facto concept for designing private algorithms, which defines a rigorous attack model independent of background knowledge and gives a quantitative representation of the degree of privacy leakage. There is a considerable amount of work [8,13–21] on analyzing the utility guarantee (i.e., statistical generalization performance) of DP-SGD algorithms. In particular, [8,13,15,17,19] have shown that private SGD algorithms can achieve the optimal excess population risk bound $\mathcal{O}\big(\frac{1}{\sqrt{n}}+\frac{1}{n\epsilon}\sqrt{d\log(1/\delta)}\big)$ for solving convex problems in different settings. Here, n is the size of the training dataset, d is the dimension, and (ϵ,δ) are privacy parameters. One nature question then arises: can DP-SGD algorithms achieve faster utility rates beyond $\mathcal{O}\big(\frac{1}{\sqrt{n}}+\frac{1}{n\epsilon}\sqrt{d\log(1/\delta)}\big)$?

We provide an affirmative answer to the above question under a low-noise condition (also referred as a realizability condition in the literature) [23–27], which assumes that there exists a model within the considered hypothesis space perfectly fits the underlying data distribution. Under this condition, we conduct a comprehensive study of DP-SGD for both pointwise and pairwise learning as well as both smooth and non-smooth losses, which is able to provide faster utility bounds in terms of the excess population risk. Our main contributions are listed as follows:

- Firstly, we are concerned with the standard pointwise learning problems where the loss function $f(\cdot;z)$ on a single datum $z=(x,y)$. For this case, we show that DP-SGD with gradient perturbation algorithm can achieve the rate $\mathcal{O}\big(\frac{1}{\sqrt{n}}+\frac{1}{n\epsilon}\sqrt{d\log(1/\delta)}\big)$ for

* Corresponding author.
E-mail address: yiming.ying@sydney.edu.au (Y. Ying).
[1] School of Mathematics and Statistics, University of Sydney, Sydney, NSW, Australia.

https://doi.org/10.1016/j.neucom.2024.127557
Received 4 March 2023; Received in revised form 8 December 2023; Accepted 10 March 2024
Available online 25 March 2024

Table 1
Comparison of different (ϵ, δ)-DP algorithms for pointwise learning. Here, α-Hölder denotes α-Hölder smooth losses.

Work	Lipschitz	Smooth	Low-noise	Gradient complexity	Utility
[14]	✓	✓	×	$\mathcal{O}\big(n^{1.5}\sqrt{\epsilon}+(n\epsilon)^{2.5}(d\log(1/\delta))^{-1}\big)$	$\mathcal{O}\big(\frac{1}{\sqrt{n}}+\frac{1}{n\epsilon}\sqrt{d\log(1/\delta)}\big)$
	✓	×	×	$\mathcal{O}\big(n^{4.5}\sqrt{\epsilon}+n^{6.5}\epsilon^{4.5}(d\log(1/\delta))^{-2}\big)$	$\mathcal{O}\big(\frac{1}{\sqrt{n}}+\frac{1}{n\epsilon}\sqrt{d\log(1/\delta)}\big)$
[13]	✓	×	×	$\mathcal{O}(n^2)$	$\mathcal{O}\big(\frac{1}{\sqrt{n}}+\frac{1}{n\epsilon}\sqrt{d\log(1/\delta)}\big)$
[19]	×	✓	×	$\mathcal{O}(n)$	$\mathcal{O}\big(\frac{1}{\sqrt{n}}+\frac{1}{n\epsilon}\sqrt{d\log(1/\delta)}\big)$
	×	α-Hölder	×	$\mathcal{O}\big(n^{\frac{2-\alpha}{1+\alpha}}+n\big)$	$\mathcal{O}\big(\frac{1}{\sqrt{n}}+\frac{1}{n\epsilon}\sqrt{d\log(1/\delta)}\big)$
Ours	✓	✓	×	$\mathcal{O}(n)$	$\mathcal{O}\big(\frac{1}{\sqrt{n}}+\frac{1}{n\epsilon}\sqrt{d\log(1/\delta)}\big)$
	✓	✓	✓	$\mathcal{O}(n)$	$\mathcal{O}\big(\frac{1}{n\epsilon}\sqrt{d\log(1/\delta)}\big)$
	✓	α-Hölder	×	$\mathcal{O}\big(n^{\frac{2-\alpha}{1+\alpha}}+n\big)$	$\mathcal{O}\big(\frac{1}{\sqrt{n}}+\frac{1}{n\epsilon}\sqrt{d\log(1/\delta)}\big)$
	✓	α-Hölder	✓	$\mathcal{O}\big(n^{\frac{2}{1+\alpha}}\big)$	$\mathcal{O}\big(n^{-\frac{1+\alpha}{2}}+\frac{1}{n\epsilon}\sqrt{d\log(1/\delta)}\big)$

Table 2
Comparison of different (ϵ, δ)-DP algorithms for pairwise learning. We report the results for Gradient descent with output perturbation (Output GD), Localized Gradient descent (Localized GD) and SGD with gradient perturbation (Gradient SGD).

Work	Method	Lipschitz	Smooth	Low-noise	Gradient complexity	Utility
[22]	Output GD	✓	✓	×	$\mathcal{O}\big(n^2\big)$	$\mathcal{O}\big(\frac{1}{\sqrt{n\epsilon}}\sqrt{d\log(1/\delta)}\big)$
[21]	Localized GD	✓	✓	×	$\mathcal{O}\big(n^3\log(1/\delta)\big)$	$\mathcal{O}\big(\frac{1}{\sqrt{n}}+\frac{1}{n\epsilon}\sqrt{d\log(1/\delta)}\big)$
[8]	Localized SGD	✓	✓	×	$\mathcal{O}\big(n\log(1/\delta)\big)$	$\mathcal{O}\big(\frac{1}{\sqrt{n}}+\frac{1}{n\epsilon}\sqrt{d}\log^{\frac{3}{2}}(1/\delta)\big)$
	Localized SGD	✓	×	×	$\mathcal{O}\big(n^2\log(1/\delta)\big)$	$\mathcal{O}\big(\frac{1}{\sqrt{n}}+\frac{1}{n\epsilon}\sqrt{d\log(1/\delta)}\big)$
Ours	Gradient SGD	✓	✓	×	$\mathcal{O}(n)$	$\mathcal{O}\big(\frac{1}{\sqrt{n}}+\frac{1}{n\epsilon}\sqrt{d\log(1/\delta)}\big)$
	Gradient SGD	✓	✓	✓	$\mathcal{O}(n)$	$\mathcal{O}\big(\frac{1}{n\epsilon}\sqrt{d\log(1/\delta)}\big)$
	Gradient SGD	✓	α-Hölder	×	$\mathcal{O}\big(n^{\frac{2-\alpha}{1+\alpha}}+n\big)$	$\mathcal{O}\big(\frac{1}{\sqrt{n}}+\frac{1}{n\epsilon}\sqrt{d\log(1/\delta)}\big)$
	Gradient SGD	✓	α-Hölder	✓	$\mathcal{O}\big(n^{\frac{2}{1+\alpha}}\big)$	$\mathcal{O}\big(n^{-\frac{1+\alpha}{2}}+\frac{1}{n\epsilon}\sqrt{d\log(1/\delta)}\big)$

both strongly smooth and α-Hölder smooth losses, which match the results in the recently work [19]. Under a low-noise condition, we remove the term $\mathcal{O}\big(\frac{1}{\sqrt{n}}\big)$ and achieve the excess risk bound of the order $\mathcal{O}\big(\frac{1}{n\epsilon}\sqrt{d\log(1/\delta)}\big)$ for strongly smooth losses. Further, a better excess risk rate $\mathcal{O}\big(n^{-\frac{1+\alpha}{2}}+\frac{1}{n\epsilon}\sqrt{d\log(1/\delta)}\big)$ is established for α-Hölder smooth losses.

- Secondly, we study the pairwise learning setting where the loss $f(\cdot; z, z')$ involves a pair of examples (z, z'). In this learning setting, we propose a simple differentially private SGD algorithm for pairwise learning with utility guarantees. Specifically, for strongly smooth losses, our algorithm only requires gradient complexity $\mathcal{O}(n)$ to achieve the optimal excess risk rate, while [8,21] require $\mathcal{O}(n^3\log(1/\delta))$ and $\mathcal{O}(n\log(1/\delta))$, respectively. We also show that this rate can be achieved even if the loss is non-smooth. Further, for both strongly smooth and non-smooth pairwise losses, we establish faster excess risk bounds under a low-noise condition. To the best of our knowledge, this is the first utility analysis which provides the excess risk bounds better than $\mathcal{O}\big(\frac{1}{\sqrt{n}}+\frac{1}{n\epsilon}\sqrt{d\log(1/\delta)}\big)$ for privacy-preserving pairwise learning.

1.1. Related work

In this subsection, we review the relevant work on DP-SGD which are close to our work. We discuss them in the pointwise and pairwise learning settings, respectively.

For pointwise learning, [14] established the excess population risk bounds in the order of $\mathcal{O}\big(\frac{1}{\sqrt{n}}+\frac{1}{n\epsilon}\sqrt{d\log(1/\delta)}\big)$ for (ϵ, δ)-differentially private stochastic convex optimization algorithms for both strongly smooth and non-smooth losses, which match the lower bound given in [15]. However, their algorithms have a large gradient complexity (measured by the total number of computing the gradient). Specifically, their analysis establishes gradient complexity $\mathcal{O}(n^{1.5}\sqrt{\epsilon}+(n\epsilon)^{2.5}(d\log(1/\delta))^{-1})$ and $\mathcal{O}(n^{4.5}\sqrt{\epsilon}+(n)^{6.5}\epsilon^{4.5}(d\log(1/\delta))^{-2})$ for strongly smooth and non-smooth losses, respectively. [17] proposed a private phased SGD algorithm for strongly smooth losses, which can achieve the optimal excess risk rate with a linear gradient complexity $\mathcal{O}(n)$. The work [13] developed a DP-SGD algorithm with gradient perturbation which improved the gradient complexity to $\mathcal{O}(n^2)$ for non-smooth losses. The work most related to our paper is [19], which studied DP-SGD with gradient perturbation. They established the optimal excess risk bounds for strongly smooth and α-Hölder smooth losses with gradient complexity $\mathcal{O}(n)$ and $\mathcal{O}(n^{\frac{2-\alpha}{1+\alpha}}+n)$, respectively, which recover the results in [13,17]. However, they did not obtain the fast rates in the low-noise case which is the main focus of our paper. For clarity, we list in Table 1 the comparison of our work again other existing work in terms of utility (excess risk) bounds, assumptions on loss function and the gradient complexity of DP-SGD in the pointwise learning setting.

For pairwise learning, [22] studied private gradient descent (GD) with output perturbation and proved that the proposed algorithm can achieve the excess risk rate $\mathcal{O}(\frac{1}{\sqrt{n\epsilon}}\sqrt{d\log(1/\delta)})$ for Lipschitz and strongly smooth losses. [21] proposed a private localized GD algorithm, which can achieve the optimal excess risk rate with gradient complexity $\mathcal{O}(n^3\log(1/\delta))$ for Lipschitz and strongly smooth losses. The work [8] developed a DP-SGD algorithm with an iterative localization technique and derived the (nearly) optimal excess risk bounds for strongly smooth and non-smooth losses with gradient complexity $\mathcal{O}(n\log(1/\delta))$ and $\mathcal{O}(n^2\log(1/\delta))$, respectively. In this work, we are interested in DP-SGD for both strongly smooth and α-Hölder smooth losses as well as the low-noise case. Table 2 summarizes the comparison of our work against the existing methods in terms of the utility (excess risk) bounds, assumptions on losses and the gradient of DP-SGD in the pairwise learning setting.

Organization of the paper. The remaining parts of the paper are organized as follows. In Section 2, we present the formulations of

pointwise and pairwise learning together with basic concepts of differential privacy. In Section 3, we introduce the DP-SGD algorithms in the settings of pointwise learning and pairwise learning and present our main results. The main proofs are given in Section 4. Section 6 concludes the paper.

2. Learning setting and preliminaries

Let ρ be a probability distribution defined on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} \subseteq \mathbb{R}$. In the standard framework of statistical learning theory [28,29], we consider the problem of learning from a training dataset $S = \{z_i\}_{i=1}^n$, where z_i is independently drawn from ρ. In the subsequent subsections, we describe the settings of pointwise and pairwise learning, the definition of differential privacy, and illustrate the goal of utility analysis.

2.1. Pointwise and pairwise learning settings

In the task of pointwise learning such as classification and regression, we aim to learn a model $\mathbf{w} \in \mathcal{W} \subseteq \mathbb{R}^d$ from training data S and measure the quality of $\mathbf{w}$ using a pointwise loss function $f(\mathbf{w}; z)$ on a single datum $z = (x, y)$. The expected population risk for pointwise learning is given by $F(\mathbf{w}) = \mathbb{E}_{z\sim\rho}[f(\mathbf{w}; z)]$. Based on training dataset S, the empirical risk minimization (ERM) problem can be formulated as follows

$$\min_{\mathbf{w}\in\mathcal{W}} \Big\{ F_S(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{w}; z_i) \Big\}. \tag{1}$$

In contrast to pointwise learning, the performance of a model $\mathbf{w}$ for pairwise learning is measured on a pair of examples (z, z') by a loss function $f(\mathbf{w}; z, z')$ [8,30–32]. Many machine learning problems can be formulated as learning with pairwise loss functions including AUC maximization [33–37], metric learning [38–40], a minimum error entropy principle [41] and ranking [42,43]. Let $[n] := \{1, \dots, n\}$ and let $\mathbf{w}^* = \arg\min_{\mathbf{w}\in\mathcal{W}}\{\bar{F}(\mathbf{w}) := \mathbb{E}_{z,z'\sim\rho}[f(\mathbf{w}; z, z')]\}$ be the best model. The ERM problem on training data S is given by

$$\min_{\mathbf{w}\in\mathcal{W}} \Big\{ \bar{F}_S(\mathbf{w}) = \frac{1}{n(n-1)} \sum_{i,j\in[n], i\neq j} f(\mathbf{w}; z_i, z_j) \Big\}. \tag{2}$$

2.2. Definition and property of differential privacy

As a privacy-preserving technology with a rigorous mathematical guarantee, DP has been widely used in several areas [8,14,15]. Its definition is stated formally as follows. Here, datasets S and $\tilde{S}$ differing by only one datum are denoted as $S \sim \tilde{S}$.

Definition 1 (*DP [12]*)**.** Given a randomized algorithm $\mathcal{M}$. If for any $S \sim \tilde{S}$ and any event $B \in \text{Range}(\mathcal{M})$, there holds $\mathbb{P}(\mathcal{M}(S) \in B) \leq e^{\epsilon}\mathbb{P}(\mathcal{M}(\tilde{S}) \in B) + \delta$, then $\mathcal{M}$ satisfies (ϵ, δ)-DP.

To show a randomized algorithm satisfies DP, we need the following concept. Let $\|\cdot\|_2$ denote the Euclidean norm.

Definition 2. A function $\mathcal{G} : \mathcal{Z}^n \to \mathcal{W}$ has the ℓ_2-sensitivity $\Delta_{\mathcal{G}}$ if for all $S \sim \tilde{S}$, there holds $\|\mathcal{G}(S) - \mathcal{G}(\tilde{S})\|_2 \leq \Delta_{\mathcal{G}}$.

We resort to the Gaussian mechanism to achieve (ϵ, δ)-DP.

Lemma 1 (*[44]*)**.** *Given a function $\mathcal{G} : \mathcal{Z}^n \to \mathcal{W}$ and a dataset S. Assume $\mathcal{G}$ has ℓ_2-sensitivity $\Delta_{\mathcal{G}}$. The Gaussian mechanism which satisfies (ϵ, δ)-DP is defined as follows:*

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

Here $\sigma = \frac{\sqrt{2\log(1.25/\delta)}\Delta_{\mathcal{G}}}{\epsilon}$, and $\mathbf{I}_d$ is the identity matrix.

We are interested in DP-SGD with strongly smooth and α-Hölder smooth losses, respectively.

Definition 3. Let $\partial f(\cdot)$ denotes a (sub)gradient of f. We say a function $\mathbf{w} \to g(\mathbf{w})$ is γ-strongly smooth with $\gamma > 0$ if, for any $\mathbf{w}_1, \mathbf{w}_2 \in \mathcal{W}$, there holds $g(\mathbf{w}_1) \leq g(\mathbf{w}_2) + \langle \partial g(\mathbf{w}_2), \mathbf{w}_1 - \mathbf{w}_2\rangle + \frac{\gamma}{2}\|\mathbf{w}_1 - \mathbf{w}_2\|_2^2$. We say a function $\mathbf{w} \to g(\mathbf{w})$ is α-Hölder smooth with $\alpha \in [0, 1)$ and parameter γ if for any $\mathbf{w}_1, \mathbf{w}_2 \in \mathcal{W}$, there holds $\|\partial g(\mathbf{w}_1) - \partial g(\mathbf{w}_2)\|_2 \leq \gamma\|\mathbf{w}_1 - \mathbf{w}_2\|_2^{\alpha}$.

For a α-Hölder smooth loss g, the smoothness parameter $\alpha \in [0, 1)$ characterizes its smoothness. In particular, $\alpha = 0$ implies that g is Lipschitz continuous (see Definition 4 below). This definition also covers many non-smooth losses including the q-norm loss $|y - \mathbf{w}^\top\mathbf{x}|^q$ [45] and the q-norm hinge loss $\left(1 - y\mathbf{w}^\top\mathbf{x}\right)_+^q$.

2.3. Target of utility analysis

We move on to describing the target of utility analysis of a randomized algorithm $\mathcal{M}$ to solve the ERM problems (1) or (2). For simplicity, we elaborate this by taking pointwise learning as example and the same procedure can apply to the case of pairwise learning.

To this end, let $\mathcal{M}(S)$ be the output produced by running $\mathcal{M}$ over S for pointwise learning. Let $\mathbf{w}^* = \arg\min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w})$. The utility of $\mathcal{M}$ is measured by the *excess population risk* $F(\mathcal{M}(S)) - F(\mathbf{w}^*)$. To study the excess population risk, we resort to the following error decomposition:

$$\begin{aligned}\mathbb{E}[F(\mathcal{M}(S)) - F(\mathbf{w}^*)] = \ & \mathbb{E}[F(\mathcal{M}(S)) - F_S(\mathcal{M}(S))] \\ & + \mathbb{E}[F_S(\mathcal{M}(S)) - F_S(\mathbf{w}^*)],\end{aligned} \tag{3}$$

where the expectation $\mathbb{E}[\cdot]$ is taken with respect to both the randomness of S and the internal randomness of $\mathcal{M}$. The first term $\mathbb{E}[F(\mathcal{M}(S)) - F_S(\mathcal{M}(S))]$ is called the generalization error. We will use algorithmic stability to control this term [13,26,28,46,47]. The second term $\mathbb{E}[F_S(\mathcal{M}(S)) - F_S(\mathbf{w}^*)]$ is the optimization error, we can use tools in optimization theory to handle it.

Definition 4. We say a function $\mathbf{w} \to g(\mathbf{w})$ is convex if, for any $\mathbf{w}_1, \mathbf{w}_2 \in \mathcal{W}$, there holds $g(\mathbf{w}_1) \geq g(\mathbf{w}_2) + \langle \partial g(\mathbf{w}_2), \mathbf{w}_1 - \mathbf{w}_2\rangle$. We say a function $\mathbf{w} \to g(\mathbf{w})$ is G-Lipschitz continuous with $G > 0$ if, for any $\mathbf{w}_1, \mathbf{w}_2 \in \mathcal{W}$, there holds $|g(\mathbf{w}_1) - g(\mathbf{w}_2)| \leq G\|\mathbf{w}_1 - \mathbf{w}_2\|_2$.

Assumption 1. Assume the loss f is nonnegative and convex with respect to the first argument.

Assumption 2. Assume the loss f is G-Lipschitz with respect to the first argument.

3. Main results

The main results of the paper are presented in this section. In Section 3.1, we first develop the privacy-preserving SGD algorithm for pointwise learning, and then present a comprehensive study of its privacy and utility guarantees. In Section 3.2, we present a computation-efficient differentially private SGD algorithm for pairwise learning and provide its privacy and utility guarantees.

3.1. DP-SGD for pointwise learning

In this subsection, we are interested in differentially private SGD for pointwise learning. To achieve (ϵ, δ)-differential privacy, we resort to the gradient perturbation mechanism, i.e., noising the stochastic gradient by Gaussian noise in each iteration of the algorithm. The detailed algorithm is given by Algorithm 1. The following theorem provides the privacy guarantee of Algorithm 1.

Theorem 2 (*Privacy Guarantee*)**.** *Suppose Assumptions 1 and 2 hold. Then Algorithm 1 with some $\beta \in (0, 1)$ satisfies (ϵ, δ)-DP if $\sigma^2 \geq 2.68G^2$ and $\lambda - 1 \leq \frac{\sigma^2}{6G^2}\log\Big(\frac{n}{\lambda\left(1+\frac{\sigma^2}{4G^2}\right)}\Big)$ with $\lambda = \frac{\log(1/\delta)}{(1-\beta)\epsilon} + 1$.*

Algorithm 1 DP-SGD for pointwise learning

Inputs: Data $S = \{z_i \in \mathcal{Z} : i = 1, \dots, n\}$, loss function $f(\mathbf{w}; z)$ with Lipschitz parameter G, the convex set $\mathcal{W} \subseteq \mathbb{R}^d$, step size $\{\eta_t\}$, privacy parameters ϵ, δ, and constant β.
Set: $\mathbf{w}_1 = \mathbf{0}$
for $t = 1$ to T **do**
  Sample $i_t \sim \text{Unif}([n])$
  $\mathbf{w}_{t+1} = \text{Proj}_{\mathcal{W}}\big(\mathbf{w}_t - \eta_t(\partial f(\mathbf{w}_t; z_{i_t}) + \mathbf{b}_t)\big)$, where $\mathbf{b}_t \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_d)$ with $\sigma^2 = \frac{14G^2T}{\beta n^2 \epsilon}\big(\frac{\log(1/\delta)}{(1-\beta)\epsilon} + 1\big)$
end for
return: $\mathbf{w}_{\text{priv}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{w}_t$

Algorithm 2 DP-SGD for pairwise learning (DP-SGD-pairwise)

Inputs: Data $S = \{z_i \in \mathcal{Z} : i = 1, \dots, n\}$, loss function $f(\mathbf{w}; z, z')$ with Lipschitz parameter G, convex set $\mathcal{W} \subseteq \mathbb{R}^d$, step size $\{\eta_t\}$, privacy parameters ϵ, δ, constant β.
Set: $\mathbf{w}_1 = \mathbf{0}$
for $t = 1$ to T **do**
  Sample (i_t, j_t) uniformly over all pairs $\{(i, j) : i, j \in [n], i \neq j\}$
  $\mathbf{w}_{t+1} = \text{Proj}_{\mathcal{W}}\big(\mathbf{w}_t - \eta_t(\partial f(\mathbf{w}_t; z_{i_t}, z_{j_t}) + \mathbf{b}_t)\big)$, where $\mathbf{b}_t \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_d)$ with $\sigma^2 = \frac{56G^2T}{\beta n^2 \epsilon}\big(\frac{\log(1/\delta)}{(1-\beta)\epsilon} + 1\big)$
end for
return: $\mathbf{w}_{\text{priv}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{w}_t$

Remark 1. In Algorithm 1, the variance σ^2 of the Gaussian noise $\mathbf{b}_t$ depends on a constant $\beta \in (0, 1)$, which should satisfy the conditions $\sigma^2 \geq 2.68G^2$ and $\lambda - 1 \leq \frac{\sigma^2}{6G^2}\log\big(\frac{n}{\lambda(1+\frac{\sigma^2}{4G^2})}\big)$. [19] studied DP-SGD with gradient perturbation for α-Hölder smooth losses and gave a sufficient condition for the existence of β under a certain setting. Specifically, they proved that under the setting $n > 18$, $T = n$ and $\delta = 1/n^2$, if $\epsilon \geq 7(n^{\frac{1}{3}} - 1) + 4\log(n)n + 7/(2n(n^{\frac{1}{3}} - 1))$, then there exists at least one β such that the privacy guarantee of DP-SGD can be promised. Indeed, our algorithm can be considered as a special case of their algorithm with $\alpha = 0$. Hence, we can also show the existence of β under the same setting.

Now, we establish the utility guarantee for strongly smooth losses. Part (a) in the following theorem provides the optimal utility bound for a general setting, i.e., the "pessimistic" case $F(\mathbf{w}^*) > 0$. Part (b) of Theorem 3 focuses on the low-noise setting, i.e., the optimistic case $F(\mathbf{w}^*) = 0$, where the best possible model $\mathbf{w}^*$ can achieve zero error. This setting is particularly intriguing in the context of deep learning, where models may possess more parameters than training examples.

Theorem 3 (*Utility Guarantee for Smooth Losses*). *Suppose Assumptions 1 and 2 hold and f is L-smooth. Let $\mathbf{w}_{priv}$ be the output of Algorithm 1.*

(a) For some constant $c > 0$, selecting $\eta_t = c\min\big\{\frac{1}{\sqrt{n}}, \frac{\epsilon}{\sqrt{d\log(1/\delta)}}\big\} \leq \min\{2/L, 1\}$ and $T \asymp n$, there holds

$$\mathbb{E}[F(\mathbf{w}_{priv})] - F(\mathbf{w}^*) = \mathcal{O}\Big(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\Big).$$

(b) Assume $F(\mathbf{w}^*) = 0$. For some constant $c > 0$, selecting $\eta_t = \frac{c\epsilon}{\sqrt{d\log(1/\delta)}} \leq \min\{2/L, 1\}$ and $T \asymp n$, there holds

$$\mathbb{E}[F(\mathbf{w}_{priv})] - F(\mathbf{w}^*) = \mathcal{O}\Big(\frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\Big).$$

Remark 2 (*[19]*). established the optimal rate for DP-SGD algorithm with gradient complexity $\mathcal{O}(n)$ when the loss is strongly smooth with a bounded parameter space. Part (a) in Theorem 3 recovers their result when the loss is strongly smooth and Lipschitz continuous. Compared with [19], we need a further Lipschitz continuous assumption. However, this assumption can be removed when we assume the parameter domain is bounded in our setting. Indeed, the smoothness of f implies that the upper bound of the gradient can be controlled by the diameter of parameter domain R, i.e., $\|\partial f(\mathbf{w}; z)\|_2 \leq M + L\|\mathbf{w}\|_2 \leq M + LR$, where L is the smoothness parameter and $M = \sup_z \|\partial f(0; z)\|_2$. Hence, our result can achieve the optimal rate under the same assumptions as [19]. For the optimistic setting, i.e., $F(\mathbf{w}^*) = 0$, Part (b) in Theorem 3 removes the term $\mathcal{O}\big(\frac{1}{\sqrt{n}}\big)$ and further improves the rate to $\mathcal{O}\big(\frac{1}{n\epsilon}\sqrt{d\log(1/\delta)}\big)$ for strongly smooth losses under a low-noise condition. A very recent work [48] provided the same rate for the private gradient descent algorithm, while they focused on the non-convex setting and assumed Polyak-Łojasiewicz condition holds.

Now, we turn to the non-smooth case. The following theorem presents the utility bound of Algorithm 1 for α-Hölder smooth losses.

Theorem 4 (*Utility Guarantee for Non-smooth Losses*). *Let $\alpha \in [0, 1)$. Suppose Assumptions 1 and 2 hold and f is α-Hölder smooth with parameter L. Let $\mathbf{w}_{priv}$ be the output of Algorithm 1*

(a) Let $c > 0$ be a constant. If $\alpha \geq 1/2$, selecting $\eta_t = c\min\big\{\frac{1}{\sqrt{n}}, \frac{\epsilon}{\sqrt{d\log(1/\delta)}}\big\} \leq \min\{2/L, 1\}$ and $T \asymp n$. If $\alpha < 1/2$, selecting $\eta_t = c\min\big\{n^{\frac{3(\alpha-1)}{2(1+\alpha)}}, \frac{\epsilon}{\sqrt{d\log(1/\delta)}}\big\} \leq \min\{2/L, 1\}$ and $T \asymp n^{\frac{2-\alpha}{1+\alpha}}$. There holds

$$\mathbb{E}[F(\mathbf{w}_{priv})] - F(\mathbf{w}^*) = \mathcal{O}\Big(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\Big).$$

(b) Assume $F(\mathbf{w}^*) = 0$. Selecting $\eta_t = c\min\Big\{n^{\frac{\alpha^2+2\alpha-3}{2(1+\alpha)}}, \frac{n\epsilon}{T\sqrt{d\log(1/\delta)}}\Big\} \leq \min\{2/L, 1\}$ and $T \asymp n^{\frac{2}{1+\alpha}}$, where $c > 0$ is a constant. There holds

$$\mathbb{E}[F(\mathbf{w}_{priv})] - F(\mathbf{w}^*) = \mathcal{O}\Big(\frac{1}{n^{\frac{1+\alpha}{2}}} + \frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\Big).$$

Remark 3 (*[19]*). studied DP-SGD with gradient perturbation for α-Hölder smooth losses and established the optimal excess risk bound with gradient complexity $\mathcal{O}\big(n^{\frac{2-\alpha}{1+\alpha}} + n\big)$. Since gradient complexity is the total number of times the algorithm computes the gradient, then the gradient complexity of the SGD-type algorithm is equivalent to the number of iterations. Our result (Part (a) in Theorem 4) matches their bounds with the same gradient complexity. As discussed in Remark 2, although we need a further Lipschitz condition, we can also recover their result under the same setting when the parameter domain is bounded. Analogous to the smooth case, Part (b) in Theorem 4 derives the excess population risk bound better than $\mathcal{O}\big(\frac{1}{\sqrt{n}} + \frac{1}{n\epsilon}\sqrt{d\log(1/\delta)}\big)$. To the best of our knowledge, this is the first excess population risk bound of the order $\mathcal{O}\big(n^{-\frac{1+\alpha}{2}} + \frac{1}{n\epsilon}\sqrt{d\log(1/\delta)}\big)$ for private SGD when the loss is non-smooth.

3.2. DP-SGD for pairwise learning

In this subsection, we focus on the differentially private SGD algorithm for pairwise learning. The proposed differentially private SGD algorithm is described in Algorithm 2. In particular, in iteration t, the algorithm draws a pair $\{(i_t, j_t)\}$ from the uniform distribution over all pairs $\{(i, j) : i, j \in [n], i \neq j\}$. Then the parameter is updated by the noised gradient $\partial f(\mathbf{w}_t; z_{i_t}, z_{j_t}) + \mathbf{b}_t$ with $\mathbf{b}_t \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_d)$. The following theorem establishes the privacy guarantee for Algorithm 2.

Theorem 5 (*Privacy Guarantee*). *Suppose Assumptions 1 and 2 hold. Then Algorithm 2 with some $\beta \in (0, 1)$ satisfies (ϵ, δ)-DP if $\sigma^2 \geq 2.68G^2$ and $\lambda - 1 \leq \frac{\sigma^2}{6G^2}\log\Big(\frac{n}{2\lambda\big(1+\frac{\sigma^2}{4G^2}\big)}\Big)$ with $\lambda = \frac{\log(1/\delta)}{(1-\beta)\epsilon} + 1$.*

By combining the stability results and the optimization error bounds (Lemmas 19 and 20 below) together, we establish the following utility guarantees for Algorithm 2 for strongly smooth and non-smooth losses, respectively.

Theorem 6 (*Utility Guarantee for Smooth Losses*). *Suppose Assumptions 1 and 2 hold and f is L-smooth. Let $\mathbf{w}_{priv}$ be the output of Algorithm 2.*

(a) Let $c > 0$ be a constant. Selecting $\eta_t = c\min\Big\{\frac{1}{\sqrt{n}}, \frac{\epsilon}{\sqrt{d\log(1/\delta)}}\Big\} \leq \min\{2/L, 1\}$ and $T \asymp n$, there holds

$$\mathbb{E}[\bar{F}(\mathbf{w}_{priv})] - \bar{F}(\mathbf{w}^*) = \mathcal{O}\Big(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\Big).$$

(b) Let $c > 0$ be a constant. Assume $\bar{F}(\mathbf{w}^*) = 0$. Selecting $\eta_t = \frac{c\epsilon}{\sqrt{d\log(1/\delta)}} \leq \min\{2/L, 1\}$ and $T \asymp n$, there holds

$$\mathbb{E}[\bar{F}(\mathbf{w}_{priv})] - \bar{F}(\mathbf{w}^*) = \mathcal{O}\Big(\frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\Big).$$

Remark 4. Under the strongly smooth and Lipschitz continuous assumptions, [22] proposed the gradient descent with output perturbation algorithm to achieve DP and provided the excess population risk bound in the order of $\mathcal{O}\big(\frac{1}{\sqrt{n\epsilon}}\sqrt{d\log(1/\delta)}\big)$ with gradient complexity $\mathcal{O}(n^2)$. [21] proposed a localized gradient descent algorithm and proved that with a large gradient complexity $\mathcal{O}\big(n^3\log(1/\delta)\big)$, the optimal risk rate $\mathcal{O}\big(\frac{1}{\sqrt{n}} + \frac{1}{n\epsilon}\sqrt{d\log(1/\delta)}\big)$ can be achieved. [8] presented a simple localized DP-SGD algorithm and showed that their algorithm achieves the optimal rate up to a $\log(1/\delta)$ term. Their algorithm needs the gradient complexity $\mathcal{O}\big(n\log(1/\delta)\big)$. Our result (Part (a) in Theorem 6) shows that Algorithm 2 exactly achieves the optimal rate only with $T \asymp n$ iterations for strongly smooth losses, which significantly reduces the computational complexity of the algorithm. Under a low-noise condition, Part (b) removes the term $\mathcal{O}\big(\frac{1}{\sqrt{n}}\big)$ and derives the excess population risk bound of the order $\mathcal{O}\big(\frac{1}{n\epsilon}\sqrt{d\log(1/\delta)}\big)$ with $T \asymp n$ iterations. To the best of our knowledge, this is the first excess population risk bound in the order of $\mathcal{O}\big(\frac{1}{n\epsilon}\sqrt{d\log(1/\delta)}\big)$ for privacy-preserving pairwise learning.

The following theorem establishes the utility bounds for Algorithm 2 when the loss is non-smooth.

Theorem 7 (*Utility Guarantee for Non-smooth Losses*). *Let $\alpha \in [0, 1)$. Suppose Assumptions 1 and 2 hold and f is α-Hölder smooth with parameter L. Let $\mathbf{w}_{priv}$ be the output of Algorithm 2.*

(a) Let $c > 0$ be a constant. If $\alpha \geq 1/2$, we choose $\eta_t = c\min\Big\{\frac{1}{\sqrt{n}}, \frac{\epsilon}{\sqrt{d\log(1/\delta)}}\Big\} \leq \min\{2/L, 1\}$ and $T \asymp n$. If $\alpha < 1/2$, we choose $\eta_t = c\min\Big\{n^{\frac{3(\alpha-1)}{2(1+\alpha)}}, \frac{\epsilon}{\sqrt{d\log(1/\delta)}}\Big\} \leq \min\{2/L, 1\}$ and $T \asymp n^{\frac{2-\alpha}{1+\alpha}}$. There holds

$$\mathbb{E}[\bar{F}(\mathbf{w}_{priv})] - \bar{F}(\mathbf{w}^*) = \mathcal{O}\Big(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\Big).$$

(b) Assume $\bar{F}(\mathbf{w}^*) = 0$. Selecting $\eta_t = c\min\Big\{n^{\frac{\alpha^2+2\alpha-3}{2(1+\alpha)}}, \frac{n\epsilon}{T\sqrt{d\log(1/\delta)}}\Big\} \leq \min\{2/L, 1\}$ and $T \asymp n^{\frac{2}{1+\alpha}}$, where $c > 0$ is a constant. There holds

$$\mathbb{E}[\bar{F}(\mathbf{w}_{priv})] - \bar{F}(\mathbf{w}^*) = \mathcal{O}\Big(\frac{1}{n^{\frac{1+\alpha}{2}}} + \frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\Big).$$

Remark 5. Part (a) shows that the optimal excess risk rate can be achieved with $T \asymp n$ if $\alpha \geq 1/2$. For the case $\alpha < 1/2$, the same rate can be also achieved with $T \asymp n^{\frac{2-\alpha}{1+\alpha}}$. For non-smooth losses (i.e., $\alpha = 0$), [8] established the optimal excess population risk rate for localized DP-SGD algorithm with $T \asymp n^2\log(1/\delta)$ for Lipschitz continuity losses. Under the same assumptions, Part (a) with $\alpha = 0$ implies that the optimal rate can be achieved with $T \asymp n^2$. Our result reduces the computational cost by a factor of $\log(1/\delta)$ for this case. Part (b) establishes the first excess population risk bounds better than $\mathcal{O}\big(\frac{1}{\sqrt{n}} + \frac{1}{n\epsilon}\sqrt{d\log(1/\delta)}\big)$ in the case with low-noise for privacy-preserving pairwise learning.

4. Proofs of main results

We first introduce some definitions and useful lemmas. To establish a tighter privacy analysis of DP-SGD, we introduce the following definition of Rényi differential privacy (RDP). Given an algorithm $\mathcal{M}$ and a dataset S, let $P_{\mathcal{M}(S)}(\theta)$ denote the density of $\mathcal{M}(S)$.

Definition 5 (*RDP [49]*). Given a randomized algorithm $\mathcal{M}$. For $\lambda > 1$, $\kappa > 0$, if for all $S \sim \tilde{S}$, there holds $D_\lambda\big(\mathcal{M}(S), \mathcal{M}(\tilde{S})\big) := \frac{1}{\lambda-1}\log\int\Big(\frac{P_{\mathcal{M}(S)}(\theta)}{P_{\mathcal{M}(\tilde{S})}(\theta)}\Big)^\lambda dP_{\mathcal{M}(\tilde{S})}(\theta) \leq \kappa$, then $\mathcal{M}$ satisfies (λ, κ)-RDP.

The following lemma will help us develop privacy guarantees for DP-SGD algorithms.

Lemma 8 (*[50]*). *Given a function $\mathcal{G} : \mathcal{Z}^n \to \mathcal{W}$. Let $\Delta_\mathcal{G}$ be the ℓ_2-sensitivity of $\mathcal{G}$. For a dataset $S \subset \mathcal{Z}^n$ with size n, let $\mathcal{M}_\mathcal{G}(S)$ be a Gaussian mechanism with $\mathcal{M}_\mathcal{G}(S) = \mathcal{G}(S) + \mathbf{b}$, where $\mathbf{b} \sim \mathcal{N}(0, \sigma^2\mathbf{I}_d)$. Denote by S_{sub} the subset of S randomly selected from the uniform distribution over all subsets of S with size k. Let $\gamma = k/n$ be the subsampling rate. Then applying $\mathcal{M}_\mathcal{G}$ to S_{sub} satisfies $(\lambda, 3.5\gamma^2\lambda\Delta_\mathcal{G}^2/\sigma^2)$-RDP if $\sigma^2 \geq 0.67\Delta_\mathcal{G}^2$ and $\lambda - 1 \leq \frac{2\sigma^2}{3\Delta_\mathcal{G}^2}\log\big(\frac{1}{\lambda\gamma(1+\sigma^2/\Delta_\mathcal{G}^2)}\big)$.*

The following lemma provides the composition result of RDP, which is useful for analyzing the total privacy guarantee of an iterative algorithm.

Lemma 9 (*[49]*). *Let $\{\mathcal{M}_k\}_{k=1}^t$ be a sequence of mechanisms, where each $\mathcal{M}_k$ is chosen based on the outputs of $\mathcal{M}_1(S), \ldots, \mathcal{M}_{k-1}(S)$. If $\mathcal{M}_k$ satisfies (λ, κ_k)-RDP for $k \in [t]$, then a mechanism $\mathcal{M}$ consists of $\{\mathcal{M}_k\}_{k=1}^t$ is $(\lambda, \sum_{k=1}^t \kappa_k)$-RDP.*

The connection between RDP and DP is given as follows.

Lemma 10 (*[49]*). *If $\mathcal{M}$ is (λ, κ)-RDP, then $\mathcal{M}$ is $(\kappa + \log(1/\nu)/(\lambda - 1), \nu)$-DP for any $\nu \in (0, 1)$.*

A fundamental property of DP is called the post-processing property, which implies that analyzing the output of an algorithm satisfies differential privacy will not increase privacy leakage.

Lemma 11 (*[49]*). *Let $\mathcal{M} : \mathcal{Z}^n \to \mathcal{W}_1$ satisfies (λ, κ)-RDP. For any $f : \mathcal{W}_1 \to \mathcal{W}_2$, there holds $f \circ \mathcal{M} : \mathcal{Z}^n \to \mathcal{W}_2$ satisfies (λ, κ)-RDP.*

Let $M = \sup_{z\in\mathcal{Z}} f(0; z)$. Define

$$M_{\alpha,1} = \begin{cases} (1 + 1/\alpha)^{\frac{\alpha}{1+\alpha}} L^{\frac{1}{1+\alpha}}, & \text{if } \alpha > 0, \\ M + L, & \text{if } \alpha = 0. \end{cases} \tag{4}$$

Our analysis needs the following self-bounding property for strongly smooth and α-Hölder smooth losses [24,51].

Lemma 12 (*[24,51]*). *Suppose f is nonnegative. For any $\mathbf{w} \in \mathbb{R}^d, z \in \mathcal{Z}$, if f is L-strongly smooth, then $\|\partial f(\mathbf{w}; z)\|_2 \leq \sqrt{2Lf(\mathbf{w}; z)}$. Let $\alpha \in [0, 1)$. If f is α-Hölder smooth with $L > 0$, then for $M_{\alpha,1}$ defined in (4) there holds $\|\partial f(\mathbf{w}; z)\|_2 \leq M_{\alpha,1} f^{\frac{\alpha}{1+\alpha}}(\mathbf{w}; z)$.*

We will use the concept of on-average argument stability to study the generalization error.

Definition 6 (*[26]*). Let $S = \{z_i\}_{i=1}^n$ and $\tilde{S} = \{z_i'\}_{i=1}^n$, where z_i, z_i' are independently drawn from ρ. For any $i \in [n]$, let $S^{(i)} = \{z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_n\}$. We say $\mathcal{M}$ is on-average ν-argument-stable if $\mathbb{E}_{S,\tilde{S},\mathcal{M}}\big[\frac{1}{n}\sum_{i=1}^n \|\mathcal{M}(S) - \mathcal{M}(S^{(i)})\|_2^2\big] \leq \nu$,

4.1. Main proofs for pointwise learning

The proof of Theorem 2 (privacy analysis of Algorithm 1) is given as follows.

Proof of Theorem 2. To show $\mathbf{w}_{\text{priv}}$ satisfies DP, we first prove that the output of each iteration, i.e., $\mathbf{w}_t$, satisfies RDP. Specifically, for any $t \in [T]$, consider the mechanism $\mathcal{M}_t = \mathcal{G}_t + \mathbf{b}_t$, where $\mathcal{G}_t = \partial f(\mathbf{w}_t; z_{i_t})$. For any $z_{i_t}, z'_{i_t} \in \mathcal{Z}$, from the Lipschitz continuity of f we know that

$$\|\partial f(\mathbf{w}_t; z_{i_t}) - \partial f(\mathbf{w}_t; z'_{i_t})\|_2 \le \|\partial f(\mathbf{w}_t; z_{i_t})\|_2 + \|\partial f(\mathbf{w}_t; z'_{i_t})\|_2 \le 2G.$$

Then from the definition of sensitivity (see Definition 2), we can show that the ℓ_2-sensitivity of $\mathcal{G}_t$ is bounded by $2G$. Note that

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

According to Lemma 8 with subsampling rate $\gamma = 1/n$, we know $\mathcal{M}_t$ is $\big(\lambda, \frac{\lambda\beta\epsilon}{T(\frac{\log(1/\delta)}{(1-\beta)\epsilon}+1)}\big)$-RDP as long as $\sigma^2 \ge 2.68G^2$ and $\lambda - 1 \le \frac{\sigma^2}{6G^2}\log\Big(\frac{n}{\lambda\big(1+\frac{\sigma^2}{4G^2}\big)}\Big)$ hold. Then by Lemma 11 we can show that $\mathbf{w}_{t+1}$ is $\big(\lambda, \frac{\lambda\beta\epsilon}{T(\frac{\log(1/\delta)}{(1-\beta)\epsilon}+1)}\big)$-RDP. Further, we can use Lemma 9 to composite the output of T iterations and get that Algorithm 1 is $\Big(\lambda, \frac{\lambda\beta\epsilon}{\frac{\log(1/\delta)}{(1-\beta)\epsilon}+1}\Big)$-RDP. By choosing $\lambda = \frac{\log(1/\delta)}{(1-\beta)\epsilon} + 1$, we can obtain that Algorithm 1 is $\Big(\frac{\log(1/\delta)}{(1-\beta)\epsilon} + 1, \beta\epsilon\Big)$-RDP. Finally, the relationship between RDP and DP (Lemma 10) implies that Algorithm 1 is (ϵ, δ)-DP if $\sigma^2 \ge 2.68G^2$ and $\lambda - 1 \le \frac{\sigma^2}{6G^2}\log\Big(\frac{n}{\lambda\big(1+\frac{\sigma^2}{4G^2}\big)}\Big)$ hold, which completes the proof. □

To study the utility guarantee of Algorithm 1, we need to estimate the generalization error $\mathbb{E}[F(\mathbf{w}_{\text{priv}}) - F_S(\mathbf{w}_{\text{priv}})]$ and the optimization error $\mathbb{E}[F_S(\mathbf{w}_{\text{priv}}) - F(\mathbf{w}^*)]$, respectively. We consider using on-average argument stability to control the generalization error. The relationship between on-average argument stability and the generalization error is given as follows [26].

Lemma 13 ([26]). *Let $\mathcal{M}$ be on-average ν-argument-stable. Let $\kappa > 0$.*

(a) Assume f is L-smooth. There holds

$$\mathbb{E}[F(\mathcal{M}(S)) - F_S(\mathcal{M}(S))] \le \frac{L}{\kappa}\mathbb{E}[F_S(\mathcal{M}(S))] + \frac{(L+\kappa)\nu}{2}.$$

(b) If Assumption 1 holds and f is α-Hölder smooth with parameter L and $\alpha \in [0, 1)$, then

$$\mathbb{E}[F(\mathcal{M}(S)) - F_S(\mathcal{M}(S))] \le \frac{M_{\alpha,1}^2}{2\kappa}\mathbb{E}[F^{\frac{2\alpha}{1+\alpha}}(\mathcal{M}(S))] + \frac{\kappa\nu}{2}.$$

When analyzing the stability of DP-SGD, we consider perturbing the stochastic gradient by the same Gaussian noise sequence for the neighboring datasets. Then the on-average argument stability of non-private SGD equals that of private SGD. We can use the following lemma directly to give the stability bounds of Algorithm 3.1 for both strongly smooth and non-smooth losses [26].

Lemma 14. *Suppose Assumption 1 holds. Let $\mathcal{M}$ be Algorithm 1 with T iterations.*

(a) If f is L-smooth and $\eta_j \le 2/L$, then $\mathcal{M}$ is on-average ν-argument-stable with

$$\nu \le \frac{8e(1+T/n)L}{n}\sum_{j=1}^{T}\eta_j^2\mathbb{E}[F_S(\mathbf{w}_j)].$$

(b) Let $\alpha \in [0, 1)$. If f is α-Hölder smooth with parameter L, then $\mathcal{M}$ is on-average ν-argument-stable with

$$\nu \le M_{\alpha,3}^2 e\sum_{j=1}^{T}\eta_j^{\frac{2}{1-\alpha}} + \frac{4eM_{\alpha,1}^2(1+T/n)}{n}\sum_{j=1}^{T}\eta_j^2\mathbb{E}\big[F_S^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_j)\big],$$

where $M_{\alpha,3} = \sqrt{\frac{1-\alpha}{1+\alpha}}(2^{-\alpha}L)^{\frac{1}{1-\alpha}}$.

The following theorem presents the generalization bounds of DP-SGD for both smooth and non-smooth losses, which directly follows from Lemmas 13 and 14.

Theorem 15. *Suppose Assumption 1 holds. Let $\mathbf{w}_{priv}$ be the output of Algorithm 1. Let $\kappa > 0$.*

(a) If f is L-smooth and $\eta_t \le 2/L$ for all $t \in [T]$, there holds

$$\mathbb{E}[F(\mathbf{w}_{priv}) - F_S(\mathbf{w}_{priv})] \le \frac{L}{\kappa}\mathbb{E}[F_S(\mathbf{w}_{priv})] + \frac{4e(L+\kappa)(1+T/n)L}{n}\sum_{t=1}^{T}\eta_t^2\mathbb{E}[F_S(\mathbf{w}_t)].$$

(b) Let $\alpha \in [0, 1)$. If f is α-Hölder smooth with parameter L, there holds

$$\mathbb{E}[F(\mathbf{w}_{priv}) - F_S(\mathbf{w}_{priv})] \le \frac{M_{\alpha,1}^2}{2\kappa}\mathbb{E}[F^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_{priv})] + \frac{\kappa}{2}\Big(M_{\alpha,3}^2 e\sum_{t=1}^{T}\eta_t^{\frac{2}{1-\alpha}} + \frac{4eM_{\alpha,1}^2(1+T/n)}{n}\sum_{t=1}^{T}\eta_t^2\mathbb{E}\Big[F_S^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_t)\Big]\Big).$$

Recall $\mathbf{w}^* = \arg\min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w})$. Let

$$M_{\alpha,2} = \begin{cases} \frac{1-\alpha}{1+\alpha}\big(2\alpha/(1+\alpha)\big)^{\frac{2\alpha}{1-\alpha}}M_{\alpha,1}^{\frac{2+2\alpha}{1-\alpha}}, & \text{if } \alpha > 0 \\ M_{\alpha,1}^2, & \text{if } \alpha = 0. \end{cases} \tag{5}$$

The following theorem establishes the optimization error for both smooth and non-smooth losses.

Theorem 16. *Suppose Assumption 1 holds. Let $\{\mathbf{w}_j\}$ be produced by Algorithm 1. Assume the step size η_j is nonincreasing.*

(a) If f is L-smooth, then

$$\sum_{j=1}^{t}\eta_j\mathbb{E}[F_S(\mathbf{w}_j) - F_S(\mathbf{w}^*)] \le \big(\frac{1}{2} + 3L\eta_1\big)\|\mathbf{w}^*\|_2^2 + 3L\sum_{j=1}^{t}\big(3\eta_j^3\sigma^2 d + 2\eta_j^2 F(\mathbf{w}^*)\big) + \sum_{j=1}^{t}3\eta_j^2\sigma^2 d.$$

(b) Let $\alpha \in [0, 1)$. If f is α-Hölder smooth with parameter L, there holds

$$\sum_{j=1}^{t}\eta_j\mathbb{E}[F_S(\mathbf{w}_j) - F_S(\mathbf{w}^*)] \le \frac{3}{4}M_{\alpha,1}^2\Big(\sum_{j=1}^{t}\eta_j^2\Big)^{\frac{1-\alpha}{1+\alpha}}\Big[2\eta_1\|\mathbf{w}^*\|_2^2 + \sum_{j=1}^{t}\big(6\eta_j^3\sigma^2 d + 4\eta_j^2 F(\mathbf{w}^*) + 3M_{\alpha,2}\eta_j^{\frac{3-\alpha}{1-\alpha}}\big)\Big]^{\frac{2\alpha}{1+\alpha}} + \frac{1}{2}\|\mathbf{w}^*\|_2^2 + \sum_{j=1}^{t}3\eta_j^2\sigma^2 d.$$

Proof. Note the projection operator Proj is non-expansive. Then for any $\alpha \in [0, 1]$, we have

$$\begin{aligned}
&\|\mathbf{w}_{t+1} - \mathbf{w}^*\|_2^2 \le \|\mathbf{w}_t - \eta_t(\partial f(\mathbf{w}_t; z_{i_t}) + \mathbf{b}_t) - \mathbf{w}^*\|_2^2 \\
&= \|\mathbf{w}_t - \mathbf{w}^*\|_2^2 + \eta_t^2\|\partial f(\mathbf{w}_t; z_{i_t}) + \mathbf{b}_t\|_2^2 + 2\eta_t\langle\mathbf{w}^* - \mathbf{w}_t, \partial f(\mathbf{w}_t; z_{i_t}) + \mathbf{b}_t\rangle \\
&\le \|\mathbf{w}_t - \mathbf{w}^*\|_2^2 + \frac{3}{2}\eta_t^2\|\partial f(\mathbf{w}_t; z_{i_t})\|_2^2 + 3\eta_t^2\|\mathbf{b}_t\|_2^2 \\
&\quad + 2\eta_t\langle\mathbf{w}^* - \mathbf{w}_t, \partial f(\mathbf{w}_t; z_{i_t}) + \mathbf{b}_t\rangle \\
&\le \|\mathbf{w}_t - \mathbf{w}^*\|_2^2 + \frac{3}{2}M_{\alpha,1}^2\eta_t^2 f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_t; z_{i_t}) + 3\eta_t^2\|\mathbf{b}_t\|_2^2 \\
&\quad + 2\eta_t\big(f(\mathbf{w}^*; z_{i_t}) - f(\mathbf{w}_t; z_{i_t})\big) + 2\eta_t\langle\mathbf{w}^* - \mathbf{w}_t, \mathbf{b}_t\rangle,
\end{aligned} \tag{6}$$

where in the second inequality we used $(a+b)^2 \le (1+\mu)a^2 + (1+1/\mu)b^2$ with $\mu = 1/2$, and the last inequality is due to the self-bounding property (Lemma 12) and the convexity of f. Rearranging the above inequality, we get

$$2\eta_t[f(\mathbf{w}_t; z_{i_t}) - f(\mathbf{w}^*; z_{i_t})] \le \|\mathbf{w}_t - \mathbf{w}^*\|_2^2 - \|\mathbf{w}_{t+1} - \mathbf{w}^*\|_2^2$$

$$+\frac{3}{2}M_{\alpha,1}^2\eta_t^2 f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_t;z_{i_t})+3\eta_t^2\|\mathbf{b}_t\|_2^2+2\eta_t\langle\mathbf{w}^*-\mathbf{w}_t,\mathbf{b}_t\rangle.$$

Taking a summation over j and noting $\mathbf{w}_1=\mathbf{0}$, we know

$$2\sum_{j=1}^t\eta_j[f(\mathbf{w}_j;z_{i_j})-f(\mathbf{w}^*;z_{i_j})]\le\|\mathbf{w}^*\|_2^2+\frac{3}{2}M_{\alpha,1}^2\sum_{j=1}^t\eta_j^2 f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_j;z_{i_j})+\sum_{j=1}^t\big(3\eta_j^2\|\mathbf{b}_j\|_2^2+2\eta_j\langle\mathbf{w}^*-\mathbf{w}_j,\mathbf{b}_j\rangle\big).$$

Note that $\mathbf{w}_j$ is independent of i_j, we can take an expectation and get

$$\sum_{j=1}^t\eta_j\mathbb{E}[F_S(\mathbf{w}_j)-F_S(\mathbf{w}^*)]=\sum_{j=1}^t\eta_j\mathbb{E}[f(\mathbf{w}_j;z_{i_j})-f(\mathbf{w}^*;z_{i_j})]$$
$$\le\frac{1}{2}\|\mathbf{w}^*\|_2^2+\frac{3}{4}M_{\alpha,1}^2\sum_{j=1}^t\eta_j^2\mathbb{E}[f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_j;z_{i_j})]+\sum_{j=1}^t3\eta_j^2\sigma^2d,\tag{7}$$

where we used $\mathbb{E}[\|\mathbf{b}_j\|_2^2]=\sigma^2d$ and $\mathbb{E}[\langle\mathbf{w}^*-\mathbf{w}_j,\mathbf{b}_j\rangle]=0$ since $\mathbf{b}_j$ is a Gaussian vector with mean 0 and variance σ^2, and $\mathbf{w}^*-\mathbf{w}_j$ is independent of $\mathbf{b}_j$.

To control the right hand side of (7), we have to estimate $\sum_{j=1}^t\eta_j^2\mathbb{E}[f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_j;z_{i_j})]$. By Young's inequality $ab\le\mu^{-1}|a|^\mu+\nu^{-1}|b|^\nu$ with $a,b\in\mathbb{R}$ and $\mu^{-1}+\nu^{-1}=1$, for any $t\in[T]$ we have

$$\eta_tM_{\alpha,1}^2f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_t;z_{i_t})=\Big(\frac{1+\alpha}{2\alpha}f(\mathbf{w}_t;z_{i_t})\Big)^{\frac{2\alpha}{1+\alpha}}\Big(\frac{2\alpha}{1+\alpha}\Big)^{\frac{2\alpha}{1+\alpha}}M_{\alpha,1}^2\eta_t$$
$$\le\frac{2\alpha}{1+\alpha}\Big(\frac{1+\alpha}{2\alpha}f(\mathbf{w}_t;z_{i_t})\Big)+\frac{1-\alpha}{1+\alpha}\Big(\big(\frac{2\alpha}{1+\alpha}\big)^{\frac{2\alpha}{1+\alpha}}M_{\alpha,1}^2\eta_t\Big)^{\frac{1+\alpha}{1-\alpha}}$$
$$=f(\mathbf{w}_t;z_{i_t})+M_{\alpha,2}\eta_t^{\frac{1+\alpha}{1-\alpha}}.$$

Putting the above inequality back into (6) yields

$$\|\mathbf{w}_{t+1}-\mathbf{w}^*\|_2^2\le\|\mathbf{w}_t-\mathbf{w}^*\|_2^2+3\eta_t^2\|\mathbf{b}_t\|_2^2+2\eta_tf(\mathbf{w}^*;z_{i_t})-\frac{1}{2}\eta_tf(\mathbf{w}_t;z_{i_t})+\frac{3}{2}M_{\alpha,2}\eta_t^{\frac{2}{1-\alpha}}+2\eta_t\langle\mathbf{w}^*-\mathbf{w}_t,\mathbf{b}_t\rangle.$$

Rearranging the above inequality and multiplying both sides by η_t, we get

$$\eta_t^2f(\mathbf{w}_t;z_{i_t})$$
$$\le2\eta_t\big(\|\mathbf{w}_t-\mathbf{w}^*\|_2^2-\|\mathbf{w}_{t+1}-\mathbf{w}^*\|_2^2\big)+6\eta_t^3\|\mathbf{b}_t\|_2^2+4\eta_t^2f(\mathbf{w}^*;z_{i_t})+3M_{\alpha,2}\eta_t^{\frac{3-\alpha}{1-\alpha}}+4\eta_t^2\langle\mathbf{w}^*-\mathbf{w}_t,\mathbf{b}_t\rangle$$
$$\le2\eta_t\|\mathbf{w}_t-\mathbf{w}^*\|_2^2-2\eta_{t+1}\|\mathbf{w}_{t+1}-\mathbf{w}^*\|_2^2+6\eta_t^3\|\mathbf{b}_t\|_2^2+4\eta_t^2f(\mathbf{w}^*;z_{i_t})+3M_{\alpha,2}\eta_t^{\frac{3-\alpha}{1-\alpha}}+4\eta_t^2\langle\mathbf{w}^*-\mathbf{w}_t,\mathbf{b}_t\rangle,$$

where we assume $\eta_t\ge\eta_{t+1}$ for all $t\in[T-1]$.

Taking a summation over j and noting $\mathbf{w}_1=\mathbf{0}$, we know

$$\sum_{j=1}^t\eta_j^2f(\mathbf{w}_j;z_{i_t})\le2\eta_1\|\mathbf{w}^*\|_2^2+\sum_{j=1}^t\big(6\eta_j^3\|\mathbf{b}_j\|_2^2+4\eta_j^2f(\mathbf{w}^*;z_{i_j})+3M_{\alpha,2}\eta_j^{\frac{3-\alpha}{1-\alpha}}+4\eta_j^2\langle\mathbf{w}^*-\mathbf{w}_j,\mathbf{b}_j\rangle\big).\tag{8}$$

Note $x\mapsto x^{\frac{2\alpha}{1+\alpha}}$ is concave. Then Jensen's inequality implies

$$\sum_{j=1}^t\eta_j^2f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_j;z_{i_j})\le\sum_{j=1}^t\eta_j^2\Big(\frac{\sum_{j=1}^t\eta_j^2f(\mathbf{w}_j;z_{i_j})}{\sum_{j=1}^t\eta_j^2}\Big)^{\frac{2\alpha}{1+\alpha}}$$
$$=\Big(\sum_{j=1}^t\eta_j^2\Big)^{\frac{1-\alpha}{1+\alpha}}\Big[\sum_{j=1}^t\eta_j^2f(\mathbf{w}_j;z_{i_j})\Big]^{\frac{2\alpha}{1+\alpha}}$$
$$\le\Big(\sum_{j=1}^t\eta_j^2\Big)^{\frac{1-\alpha}{1+\alpha}}\Big[2\eta_1\|\mathbf{w}^*\|_2^2+\sum_{j=1}^t\big(6\eta_j^3\|\mathbf{b}_j\|_2^2+4\eta_j^2f(\mathbf{w}^*;z_{i_j})+3M_{\alpha,2}\eta_j^{\frac{3-\alpha}{1-\alpha}}+4\eta_j^2\langle\mathbf{w}^*-\mathbf{w}_j,\mathbf{b}_j\rangle\big)\Big]^{\frac{2\alpha}{1+\alpha}}.\tag{9}$$

Plugging the above inequality back into (7) and noting that $\mathbb{E}[F_S(\mathbf{w}^*)]=F(\mathbf{w}^*)$, we have

$$\sum_{j=1}^t\eta_j\mathbb{E}[F_S(\mathbf{w}_j)-F_S(\mathbf{w}^*)]$$
$$\le\frac{1}{2}\|\mathbf{w}^*\|_2^2+\sum_{j=1}^t3\eta_j^2\sigma^2d+\frac{3}{4}M_{\alpha,1}^2\Big(\sum_{j=1}^t\eta_j^2\Big)^{\frac{1-\alpha}{1+\alpha}}\mathbb{E}\Big[2\eta_1\|\mathbf{w}^*\|_2^2+\sum_{j=1}^t\big(6\eta_j^3\|\mathbf{b}_j\|_2^2+4\eta_j^2f(\mathbf{w}^*;z_{i_j})+3M_{\alpha,2}\eta_j^{\frac{3-\alpha}{1-\alpha}}+4\eta_j^2\langle\mathbf{w}^*-\mathbf{w}_j,\mathbf{b}_j\rangle\big)\Big]^{\frac{2\alpha}{1+\alpha}}$$
$$\le\frac{1}{2}\|\mathbf{w}^*\|_2^2+\frac{3}{4}M_{\alpha,1}^2\Big(\sum_{j=1}^t\eta_j^2\Big)^{\frac{1-\alpha}{1+\alpha}}\Big[2\eta_1\|\mathbf{w}^*\|_2^2+\sum_{j=1}^t\big(6\eta_j^3\sigma^2d+4\eta_j^2F(\mathbf{w}^*)+3M_{\alpha,2}\eta_j^{\frac{3-\alpha}{1-\alpha}}\big)\Big]^{\frac{2\alpha}{1+\alpha}}+\sum_{j=1}^t3\eta_j^2\sigma^2d,$$

where the last inequality used Jensen's inequality for concave mapping and $\mathbb{E}_{\mathcal{A}}[\langle\mathbf{w}^*-\mathbf{w}_j,\mathbf{b}_j\rangle]=0$. Part (b) is proved. From the definition we know that α-Hölder smoothness with $\alpha=1$ corresponds to the strongly smoothness of f. Hence, Part (a) in the theorem directly follows by setting $\alpha=1$ in the above inequality. □

Now, we can establish the proofs of the excess population risk bounds of DP-SGD for pointwise learning by combining Theorems 15 and 16 together.

Proof of Theorem 3. Part (a) in Lemma 14 implies

$$\frac{1}{n}\sum_{i=1}^n\mathbb{E}\big[\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2^2\big]\le\frac{8e(1+t/n)L}{n}\sum_{j=1}^t\eta_j^2\mathbb{E}[F_S(\mathbf{w}_j)].$$

Plugging the above stability bounds back into Part (a) of Lemma 13, we get

$$\mathbb{E}[F(\mathbf{w}_{t+1})]\le\Big(1+\frac{L}{\kappa}\Big)\mathbb{E}[F_S(\mathbf{w}_{t+1})]+\frac{4e(L+\kappa)(1+t/n)L}{n}\sum_{j=1}^t\eta_j^2\mathbb{E}[F_S(\mathbf{w}_j)].\tag{10}$$

Note that $\mathbf{w}_j$ is independent of $\mathbf{b}_j$ and i_j. Eq. (8) implies

$$\sum_{j=1}^t\eta_j^2\mathbb{E}[F_S(\mathbf{w}_j)]=\sum_{j=1}^t\eta_j^2\mathbb{E}[f(\mathbf{w}_j;z_{i_t})]$$
$$\le2\eta_1\|\mathbf{w}^*\|_2^2+\sum_{j=1}^t\big(6\eta_j^3\mathbb{E}[\|\mathbf{b}_j\|_2^2]+4\eta_j^2\mathbb{E}[f(\mathbf{w}^*;z_{i_j})]+4\eta_j^2\mathbb{E}[\langle\mathbf{w}^*-\mathbf{w}_j,\mathbf{b}_j\rangle]\big)$$
$$\le2\eta_1\|\mathbf{w}^*\|_2^2+\sum_{j=1}^t\big(6\eta_j^3\sigma^2d+4\eta_j^2\ F(\mathbf{w}^*)\big),\tag{11}$$

where we used $\mathbb{E}[\|\mathbf{b}_j\|_2^2]=\sigma^2d$, $\mathbb{E}[f(\mathbf{w}^*;z_{i_j})]=F(\mathbf{w}^*)$ and $\mathbb{E}[\langle\mathbf{w}^*-\mathbf{w}_j,\mathbf{b}_j\rangle]=0$.

Plugging (11) back into (10) and summing over t yield

$$\sum_{t=1}^T\eta_t\mathbb{E}[F(\mathbf{w}_t)]\le\Big(1+\frac{L}{\kappa}\Big)\sum_{t=1}^T\eta_t\mathbb{E}[F_S(\mathbf{w}_t)]+\frac{8e(L+\kappa)(1+T/n)L}{n}\times\sum_{t=1}^T\eta_t\Big[\eta_1\|\mathbf{w}^*\|_2^2+\sum_{j=1}^t\big(3\eta_j^3\sigma^2d+2\eta_j^2\ F(\mathbf{w}^*)\big)\Big].\tag{12}$$

Combining (12) with Part (a) in Theorem 16 we can get

$$\sum_{t=1}^T\eta_t\mathbb{E}[F(\mathbf{w}_t)]$$
$$\le\Big(1+\frac{L}{\kappa}\Big)\Big(\sum_{t=1}^T\eta_tF(\mathbf{w}^*)+\Big(\frac{1}{2}+3L\eta_1\Big)\|\mathbf{w}^*\|_2^2+3\sum_{j=1}^t\big(3L\eta_j+1\big)\eta_j^2\sigma^2d+6\sum_{j=1}^t\eta_j^2F(\mathbf{w}^*)\Big)+\frac{8e(L+\kappa)(1+T/n)L}{n}\sum_{t=1}^T\eta_t\Big[\eta_1\|\mathbf{w}^*\|_2^2$$

$$+\sum_{j=1}^{t}\big(3\eta_j^3\sigma^2 d+2\eta_j^2\ F(\mathbf{w}^*)\big)\Big].$$

Let $\eta_t=\eta\le\min\{2/L,1\}$. Further, we assume $T\ge n$. Note $\mathbf{w}_{\text{priv}}=\frac{1}{T}\sum_{t=1}^{T}\mathbf{w}_t$. Then according to Jensen's inequality, there holds

$$\begin{aligned}&\mathbb{E}[F(\mathbf{w}_{\text{priv}})-F(\mathbf{w}^*)]\\&=\mathcal{O}\Big(\Big(\frac{(1+\kappa^{-1})}{T\eta}+\frac{(1+\kappa)T\eta}{n^2}\Big)\|\mathbf{w}^*\|_2^2+\big(1+\kappa^{-1}\big)\sigma^2 d\eta+\frac{(1+\kappa)T^2\eta^3\sigma^2 d}{n^2}\\&\quad+\Big(\kappa^{-1}+\big(1+\kappa^{-1}\big)\eta+\frac{(1+\kappa)T^2\eta^2}{n^2}\Big)F(\mathbf{w}^*)\Big).\end{aligned}$$

Recalling that $\sigma^2 d=\frac{14G^2Td}{\beta n^2\epsilon}\Big(\frac{\log(1/\delta)}{(1-\beta)\epsilon}+1\Big)$, we further have

$$\begin{aligned}&\mathbb{E}[F(\mathbf{w}_{\text{priv}})-F(\mathbf{w}^*)]\\&=\mathcal{O}\Big(\Big(\frac{(1+\kappa^{-1})}{T\eta}+\frac{(1+\kappa)T\eta}{n^2}\Big)\|\mathbf{w}^*\|_2^2\\&\quad+\Big(\kappa^{-1}+\frac{T^2\eta^2(1+\kappa)}{n^2}+\big(\kappa^{-1}+1\big)\eta\Big)\\&\quad\times F(\mathbf{w}^*)+\Big(\big(1+\kappa^{-1}\big)\eta+\frac{T^2\eta^3(1+\kappa)}{n^2}\Big)\frac{Td\log(1/\delta)}{n^2\epsilon^2}\Big).\end{aligned}\tag{13}$$

(a) If we set $T\asymp n$ and $\kappa=\sqrt{n}$, then Eq. (13) implies

$$\begin{aligned}&\mathbb{E}[F(\mathbf{w}_{\text{priv}})-F(\mathbf{w}^*)]\\&=\mathcal{O}\Big(\Big(\frac{1}{\sqrt{n}}+\eta^2\sqrt{n}+\eta\Big)F(\mathbf{w}^*)+\big(\frac{1}{n\eta}+\frac{\eta}{\sqrt{n}}\big)\|\mathbf{w}^*\|_2^2\\&\quad+\big(\eta+\eta^3\sqrt{n}\big)\frac{d\log(1/\delta)\eta}{n\epsilon^2}\Big).\end{aligned}$$

Further let $\eta_t=c/\max\Big\{\sqrt{n},\frac{\sqrt{d\log(1/\delta)}}{\epsilon}\Big\}\le\min\{2/L,1\}$ for some constant $c>0$, then there holds

$$\mathbb{E}[F(\mathbf{w}_{\text{priv}})-F(\mathbf{w}^*)]=\mathcal{O}\Big(\frac{1}{\sqrt{n}}+\frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\Big),$$

where we assume $\sqrt{d\log(1/\delta)}=\mathcal{O}(n\epsilon)$ (otherwise the bound will not converge).

(b) Consider the low noise case, i.e, $F(\mathbf{w}^*)=0$. Let $\kappa=1$. We can choose $T\asymp n$ and get

$$\mathbb{E}[F(\mathbf{w}_{\text{priv}})-F(\mathbf{w}^*)]=\mathcal{O}\Big(\big(\frac{1}{n\eta}+\frac{\eta}{n}\big)\|\mathbf{w}^*\|_2^2+\frac{d\log(1/\delta)\eta}{n\epsilon^2}\Big).$$

Let $\eta_t=\frac{c\epsilon}{\sqrt{d\log(1/\delta)}}\le\min\{2/L,1\}$ for some constant $c>0$, then

$$\mathbb{E}[F(\mathbf{w}_{\text{priv}})-F(\mathbf{w}^*)]=\mathcal{O}\Big(\frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\Big).$$

The proof is completed. □

Now, we establish the utility guarantee of Algorithm 1 for non-smooth losses.

Proof of Theorem 4. Recall that $\mathbf{w}_{\text{priv}}=\frac{1}{T}\sum_{t=1}^{T}\mathbf{w}_t$. Jensen's inequality implies that

$$\begin{aligned}&\mathbb{E}[F(\mathbf{w}_{\text{priv}})]-F(\mathbf{w}^*)=\big(\sum_{t=1}^{T}\eta_t\big)^{-1}\sum_{t=1}^{T}\eta_t\mathbb{E}[F(\mathbf{w}_t)-F(\mathbf{w}^*)]\\&=\big(\sum_{t=1}^{T}\eta_t\big)^{-1}\sum_{t=1}^{T}\eta_t\mathbb{E}[F(\mathbf{w}_t)-F_S(\mathbf{w}_t)]\\&\quad+\big(\sum_{t=1}^{T}\eta_t\big)^{-1}\sum_{t=1}^{T}\eta_t\mathbb{E}[F_S(\mathbf{w}_t)-F(\mathbf{w}^*)].\end{aligned}\tag{14}$$

We first estimate $\big(\sum_{t=1}^{T}\eta_t\big)^{-1}\sum_{t=1}^{T}\eta_t\mathbb{E}[F(\mathbf{w}_t)-F_S(\mathbf{w}_t)]$. Putting part (b) in Lemma 14 back into part (b) of Lemma 13, we get

$$\begin{aligned}&\mathbb{E}[F(\mathbf{w}_{t+1})-F_S(\mathbf{w}_{t+1})]\\&\le\frac{M_{\alpha,1}^2}{2\kappa}\mathbb{E}[F^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_{t+1})]+\frac{2eM_{\alpha,1}^2\kappa(1+t/n)}{n}\sum_{j=1}^{t}\eta_j^2\mathbb{E}\Big[F_S^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_j)\Big]\\&\quad+\frac{eM_{\alpha,3}^2\kappa}{2}\sum_{j=1}^{t}\eta_j^{\frac{2}{1-\alpha}}.\end{aligned}$$

Let $\xi_j=\max\big\{\mathbb{E}[F(\mathbf{w}_j)]-\mathbb{E}[F_S(\mathbf{w}_j)],0\big\}$. Since $x\mapsto x^{\frac{2\alpha}{1+\alpha}}$ is concave, then there holds

$$\begin{aligned}\mathbb{E}[F^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_{t+1})]&\le\big(\mathbb{E}[F(\mathbf{w}_{t+1})]-\mathbb{E}[F_S(\mathbf{w}_{t+1})]+\mathbb{E}[F_S(\mathbf{w}_{t+1})]\big)^{\frac{2\alpha}{1+\alpha}}\\&\le\xi_{t+1}^{\frac{2\alpha}{1+\alpha}}+\big(\mathbb{E}[F_S(\mathbf{w}_{t+1})]\big)^{\frac{2\alpha}{1+\alpha}}.\end{aligned}$$

Combining the above two inequalities together yields

$$\begin{aligned}\xi_{t+1}\le&\frac{M_{\alpha,1}^2}{2\kappa}\big(\xi_{t+1}^{\frac{2\alpha}{1+\alpha}}+\big(\mathbb{E}[F_S(\mathbf{w}_{t+1})]\big)^{\frac{2\alpha}{1+\alpha}}\big)+e\kappa\sum_{j=1}^{t}\Big(\frac{M_{\alpha,3}^2}{2}\eta_j^{\frac{2}{1-\alpha}}\\&+\frac{2M_{\alpha,1}^2(1+\frac{t}{n})\eta_j^2}{n}\big(\mathbb{E}[F_S(\mathbf{w}_j)]\big)^{\frac{2\alpha}{1+\alpha}}\Big).\end{aligned}$$

The solution of the above inequality is

$$\begin{aligned}\xi_{t+1}=&\mathcal{O}\Big(\kappa^{\frac{1+\alpha}{\alpha-1}}+\frac{1}{\kappa}\big(\mathbb{E}[F_S(\mathbf{w}_{t+1})]\big)^{\frac{2\alpha}{1+\alpha}}+\kappa\sum_{j=1}^{t}\eta_j^{\frac{2}{1-\alpha}}\\&+\kappa\Big(\frac{1}{n}+\frac{T}{n^2}\Big)\sum_{j=1}^{t}\eta_j^2\big(\mathbb{E}[F_S(\mathbf{w}_j)]\big)^{\frac{2\alpha}{1+\alpha}}\Big).\end{aligned}$$

We assume $T\ge n$ and set $\eta_t=\eta$. From the definition of ξ_{t+1} we can get

$$\begin{aligned}&\big(T\eta\big)^{-1}\sum_{t=1}^{T}\eta_t\mathbb{E}[F(\mathbf{w}_t)-F_S(\mathbf{w}_t)]=\mathcal{O}\Big(\kappa^{\frac{1+\alpha}{\alpha-1}}+\kappa T\eta^{\frac{2}{1-\alpha}}\\&+\kappa Tn^{-2}\sum_{t=1}^{T}\eta^2\big(\mathbb{E}[F_S(\mathbf{w}_t)]\big)^{\frac{2\alpha}{1+\alpha}}+\big(\kappa T\eta\big)^{-1}\sum_{t=1}^{T}\eta\big(\mathbb{E}[F_S(\mathbf{w}_{t+1})]\big)^{\frac{2\alpha}{1+\alpha}}\Big).\end{aligned}\tag{15}$$

Since $\mathbb{E}[\langle\mathbf{w}^*-\mathbf{w}_t,\mathbf{b}_t\rangle]=0$, Eq. (8) with $\eta_t=\eta$ implies

$$\begin{aligned}&\sum_{t=1}^{T}\eta^2\big(\mathbb{E}[F_S(\mathbf{w}_t)]\big)^{\frac{2\alpha}{1+\alpha}}\le\sum_{t=1}^{T}\eta^2\Big(\frac{\sum_{t=1}^{T}\eta^2\mathbb{E}[F_S(\mathbf{w}_t)]}{\sum_{t=1}^{T}\eta^2}\Big)^{\frac{2\alpha}{1+\alpha}}\\&\le\big(T\eta^2\big)^{\frac{1-\alpha}{1+\alpha}}\Big(2\eta\|\mathbf{w}^*\|_2^2+6T\eta^3\sigma^2 d+4T\eta^2F(\mathbf{w}^*)+3M_{\alpha,2}T\eta^{\frac{3-\alpha}{1-\alpha}}\Big)^{\frac{2\alpha}{1+\alpha}}\\&=\mathcal{O}\Big(\big(T\eta^2\big)^{\frac{1-\alpha}{1+\alpha}}\Big(\eta+T\eta^3\sigma^2 d+T\eta^2F(\mathbf{w}^*)+T\eta^{\frac{3-\alpha}{1-\alpha}}\Big)^{\frac{2\alpha}{1+\alpha}}\Big).\end{aligned}$$

Dividing both sides by η, we get

$$\begin{aligned}&\sum_{t=1}^{T}\eta\big(\mathbb{E}[F_S(\mathbf{w}_t)]\big)^{\frac{2\alpha}{1+\alpha}}\\&=\mathcal{O}\Big(T^{\frac{1-\alpha}{1+\alpha}}\eta^{\frac{1-3\alpha}{1+\alpha}}\Big(\eta+T\eta^3\sigma^2 d+T\eta^2F(\mathbf{w}^*)+T\eta^{\frac{3-\alpha}{1-\alpha}}\Big)^{\frac{2\alpha}{1+\alpha}}\Big).\end{aligned}$$

Now, plugging the above two inequalities back into (15), we have

$$\begin{aligned}&\big(T\eta\big)^{-1}\sum_{t=1}^{T}\eta_t\mathbb{E}[F(\mathbf{w}_t)-F_S(\mathbf{w}_t)]\\&=\mathcal{O}\Big(\big(\kappa T\eta\big)^{-1}T^{\frac{1-\alpha}{1+\alpha}}\eta^{\frac{1-3\alpha}{1+\alpha}}\big(\eta+T\eta^3\sigma^2 d+\eta^2F(\mathbf{w}^*)+T\eta^{\frac{3-\alpha}{1-\alpha}}\big)^{\frac{2\alpha}{1+\alpha}}+\kappa^{\frac{1+\alpha}{\alpha-1}}\\&+\kappa Tn^{-2}\big(T\eta^2\big)^{\frac{1-\alpha}{1+\alpha}}\big(\eta+T\eta^3\sigma^2 d+T\eta^2F(\mathbf{w}^*)+T\eta^{\frac{3-\alpha}{1-\alpha}}\big)^{\frac{2\alpha}{1+\alpha}}+\kappa T\eta^{\frac{2}{1-\alpha}}\Big)\\&=\mathcal{O}\Big(\big(\kappa^{-1}T^{\frac{-2\alpha}{1+\alpha}}\eta^{\frac{-4\alpha}{1+\alpha}}+\kappa n^{-2}T^{\frac{2}{1+\alpha}}\eta^{\frac{2-2\alpha}{1+\alpha}}\big)\big(\eta+T\eta^3\sigma^2 d+T\eta^2F(\mathbf{w}^*)\\&+T\eta^{\frac{3-\alpha}{1-\alpha}}\big)^{\frac{2\alpha}{1+\alpha}}+\kappa^{\frac{1+\alpha}{\alpha-1}}+\kappa T\eta^{\frac{2}{1-\alpha}}\Big).\end{aligned}\tag{16}$$

Part (b) in Theorem 16 with $\eta_t=\eta$ implies

$$\begin{aligned}&\big(T\eta\big)^{-1}\sum_{t=1}^{T}\eta\mathbb{E}[F_S(\mathbf{w}_t)-F(\mathbf{w}^*)]=\big(T\eta\big)^{-1}\sum_{t=1}^{T}\eta\mathbb{E}[F_S(\mathbf{w}_t)-F_S(\mathbf{w}^*)]\\&=\mathcal{O}\Big(T^{\frac{-2\alpha}{1+\alpha}}\eta^{\frac{1-3\alpha}{1+\alpha}}\big(\eta+T\eta^3\sigma^2 d+T\eta^2F(\mathbf{w}^*)+T\eta^{\frac{3-\alpha}{1-\alpha}}\big)^{\frac{2\alpha}{1+\alpha}}+\frac{1}{T\eta}+\eta\sigma^2 d\Big).\end{aligned}\tag{17}$$

Plugging (16) and (17) back into (14) yields

$$\mathbb{E}[F(\mathbf{w}_{\text{priv}})]-F(\mathbf{w}^*)$$

$$= \mathcal{O}\Big(\kappa^{\frac{1+\alpha}{\alpha-1}} + \kappa T\eta^{\frac{2}{1-\alpha}} + \big(\kappa n^{-2}T^{\frac{2}{1+\alpha}}\eta^{\frac{2-2\alpha}{1+\alpha}} + T^{\frac{-2\alpha}{1+\alpha}}\eta^{\frac{1-3\alpha}{1+\alpha}} + \kappa^{-1}T^{\frac{-2\alpha}{1+\alpha}}\eta^{\frac{-4\alpha}{1+\alpha}}\big)$$
$$\times \big(\eta + T\eta^3\sigma^2 d + T\eta^2 F(\mathbf{w}^*) + T\eta^{\frac{3-\alpha}{1-\alpha}}\big)^{\frac{2\alpha}{1+\alpha}} + \frac{1}{T\eta} + \eta\sigma^2 d\Big). \quad (18)$$

Now, we can prove part (a) by choosing suitable κ, η and T. Let $\kappa = \sqrt{n}$ and $\eta = c\min\Big\{\frac{1}{\sqrt{n}}, \frac{\epsilon}{\sqrt{d\log(1/\delta)}}\Big\}$. Recall that $\sigma^2 d = \mathcal{O}\Big(\frac{Td\log(1/\delta)}{n^2\epsilon^2}\Big)$. Note we assume $\eta T \geq 1$. Then

$$\eta + T\eta^3\sigma^2 d + T\eta^2 + T\eta^{\frac{3-\alpha}{1-\alpha}} = \mathcal{O}\Big(\frac{T^2\eta^3 d\log(1/\delta)}{n^2\epsilon^2} + T\eta^2\Big)$$
$$= \mathcal{O}\Big(T^2n^{-2}\eta + T\eta^2\Big).$$

Combining the above equation with Eq. (18), we get

$$\mathbb{E}[F(\mathbf{w}_{\text{priv}})] - F(\mathbf{w}^*)$$
$$= \mathcal{O}\Big(\big(T^{\frac{-2\alpha}{1+\alpha}}\eta^{\frac{1-3\alpha}{1+\alpha}} + n^{-\frac{1}{2}}T^{\frac{-2\alpha}{1+\alpha}}\eta^{\frac{-4\alpha}{1+\alpha}} + n^{-\frac{3}{2}}T^{\frac{2}{1+\alpha}}\eta^{\frac{2-2\alpha}{1+\alpha}}\big)$$
$$\times \big(T^2n^{-2}\eta + T\eta^2\big)^{\frac{2\alpha}{1+\alpha}} + n^{\frac{1+\alpha}{2(\alpha-1)}} + \frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\Big).$$

If we further choose $T \asymp n$, then for any $\alpha \in [1/2, 1)$ there holds

$$\mathbb{E}[F(\mathbf{w}_{\text{priv}})] - F(\mathbf{w}^*) = \mathcal{O}\Big(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\Big).$$

For the case $\alpha \in [0, 1/2)$, let $\kappa = \sqrt{n}$ and $\eta = c\min\Big\{n^{\frac{3(\alpha-1)}{2(1+\alpha)}}, \frac{\epsilon}{\sqrt{d\log(1/\delta)}}\Big\} \leq \min\{2/L, 1\}$, where $c > 0$ is a constant. Similar to the discussion of Part (a), this choice of η implies

$$\eta + T\eta^3\sigma^2 d + T\eta^2 + T\eta^{\frac{3-\alpha}{1-\alpha}} = \mathcal{O}\Big(\frac{T^2\eta^3 d\log(1/\delta)}{n^2\epsilon^2} + T\eta^2\Big)$$
$$= \mathcal{O}\Big(\frac{T^2\eta^2\sqrt{d\log(1/\delta)}}{n(n\epsilon)} + T\eta^2\Big).$$

Further setting $T \asymp n^{\frac{2-\alpha}{1+\alpha}}$, then combining the above equation with Eq. (18) implies

$$\mathbb{E}[F(\mathbf{w}_{\text{priv}})] - F(\mathbf{w}^*)$$
$$= \mathcal{O}\Big(\frac{n^{\frac{-5\alpha^2+4\alpha-3}{2(1+\alpha)^2}}\sqrt{d\log(1/\delta)}}{n\epsilon} + n^{\frac{1+\alpha}{2(\alpha-1)}} + \frac{\sqrt{d\log(1/\delta)}}{n^{\frac{2-\alpha}{1+\alpha}}\epsilon}$$
$$+ \frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\Big)$$
$$= \mathcal{O}\Big(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\Big),$$

where the last equality used $\alpha < 1/2$. The proof of part (a) is completed.

Finally, we consider the low noise case. Let $\eta = c\min\Big\{n^{\frac{\alpha^2+2\alpha-3}{2(1+\alpha)}}, \frac{n\epsilon}{T\sqrt{d\log(1/\delta)}}\Big\}$ such that $\eta \leq \min\{2/L, 1\}$, where c is a positive constant. Then (18) with $F(\mathbf{w}^*) = 0$ implies

$$\mathbb{E}[F(\mathbf{w}_{\text{priv}})] - F(\mathbf{w}^*)$$
$$= \mathcal{O}\Big(\big(\kappa n^{-2}T^{\frac{2}{1+\alpha}}\eta^{\frac{2-2\alpha}{1+\alpha}} + T^{\frac{-2\alpha}{1+\alpha}}\eta^{\frac{1-3\alpha}{1+\alpha}} + \kappa^{-1}T^{\frac{-2\alpha}{1+\alpha}}\eta^{\frac{-4\alpha}{1+\alpha}}\big)$$
$$\times \big(\eta + T\eta^3\sigma^2 d + T\eta^{\frac{3-\alpha}{1-\alpha}}\big)^{\frac{2\alpha}{1+\alpha}} + \kappa^{\frac{1+\alpha}{\alpha-1}} + \kappa T\eta^{\frac{2}{1-\alpha}} + \frac{1}{T\eta} + \eta\sigma^2 d\Big).$$

Note

$$\eta + T\eta^3\sigma^2 d + T\eta^{\frac{3-\alpha}{1-\alpha}} = \mathcal{O}\Big(\eta\Big(1 + \frac{T^2\eta^2 d\log(1/\delta)}{n^2\epsilon^2}\Big)\Big) = \mathcal{O}(\eta),$$

where we used $T^2\eta^2 = \mathcal{O}\big(n^2\epsilon^2/(d\log(1/\delta))\big)$. Further, if we choose $\gamma = n^{\frac{1-\alpha}{2}}$ and $T \asymp n^{\frac{2}{1+\alpha}}$, there holds

$$\mathbb{E}[F(\mathbf{w}_{\text{priv}})] - F(\mathbf{w}^*) = \mathcal{O}\Big(\frac{1}{n^{\frac{1+\alpha}{2}}} + \frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\Big),$$

which completes the proof. □

4.2. Proofs for pairwise learning

Now, we give the proofs for the pairwise learning algorithm (Algorithm 2). We first give the proof of its privacy guarantee.

Proof of Theorem 5. For each $t \in [T]$, let $\mathcal{M}_t = \mathcal{G}_t + \mathbf{b}_t$ with $\mathcal{G}_t = \partial f(\mathbf{w}_t; z_{i_t}, z_{j_t})$. Similar to before, we can show that the ℓ_2-sensitivity of $\mathcal{G}_t$ is upper bounded by 2G by using Lipschitz continuity of f. Notice that

$$\sigma^2 = \frac{56G^2T}{\beta n^2\epsilon}\Big(\frac{\log(1/\delta)}{(1-\beta)\epsilon} + 1\Big).$$

Since z_{i_t} and z_{j_t} are drawn uniformly without replacement from the training set S, then according to Lemma 8 with $\gamma = 2/n$, we know $\mathcal{M}_t$ satisfies $\Big(\lambda, \frac{\lambda\beta\epsilon}{T\left(\frac{\log(1/\delta)}{(1-\beta)\epsilon}+1\right)}\Big)$-RDP if $\sigma^2 \geq 2.68G^2$ and $\lambda - 1 \leq \frac{\sigma^2}{6G^2}\log\Big(\frac{n}{2\lambda\left(1+\frac{\sigma^2}{4G^2}\right)}\Big)$ hold. Now, let $\lambda = \frac{\log(1/\delta)}{(1-\beta)\epsilon} + 1$. Then we get $\mathcal{M}_t$ satisfies $\Big(\frac{\log(1/\delta)}{(1-\beta)\epsilon} + 1, \frac{\beta\epsilon}{T}\Big)$-RDP. According to Lemmas 11 and 9, we can show that Algorithm 2 is $\Big(\frac{\log(1/\delta)}{(1-\beta)\epsilon} + 1, \beta\epsilon\Big)$-RDP. Finally, Lemma 10 implies Algorithm 2 is (ϵ, δ)-DP if $\sigma^2 \geq 2.68G^2$ and $\lambda - 1 \leq \frac{\sigma^2}{6G^2}\log\Big(\frac{n}{\lambda\left(1+\frac{\sigma^2}{4G^2}\right)}\Big)$ hold. The proof is completed. □

To establish the generalization analysis of Algorithm 2, we first introduce the following lemma which addresses the connection between stability and generalization error.

Lemma 17. *Let $\mathcal{M}$ be on-average ν-argument-stable. Let $\kappa > 0$.*

(a) Assume f is nonnegative. If f is L-smooth, there holds

$$\mathbb{E}[\bar{F}(\mathcal{M}(S)) - \bar{F}_S(\mathcal{M}(S))] \leq \frac{L}{\kappa}\mathbb{E}[\bar{F}_S(\mathcal{M}(S))] + 2(L+\kappa)\nu.$$

(b) Let $\alpha \in [0, 1)$. If Assumption 1 holds and f is α-Hölder smooth with parameter L, there holds

$$\mathbb{E}[\bar{F}(\mathcal{M}(S)) - \bar{F}_S(\mathcal{M}(S))] \leq \frac{M_{\alpha,1}^2}{2\kappa}\mathbb{E}\big[\bar{F}^{\frac{2\alpha}{1+\alpha}}(\mathcal{M}(S))\big] + 2\kappa\nu.$$

Proof. Part (a) can be found in [32]. We give the proof of Part (b). Recall that $S = \{z_1, \ldots, z_n\}$ and $\tilde{S} = \{z_1', \ldots, z_n'\}$ are drawn independently from ρ. For any $i \in [n]$, denote $S^{(i)} = \{z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_n\}$. Further, let

$$S^{(i,j)} = \{z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_{j-1}, z_j', z_{j+1}, \ldots, z_n\}.$$

From the symmetry between z_i, z_j and z_i', z_j', there holds

$$\mathbb{E}_{S,\tilde{S},\mathcal{M}}[\bar{F}(\mathcal{M}(S)) - \bar{F}_S(\mathcal{M}(S))]$$
$$= \frac{1}{n(n-1)}\sum_{i,j\in[n]:i\neq j}\mathbb{E}_{S,\tilde{S},\mathcal{M}}[\bar{F}(\mathcal{M}(S^{(i,j)})) - \bar{F}_S(\mathcal{M}(S))]$$
$$= \frac{1}{n(n-1)}\sum_{i,j\in[n]:i\neq j}\mathbb{E}_{S,\tilde{S},\mathcal{M}}[f(\mathcal{M}(S^{(i,j)}); z_i, z_j)) - f(\mathcal{M}(S); z_i, z_j)]$$
$$\leq \frac{1}{n(n-1)}\sum_{i,j\in[n]:i\neq j}\mathbb{E}_{S,\tilde{S},\mathcal{M}}\big[\big\langle\partial f(\mathcal{M}(S^{(i,j)}; z_i, z_j)), \mathcal{M}(S^{(i,j)}) - \mathcal{M}(S)\big\rangle\big], \quad (19)$$

where the second equality follows from $\mathbb{E}_{z_i,z_j}[f(\mathcal{M}(S^{(i,j)}); z_i, z_j)] = \bar{F}(\mathcal{M}(S^{(i,j)}))$ by noting that z_i, z_j are independent of $\mathcal{M}(S^{(i,j)})$, and in the last inequality we used the convexity of f.

By the Schwartz's inequality and self-bounding property (Lemma 12) we know

$$\langle\partial f(\mathcal{M}(S^{(i,j)}; z_i, z_j)), \mathcal{M}(S^{(i,j)}) - \mathcal{M}(S)\rangle$$
$$\leq \frac{1}{2\kappa}\|\partial f(\mathcal{M}(S^{(i,j)}; z_i, z_j))\|_2^2 + \frac{\kappa}{2}\|\mathcal{M}(S^{(i,j)}) - \mathcal{M}(S)\|_2^2$$
$$\leq \frac{M_{\alpha,1}^2}{2\kappa}f^{\frac{2\alpha}{1+\alpha}}(\mathcal{M}(S^{(i,j)}; z_i, z_j)) + \kappa\|\mathcal{M}(S^{(i,j)}) - \mathcal{M}(S^i)\|_2^2$$
$$+ \kappa\|\mathcal{M}(S^{(i)}) - \mathcal{M}(S)\|_2^2.$$

Plugging the above inequality back into Eq. (19) we get

$$\begin{aligned}
&\mathbb{E}_{S,\tilde{S},\mathcal{M}}[\bar{F}(\mathcal{M}(S)) - \bar{F}_S(\mathcal{M}(S))] \\
&\le \frac{1}{n(n-1)} \sum_{i,j\in[n]:i\neq j} \mathbb{E}_{S,\tilde{S},\mathcal{M}}\Big[\frac{M_{\alpha,1}^2}{2\kappa} f^{\frac{2\alpha}{1+\alpha}}(\mathcal{M}(S^{(i,j)}; z_i, z_j)) \\
&\quad + \kappa\|\mathcal{M}(S^{(i,j)}) - \mathcal{M}(S^i)\|_2^2 + \kappa\|\mathcal{M}(S^{(i)}) - \mathcal{M}(S)\|_2^2\Big] \\
&= \frac{M_{\alpha,1}^2}{2\kappa n(n-1)} \sum_{i,j\in[n]:i\neq j} \mathbb{E}_{S,\tilde{S},\mathcal{M}}\big[f^{\frac{2\alpha}{1+\alpha}}(\mathcal{M}(S^{(i,j)}; z_i, z_j))\big] \\
&\quad + \frac{2\kappa}{n(n-1)} \sum_{i,j\in[n]:i\neq j} \mathbb{E}_{S,\tilde{S},\mathcal{M}}\big[\|\mathcal{M}(S^{(i)}) - \mathcal{M}(S)\|_2^2\big],
\end{aligned}$$

where the last equality is due to $\mathbb{E}_{S,\tilde{S},\mathcal{M}}\big[\|\mathcal{M}(S^{(i,j)}) - \mathcal{M}(S^i)\|_2^2\big] = \mathbb{E}_{S,\tilde{S},\mathcal{M}}\big[\|\mathcal{M}(S^{(j)}) - \mathcal{M}(S)\|_2^2\big]$.

Since $x \mapsto x^{\frac{2\alpha}{1+\alpha}}$ is concave and z_i, z_j are independent of $\mathcal{M}(S^{(i,j)})$, we know

$$\begin{aligned}
&\mathbb{E}_{S,\tilde{S},\mathcal{M}}\big[f^{\frac{2\alpha}{1+\alpha}}(\mathcal{M}(S^{(i,j)}; z_i, z_j))\big] \\
&\le \mathbb{E}_{S,\tilde{S},\mathcal{M}}\Big[\big(\mathbb{E}_{z_i,z_j}[f(\mathcal{M}(S^{(i,j)}; z_i, z_j))]\big)^{\frac{2\alpha}{1+\alpha}}\Big] = \mathbb{E}_{S,\mathcal{M}}\big[\bar{F}^{\frac{2\alpha}{1+\alpha}}(\mathcal{M}(S))\big].
\end{aligned}$$

Combining the above two inequalities together implies

$$\begin{aligned}
&\mathbb{E}_{S,\tilde{S},\mathcal{M}}[\bar{F}(\mathcal{M}(S)) - \bar{F}_S(\mathcal{M}(S))] \\
&= \frac{M_{\alpha,1}^2}{2\kappa n(n-1)} \sum_{i,j\in[n]:i\neq j} \mathbb{E}_{S,\mathcal{M}}\big[\bar{F}^{\frac{2\alpha}{1+\alpha}}(\mathcal{M}(S))\big] + \frac{2\kappa}{n(n-1)} \\
&\quad \times \sum_{i,j\in[n]:i\neq j} \mathbb{E}_{S,\tilde{S},\mathcal{M}}[\|\mathcal{M}(S^{(i)}) - \mathcal{M}(S)\|_2^2] \\
&= \frac{M_{\alpha,1}^2}{2\kappa}\mathbb{E}_{S,\mathcal{M}}\big[\bar{F}^{\frac{2\alpha}{1+\alpha}}(\mathcal{M}(S))\big] + \frac{2\kappa}{n}\sum_{i=1}^{n}\mathbb{E}_{S,\tilde{S},\mathcal{M}}[\|\mathcal{M}(S^{(i)}) - \mathcal{M}(S)\|_2^2].
\end{aligned}$$

The proof of Part (b) is completed. □

Our stability analysis for α-Hölder smooth losses requires the following lemma.

Lemma 18 ([26]). *For all $z, z' \in \mathcal{Z}$, suppose $\mathbf{w} \mapsto f(\mathbf{w}; z, z')$ is convex, $\mathbf{w} \mapsto \partial f(\mathbf{w}; z, z')$ is α-Hölder smooth with parameter L and $\alpha \in [0, 1)$. Then for any $\mathbf{w}, \mathbf{w}'$ and $\eta > 0$, there holds*

$$\|\mathbf{w} - \eta\partial f(\mathbf{w}; z, z') - \mathbf{w}' + \eta\partial f(\mathbf{w}'; z, z')\|_2^2 \le \|\mathbf{w} - \mathbf{w}'\|_2^2 + \bar{M}_{\alpha,3}^2\eta^{\frac{2}{1-\alpha}}.$$

The stability results of Algorithm 2 are established in the following lemma.

Lemma 19. *Suppose Assumption 2 holds. Let $\mathcal{M}$ be Algorithm 2 with T iterations.*

(a) If f is L-smooth and $\eta_t \le 2/L$ for all $t \in [T]$, then $\mathcal{M}$ is on-average ν-argument-stable with

$$\nu \le \frac{16L(1+2T/n)e}{n}\sum_{j=1}^{T}\eta_j^2\mathbb{E}[F_S(\mathbf{w}_j)].$$

(b) Let $\alpha \in [0, 1)$. If f is α-Hölder smooth with parameter L, then $\mathcal{M}$ is on-average ν-argument-stable with

$$\nu \le \frac{8eM_{\alpha,1}^2(1+2T/n)}{n}\sum_{j=1}^{T}\eta_j^2\mathbb{E}\big[F_S^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_j)\big] + \bar{M}_{\alpha,3}^2 e\sum_{j=1}^{T}\eta_j^{\frac{2}{1-\alpha}},$$

where $\bar{M}_{\alpha,3} = \sqrt{\frac{e(1-\alpha)}{1+\alpha}}(2^{-\alpha}L)^{\frac{1}{1-\alpha}}$.

Proof. Since adding noise to gradient will not change stability results, then we only need to address the on-average stability bounds of non-private SGD for pairwise learning. The proof of part (a) can be found in [32]. We only give the proof of part (b). For any $i \in [n]$, let $S, S^{(i)}$ and $\tilde{S}$ be constructed as Definition 6. For any S and $i \in [n]$, we discuss three different cases.

Case 1. If $i_t \neq i$ and $j_t \neq i$, it then follows from the update rule of $\mathbf{w}_{t+1}$ and Lemma 18 that

$$\begin{aligned}
\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2^2 &\le \|\mathbf{w}_t - \eta_t\partial f(\mathbf{w}_t; z_{i_t}, z_{j_t}) - \mathbf{w}_t^{(i)} + \eta_t\partial f(\mathbf{w}_t^{(i)}; z_{i_t}, z_{j_t})\|_2^2 \\
&\le \|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2 + \bar{M}_{\alpha,3}^2\eta_t^{\frac{2}{\alpha-1}}.
\end{aligned}$$

Case 2. If $i_t = i$, the update rule and the standard inequality $(a+b)^2 \le (1+\mu)a^2 + (1+1/\mu)b^2$ imply that

$$\begin{aligned}
&\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2^2 \\
&\le (1+\mu)\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2 + (1+1/\mu)\eta_t^2\big(\|\partial f(\mathbf{w}_t; z_i, z_{j_t}) \\
&\quad - \partial f(\mathbf{w}_t^{(i)}; z_i', z_{j_t})\|_2^2\big) \\
&\le (1+\mu)\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2 + 2(1+1/\mu)\eta_t^2\big(\|\partial f(\mathbf{w}_t; z_i, z_{j_t})\|_2^2 \\
&\quad + \|\partial f(\mathbf{w}_t^{(i)}; z_i', z_{j_t})\|_2^2\big) \\
&\le (1+\mu)\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2 + 2M_{\alpha,1}^2(1+1/\mu)\eta_t^2\big(f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_t; z_i, z_{j_t}) \\
&\quad + f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_t^{(i)}; z_i', z_{j_t})\big).
\end{aligned}$$

Case 3. If $j_t = i$, similar to Case 2, we have

$$\begin{aligned}
&\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2^2 \\
&\le (1+\mu)\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2 + 2M_{\alpha,1}^2(1+1/\mu)\eta_t^2\big(f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_t; z_{i_t}, z_i) \\
&\quad + f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_t^{(i)}; z_{i_t}, z_i')\big).
\end{aligned}$$

Note $\mathbf{Pr}(i_t \neq i \text{ and } j_t \neq i) = \frac{(n-1)(n-2)}{n(n-1)}$ and $\mathbf{Pr}(i_t = i \text{ and } j_t = j) = \frac{1}{n(n-1)}$ for any $j \neq i$. Combining Cases 1–3 together yields

$$\begin{aligned}
&\mathbb{E}_{i_t,j_t}\big[\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2^2\big] \\
&\le \frac{(n-1)(n-2)}{n(n-1)}\Big(\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2 + \bar{M}_{\alpha,3}^2\eta_t^{\frac{2}{\alpha-1}}\Big) \\
&\quad + \frac{1}{n(n-1)}\sum_{j\in[n]:j\neq i}\big((1+\mu)\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2 + 2M_{\alpha,1}^2(1+1/\mu)\eta_t^2 \\
&\quad \times \big(f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_t; z_i, z_j) + f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_t^{(i)}; z_i', z_j)\big)\big) \\
&\quad + \frac{1}{n(n-1)}\sum_{j\in[n]:j\neq i}\big((1+\mu)\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2 + 2M_{\alpha,1}^2(1+1/\mu)\eta_t^2 \\
&\quad \times \big(f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_t; z_j, z_i) + f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_t^{(i)}; z_j, z_i')\big)\big) \\
&\le \Big(1 + \frac{2\mu}{n}\Big)\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2 + \bar{M}_{\alpha,3}^2\eta_t^{\frac{2}{\alpha-1}} + \frac{2(1+1/\mu)M_{\alpha,1}^2\eta_t^2}{n(n-1)} \\
&\quad \times \sum_{j\in[n]:j\neq i}\big[f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_t; z_i, z_j) + f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_t^{(i)}; z_i', z_j) + f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_t; z_j, z_i) \\
&\quad + f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_t^{(i)}; z_j, z_i')\big].
\end{aligned}$$

Taking an average over i we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{i_t,j_t}\big[\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2^2\big] \\
&\le \Big(1 + \frac{2\mu}{n}\Big)\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2 + \bar{M}_{\alpha,3}^2\eta_t^{\frac{2}{\alpha-1}} \\
&\quad + \frac{2(1+1/\mu)M_{\alpha,1}^2\eta_t^2}{n^2(n-1)}\sum_{i=1}^{n}\sum_{j\in[n]:j\neq i}\big[f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_t; z_i, z_j) \\
&\quad + f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_t^{(i)}; z_i', z_j) + f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_t; z_j, z_i) + f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_t^{(i)}; z_j, z_i')\big].
\end{aligned}$$

Further, taking an expectation over both sides yields

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{S,\tilde{S},\mathcal{M}}\big[\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2^2\big] \\
&\le \Big(1 + \frac{2\mu}{n}\Big)\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{S,\tilde{S},\mathcal{M}}\big[\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2\big] + \bar{M}_{\alpha,3}^2\eta_t^{\frac{2}{\alpha-1}}
\end{aligned}$$

$$+\frac{2(1+1/\mu)M_{\alpha,1}^2\eta_t^2}{n^2(n-1)}\sum_{i=1}^{n}\mathbb{E}_{S,\tilde{S},\mathcal{M}}\Big[\sum_{j\in[n]:j\neq i}\big[f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_t;z_i,z_j)+f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_t^{(i)};z_i',z_j)+f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_t;z_j,z_i)+f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_t^{(i)};z_j,z_i')\big]\Big].$$

Due to the symmetry between z_i and z_i' we know

$$\mathbb{E}_{S,\mathcal{M}}\Big[\sum_{j\in[n]:j\neq i}\Big[f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_t;z_i,z_j)+f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_t;z_j,z_i)\Big]\Big]$$
$$=\mathbb{E}_{S,\tilde{S},\mathcal{M}}\Big[\sum_{j\in[n]:j\neq i}\Big[f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_t^{(i)};z_i',z_j)+f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_t^{(i)};z_j,z_i')\Big]\Big].$$

It then follows that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{S,\tilde{S},\mathcal{M}}\big[\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2^2\big]$$
$$\le\Big(1+\frac{2\mu}{n}\Big)\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{S,\tilde{S},\mathcal{M}}\big[\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2^2\big]+\bar{M}_{\alpha,3}^2\eta_t^{\frac{2}{\alpha-1}}$$
$$+\frac{4(1+1/\mu)M_{\alpha,1}^2\eta_t^2}{n^2(n-1)}\sum_{i=1}^{n}\mathbb{E}_{S,\mathcal{M}}\Big[\sum_{j\in[n]:j\neq i}\Big[f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_t;z_i,z_j)+f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_t;z_j,z_i)\Big]\Big]$$
$$=\Big(1+\frac{2\mu}{n}\Big)\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{S,\tilde{S},\mathcal{M}}\big[\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2^2\big]+\bar{M}_{\alpha,3}^2\eta_t^{\frac{2}{\alpha-1}}$$
$$+\frac{8(1+1/\mu)M_{\alpha,1}^2\eta_t^2}{n}\mathbb{E}_{S,\mathcal{M}}\Big[\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\in[n]:j\neq i}f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_t;z_i,z_j)\Big],$$

where we used

$$\sum_{i=1}^{n}\sum_{j\in[n]:j\neq i}f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_t;z_j,z_i)=\sum_{i=1}^{n}\sum_{j\in[n]:j\neq i}f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_t;z_i,z_j).$$

Further, according to Jensen's inequality and $\mathbf{w}_1=\mathbf{w}_1'$, we know

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{S,\tilde{S},\mathcal{M}}\big[\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2^2\big]$$
$$\le\Big(1+\frac{2\mu}{n}\Big)\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{S,\tilde{S},\mathcal{M}}\big[\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2^2\big]+\bar{M}_{\alpha,3}^2\eta_t^{\frac{2}{\alpha-1}}$$
$$+\frac{8(1+1/\mu)M_{\alpha,1}^2\eta_t^2}{n}\mathbb{E}_{S,\mathcal{M}}\Big[F_S^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_t)\Big],$$

Now, we can apply the above inequality recursively and get

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{S,\tilde{S},\mathcal{M}}\big[\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2^2\big]$$
$$\le\frac{8(1+1/\mu)M_{\alpha,1}^2}{n}\sum_{j=1}^{t}\Big(1+\frac{2\mu}{n}\Big)^{t-j}\eta_t^2\mathbb{E}_{S,\mathcal{M}}\Big[F_S^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_j)\Big]$$
$$+\bar{M}_{\alpha,3}^2\sum_{j=1}^{t}\Big(1+\frac{2\mu}{n}\Big)^{t+1-j}\eta_t^{\frac{2}{\alpha-1}}.$$

Finally, we can set $\mu=\frac{n}{2t}$ and use $(1+1/t)^t\le e$ to get

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{S,\tilde{S},\mathcal{M}}\big[\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2^2\big]$$
$$\le\frac{8e(1+2t/n)M_{\alpha,1}^2}{n}\sum_{j=1}^{t}\eta_t^2\mathbb{E}_{S,\mathcal{M}}\Big[F_S^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_j)\Big]+\bar{M}_{\alpha,3}^2e\sum_{j=1}^{t}\eta_t^{\frac{2}{\alpha-1}},$$

which completes the proof. □

As discussed in [32], the proof of optimization error for Algorithm 2 is the same as Algorithm 1. Hence, the optimization error of Algorithm 2 directly follows from Theorem 16, which is established in the following lemma. Here, $\alpha=1$ corresponds to the strongly smooth case.

Lemma 20. *Let $\alpha\in[0,1]$. Suppose Assumption 1 holds and f is α-Hölder smooth with parameter $L>0$. Let $\{\mathbf{w}_j\}$ be produced by Algorithm 2. Then*

$$\sum_{j=1}^{t}\eta_j\mathbb{E}[\bar{F}_S(\mathbf{w}_j)-\bar{F}_S(\mathbf{w}^*)]\le\frac{1}{2}\|\mathbf{w}^*\|_2^2+\frac{3M_{\alpha,1}^2}{4}\Big(\sum_{j=1}^{t}\eta_j^2\Big)^{\frac{1-\alpha}{1+\alpha}}\Big[2\eta_1\|\mathbf{w}^*\|_2^2$$
$$+\sum_{j=1}^{t}\big(6\eta_j^3\sigma^2d+4\eta_j^2\bar{F}(\mathbf{w}^*)+3M_{\alpha,2}\eta_j^{\frac{3-\alpha}{1-\alpha}}\big)\Big]^{\frac{2\alpha}{1+\alpha}}+\sum_{j=1}^{t}3\eta_j^2\sigma^2d$$

and

$$\sum_{j=1}^{t}\eta_j^2\mathbb{E}[\bar{F}_S(\mathbf{w}_t)]\le2\eta_1\|\mathbf{w}^*\|_2^2+\sum_{j=1}^{t}\big(6\eta_j^3\sigma^2d+4\eta_j^2\bar{F}(\mathbf{w}^*)+3M_{\alpha,2}\eta_j^{\frac{3-\alpha}{1-\alpha}}\big).$$

Now, we move on to utility guarantees for strongly smooth and non-smooth cases. We first present the proof for strongly smooth case.

Proof of Theorem 6. Combining part (a) in Lemmas 17, 19 and 20 together we have

$$\mathbb{E}[\bar{F}(\mathbf{w}_{t+1})]$$
$$\le\Big(1+\frac{L}{\kappa}\Big)\mathbb{E}[\bar{F}_S(\mathbf{w}_{t+1})]+\frac{32e(L+\kappa)(1+2t/n)L}{n}\Big[2\eta_1\|\mathbf{w}^*\|_2^2$$
$$+\sum_{j=1}^{t}\big(6\eta_j^3\sigma^2d+4\eta_j^2\bar{F}(\mathbf{w}^*)\big)\Big].$$

Multiplying η_{t+1} and taking a summation yield

$$\sum_{t=1}^{T}\eta_t\mathbb{E}[\bar{F}(\mathbf{w}_t)]$$
$$\le\Big(1+\frac{L}{\kappa}\Big)\sum_{t=1}^{T}\eta_t\mathbb{E}[\bar{F}_S(\mathbf{w}_t)]+\frac{32e(L+\kappa)(1+2T/n)L}{n}\sum_{t=1}^{T}\eta_t\Big[2\eta_1\|\mathbf{w}^*\|_2^2$$
$$+\sum_{j=1}^{t}\big(6\eta_j^3\sigma^2d+4\eta_j^2\bar{F}(\mathbf{w}^*)\big)\Big].\tag{20}$$

Lemma 20 with $\alpha=1$ implies

$$\sum_{t=1}^{T}\eta_t\mathbb{E}[\bar{F}_S(\mathbf{w}_t)]\le\sum_{t=1}^{T}\eta_t\mathbb{E}[\bar{F}_S(\mathbf{w}^*)]+\frac{1}{2}\|\mathbf{w}^*\|_2^2+3L\Big(\eta_1\|\mathbf{w}^*\|_2^2$$
$$+\sum_{t=1}^{T}\big(3\eta_t^3\sigma^2d+2\eta_t^2F(\mathbf{w}^*)\big)\Big)+\sum_{t=1}^{T}3\eta_t^2\sigma^2d.\tag{21}$$

Plugging (21) back into (20) we can get

$$\sum_{t=1}^{T}\eta_t\mathbb{E}[\bar{F}(\mathbf{w}_t)]$$
$$\le\Big(1+\frac{L}{\kappa}\Big)\Big(\sum_{t=1}^{T}\eta_t\bar{F}(\mathbf{w}^*)+\Big(\frac{1}{2}+3L\eta_1\Big)\|\mathbf{w}^*\|_2^2+3\sum_{j=1}^{t}(3L\eta_j+1)\eta_j^2\sigma^2d$$
$$+4\sum_{j=1}^{t}\eta_j^2\bar{F}(\mathbf{w}^*)\Big)+\frac{32e(L+\kappa)(1+2T/n)L}{n}\sum_{t=1}^{T}\eta_t\Big[2\eta_1\|\mathbf{w}^*\|_2^2$$
$$+\sum_{j=1}^{t}\big(6\eta_j^3\sigma^2d+4\eta_j^2\bar{F}(\mathbf{w}^*)\big)\Big].$$

Let $\eta_t=\eta\le\min\{2/L,1\}$ and assume $T\ge n$. Recall that $\sigma^2d=\mathcal{O}\big(\frac{Td\log(1/\delta)}{n^2\epsilon^2}\big)$. According to Jensen's inequality, there holds

$$\mathbb{E}[\bar{F}(\mathbf{w}_{\text{priv}})-\bar{F}(\mathbf{w}^*)]$$
$$=\mathcal{O}\Big(\Big(\kappa^{-1}+\frac{T^2\eta^2(1+\kappa)}{n^2}+(\kappa^{-1}+1)\eta\Big)\bar{F}(\mathbf{w}^*)+\Big(\frac{(1+\kappa^{-1})}{T\eta}$$
$$+\frac{(1+\kappa)T\eta}{n^2}\Big)\|\mathbf{w}^*\|_2^2+\Big((1+\kappa^{-1})\eta+\frac{T^2\eta^3(1+\kappa)}{n^2}\Big)\frac{Td\log(1/\delta)}{n^2\epsilon^2}\Big).\tag{22}$$

Now, we give the proof of part (a). We set $T\asymp n$, $\kappa=\sqrt{n}$ and $\eta=c/\max\big\{\sqrt{n},\frac{\sqrt{d\log(1/\delta)}}{\epsilon}\big\}\le\min\{2/L,1\}$, where $c>0$ is a constant. Then from (22) we obtain

$$\mathbb{E}[\bar{F}(\mathbf{w}_{\text{priv}})-\bar{F}(\mathbf{w}^*)]=\mathcal{O}\Big(\frac{1}{\sqrt{n}}+\frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\Big),$$

where we also assume $\sqrt{d\log(1/\delta)}=\mathcal{O}(n\epsilon)$.

(b) For the low-noise case, setting $\kappa\ge1$, $T\asymp n$ and $\eta=\frac{c\epsilon}{\sqrt{d\log(1/\delta)}}\le\min\{2/L,1\}$ yields the desired result. □

Finally, we establish the utility guarantee for the non-smooth case.

Proof of Theorem 7. Similar to Theorem 4, we can plug part (b) in Lemma 19 back into part (b) in Lemma 17 and get that

$$\Big(\sum_{t=1}^{T}\eta_t\Big)^{-1}\sum_{t=1}^{T}\eta_t\mathbb{E}[\bar{F}(\mathbf{w}_t)-\bar{F}_S(\mathbf{w}_t)] = \mathcal{O}\Big(\kappa^{\frac{1+\alpha}{\alpha-1}} + \kappa T\eta^{\frac{2}{1-\alpha}} + \big(\kappa T\eta\big)^{-1}\sum_{t=1}^{T}\eta\big(\mathbb{E}[\bar{F}_S(\mathbf{w}_{t+1})]\big)^{\frac{2\alpha}{1+\alpha}} + \kappa T n^{-2}\sum_{t=1}^{T}\eta^2\big(\mathbb{E}[\bar{F}_S(\mathbf{w}_t)]\big)^{\frac{2\alpha}{1+\alpha}}\Big). \tag{23}$$

Further, combining Eq. (23) and Lemma 20 together we can obtain

$$\Big(\sum_{t=1}^{T}\eta\Big)^{-1}\sum_{t=1}^{T}\eta\mathbb{E}[\bar{F}_S(\mathbf{w}_t)-\bar{F}(\mathbf{w}^*)] = \mathcal{O}\Big(T^{\frac{-2\alpha}{1+\alpha}}\eta^{\frac{1-3\alpha}{1+\alpha}}\Big(\eta+T\eta^3\sigma^2 d+T\eta^2\bar{F}(\mathbf{w}^*)+T\eta^{\frac{3-\alpha}{1-\alpha}}\Big)^{\frac{2\alpha}{1+\alpha}} + \frac{1}{T\eta}+\eta\sigma^2 d\Big). \tag{24}$$

Plugging (23) and (24) back into (14) we can get

$$\mathbb{E}[\bar{F}(\mathbf{w}_{\text{priv}})-\bar{F}(\mathbf{w}^*)] = \mathcal{O}\Big(\big(\frac{(1+\kappa^{-1})}{T\eta}+\frac{(1+\kappa)T\eta}{n^2}\big)\|\mathbf{w}^*\|_2^2+\Big(\kappa^{-1}+\frac{T^2\eta^2(1+\kappa)}{n^2} + \big(\kappa^{-1}+1\big)\eta\Big)\bar{F}(\mathbf{w}^*)+\Big(\big(1+\kappa^{-1}\big)\eta+\frac{T^2\eta^3(1+\kappa)}{n^2}\Big)\frac{Td\log(1/\delta)}{n^2\epsilon^2}\Big). \tag{25}$$

The rest of the proof is similar to Theorem 4. We omit it for simplicity. □

5. Simulations

In this section, we present experimental results to verify our theoretical results. Especially, we take Algorithm 1 with f being either the least square loss or the logistic loss as examples to show that DP-SGD algorithm performs better under the low noise setting compared to general settings.

5.1. Experimental setting and datasets

For the least square loss, consider coefficient $\mathbf{w}^* = [0.5, 0.3, 0, 0.1, 0.2, 0, 0, 0, 0, 0.1] \in \mathbb{R}^d$ with dimension $d = 10$. Let data size $n = 5000$. For $i = 1, \ldots, n$, we simulated $z_i = (\mathbf{x}_i, y_i)$ with $\mathbf{x} \in \mathbb{R}^d$ such that $\|\mathbf{x}_i\|_2 = 1$ and

$$y_i = \mathbf{x}_i^\top \mathbf{w}^* + \gamma_i$$

such that $|y_i| \leq 0.5$, where $\gamma_i \sim \mathcal{N}(\mathbf{0}, \kappa^2\mathbf{I}_{d\times d})$ is the Gaussian error with variance κ^2. Here, $\mathbf{I}_{d\times d} \in \mathbb{R}^{d\times d}$ is the identity matrix and we consider $\kappa = 0, 0.1$ and 0.15. The setting $\kappa = 0$ corresponds to the low noise setting, i.e., $L(\mathbf{w}^*) = 0$, and the settings $\kappa = 0.1$ and $\kappa = 0.15$ correspond to the noise setting. We run non-private SGD and DP-SGD with step size $\eta = 0.01$, $\delta = 1/n^2$ and $T = 1000$. Note that for DP-SGD we need to choose a constant $\beta \in (0, 1)$ such that the algorithm satisfies (ϵ, δ)-DP. We search β from 0.0001 to 0.9999 and choose the best one that corresponds to the smallest variance σ^2. We aim to clarify how the behavior of DP-SGD would change w.r.t. the privacy parameter ϵ under the low-noise and noise settings. To this aim, we very ϵ over the set $\{0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5\}$. Due to the randomness of the privacy algorithm, we repeat 500 runs of the randomized training procedure for each parameter setting and report the average of the mean square error as the experimental results.

For the logistic loss, we simulated $\{z_i = (\mathbf{x}_i, y_i)\}_{i=1}^n$ with $n = 5000$, $\mathbf{w}^* = [5, 3, 0, 0.1, 0.2, 0, 0, 0, 0, 0.1] \in \mathbb{R}^{10}$ such that $\|\mathbf{x}_i\|_2 = 1$ and $y_i = 1$ if $p(x) = \frac{1}{1+\exp(-y_i(\mathbf{w}^*)^\top\mathbf{x}_i+\gamma_i)} > 0.5$ and $y_i = -1$ else. Here, $\gamma_i \sim \mathcal{N}(0, \kappa^2\mathbf{I}_{10\times 10})$ is the Gaussian noise, and we consider $\kappa = 0, 0.4$ and 0.6 for this case. Similarly, we run non-private SGD and DP-SGD with $\delta = 1/n^2$, step size $\eta = 0.01$ and $T = 1000$. We search β from 0.0001 to 0.9999 and choose the best one with the smallest variance. The privacy parameter ϵ is varied over the set $\{1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5, 6\}$. We repeat 500 runs of the randomized training procedure for each parameter setting and report the average misclassification rate as the experimental results.

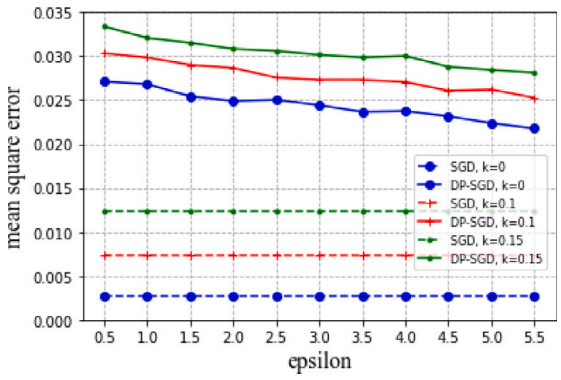


Fig. 1. Mean square error versus privacy parameter ϵ with the least square loss and $\delta = 1/n^2$ for different noise setting.

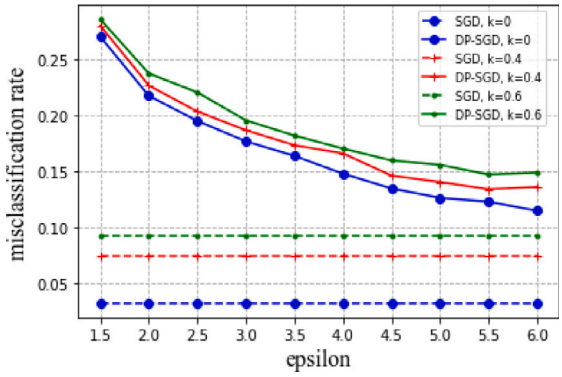


Fig. 2. Misclassification rate versus privacy parameter ϵ with the logistic loss and $\delta = 1/n^2$ for different noise setting.

5.2. Experimental results

In Fig. 1, we compare the behavior of non-private SGD and DP-SGD under different privacy levels and the low-noise/noise settings for the least square loss. The dotted lines correspond to the results of non-privacy SGD under different noise settings, and the solid lines correspond to the results of DP-SGD under different noise settings. Fig. 1 shows that under the same privacy level, i.e., the same ϵ, DP-SGD performs better in the low-noise setting ($\kappa = 0$) compared with the noise settings ($\kappa = 0.1$ and $\kappa = 0.15$). Our theoretical analysis (see Theorem 3) also indicates that DP-SGD under the low-noise case enjoys a stronger theoretical guarantee than DP-SGD under the general case, which is consistent with the experimental results. In addition, we can see clearly the trade-off between the utility and the privacy parameter. The average mean square error decreases as we relax the privacy requirements when ϵ becomes larger. Fig. 2 presents the behavior of non-private SGD and DP-SGD under different privacy levels and the low-noise/noise settings for the logistic loss. Similar to Fig. 1, Fig. 2 shows that both non-privacy SGD and DP-SGD have better performances under the low-noise case ($\kappa = 0$) than those under the general settings ($\kappa = 0.4$ and $\kappa = 0.6$), which are consistent with our theoretical results. Besides, the misclassification rate decreases as the privacy parameter ϵ increases.

6. Conclusion

In this paper, we conducted a systematic analysis of DP-SGD with gradient perturbation for both pointwise and pairwise learning problems. For pointwise learning, we introduced a low-noise condition and derived sharper excess population risk bounds. Specifically, we achieved bounds in the order of $\mathcal{O}\big(\frac{1}{n\epsilon}\sqrt{d\log(1/\delta)}\big)$ and $\mathcal{O}\big(n^{-\frac{1+\alpha}{2}} + \frac{1}{n\epsilon}\sqrt{d\log(1/\delta)}\big)$ for strongly smooth and α-Hölder smooth losses, respectively.

Regarding pairwise learning, we presented a computationally efficient DP-SGD algorithm with utility guarantees. Our analysis demonstrated that our algorithm achieves the optimal excess risk bounds of the order $\mathcal{O}\big(\frac{1}{\sqrt{n}} + \frac{1}{n\epsilon}\sqrt{d\log(1/\delta)}\big)$ for both strongly smooth and α-Hölder smooth losses. Furthermore, we established faster excess risk bounds for both strongly smooth and α-Hölder smooth losses under a low-noise condition. Notably, our work represents the first utility analysis for privacy-preserving pairwise learning that provides excess risk rates tighter than $\mathcal{O}\big(\frac{1}{\sqrt{n}} + \frac{1}{n\epsilon}\sqrt{d\log(1/\delta)}\big)$.

There are several open questions that remain for further study. Firstly, it would be interesting to explore whether our analysis of DP-SGD with uniform sampling can be extended to DP-SGD with Markov sampling, which poses a more challenging task. Secondly, an unexplored area for us is to investigate the utility analysis of DP-SGD with a neural network structure. Addressing these questions would contribute to a deeper understanding of privacy-preserving machine learning algorithms.

CRediT authorship contribution statement

Puyu Wang: Formal analysis, Methodology, Writing – original draft, Writing – review & editing. **Yunwen Lei:** Methodology, Writing – original draft, Writing – review & editing. **Yiming Ying:** Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing. **Ding-Xuan Zhou:** Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

The work described in this paper is partially done when the first and the last author, Puyu Wang and Ding-Xuan Zhou, worked at City University of Hong Kong. Ding-Xuan Zhou's work is supported by the Laboratory for AI-Powered Financial Technologies under the InnoHK scheme, the Research Grants Council of Hong Kong [Projects No. CityU 11308121, No. N_CityU102/20, and No. C1013-21GF], the National Science Foundation of China [Project No. 12061160462], and the Hong Kong Institute for Data Science. Yiming's work is supported by National Science Foundation grants (IIS-2103450, IIS-2110546 and DMS-2110836). The work of Yunwen Lei is partially supported by the Research Grants Council of Hong Kong [Project No. 22303723].

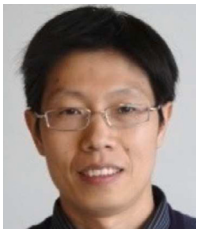
References

[1] J. Duchi, Y. Singer, Efficient online and batch learning using forward backward splitting, J. Mach. Learn. Res. 10 (Dec) (2009) 2899–2934.
[2] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016.
[3] X. Li, F. Orabona, On the convergence of stochastic gradient descent with adaptive stepsizes, in: The 22nd International Conference on Artificial Intelligence and Statistics, PMLR, 2019, pp. 983–992.
[4] Y. Li, Y. Liang, Learning overparameterized neural networks via stochastic gradient descent on structured data, in: Advances in Neural Information Processing Systems, Vol. 31, 2018.
[5] J. Lin, L. Rosasco, Optimal rates for multi-pass stochastic gradient methods, J. Mach. Learn. Res. 18 (1) (2017) 3375–3421.
[6] A. Rakhlin, O. Shamir, K. Sridharan, Making gradient descent optimal for strongly convex stochastic optimization, in: ICML, 2012.
[7] N. Roux, M. Schmidt, F. Bach, A stochastic gradient method with an exponential convergence _rate for finite training sets, Adv. Neural Inf. Process. Syst. 25 (2012).
[8] Z. Yang, Y. Lei, P. Wang, T. Yang, Y. Ying, Simple stochastic and online gradient descent algorithms for pairwise learning, Adv. Neural Inf. Process. Syst. (2021) 20160–20171.
[9] T. Zhang, Solving large scale linear prediction problems using stochastic gradient descent algorithms, in: Proceedings of the Twenty-First International Conference on Machine Learning, 2004, p. 116.
[10] L. Zhu, Z. Liu, S. Han, Deep leakage from gradients, Adv. Neural Inf. Process. Syst. 32 (2019).
[11] R. Shokri, M. Stronati, C. Song, V. Shmatikov, Membership inference attacks against machine learning models, in: 2017 IEEE Symposium on Security and Privacy, SP, IEEE, 2017, pp. 3–18.
[12] C. Dwork, F. McSherry, K. Nissim, A. Smith, Calibrating noise to sensitivity in private data analysis, in: Theory of Cryptography Conference, Springer, 2006, pp. 265–284.
[13] R. Bassily, V. Feldman, C. Guzmán, K. Talwar, Stability of stochastic gradient descent on nonsmooth convex losses, Adv. Neural Inf. Process. Syst. 33 (2020).
[14] R. Bassily, V. Feldman, K. Talwar, A.G. Thakurta, Private stochastic convex optimization with optimal rates, in: Advances in Neural Information Processing Systems, 2019, pp. 11279–11288.
[15] R. Bassily, A. Smith, A. Thakurta, Private empirical risk minimization: Efficient algorithms and tight error bounds, in: 2014 IEEE 55th Annual Symposium on Foundations of Computer Science, IEEE, 2014, pp. 464–473.
[16] R. Bassily, C. Guzmán, M. Menart, Differentially private stochastic optimization: New results in convex and non-convex settings, in: Advances in Neural Information Processing Systems, Vol. 34, 2021, pp. 9317–9329.
[17] V. Feldman, T. Koren, K. 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.
[18] J. Su, L. Hu, D. Wang, Faster rates of private stochastic convex optimization, in: International Conference on Algorithmic Learning Theory, PMLR, 2022, pp. 995–1002.
[19] P. Wang, Y. Lei, Y. Ying, H. Zhang, Differentially private SGD with non-smooth losses, Appl. Comput. Harmon. Anal. 56 (2022) 306–336.
[20] P. Wang, Z. Yang, Y. Lei, Y. Ying, H. Zhang, Differentially private empirical risk minimization for AUC maximization, Neurocomputing 461 (2021) 419–437.
[21] Z. Xue, S. Yang, M. Huai, D. Wang, Differentially private pairwise learning revisited., in: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, 2021, pp. 3242–3248.
[22] M. Huai, D. Wang, C. Miao, J. Xu, A. Zhang, Pairwise learning with differential privacy guarantees, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 694–701.
[23] M. Schliserman, T. Koren, Stability vs implicit bias of gradient methods on separable data and beyond, in: Conference on Learning Theory, PMLR, 2022, pp. 3380–3394.
[24] N. Srebro, K. Sridharan, A. Tewari, Smoothness, low noise and fast rates, Adv. Neural Inf. Process. Syst. 23 (2010).
[25] O. Shamir, Gradient methods never overfit on separable data, J. Mach. Learn. Res. 22 (1) (2021) 3847–3866.
[26] Y. Lei, Y. Ying, Fine-grained analysis of stability and generalization for stochastic gradient descent, in: International Conference on Machine Learning, 2020, pp. 5809–5819.
[27] S. Nagayasu, S. Watanbe, Asymptotic behavior of free energy when optimal probability distribution is not unique, Neurocomputing 500 (2022) 528–536.
[28] O. Bousquet, A. Elisseeff, Stability and generalization, J. Mach. Learn. Res. 2 (Mar) (2002) 499–526.
[29] V. Vapnik, The Nature of Statistical Learning Theory, Springer science & business media, 1999.

[30] S. Wang, B. Sheng, Error analysis of kernel regularized pairwise learning with a strongly convex loss, Math. Found. Comput. (2022).
[31] Y. Lei, A. Ledent, M. Kloft, Sharper generalization bounds for pairwise learning, in: Advances in Neural Information Processing Systems, Vol. 33, 2020, pp. 21236–21246.
[32] Y. Lei, M. Liu, Y. Ying, Generalization guarantee of SGD for pairwise learning, in: Advances in Neural Information Processing Systems, Vol. 34, 2021, pp. 21216–21228.
[33] C. Cortes, M. Mohri, AUC optimization vs. Error rate minimization, in: Advances in Neural Information Processing Systems, 2003.
[34] W. Gao, R. Jin, S. Zhu, Z.-H. Zhou, One-pass AUC optimization, in: International Conference on Machine Learning, 2013, pp. 906–914.
[35] M. Liu, X. Zhang, Z. Chen, X. Wang, T. Yang, Fast stochastic AUC maximization with o (1/n)-convergence rate, in: International Conference on Machine Learning, 2018, pp. 3195–3203.
[36] Y. Ying, L. Wen, S. Lyu, Stochastic online AUC maximization, in: Advances in Neural Information Processing Systems, Vol. 29, 2016.
[37] P. Zhao, S.C. Hoi, R. Jin, T. Yang, Online AUC maximization, in: International Conference on Machine Learning, 2011, pp. 233–240.
[38] A. Bellet, A. Habrard, M. Sebban, A survey on metric learning for feature vectors and structured data, 2013, arXiv preprint arXiv:1306.6709.
[39] Q. Cao, Z.-C. Guo, Y. Ying, Generalization bounds for metric and similarity learning, Mach. Learn. 102 (1) (2016) 115–132.
[40] R. Jin, S. Wang, Y. Zhou, Regularized distance metric learning: Theory and algorithm, in: Advances in Neural Information Processing Systems, Vol. 22, 2009, pp. 862–870.
[41] T. Hu, J. Fan, Q. Wu, D.-X. Zhou, Regularization schemes for minimum error entropy principle, Anal. Appl. (Singap.) 13 (04) (2015) 437–455.
[42] S. Agarwal, P. Niyogi, Generalization bounds for ranking algorithms via algorithmic stability., J. Mach. Learn. Res. 10 (2) (2009) 441–474.
[43] S. Clémençon, G. Lugosi, N. Vayatis, Ranking and empirical minimization of U-statistics, Ann. Statist. 36 (2) (2008) 844–874.
[44] C. Dwork, A. Roth, et al., The algorithmic foundations of differential privacy, Found. Trends Theor. Comput. Sci. 9 (3–4) (2014) 211–407.
[45] I. Steinwart, A. Christmann, Support Vector Machines, Springer Science & Business Media, 2008.
[46] M. Hardt, B. Recht, Y. Singer, Train faster, generalize better: Stability of stochastic gradient descent, in: International Conference on Machine Learning, 2016, pp. 1225–1234.
[47] I. Kuzborskij, C. Lampert, Data-dependent stability of stochastic gradient descent, in: International Conference on Machine Learning, 2018, pp. 2820–2829.
[48] Y. Kang, Y. Liu, J. Li, W. Wang, Sharper utility bounds for differentially private models, 2022, arXiv preprint arXiv:2204.10536.
[49] I. Mironov, Rényi differential privacy, in: 2017 IEEE 30th Computer Security Foundations Symposium, CSF, IEEE, 2017, pp. 263–275.
[50] Z. Liang, B. Wang, Q. Gu, S. Osher, Y. Yao, Exploring private federated learning with Laplacian smoothing, 2020, arXiv preprint arXiv:2005.00218.
[51] Y. Ying, D.-X. Zhou, Unregularized online learning algorithms with general loss functions, Appl. Comput. Harmon. Anal. 42 (2) (2017) 224–244.

Puyu Wang received her Ph.D. degree in Statistics from Northwest University, Xi'an, China, in 2021. From November 2019 to November 2020, she was a visiting student at the Department of Mathematics and Statistics, State University of New York at Albany. She joined Hong Kong Baptist University as a postdoc in July 2023. Previously, she was a postdoc at the City University of Hong Kong from July 2021 to July 2023. Her main research interests include statistical machine learning and differential privacy.

Yunwen Lei received his B.S. degree from Hunan University and Ph.D. degree from Wuhan University. He is an assistant professor at the Department of Mathematics, The University of Hong Kong. His research interests lie in the areas of stochastic optimization and learning theory.



Yiming Ying is a Professor at the Department of Mathematics and Statistics, SUNY Albany. He earned his Ph.D. in mathematics from Zhejiang University in 2002 and previously served as an Assistant Professor at the University of Exeter, England. His research interests are Statistical Learning Theory, Trustworthy Machine Learning, and Optimization. Yiming Ying is actively involved in the academic community, serving as an associate editor for esteemed journals such as Transactions on Machine Learning Research, Neurocomputing, Mathematical Foundation of Computing, and Mathematics of Computation and Data Science. Additionally, he serves as a Senior Program Member/Area Chair for NeurIPS, ICML, and AISTATS.

Ding-Xuan Zhou received his B.Sc. and Ph.D. degrees in mathematics from Zhejiang University, China, in 1988 and 1991, respectively. He worked as a faculty member at the City University of Hong Kong from 1996 to 2022 and joined the University of Sydney in 2022 as a professor, serving also as Head of the School of Mathematics and Statistics. His current research interests include theory of deep learning, deep neural networks, wavelet analysis, and approximation theory. Prof. Zhou is serving on the Editorial Board of over 10 international journals and is Editor-in-Chief of the journal "Analysis and Application". He was rated in 2014-17 by Thomson Reuters/Clarivate Analytics as a highly-cited researcher.