Stability and Differential Privacy of Stochastic Gradient Descent for Pairwise Learning with Non-Smooth Loss

Zhenhuan Yang[1] **Yunwen Lei**[2] **Siwei Lyu**[3] **Yiming Ying**[1,*]
University at Albany, SUNY, USA[1] University of Birmingham, UK[2] University at Buffalo, SUNY, USA[3]

Abstract

Pairwise learning has recently received increasing attention since it subsumes many important machine learning tasks (e.g. AUC maximization and metric learning) into a unifying framework. In this paper, we give the *first-ever-known* stability and generalization analysis of stochastic gradient descent (SGD) for pairwise learning with *non-smooth* loss functions, which are widely used (e.g. Ranking SVM with the hinge loss). We introduce a novel decomposition in its stability analysis to decouple the pairwisely dependent random variables, and derive generalization bounds which are consistent with the setting of pointwise learning. Furthermore, we apply our stability analysis to develop differentially private SGD for pairwise learning, for which our utility bounds match with the state-of-the-art output perturbation method (Huai et al., 2020) with smooth losses. Finally, we illustrate the results using specific examples of AUC maximization and similarity metric learning. As a byproduct, we provide an affirmative solution to an open question on the advantage of the nuclear-norm constraint over the Frobenius-norm constraint in similarity metric learning.

1 Introduction

Let the input space $\mathcal{X}$ be a compact domain of $\mathbb{R}^d$, the output space $\mathcal{Y} \subseteq \mathbb{R}$, and the domain of model parameters $\mathcal{W} \subseteq \mathbb{R}^d$. In the standard supervised learning, one aims to learn the relation between the input and output variables from a training dataset $S = \{z_i = (\mathbf{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y} : i = 1, 2, \ldots, n\}$ which is i.i.d. from an unknown distribution P on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. In such cases, the quality of a model parameter $\mathbf{w}$ is often measured by a *pointwise loss* function $\ell(\mathbf{w}, z)$.

In this paper, we are concerned with another important class of learning tasks called *pairwise learning* where the quality of a model parameter $\mathbf{w}$ is measured by a pairwise loss $\ell(\mathbf{w}, z, z')$ on pairs of examples (z, z') as opposed to the pointwise loss $\ell(\mathbf{w}, z)$ in standard classification and regression. This pairwise learning framework instantiates many important learning tasks such as similarity and metric learning (Weinberger and Saul, 2009; Xing et al., 2003; Ying and Li, 2012), AUC maximization and bipartite ranking (Agarwal and Niyogi, 2009; Clémençon et al., 2008; Gao et al., 2013; Ying et al., 2016; Zhao et al., 2011), gradient learning (Mukherjee and Wu, 2006; Mukherjee and Zhou, 2006), and minimum error entropy principle (Hu et al., 2013).

Stochastic gradient descent (SGD) has become the workhorse behind many machine learning algorithms for large-scale data analysis. SGD and its variants have been widely studied in the pointwise learning case (Bach and Moulines, 2013; Bottou and Cun, 2004; Lacoste-Julien et al., 2012; Rakhlin et al., 2012; Shalev-Shwartz et al., 2009; Ying and Zhou, 2006) as well as the pairwise learning case (Kar et al., 2013; Lin et al., 2017; Wang et al., 2012; Ying and Zhou, 2016). In particular, Kar et al. (2013); Wang et al. (2012) studied the online-to-batch conversion bounds for online pairwise learning. The work of Shen et al. (2020) studied the stability and generalization of SGD in pairwise learning and derived lower bounds for their optimization error over a class of pairwise losses. This work used the uniform stability (Agarwal et al., 2010) which was largely motivated by Hardt et al. (2016) in the pointwise case. However, there are some fundamental limitations in the work by Shen et al. (2020): it requires the pairwise loss to be both Lipschitz continuous and strongly smooth, and the parameter domain $\mathcal{W}$ is assumed to be bounded. Such assumptions are very restrictive which are violated in many cases

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. * Corresponding author.

such as the least square loss for AUC maximization $(1 - \mathbf{w}^\top(\mathbf{x} - \mathbf{x}'))^2 \mathbb{I}_{[y=1 \wedge y'=-1]}$ with $\mathbf{w} \in \mathbb{R}^d$ ($\mathbb{I}$ is the indicator function) and the hinge loss for metric learning $(1 + \tau(y,y')(\mathbf{x} - \mathbf{x}')^\top \mathbf{w}(\mathbf{x} - \mathbf{x}'))_+$ where $\mathbf{w}$ is a positive semi-definite matrix, and $\tau(y,y') = 1$ if $\mathbf{x}, \mathbf{x}'$ are from the same class and -1 otherwise.

On the other important front, the concept of stability is closely related to differential privacy (DP) (Dwork et al., 2006, 2014) which is a well accepted mathematical definition for privacy protection. While private SGD has been extensively studied (Bassily et al., 2020, 2019; Wu et al., 2017) in pointwise learning, there is litter work on differentially private SGD for pairwise learning except the very recent work of Huai et al. (2020). However, the study (Huai et al., 2020) again requires the loss to be both Lipschitz continuous and strongly smooth.

In this paper, we study the stability, generalization, and differential privacy of SGD for pairwise learning with non-smooth losses. Our contributions can be summarized as follows.

- We establish the first-ever-known stability bounds of SGD for pairwise learning with non-smooth loss functions. Our results hold true for both bounded and unbounded parameter domains. The proof techniques are mainly motivated by the recent work (Bassily et al., 2020; Lei and Ying, 2020) where stability of SGD was established in the pointwise case. The main challenge here is that pairs of examples involved in pairwise learning are not statistically independent. To overcome this hurdle, we develop a novel approach for decoupling such pairwisely dependent random variables in the analysis. We also derive the first generalization bound in high probability for SGD in pairwise learning using the stability approach.
- We study the differential privacy guarantee and utility bounds of private SGD for pairwise learning by output perturbation method. Our idea is to use our stability results to derive its sensitivity with high probability w.r.t. the randomness of algorithm, and hence guarantee its differential privacy with smaller added noise. The resulting utility bound matches with the output perturbation method in Huai et al. (2020) for private SGD in pairwise learning with smooth losses.
- We provide concrete examples of pairwise learning including AUC maximization and similarity metric learning to illustrate our stability and differential privacy results. In particular, we give an affirmative solution to the open question raised in Cao et al. (2016) that whether similarity metric learning with nuclear-norm constraint can yield milder dependence on the dimensionality than the Frobenius-norm constraint.

Other Related Work. Generalization analysis for the ERM formulation in pairwise learning was studied using U-Statistics (e.g. De la Pena and Giné (2012)) for ranking Clémençon et al. (2008); Rejchel (2012) and metric learning (Cao et al., 2016; Verma and Branson, 2015). There are a considerable amount of work on studying SGD and online learning algorithms in pairwise learning. In particular, generalization bounds for online pairwise learning algorithms were established in Kar et al. (2013); Wang et al. (2012) using online-to-batch conversion techniques (Cesa-Bianchi et al., 2004) which involves the Rademacher complexity or the covering number. The convergence (optimization error) of SGD type algorithms for pairwise learning was obtained in Lin et al. (2017); Ying and Zhou (2016) where the algorithms there directly minimize the population risk. In this setting, there is no need to consider generalization (estimation error) i.e. the difference between the empirical risk and the true population risk.

Algorithmic stability and generalization bounds were established in Agarwal and Niyogi (2009) for ranking problems, and in Jin et al. (2009) for regularized metric learning with a strongly convex objective function, and both studies considered the ERM formulation with a strongly convex objective function. Recently, the uniform stability and its trade-off with optimization errors were studied in Shen et al. (2020) for SGD in pairwise learning, which is inspired by the recent work in pointwise learning (Charles and Papailiopoulos, 2018; Hardt et al., 2016; Kuzborskij and Lampert, 2018). However, the loss there is assumed to be Lipschitz and strongly smooth and the domain $\mathcal{W}$ needs to be bounded.

The concept of stability was recently used to study the generalization (utility) of differentially private SGD algorithms, particularly in pointwise learning. Specifically, the work of Wu et al. (2017) studied the output perturbation using sensitivity analysis which is very close to the concept of uniform stability. In Bassily et al. (2019), using stability approach, the optimal excess generalization bound $\tilde{\mathcal{O}}\big(\max\{1/\sqrt{n}, \sqrt{d}/(n\epsilon)\}\big)$ was established for (ϵ,δ)-DP algorithms which, however, requires the loss function to be Lipschitz and strongly smooth, and the domain $\mathcal{W}$ be bounded. For the non-smooth loss, it proposed to smooth the loss by its Moreau envelope function which is not an ideal solution as the Moreau envelope function is not easy to compute for a general loss. In Feldman et al. (2020), multi-phrased SGD were proposed with the optimal population risk in which, for the non-smooth case, their algorithm is significantly more involved than the noisy SGD algorithm. In regard to the differential private SGD in the pairwise case, the only work that we are aware of is Huai et al. (2020) which studied both gradient perturbation and output perturbation with Gaussian noise. They derive the rate $\tilde{\mathcal{O}}\big(\sqrt{d}/(\sqrt[4]{n}\epsilon)\big)$ for gradient perturbation and $\tilde{\mathcal{O}}\big(\sqrt{d}/(\sqrt{n}\epsilon)\big)$ for output perturbation. Note that the loss function there needs to be both Lipschitz continuous and strongly smooth.

2 Main Results

Before stating our main results, we first introduce necessary materials and notations. Given a pairwise loss function $\ell : \mathcal{W} \times \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$, we aim to minimize the following population risk

$$R(\mathbf{w}) = \mathbb{E}_{\mathbf{z},\mathbf{z}'}[\ell(\mathbf{w}, \mathbf{z}, \mathbf{z}')],$$

where $\mathbf{z}$ and $\mathbf{z}'$ are drawn independently from the population distribution P on $\mathcal{Z}$. The population distribution is often unknown and we only have access to a set of i.i.d. training data $S = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n\} \in \mathcal{Z}^n$. The task then reduces to minimizing the empirical risk

$$\min_{\mathbf{w}\in\mathcal{W}} R_S(\mathbf{w}) := \frac{1}{n(n-1)} \sum_{i,j=1, i\neq j}^{n} \ell(\mathbf{w}, \mathbf{z}_i, \mathbf{z}_j). \quad (1)$$

Randomized optimization algorithm $\mathcal{A} : \mathcal{Z}^n \to \mathcal{W}$ provides an efficient approach to find an approximate solution to problem (1), which takes S as input and produces an output $\mathcal{A}(S) \in \mathcal{W}$. The randomized algorithm $\mathcal{A}$ here can be either SGD for pairwise learning or its noisy variant for differential privacy. The performance of $\mathcal{A}$ is quantified by the excess population risk: $\epsilon_{\text{risk}}(\mathcal{A}(S)) = R(\mathcal{A}(S)) - \inf_{\mathbf{w}\in\mathcal{W}} R(\mathbf{w})$. We can decompose $\epsilon_{\text{risk}}(\mathcal{A}(S))$ as follows:

$$\begin{aligned}\epsilon_{\text{risk}}(\mathcal{A}(S)) =& [R(\mathcal{A}(S)) - R_S(\mathcal{A}(S))] + [R_S(\mathbf{w}_*) - R(\mathbf{w}_*)] \\ &+ [R_S(\mathcal{A}(S)) - R_S(\mathbf{w}_*)], \quad (2)\end{aligned}$$

where $\mathbf{w}_* \in \arg\min_{\mathbf{w}\in\mathcal{W}} R(\mathbf{w})$. The first term on the right hand side of (2) is called the estimation error. Since $\mathbf{w}_*$ is fixed, the term $R_S(\mathbf{w}_*) - R(\mathbf{w}_*)$ can be trivially handled by the standard Hoeffding inequality. As a comparison, the estimation of the term $R(\mathcal{A}(S)) - R_S(\mathcal{A}(S))$, also called the generalization error, is much more challenging since $\mathcal{A}(S)$ depends on S. We will develop novel stability analysis to handle this term. The last term $R_S(\mathcal{A}(S)) - R_S(\mathbf{w}_*)$ is called the optimization error and we can bound it by applying optimization theory.

We now introduce some necessary assumptions and definitions. Let $\|\cdot\|_2$ denote the Euclidean norm on $\mathbb{R}^d$ and $\langle\cdot,\cdot\rangle$ denote the corresponding inner product. Given a function $f : \mathcal{W} \to \mathbb{R}$, let $\partial f(\mathbf{w})$ be a subgradient of f at $\mathbf{w}$. A function f is said to be convex if

Algorithm 1 SGD for Pairwise Learning

Input: Data set $S = \{\mathbf{z}_1, \cdots, \mathbf{z}_n\}$, step size η, number of iterations T, initial point $\mathbf{w}_1 = 0$ and initial sample $i_1 \in [n]$ from uniform distribution
for $t = 1$ to T **do**
  Select $i_{t+1} \in [n]$ by uniform distribution
  $\mathbf{w}_{t+1} = \Pi_{\mathcal{W}}\big(\mathbf{w}_t - \frac{\eta}{t}\sum_{k=1}^{t} \partial\ell(\mathbf{w}_t, \mathbf{z}_{i_{t+1}}, \mathbf{z}_{i_k})\big)$
end for
Output: $\bar{\mathbf{w}}_T = \frac{1}{T}\sum_{t=1}^{T} \mathbf{w}_t$

for any $\mathbf{w}, \mathbf{w}' \in \mathcal{W}$, there holds

$$f(\mathbf{w}') \geq f(\mathbf{w}) + \langle \partial f(\mathbf{w}), \mathbf{w}' - \mathbf{w}\rangle.$$

A function f is said to be G-Lipschitz continuous if, for any $\mathbf{w}, \mathbf{w}' \in \mathcal{W}$, there holds

$$|f(\mathbf{w}) - f(\mathbf{w}')| \leq G\|\mathbf{w} - \mathbf{w}'\|_2.$$

Throughout this paper, we assume that the (possibly non-smooth) loss function $\ell(\mathbf{w}, \mathbf{z}, \mathbf{z}')$ is nonnegative, convex and G-Lipschitz continuous w.r.t $\mathbf{w}$.

2.1 Stability and Excess Risk Analysis

In this subsection, we consider the stability and generalization of the SGD algorithms for pairwise learning. The SGD algorithm is described in Algorithm 1 which has been widely discussed in Lin et al. (2017); Wang et al. (2012); Ying and Zhou (2016). Note that $\Pi_{\mathcal{W}}(\cdot)$ is the projection onto the parameter space $\mathcal{W}$ and $[n] = \{1, \ldots, n\}$. In this subsection, the notation $\mathcal{A}$ denotes Algorithm 1.

In particular, we will use the uniform argument stability (UAS) (Liu et al., 2017) where its original concept was stated in expectation w.r.t. the internal randomness of $\mathcal{A}$. We will use its probabilistic version here. Specifically, let $S = \{\mathbf{z}_1, \cdots, \mathbf{z}_n\}$ and $S' = \{\mathbf{z}'_1, \cdots, \mathbf{z}'_n\}$ be two neighborhood datasets that differ only in one single example. For any $\gamma \in (0, 1)$, $\mathcal{A}$ is called ϵ_{stab}-UAS with probability $1 - \gamma$ if for any neighborhood datasets S and S',

$$\mathbb{P}_{\mathcal{A}}[\|\mathcal{A}(S) - \mathcal{A}(S')\|_2 > \epsilon_{\text{stab}}] \leq \gamma.$$

We emphasize the probability here is taken over the internal randomness of $\mathcal{A}$, i.e. the uniform distribution of generating i_t's.

The following theorem states a high-probability UAS result for Algorithm 1 with non-smooth losses. Here, $\mathbf{w}_{t+1}$ and $\mathbf{w}'_{t+1}$ denote the $(t+1)$-th iterate of Algorithm 1 based on samples S and S', respectively. And, the notation $\tilde{\mathcal{O}}(\cdot)$ indicates that the bound is up to a logarithmic term.

Theorem 1. *Suppose that we run Algorithm 1 under random selection with replacement for t iterations based on S and S'. Then, with probability $1-\gamma$ w.r.t. the internal randomness of $\mathcal{A}$, we have, for any S and S', that*

$$\|\mathbf{w}_{t+1} - \mathbf{w}'_{t+1}\|_2^2 \le 4e\eta^2 G^2 \Big[t + \ln^2(et) \times \Big(\frac{t}{n} + \ln(1/\gamma) + \sqrt{\frac{t\ln(1/\gamma)}{n}}\Big)^2\Big] \quad (3)$$

In particular, if $T \ge n$, then the output of Algorithm 1 is ϵ_{stab}-UAS with high probability where

$$\epsilon_{stab} = \tilde{\mathcal{O}}\Big(\eta\sqrt{T} + \frac{\eta T \ln(T)}{n}\Big).$$

The proof of Theorem 1 is given in Section 3.1. This bound matches the result in the pointwise learning with non-smooth losses (Bassily et al., 2020; Lei and Ying, 2020) up to a logarithmic term of T. The proof is motivated by Lei and Ying (2020) in the pointwise case but more involved in pairwise learning. Indeed, the key challenge, in comparison with pointwise learning, is that the sub-gradient estimator at the t-th step depends not only on the current example $\mathbf{z}_{i_{t+1}}$ but also on previous examples $\{\mathbf{z}_{i_k} : k = 1, \cdots, t\}$.

To our best knowledge, Shen et al. (2020) is the only available work which considered the stability of SGD in pairwise learning. However, their work required the loss to be Lipschitz continuous and strongly smooth to ensure the non-expansiveness of the gradient update, which is very critical for the proof of the main results there. The non-smoothness assumption in our paper makes the corresponding gradient update no longer non-expansive, and therefore the arguments in Shen et al. (2020) no longer apply. We bypass this obstacle by a refined control of the expansiveness between adjacent steps. To address this dependence issue, the work of Shen et al. (2020) counts the number m of different examples $\mathbf{z}_i \neq \mathbf{z}'_i$ encountered by SGD until iteration t, which obeys a binomial distribution. In contrast, high-probability analysis here for non-smooth loss is more challenging and involved because directly applying concentration inequality to similar binomial distribution yields an undesired estimation. We overcome this hurdle by decomposing the sub-gradients into sum of t pairs of dependent random variables first, and then upper bound this sum by two sums of independent random variables. From this new decomposition, we can apply the Chernoff-type tail bounds to these two sums of independent random variables to get the desired estimation. One can see Section 3.1 for more details.

Based on Theorem 1 and the error decomposition (2), we derive the excess risk bounds for bounded (Theorem 2) and unbounded domains (Theorem 3). To bound the optimization error, we need the following variant of Rademacher average (Bartlett and Mendelson, 2002)

$$\mathcal{R}_t(\ell \circ \mathcal{W}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\Big[\sup_{\mathbf{w}\in\mathcal{W}} \frac{1}{t}\sum_{k=1}^{t} \sigma_k \ell(\mathbf{w}, \mathbf{z}_i, \mathbf{z}_{i_k})\Big]. \quad (4)$$

Here σ_k are Rademacher random variables taking values in $\{\pm 1\}$ with equal probability $1/2$, and the expectation is taken over $\mathbf{z}_i$, $\mathbf{z}_{i_k}$ and σ_k.

Theorem 2. *Suppose $\mathcal{W}$ is bounded with diameter D. Denote $M = \sup_{\mathbf{z},\mathbf{z}'} \ell(0, \mathbf{z}, \mathbf{z}')$. Assume we run Algorithm 1 for $T \ge n$ iterations under random selection with replacement rule. Then for any $\gamma \in (0,1)$, with probability at least $1-\gamma$ w.r.t. the sample S and the internal randomness of $\mathcal{A}$, we have*

$$\epsilon_{risk}(\bar{\mathbf{w}}_T) \le \frac{4}{T}\sum_{t=1}^{T}\mathcal{R}_t(\ell\circ\mathcal{W}) + \frac{D^2}{2T\eta} + \frac{\eta G^2}{2} + c_2\sqrt{\frac{\ln(6T/\gamma)}{n}} + c_1\eta\lceil\ln(n)\rceil\Big(\sqrt{T} + \frac{\sqrt{3}T\ln(eT)\ln(6/\gamma)}{n}\Big),$$

where $c_1 = 100\sqrt{6}e^{3/2}G\max\{1,G\}\ln(6e/\gamma)$ and $c_2 = (6+19e)(M+GD)$.

In particular, if $\mathcal{R}_t(\ell\circ\mathcal{W}) = \mathcal{O}(1/\sqrt{t})$ and we choose $T = n^2$ and $\eta = \mathcal{O}(n^{-3/2})$ then with high probability we have

$$\epsilon_{risk}(\bar{\mathbf{w}}_T) = \tilde{\mathcal{O}}\Big(\frac{\ln^2(n)}{\sqrt{n}}\Big).$$

Theorem 2 is proved in Appendix A.2. Using standard technique (Bartlett and Mendelson, 2002), the Rademacher complexity estimation of $\mathcal{R}_t(\ell\circ\mathcal{W}) = \mathcal{O}(1/\sqrt{t})$ holds true in many cases when $\mathcal{X}$ and $\mathcal{W}$ are bounded (e.g. see Section 4 for concrete examples of AUC maximization and similarity metric learning). It is worthy of mentioning that the choice of $T = n^2$ is consistent with pointwise learning with non-smooth loss (Bassily et al., 2020; Lei and Ying, 2020).

We can also derive excess generalization bounds for Algorithm 1 even when $\mathcal{W}$ is unbounded. Specifically, let $D = \|\mathbf{w}_*\|_2$ and $\mathcal{W}_D = \{\mathbf{w} \in \mathcal{W} \,|\, \|\mathbf{w}\|_2 \le D\}$. The main idea is to show that the iterate $\mathbf{w}_t$ from Algorithm 1 has an adaptive bound, i.e. $\mathbf{w}_t \in \mathcal{W}_t = \{\mathbf{w}\in\mathcal{W} \,|\, \|\mathbf{w}\|_2^2 \le (G^2+M)\eta t\}$.

Theorem 3. *Denote $M = \sup_{\mathbf{z},\mathbf{z}'}\ell(0,\mathbf{z},\mathbf{z}')$ and $D = \|\mathbf{w}_*\|_2$. Suppose we run Algorithm 1 for $T \ge n$ iterations. For any $\gamma \in (0,1)$, with probability at least $1-\gamma$ w.r.t. the sample S and the internal randomness of $\mathcal{A}$,* *we have*

$$\epsilon_{risk}(\bar{\mathbf{w}}_T) \leq \frac{2}{T}\sum_{t=1}^{T}\big(\mathcal{R}_t(\ell\circ\mathcal{W}_t)+\mathcal{R}_t(\ell\circ\mathcal{W}_D)\big)+\frac{D^2}{2T\eta}+\frac{\eta G^2}{2}$$
$$+c_4\sqrt{\eta\ln(6T/\gamma)}+c_5\sqrt{\frac{\eta T\ln(6e/\gamma)}{n}}+c_3\sqrt{\frac{\ln(6T/\gamma)}{n}}$$
$$+c_1\eta\lceil\ln(n)\rceil\Big(\sqrt{T}+\frac{4T\ln(eT)\sqrt{\ln(6n/\gamma)}}{n}\Big),$$

where $c_1 = 100\sqrt{6}e^{3/2}G\max\{1, G\}\ln(6e/\gamma)$, $c_3 = (7+12\sqrt{2}e)M+4GD+16eG$, $c_4 = 3G\sqrt{G^2+2M}$ *and* $c_5 = 12\sqrt{2}eG\sqrt{G^2+2M}$.

In particular, if $\mathcal{R}_t(\ell\circ\mathcal{W}_t) = \mathcal{O}(\eta\sqrt{t})$ *and* $\mathcal{R}_t(\ell\circ\mathcal{W}_D) = \mathcal{O}(1/\sqrt{t})$ *and we choose* $T = n^{4/3}$ *and* $\eta = \mathcal{O}(n^{-1})$, *then with high probability we have*

$$\epsilon_{risk}(\bar{\mathbf{w}}_T) = \tilde{\mathcal{O}}\Big(\frac{\ln^2(n)}{n^{1/3}}\Big).$$

Theorem 3 is proved in Appendix A.3. In particular, one can show that the Rademacher complexity can be estimated using standard technique (Bartlett and Mendelson, 2002) such that $\mathcal{R}_t(\ell\circ\mathcal{W}_D) = \mathcal{O}(D/\sqrt{t})$ when $\mathcal{X}$ is a bounded domain. Therefore by the definition of $\mathcal{W}_t$ one can similarly show that $\mathcal{R}_t(\ell\circ\mathcal{W}_t) = \mathcal{O}(\eta t/\sqrt{t}) = \mathcal{O}(\eta\sqrt{t})$. One can see more discussion on such estimation in Section 4. Therefore, Theorem 3 mainly differs from Theorem 2 in the additional $\tilde{\mathcal{O}}\big(\sqrt{\eta T/n}\big)$ term where $T \geq n$. This is due to the unboundedness of $\mathcal{W}$. Our excess risk bound is consistent with the results in Lin et al. (2016) in the pointwise setting (up to a logarithmic term), where the authors studied SGD for non-smooth loss functions in the pointwise setting using uniform convergence. However, the bound there is given in expectation while we have provided a high-probability bound.

2.2 Differentially Private Pairwise Learning

We show the implication of stability analysis in analyzing differentially private SGD in pairwise learning. We start by introducing the notion of differential privacy.

Definition 1 (Differential Privacy (Dwork et al., 2006))**.** A (randomized) algorithm $\mathcal{A}$ is called (ϵ, δ)-differentially private (DP) if, for all neighboring datasets S, S' differing by only one example and for all events O in the output space of $\mathcal{A}$, the following holds

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

There are other forms of differential privacy such as Gaussian differential privacy (Bu et al., 2020; Dong et al., 2019). In this paper we restrict our attention to the standard DP mentioned above. In particular, we consider Gaussian mechanism (Dwork et al., 2006), i.e. given any query function $q : \mathcal{S}^n \to \mathbb{R}^d$, let $\mathcal{A}(S) = q(S)+\mathbf{u}$ where $\mathbf{u}\sim\mathcal{N}(0, \sigma^2\mathbf{I}_d)$ with $\mathbf{I}_d$ being the identical matrix. For all neighborhood datasets S, S' that differ by one example, the ℓ_2-sensitivity Δ of the query function q is defined as $\Delta(q) = \sup_{S,S'}\|q(S)-q(S')\|_2$.

Algorithm 2 Private SGD for Pairwise Learning with Output Perturbation

Input: Private dataset $S = \{\mathbf{z}_1, \cdots, \mathbf{z}_n\}$, privacy parameter ϵ, δ, stepsize η, number of iterations T, initial point $\mathbf{w}_1 = 0$ and initial sample $i_1 \in [n]$ from uniform distribution
for $t = 1$ to T **do**
  Select $i_{t+1} \in [n]$ from uniform distribution
  $\mathbf{w}_{t+1} = \Pi_{\mathcal{W}}\Big(\mathbf{w}_t - \frac{\eta}{t}\sum_{k=1}^{t}\partial\ell(\mathbf{w}_t, \mathbf{z}_{i_{t+1}}, \mathbf{z}_{i_k})\Big)$
end for
$\bar{\mathbf{w}}_T = \frac{1}{T}\sum_{t=1}^{T}\mathbf{w}_t$
Sample $\mathbf{u}\sim\mathcal{N}(0, \sigma^2\mathbf{I}_d)$ with σ^2 being given by (5)
Output: $\mathbf{w}_{\text{priv}} = \Pi_{\mathcal{W}}\big(\bar{\mathbf{w}}_T+\mathbf{u}\big)$

We develop a private version of SGD for pairwise learning. In this subsection, the notation $\mathcal{A}$ denotes Algorithm 2. The idea is to add Gaussian noise to the output of the non-private Algorithm 1. In return, Algorithm 2 is guaranteed to be (ϵ, δ)-DP by properly choosing σ as shown below.

Theorem 4. *Given the total number of iterations* T, *for any privacy budget* $\epsilon > 0$ *and* $\delta > 0$, *Algorithm 2 satisfies* (ϵ, δ)*-differential privacy with*

$$\sigma^2 = \frac{8e\eta^2G^2\ln(2.5/\delta)}{\epsilon^2}\Big(T+\frac{3T^2\ln^2(eT)\ln^2(2/\delta)}{n^2}\Big). \quad (5)$$

The proof of Theorem 4 is given in Section 3.2. The goal here is to guarantee privacy with the added noise being as small as possible. The key observation is the UAS of the non-private output $\bar{\mathbf{w}}_T$ can be used to quantify the high-probability sensitivity of the query function $q(S) = \bar{\mathbf{w}}_T$. Specifically, subsampling forms an event of probability measure $1-\delta/2$ under which a small sensitivity $\tilde{\mathcal{O}}(\eta\sqrt{T}+\eta T\ln(T)/n)$ holds true. Hence, under this event, we only need to add noise with $\sigma = \tilde{\mathcal{O}}((\eta\sqrt{T}+\eta T\ln(T))\ln(2/\delta)/(n\epsilon))$ to guarantee a slightly restrictive $(\epsilon, \delta/2)$-DP. Therefore the algorithm is $(\epsilon;\delta)$-DP over the whole event space. Wu et al. (2017) studied differential private SGD by output perturbation method in the pointwise learning setting and they also utilized the idea of bounding sensitivity by UAS. However, they considered the stability and sensitivity regardless of the randomness of the algorithm, which is not suitable for high probability analysis of utility bound later. In contrast, our technique can also be applied to derive privacy guarantee and high probability utility in pointwise learning. Huai et al. (2020) also studied the sensitivity of SGD for Pairwise learning. However, they focused on the online setting where the data arrives in a streaming manner, and hence the different example between S and S' will only appear once in the algorithm. While in our stochastic setting the different example can be used more than once by subsampling, it is more challenging to measure the sensitivity. Moreover, their analysis depends on the strong smoothness of the loss function while we allow the loss function to be non-smooth.

In order to derive the utility bound of Algorithm 2, we need a new error decomposition scheme as follow

$$\begin{aligned}\epsilon_{\text{risk}}(\mathbf{w}_{\text{priv}}) &= R(\mathbf{w}_{\text{priv}}) - R(\mathbf{w}_*)\\ &= R(\mathbf{w}_{\text{priv}}) - R(\bar{\mathbf{w}}_T) + R(\bar{\mathbf{w}}_T) - R(\mathbf{w}_*),\end{aligned} \tag{6}$$

where $R(\bar{\mathbf{w}}_T) - R(\mathbf{w}_*)$ measures the excess risk incurred by the non-private output $\bar{\mathbf{w}}_T$ (Algorithm 1) and $R(\mathbf{w}_{\text{priv}}) - R(\bar{\mathbf{w}}_T)$ measures the effect of perturbation by adding random noises. The utility bound is given as follow.

Theorem 5. *Suppose $\mathcal{W}$ is bounded with diameter D. Consider Algorithm 2 for T iterations under random selection with replacement rule. For any privacy budget $\epsilon > 0$, $\delta > 0$, and for any $\gamma \in (\max\{4\delta, \exp(-d/8)\}, 1)$, with probability at least $1-\gamma$, we have*

$$\begin{aligned}\epsilon_{risk}(\mathbf{w}_{priv}) \le \frac{4}{T}\sum_{t=1}^{T}\mathcal{R}_t(\ell\circ\mathcal{W}) + \frac{D^2}{2T\eta} + \frac{\eta G^2}{2}\\ + c_2\sqrt{\frac{\ln(6T/\gamma)}{n}} + 2G\sigma\sqrt{d}\ln^{1/4}(4/\gamma).\\ + c_1\eta\lceil\ln(n)\rceil\Big(\sqrt{T} + \frac{\sqrt{3}T\ln(eT)\ln(2/\delta)}{n}\Big),\end{aligned}$$

where $c_1 = 100\sqrt{6}e^{3/2}G\max\{1, G\}\ln(6e/\gamma)$ and $c_2 = (6+19e)(M+GD)$.

In particular, letting σ satisfy (5) and choosing $T = n^2$ and $\eta = \mathcal{O}(n^{-3/2})$, then with high probability we have

$$\epsilon_{risk}(\mathbf{w}_{priv}) = \tilde{\mathcal{O}}\Big(\frac{\sqrt{d}}{\sqrt{n}\epsilon}\Big).$$

Theorem 5 is proved in Appendix A.4. The difference compared to Theorem 2 is the additional $\tilde{\mathcal{O}}(\sigma\sqrt{d})$ term caused by $R(\mathbf{w}_{\text{priv}}) - R(\bar{\mathbf{w}}_T)$ in (6). The utility bound $\tilde{\mathcal{O}}\big(\sqrt{d}/(\sqrt{n}\epsilon)\big)$ matches that of the output perturbation for pairwise learning studied in Huai et al. (2020) which, however, requires the loss to be both strongly smooth and Lipschitz continuous. Our analysis only needs the loss to be Lipschitz continuous.

3 Main Proofs for Theorems 1 and 4

In this section, we provide technical proofs for Theorems 1 and 4. Proofs of other Theorems can be found in the Appendix. Throughout this section, we let $\hat{L}_{t+1}(\mathbf{w}_t)$ denote the accumulated loss until $\mathbf{z}_{i_{t+1}}$ is revealed. i.e. $\hat{L}_{t+1}(\mathbf{w}_t) = \frac{1}{t}\sum_{k=1}^{t}\ell(\mathbf{w}_t, \mathbf{z}_{i_{t+1}}, \mathbf{z}_{i_k})$.

3.1 Proof of Theorem 1

To prove Theorem 1, we need the following Chernoff's bound for a summation of independent Bernoulli random variables (Wainwright, 2019).

Lemma 1 (Chernoff bound for Bernoulli vector)**.** *Let $X_1, \ldots, X_t$ be independent random variables taking values in $\{0,1\}$. Let $X = \sum_{j=1}^{t} X_j$ and $\mu = \mathbb{E}[X]$. Then for any $\tilde{\gamma} > 0$, with probability at least $1 - \exp\big(-\mu\tilde{\gamma}^2/(2+\tilde{\gamma})\big)$ we have $X \le (1+\tilde{\gamma})\mu$.*

Proof of Theorem 1. Without loss of generality, assume that S and S' differs in n-th position. Denote $\delta_{t+1,k} = \partial\ell(\mathbf{w}_t, \mathbf{z}_{i_{t+1}}, \mathbf{z}_{i_k}) - \partial\ell(\mathbf{w}'_t, \mathbf{z}_{i_{t+1}}, \mathbf{z}_{i_k})$ and $\delta'_{t+1,k} = \partial\ell(\mathbf{w}_t, \mathbf{z}_{i_{t+1}}, \mathbf{z}_{i_k}) - \partial\ell(\mathbf{w}'_t, \mathbf{z}'_{i_{t+1}}, \mathbf{z}'_{i_k})$. The following recursive inequality holds

$$\begin{aligned}\|\mathbf{w}_{t+1} - \mathbf{w}'_{t+1}\|_2^2 &= \big\|\mathbf{w}_t - \eta\partial\hat{L}_{t+1}(\mathbf{w}_t) - \mathbf{w}'_t + \eta\partial\hat{L}'_{t+1}(\mathbf{w}'_t)\big\|_2^2\\ &= \Big\|\mathbf{w}_t - \mathbf{w}'_t - \frac{\eta}{t}\sum_{k=1}^{t}\delta'_{t+1,k}\Big\|_2^2\\ &\le \frac{1}{t}\sum_{k=1}^{t}\big\|\mathbf{w}_t - \mathbf{w}'_t - \eta\delta'_{t+1,k}\big\|_2^2.\end{aligned} \tag{7}$$

Now we estimate the term on the right hand side of (7) by considering two cases. For the case $i_{t+1} \neq n$ and $i_k \neq n$, we have $\mathbf{z}_{i_{t+1}} = \mathbf{z}'_{i_{t+1}}$ and $\mathbf{z}_{i_k} = \mathbf{z}'_{i_k}$. Then

$$\begin{aligned}&\big\|\mathbf{w}_t - \mathbf{w}'_t - \eta\delta_{t+1,k}\big\|^2\\ =&\big\|\mathbf{w}_t - \mathbf{w}'_t\|_2^2 + \eta^2\|\delta_{t+1,k}\big\|_2^2 - 2\eta\langle\mathbf{w}_t - \mathbf{w}'_t, \delta_{t+1,k}\rangle\\ \le&\big\|\mathbf{w}_t - \mathbf{w}'_t\big\|_2^2 + 4\eta^2G^2,\end{aligned}$$

where the last inequality holds because ℓ is G-Lipschitz and convex. If $i_{t+1} = n$ or $i_k = n$, then $\mathbf{z}_{i_{t+1}} \neq \mathbf{z}'_{i_{t+1}}$ or $\mathbf{z}_{i_k} \neq \mathbf{z}'_{i_k}$. It follows from the Young's inequality that for any $p > 0$

$$\begin{aligned}&\|\mathbf{w}_t - \mathbf{w}'_t - \eta\delta'_{t+1,k}\|^2\\ \le&(1+p)\|\mathbf{w}_t - \mathbf{w}'_t\|_2^2 + (1+1/p)\eta^2\|\delta'_{t+1,k}\|_2^2\\ \le&(1+p)\|\mathbf{w}_t - \mathbf{w}'_t\|_2^2 + 4(1+1/p)\eta^2G^2.\end{aligned}$$

Combining the above two inequalities together and let $Y_t = \frac{1}{t}\sum_{k=1}^{t}\mathbb{I}_{[i_{t+1}=n \vee i_k=n]}$, we have

$$\|\mathbf{w}_{t+1} - \mathbf{w}'_{t+1}\|_2^2 \le (1+pY_t)\|\mathbf{w}_t - \mathbf{w}'_t\|_2^2 + 4(1+Y_t/p)\eta^2G^2.$$

Applying the above inequality recursively we have

$$\|\mathbf{w}_{t+1}-\mathbf{w}'_{t+1}\|_2^2 \overset{(a)}{\leq} \sum_{j=1}^{t}\prod_{l=j+1}^{t}(1+pY_l)(4+4Y_j/p)\eta_j^2G^2$$
$$\overset{(b)}{\leq} \sum_{j=1}^{t}\prod_{l=j+1}^{t}(1+p)^{Y_l}(4+4Y_j/p)\eta_j^2G^2$$
$$\overset{(c)}{\leq} (1+p)^{\sum_{l=1}^{t}Y_l}\eta^2G^2\big(4t+4\sum_{l=1}^{t}Y_l/p\big), \tag{8}$$

where (a) is due to the recursive relation, (b) is due to $1+ax \leq (1+a)^x$ for $a>0$ and $x\geq 0$ and (c) inequality is due to $\prod_{i=a}^{b}x^i \leq x^{\sum_{i=1}^{b}i}$ for $a\geq 1$. We note that $Y_1,\cdots,Y_t$ are dependent variables, but the sum of Y_l's has the following decomposition:

$$\sum_{l=1}^{t}Y_l=\sum_{l=1}^{t}\frac{1}{l}\sum_{k=1}^{l}\mathbb{I}_{[i_{l+1}=n\vee i_k=n]}\leq\sum_{l=1}^{t}\frac{1}{l}\sum_{k=1}^{l}\big(\mathbb{I}_{[i_{l+1}=n]}+\mathbb{I}_{[i_k=n]}\big)$$
$$=\sum_{l=1}^{t}\mathbb{I}_{[i_{l+1}=n]}+\sum_{l=1}^{t}\frac{1}{l}\sum_{k=1}^{l}\mathbb{I}_{[i_k=n]}$$
$$\leq\sum_{l=1}^{t}\mathbb{I}_{[i_{l+1}=n]}+\ln(t)\sum_{k=1}^{l}\mathbb{I}_{[i_k=n]}\leq\ln(et)\sum_{k=1}^{t+1}\mathbb{I}_{[i_k=n]}.$$

Applying Lemma 1 with $X_k=\mathbb{I}_{[i_k=n]}$ and $X=\sum_{k=1}^{t}X_k$, with probability at least $1-\gamma$, we have

$$\sum_{k=1}^{t+1}\mathbb{I}_{[i_k=n]}\leq\frac{t+1}{n}+\ln(1/\gamma)+\sqrt{\frac{(t+1)\ln(1/\gamma)}{n}}.$$

For the simplicity of notation, let $c_{\gamma,t}=\ln(et)((t+1)/n+\ln(1/\gamma)+\sqrt{(t+1)\ln(1/\gamma)/n})$. Plugging the above inequality back into (8), we derive the following inequality with probability $1-\gamma$

$$\|\mathbf{w}_{t+1}-\mathbf{w}'_{t+1}\|_2^2\leq 4\eta^2G^2(1+p)^{c_{\gamma,t}}(t+c_{\gamma,t}/p).$$

By selecting $p=1/c_{\gamma,t}$ in the above equality, we have $(1+p)^{c_{\gamma,t}}\leq e$. Therefore we have proved (3) in Theorem 1. Now, since the bound on left hand side of (3) is monotonically increasing, with probability $1-\gamma$, we have

$$\|\bar{\mathbf{w}}_T-\bar{\mathbf{w}}'_T\|_2^2\leq\frac{1}{T}\sum_{t=1}^{T}\|\mathbf{w}_T-\mathbf{w}'_T\|_2^2$$
$$\leq 4e\eta^2G^2\Big(T+\frac{3T^2\ln^2(eT)\ln^2(1/\gamma)}{n^2}\Big), \tag{9}$$

where we have used the fact that $T\geq n$. Therefore the ϵ_{stab}-UAS bound holds by calling the convexity of ℓ_2-norm. □

3.2 Proof of Theorem 4

In order to establish the privacy guarantee of Algorithm 2, we need the following lemmas. The first lemma characterizes the necessary scale of σ of Gaussian mechanism (Dwork et al., 2014).

Lemma 2 (Gaussian mechanism). *For a Gaussian mechanism $\mathcal{A}(S)=q(S)+\mathbf{u}$ with $\mathbf{u}\sim\mathcal{N}(0,\sigma^2\mathbf{I}_d)$, if q has ℓ_2-sensitivity $\Delta(q)$ and assume that $\sigma\geq\sqrt{2\ln(1.25/\delta)}\Delta(q)/\epsilon$, then $\mathcal{A}$ yields (ϵ,δ)-DP.*

The next lemma indicates that differential privacy is immune to post-processing (Dwork et al., 2014).

Lemma 3 (Post-processing). *Let $\mathcal{A}:\mathcal{Z}^n\to\mathcal{W}$ be a (randomized) algorithm that is (ϵ,δ)-DP. Let $f:\mathcal{W}\to\mathcal{W}$ be an arbitrary randomized mapping. Then $f\circ\mathcal{A}:\mathcal{Z}^n\to\mathcal{W}$ is (ϵ,δ)-DP.*

Proof of Theorem 4. Consider the mechanism $\mathcal{A}'_T=\bar{\mathbf{w}}_T+\mathbf{u}$ and for any S,S', consider the ℓ_2-sensitivity $\Delta_T=\|\bar{\mathbf{w}}_T-\bar{\mathbf{w}}'_T\|_2$. Let $I=\{i_1,\cdots,i_T\}$ be the sequence of sampling after T iterations in Algorithm 2. Choosing $\gamma=\delta/2$ in Equation (9), then the event

$$E=\Big\{I|\Delta_T^2\leq 4e\eta^2G^2\Big(T+\frac{3T^2\ln^2(eT)\ln^2(2/\delta)}{n^2}\Big)\Big\}$$

satisfies $\mathbb{P}[I\in E]\geq 1-\delta/2$. When $I\in E$, Lemma 2 implies $\mathcal{A}'_T$ satisfies $(\epsilon,\delta/2)$-DP when

$$\sigma=\frac{\sqrt{2\ln(2.5/\delta)}\Delta_T}{\epsilon}.$$

Furthermore, by Lemma 3, the final output $\mathbf{w}_{\text{priv}}=\Pi_{\mathcal{W}}(\mathcal{A}'_T)$ also satisfies $(\epsilon,\delta/2)$-DP. Therefore, for any $\epsilon>0$ and any event O in the output space of $\mathbf{w}_{\text{priv}}$,

$$\mathbb{P}\big[\mathbf{w}_{\text{priv}}(S)\in O\big]=\mathbb{P}\big[\mathbf{w}_{\text{priv}}(S)\in O\cap I\in E\big]$$
$$+\mathbb{P}\big[\mathbf{w}_{\text{priv}}(S)\in O\cap I\notin E\big]$$
$$=\mathbb{P}\big[\mathbf{w}_{\text{priv}}(S)\in O|I\in E\big]\mathbb{P}\big[I\in E\big]$$
$$+\mathbb{P}\big[\mathbf{w}_{\text{priv}}(S)\in O|I\notin E\big]\mathbb{P}\big[I\notin E\big]$$
$$\leq\Big(e^{\epsilon}\mathbb{P}[\mathbf{w}_{\text{priv}}(S')\in O|I\in E]+\frac{\delta}{2}\Big)\mathbb{P}[I\in E]+\frac{\delta}{2}$$
$$\leq e^{\epsilon}\mathbb{P}\big[\mathbf{w}_{\text{priv}}(S')\in O\cap I\in E\big]+\frac{\delta}{2}+\frac{\delta}{2}$$
$$\leq e^{\epsilon}\mathbb{P}\big[\mathbf{w}_{\text{priv}}(S')\in O\big]+\delta$$

where the first inequality is because when $I\in E$, $\mathbf{w}_{\text{priv}}$ satisfies $(\epsilon,\delta/2)$-DP and the fact $\mathbb{P}[I\notin E]\leq\delta/2$, the second inequality is by the definition of conditional probability. The proof is complete. □

4 Applications

In this section, we illustrate our main results in the above sections by considering two concrete examples of pairwise learning, namely AUC maximization and similarity metric learning. According to Theorems 2 and 5, the key here is to estimate the Rademacher complexity defined by (4).

AUC Maximization. AUC maximization aims to learn a ranking function $h_{\mathbf{w}}$ defined by $h_{\mathbf{w}}(\mathbf{x},\mathbf{x}') = \mathbf{w}^\top(\mathbf{x}-\mathbf{x}')$. One expects $h_{\mathbf{w}}$ will rank positive examples higher than negative examples, i.e. $\mathbf{w}^\top(\mathbf{x}-\mathbf{x}') \geq 0$ for $y = 1$ and $y' = -1$. Using the hinge loss $\ell(\mathbf{w},\mathbf{z},\mathbf{z}') = (1-h_{\mathbf{w}}(\mathbf{x},\mathbf{x}'))_+\mathbb{I}_{[y=1\wedge y'=-1]}$, AUC maximization can be formulated as

$$\min_{\mathbf{w}\in\mathcal{W}} \mathbb{E}_{\mathbf{z},\mathbf{z}'}\left[(1-\mathbf{w}^\top(\mathbf{x}-\mathbf{x}'))_+\mathbb{I}_{[y=1\wedge y'=-1]}\right]. \quad (10)$$

Denote $\kappa = \sup_{\mathbf{x}}\|\mathbf{x}\|_2$. The Rademacher complexity defined by (4) for AUC maximization is given in the following lemma.

Lemma 4. *Given the parameter space $\mathcal{W} = \{\mathbf{w} \in \mathbb{R}^d | \|\mathbf{w}\|_2 \leq D\}$, the Rademacher complexity of $\mathcal{H} = \{h_{\mathbf{w}} | \mathbf{w} \in \mathcal{W}\}$ can be upper bounded by $\mathcal{R}_t(\mathcal{H}) \leq 2D\kappa/\sqrt{t}$.*

Note in the case of (10), it is easy to check $R_t(\ell\circ\mathcal{H}) \leq 4GD\kappa/\sqrt{t}$ by Ledoux-Talagrand inequality (Ledoux and Talagrand, 2013). Combining this lemma with Theorems 2 and 5, one can derive the following excess risk and utility bound for Algorithms 1 and 2 in the context of non-smooth AUC maximization.

Corollary 1. *Consider the problem of AUC maximization (10). If one runs Algorithm 1 with $T = n^2$ and $\eta = \mathcal{O}(n^{-3/2})$, then, with high probability we have*

$$\epsilon_{risk}(\bar{\mathbf{w}}_T) = \tilde{\mathcal{O}}\Big(\sqrt{\frac{\kappa}{n}}\Big).$$

Corollary 2. *For the problem of AUC maximization (10), if one runs Algorithm 2 with $T = n^2$, $\eta = \mathcal{O}(n^{-3/2})$ and σ given by (5), then, with high probability we have*

$$\epsilon_{risk}(\mathbf{w}_{priv}) = \tilde{\mathcal{O}}\Big(\frac{\sqrt{\kappa d}}{\sqrt{n}\epsilon}\Big).$$

Similarity Metric Learning. We now turn to another notable example of pairwise learning called similarity metric learning. It aims to learn a (squared) Mahalanobis distance metric which is defined by $h_{\mathbf{w}}(\mathbf{x},\mathbf{x}') = (\mathbf{x}-\mathbf{x}')^\top\mathbf{w}(\mathbf{x}-\mathbf{x}')$ parametrized by a positive semi-definite matrix $\mathbf{w} \in \mathbb{R}^{d\times d}$. The intuition behind similarity metric learning is that the distance between samples from the same class should be small and the distance between examples from distinct classes should be large. Using the hinge loss $\ell(\mathbf{w},\mathbf{z},\mathbf{z}') = (1+\tau(y,y')h_{\mathbf{w}}(\mathbf{x},\mathbf{x}'))_+$, it can be formulated as

$$\min_{\mathbf{w}\in\mathcal{W}} \mathbb{E}_{\mathbf{z},\mathbf{z}'}\left[(1+\tau(y,y')(\mathbf{x}-\mathbf{x}')^\top\mathbf{w}(\mathbf{x}-\mathbf{x}'))_+\right], \quad (11)$$

where $\tau(y,y') = 1$ if $y = y'$ and -1 otherwise.

Lemma 5. *Consider the parameter space defined via the nuclear norm $\mathcal{W} = \big\{\mathbf{w} \in \mathbb{R}^{d\times d}, \|\mathbf{w}\|_{S_1} \leq D\big\}$, where $\|\mathbf{w}\|_{S_1}$ denotes the nuclear norm of a matrix $\mathbf{w}$. The complexity of $\mathcal{H} = \{h_{\mathbf{w}} : \mathbf{w} \in \mathcal{W}\}$ is bounded by*

$$\mathcal{R}_t(\mathcal{H}) = \mathcal{O}\Big(\frac{D\big\|\mathbb{E}[\|X\|_2^2XX^\top]\big\|_{S_\infty}^{\frac{1}{2}}\sqrt{\log d}}{\sqrt{t}}\Big), \quad (12)$$

where $\|\cdot\|_{S_\infty}$ denotes the largest singular value.

The proof of Lemma 5 is postponed to Appendix A.5.

As direct corollaries of Lemma 5, we can derive generalization bounds for metric learning from Theorems 2 and 5. For brevity, denote $\chi = \big\|\mathbb{E}[\|X\|_2^2XX^\top]\big\|_{S_\infty}$. We derive the following results of SGD for pairwise learning in the context of non-smooth metric learning.

Corollary 3. *Consider the similarity metric learning problem (11). If one runs Algorithm 1 for $T = n^2$ and $\eta = \mathcal{O}\big(n^{-3/2}\big)$, then, with high probability we have*

$$\epsilon_{risk}(\bar{\mathbf{w}}_T) = \tilde{\mathcal{O}}\Big(\sqrt{\frac{\chi\log(d)}{n}}\Big).$$

Corollary 4. *Consider the similarity metric learning problem (11). If one runs Algorithm 2 with $T = n^2$, $\eta = \mathcal{O}(n^{-3/2})$ and σ given by (5), then, with high probability we have*

$$\epsilon_{risk}(\mathbf{w}_{priv}) = \tilde{\mathcal{O}}\Big(\frac{\sqrt{\chi d\log(d)}}{\sqrt{n}\epsilon}\Big).$$

Remark 1. We now show the advantage of our result as compared to the existing results. Based on the argument in Lei and Ying (2016), it can be shown

$$\mathcal{R}_t(\mathcal{H}) = \mathcal{O}\Big(\frac{D\sup_{\mathbf{x}}\|\mathbf{x}\|^2\sqrt{\log d}}{\sqrt{t}}\Big). \quad (13)$$

The difference between (12) and (13) is that we replace $\sup_{\mathbf{x}}\|\mathbf{x}\|^2$ by the term $\big\|\mathbb{E}[\|X\|_2^2XX^\top]\big\|_{S_\infty}^{\frac{1}{2}}$. Notice $\Big\|\mathbb{E}\big[\|X\|^2XX^\top\big]\Big\|_{S_\infty} \geq \frac{1}{d}\mathrm{tr}\Big(\mathbb{E}[XX^\top XX^\top]\Big) = \frac{1}{d}\mathbb{E}\Big[\mathrm{tr}\big(XX^\top XX^\top\big)\Big] = \frac{1}{d}\mathbb{E}\big[\|X\|_2^4\big]$. If we assume $\mathbb{E}\big[\|X\|_2^4\big] \gtrsim d^2$, then the upper bound of (12) satisfies the relation $\gtrsim \sqrt{d\log d}/\sqrt{t}$ and in the extreme case this lower bound can be achieved within a constant factor. As a comparison, the upper bound in (13) satisfies the relation $\gtrsim d\sqrt{(\log d)/t}$. That is, our argument outperforms the existing results by enjoying a milder dependency on the dimensionality for using nuclear-norm constraints, which is appealing in the high-dimensional setting. If we use Frobenius-norm constraint in defining $\mathcal{W} = \big\{\mathbf{w} \in \mathbb{R}^{d\times d}, \|\mathbf{w}\|_F \leq D_2\big\}$, then one can show that $\mathcal{R}_t(\mathcal{H}) = \mathcal{O}\big(D_2\sup_{\mathbf{x}}\|\mathbf{x}\|^2/\sqrt{t}\big)$ (Lei and Ying, 2016). This matches the bound (13) within a logarithmic factor except that D there is replaced by D_2. Since $\|\mathbf{w}\|_F \leq \|\mathbf{w}\|_{S_1}$, the argument in Lei and Ying (2016) leads to a misleading argument that Frobenius-norm constraint is always preferable to the nuclear-norm constraint. It was posed as an open question on whether one can derive a generalization bound for similarity metric learning showing the advantage of nuclear-norm constraint over Frobenius-norm constraint (Cao et al., 2016). We provide an affirmative solution to this open question in Lemma 5.

5 Conclusions

In this paper, we provide the first-ever-known stability analysis of SGD for pairwise learning with non-smooth losses and obtain optimal excess risk bounds $\tilde{\mathcal{O}}(1/\sqrt{n})$. We extend our analysis to unbounded parameter space and achieve a rate of $\tilde{\mathcal{O}}(n^{-1/3})$. We apply our stability results to study differentially private SGD algorithms in pairwise learning. Our output perturbation method achieves utility bound $\tilde{\mathcal{O}}(\sqrt{d}/(\sqrt{n}\epsilon))$, which matches the previous results in Huai et al. (2020) for smooth losses. Finally, we provide two examples to illustrate our stability and differential privacy results. In particular, the analysis for the example of metric learning shows the advantage of nuclear norm constraint over Frobenius norm constraint which solved an open question raised in Cao et al. (2016).

Here we only considered SGD with replacement. It would be interesting to extend our analysis to SGD without replacement which is drawing increasing interests. The utility bound is suboptimal as compared with pointwise learning with non-smooth losses, which is $\tilde{\mathcal{O}}\big(\max\{1/\sqrt{n}, \sqrt{d}/(n\epsilon)\}\big)$. It remains an open question to us if the same bound can be achieved in pairwise learning.

Acknowledgement

The authors would like to thank the anonymous reviewers for their constructive comments and suggestions. This work is supported by NSF grants IIS-1816227 and IIS-2008532. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agency.

References

Agarwal, S., Dugar, D., and Sengupta, S. (2010). Ranking chemical structures for drug discovery: a new machine learning approach. *Journal of chemical information and modeling*, 50(5):716–731.

Agarwal, S. and Niyogi, P. (2009). Generalization bounds for ranking algorithms via algorithmic stability. *Journal of Machine Learning Research*, 10(Feb):441–474.

Bach, F. and Moulines, E. (2013). Non-strongly-convex smooth stochastic approximation with convergence rate o (1/n). In *Advances in neural information processing systems*, pages 773–781.

Bartlett, P. L. and Mendelson, S. (2002). Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482.

Bassily, R., Feldman, V., Guzmán, C., and Talwar, K. (2020). Stability of stochastic gradient descent on nonsmooth convex losses.

Bassily, R., Feldman, V., Talwar, K., and Thakurta, A. G. (2019). Private stochastic convex optimization with optimal rates. In *Advances in Neural Information Processing Systems*, pages 11279–11288.

Bottou, L. and Cun, Y. L. (2004). Large scale online learning. In *Advances in neural information processing systems*.

Bu, Z., Dong, J., Long, Q., and Su, W. J. (2020). Deep learning with gaussian differential privacy. *Harvard data science review*, 2020(23).

Cao, Q., Guo, Z.-C., and Ying, Y. (2016). Generalization bounds for metric and similarity learning. *Machine Learning*, 102(1):115–132.

Cesa-Bianchi, N., Conconi, A., and Gentile, C. (2004). On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057.

Charles, Z. and Papailiopoulos, D. (2018). Stability and generalization of learning algorithms that converge to global optima. In *International Conference on Machine Learning*, pages 745–754. PMLR.

Clémençon, S., Lugosi, G., and Vayatis, N. (2008). Ranking and empirical minimization of u-statistics. *The Annals of Statistics*, pages 844–874.

De la Pena, V. and Giné, E. (2012). *Decoupling: from dependence to independence*. Springer Science & Business Media.

Dong, J., Roth, A., and Su, W. J. (2019). Gaussian differential privacy. *arXiv preprint arXiv:1905.02383*.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer.

Dwork, C., Roth, A., et al. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407.

Feldman, V., Koren, T., and Talwar, K. (2020). Private stochastic convex optimization: Optimal rates in linear time. *arXiv preprint arXiv:2005.04763*.

Gao, W., Jin, R., Zhu, S., and Zhou, Z.-H. (2013). One-pass auc optimization. In *International Conference on Machine Learning*, pages 906–914.

Hardt, M., Recht, B., and Singer, Y. (2016). Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 1225–1234.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30.

Hu, T., Fan, J., Wu, Q., and Zhou, D.-X. (2013). Learning theory approach to minimum error entropy criterion. *Journal of Machine Learning Research*, 14(Feb):377–397.

Huai, M., Wang, D., Miao, C., Xu, J., and Zhang, A. (2020). Pairwise learning with differential privacy guarantees. In *AAAI*.

Jin, R., Wang, S., and Zhou, Y. (2009). Regularized distance metric learning: Theory and algorithm. In *Advances in neural information processing systems*, pages 862–870.

Kar, P., Sriperumbudur, B., Jain, P., and Karnick, H. (2013). On the generalization ability of online learning algorithms for pairwise loss functions. In *International Conference on Machine Learning*, pages 441–449.

Kuzborskij, I. and Lampert, C. (2018). Data-dependent stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 2815–2824. PMLR.

Lacoste-Julien, S., Schmidt, M., and Bach, F. (2012). A simpler approach to obtaining an o (1/t) convergence rate for the projected stochastic subgradient method. *arXiv preprint arXiv:1212.2002*.

Ledoux, M. and Talagrand, M. (2013). *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media.

Lei, Y., Ledent, A., and Kloft, M. (2020). Sharper generalization bounds for pairwise learning. *Advances in Neural Information Processing Systems*, 33.

Lei, Y. and Ying, Y. (2016). Generalization analysis of multi-modal metric learning. *Analysis and Applications*, 14(04):503–521.

Lei, Y. and Ying, Y. (2020). Fine-grained analysis of stability and generalization for stochastic gradient descent. In *ICML*.

Lin, J., Camoriano, R., and Rosasco, L. (2016). Generalization properties and implicit regularization for multiple passes sgm. In *International Conference on Machine Learning*, pages 2340–2348.

Lin, J., Lei, Y., Zhang, B., and Zhou, D.-X. (2017). Online pairwise learning algorithms with convex loss functions. *Information Sciences*, 406:57–70.

Liu, T., Lugosi, G., Neu, G., and Tao, D. (2017). Algorithmic stability and hypothesis complexity. In *ICML*.

Lust-Piquard, F. and Pisier, G. (1991). Non commutative khintchine and paley inequalities. *Arkiv för matematik*, 29(1-2):241–260.

Mukherjee, S. and Wu, Q. (2006). Estimation of gradients and coordinate covariation in classification. *Journal of Machine Learning Research*, 7(Nov):2481–2514.

Mukherjee, S. and Zhou, D.-X. (2006). Learning coordinate covariances via gradients. *Journal of Machine Learning Research*, 7(Mar):519–549.

Qiu, R. and Wicks, M. (2014). *Cognitive networked sensing and big data*. Springer.

Rakhlin, A., Shamir, O., and Sridharan, K. (2012). Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Conference on Machine Learning*, pages 449–456.

Rejchel, W. (2012). On ranking and generalization bounds. *Journal of Machine Learning Research*, 13:1373–1392.

Shalev-Shwartz, S., Shamir, O., Srebro, N., and Sridharan, K. (2009). Stochastic convex optimization. In *COLT*.

Shen, W., Yang, Z., Ying, Y., and Yuan, X. (2020). Stability and optimization error of stochastic gradient descent for pairwise learning. *Analysis and Applications*, 18(05):887–927.

Tropp, J. A. (2015). An introduction to matrix concentration inequalities. *arXiv preprint arXiv:1501.01571*.

Verma, N. and Branson, K. (2015). Sample complexity of learning mahalanobis distance metrics. In *Advances in neural information processing systems*, pages 2584–2592.

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Wang, Y., Khardon, R., Pechyony, D., and Jones, R. (2012). Generalization bounds for online learning algorithms with pairwise loss functions. In *Conference on Learning Theory*, pages 13–1.

Weinberger, K. Q. and Saul, L. K. (2009). Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10:207–244.

Wu, X., Li, F., Kumar, A., Chaudhuri, K., Jha, S., and Naughton, J. (2017). Bolt-on differential privacy for scalable stochastic gradient descent-based analytics. In *Proceedings of the 2017 ACM International Conference on Management of Data*, pages 1307–1322.

Xing, E. P., Jordan, M. I., Russell, S. J., and Ng, A. Y. (2003). Distance metric learning with application to clustering with side-information. In *Advances in neural information processing systems*.

Ying, Y. and Li, P. (2012). Distance metric learning with eigenvalue optimization. *The Journal of Machine Learning Research*, 13:1–26.

Ying, Y., Wen, L., and Lyu, S. (2016). Stochastic online auc maximization. In *Advances in Neural Information Processing Systems*.

Ying, Y. and Zhou, D.-X. (2006). Online regularized classification algorithms. *IEEE Transactions on Information Theory*, 52(11):4775–4788.

Ying, Y. and Zhou, D.-X. (2016). Online pairwise learning algorithms. *Neural computation*, 28(4):743–777.

Zhao, P., Jin, R., Yang, T., and Hoi, S. C. (2011). Online auc maximization. In *Proceedings of the 28th international conference on machine learning (ICML-11)*.

A Supplementary

A.1 Supporting Theorems and Lemmas

Let us recall the excess risk of a randomized algorithm $\mathcal{A}$ defined as $\epsilon_{\text{risk}}(\mathcal{A}(S)) = R(\mathcal{A}(S)) - R(\mathbf{w}_*)$, which can be decomposed by

$$\epsilon_{\text{risk}}(\mathcal{A}(S)) = R(\mathcal{A}(S)) - R_S(\mathcal{A}(S)) + R_S(\mathbf{w}_*) - R(\mathbf{w}_*) + R_S(\mathcal{A}(S)) - R_S(\mathbf{w}_*). \tag{A.14}$$

Hence, before introducing the proofs we will give some theorems and lemmas that are repeatedly used to bound each term in Equation (A.14).

Here we simply assume bounds for $\|\mathbf{w}\|$. A simple lemma indicates that if $\mathbf{w}$ is bounded, then $\ell(\mathbf{w}, \cdot, \cdot)$ is also bounded. In the subsequent sections, we will characterize the bounds for the iterates $\{\mathbf{w}_t\}$ whenever the parameter space $\mathcal{W}$ is bounded or unbounded.

Lemma A.6. *Suppose ℓ is nonnegative, convex and G-Lipschitz. Let $M = \sup_{\mathbf{z},\mathbf{z}'} \ell(0, \mathbf{z}, \mathbf{z}')$. For any $\mathbf{w} \in W$ that $\mathbf{w} \leq B$ for some $0 \leq B < \infty$, then $\sup_{\mathbf{z},\mathbf{z}'} \ell(\mathbf{w}, \mathbf{z}, \mathbf{z}') \leq M + GB$.*

Proof. By convexity of ℓ, we have for any $\mathbf{z}, \mathbf{z}'$

$$\ell(\mathbf{w}, \mathbf{z}, \mathbf{z}') \leq \sup_{\mathbf{z},\mathbf{z}'} \ell(0, \mathbf{z}, \mathbf{z}') + \langle \mathbf{w}, \partial \ell(\mathbf{w}, \mathbf{z}, \mathbf{z}') \rangle \leq M + \|\mathbf{w}\| \|\partial \ell(\mathbf{w}, \mathbf{z}, \mathbf{z}')\|_2 \leq M + GB$$

where the second inequality is due to Cauchy-Schwarz inequality. The proof is complete by taking the supremum. □

The first theorem in this section is the the high probability generalization bound of UAS algorithms in pairwise learning. This theorem is an extension of Theorem 1 in Lei et al. (2020) for generalization bound of uniformly stable algorithms in pairwise learning.

Theorem A.6. *Suppose ℓ is nonnegative, convex and G-Lipschitz. Let $\mathcal{A}$ be a ϵ-UAS randomized algorithm for pairwise learning. Suppose the output of $\mathcal{A}$ is bounded by B and let $M = \sup_{\mathbf{z},\mathbf{z}'} \ell(0, \mathbf{z}, \mathbf{z}')$. Then we have for any $\gamma \in (0,1)$, with probability at least $1-\gamma$ with respect to the sample S and the internal randomness of $\mathcal{A}$,*

$$R(\mathcal{A}(S)) - R_S(\mathcal{A}(S)) \leq 4\epsilon + 48\sqrt{6}eG\epsilon\lceil \ln(n) \rceil \ln(e/\gamma) + 12\sqrt{2}e(M+GB)\sqrt{\frac{\ln(e/\gamma)}{n}}.$$

Proof. According to Theorem 1 in Lei et al. (2020), we only need to check the expected boundedness of $\ell(\mathcal{A}(S), \cdot, \cdot)$ and the uniform stability of $\mathcal{A}$. For the boundedness part, by Lemma A.6 we know

$$\big|\mathbb{E}[\ell(\mathcal{A}(S), \mathbf{z}, \mathbf{z}')]\big| \leq M + GB$$

for any $\mathbf{z}, \mathbf{z}'$. For the uniform stability, since $\mathcal{A}$ is ϵ-UAS, by the Lipschitz continuity of ℓ we have

$$\sup_{\mathbf{z},\mathbf{z}'} |\ell(\mathcal{A}(S), \mathbf{z}, \mathbf{z}') - \ell(\mathcal{A}(S'), \mathbf{z}, \mathbf{z}')| \leq G\|\mathcal{A}(S) - \mathcal{A}(S')\|_2 \leq G\epsilon.$$

The proof is complete. □

The next corollary is a direct application of Theorem A.6, which states if UAS holds with high probability, then so is the generalization.

Corollary A.5. *Let $\mathcal{A}$ be a randomized algorithm for pairwise learning. If for any $\gamma_0 \in (0,1)$, we have, for any neighborhood datasets S, S',*

$$\mathbb{P}_{\mathcal{A}}\Big[\|\mathcal{A}(S) - \mathcal{A}(S')\|_2 > \epsilon\Big] \leq \gamma_0.$$

Suppose ℓ is nonnegative, convex and G-Lipschitz. Suppose the output of $\mathcal{A}$ is bounded by B and let $M = \sup_{\mathbf{z},\mathbf{z}'} \ell(0, \mathbf{z}, \mathbf{z}')$. Then we have for any $\gamma \in (0,1)$,

$$\mathbb{P}_{S,\mathcal{A}}\Big[R(\mathcal{A}(S)) - R_S(\mathcal{A}(S)) > 4\epsilon + 48\sqrt{6}eG\epsilon\lceil \ln(n) \rceil \ln(e/\gamma) + 12\sqrt{2}e(M+GB)\sqrt{\frac{\ln(e/\gamma)}{n}}\Big] \leq \gamma + \gamma_0.$$

Proof. Denote $E = \{\mathcal{A} | \|\mathcal{A}(S) - \mathcal{A}(S')\|_2 > \epsilon\}$ and $F = \{S, \mathcal{A} | R(\mathcal{A}(S)) - R_S(\mathcal{A}(S)) > 4\epsilon + 48\sqrt{6}eG\epsilon\lceil\ln(n)\rceil\ln(e/\gamma) + 12\sqrt{2}e(M+GB)\sqrt{\ln(e/\gamma)/n}\}$. Then by assumption we have $\mathbb{P}_{\mathcal{A}}[\mathcal{A} \in E] \le \gamma_0$. By Theorem A.6, for any $\gamma \in (0,1)$, we have $\mathbb{P}_{S,\mathcal{A}}[S, \mathcal{A} \in F | \mathcal{A} \notin E] \le \gamma$. Then the following identity holds

$$\begin{aligned}\mathbb{P}_{S,\mathcal{A}}[S,\mathcal{A} \in F] =& \mathbb{P}_{S,\mathcal{A}}[S,\mathcal{A} \in F \cap \mathcal{A} \in E] + \mathbb{P}_{S,\mathcal{A}}[S,\mathcal{A} \in F \cap \mathcal{A} \notin E] \\ =& \mathbb{P}_{S,\mathcal{A}}[S,\mathcal{A} \in F | \mathcal{A} \in E]\mathbb{P}[\mathcal{A} \in E] + \mathbb{P}_{S,\mathcal{A}}[S,\mathcal{A} \in F | \mathcal{A} \notin E]\mathbb{P}[\mathcal{A} \notin E] \\ \le& \gamma_0 + \gamma.\end{aligned}$$

The proof is completed. □

Combining Corollary A.5 and the stability result in Theorem 1, we arrive at the following generalization bound for Algorithm 1.

Corollary A.6. *Suppose ℓ is nonnegative, convex and G-Lipschitz. Let $B_T = \|\bar{\mathbf{w}}_T\|$ and $M = \sup_{\mathbf{z},\mathbf{z}'} \ell(0, \mathbf{z}, \mathbf{z}')$. If we run Algorithm 1 for $T \ge n$ iterations under random selection with replacement rule. For any $\gamma \in (0,1)$, with probability at least $1-\gamma$ with respect to the sample S and the internal randomness of Algorithm 1, we have*

$$R(\bar{\mathbf{w}}_T) - R_S(\bar{\mathbf{w}}_T) \le 2\sqrt{e}\eta G(4 + 48\sqrt{6}eG\lceil\ln(n)\rceil\ln(2e/\gamma))\Big(\sqrt{T} + \frac{\sqrt{3}T\ln(eT)\ln(2/\gamma)}{n}\Big) + 12\sqrt{2}e(M + GB_T)\sqrt{\frac{\ln(2e/\gamma)}{n}}.$$

Proof. By Theorem 1, elementary inequality and the fact that stability is monotonically increasing, we have with probability at least $1-\gamma/2$,

$$\big\|\bar{\mathbf{w}}_T - \bar{\mathbf{w}}'_T\big\|_2^2 \le 4e\eta^2G^2\Big(T + \frac{3T^2\ln^2(eT)\ln^2(2/\gamma)}{n^2}\Big).$$

The proof is completed by convexity of $\|\cdot\|_2$ and applying Theorem A.6 with probability $1 - \frac{\gamma}{2}$. □

The next theorem gives a bound on $R_S(\mathbf{w}_*) - R(\mathbf{w}_*)$ by Hoeffding inequality of U-statistics Hoeffding (1963).

Theorem A.7. *Suppose ℓ is convex and G-Lipschitz. Let $M = \sup_{\mathbf{z},\mathbf{z}'} \ell(0,\mathbf{z},\mathbf{z}')$ and $B = \|\mathbf{w}_*\|_2$. For any $\gamma \in (0,1)$, with probability at least $1-\gamma$ with respect to the sample S, we have*

$$R_S(\mathbf{w}_*) - R(\mathbf{w}_*) \le (M + GB)\sqrt{\frac{\ln(1/\gamma)}{n}}.$$

Proof. The result is derived by applying Hoeffding inequality since $\ell(\mathbf{w}_*, \mathbf{z}, \mathbf{z}') \le M + GB$ for any $\mathbf{z}, \mathbf{z}'$ according to Lemma A.6. □

Next we give an upper bound on the optimization error $R_S(\bar{\mathbf{w}}_T) - R_S(\mathbf{w}_*)$. The results are inspired by Kar et al. (2013), where they consider the online-to-batch generalization bound for pairwise learning. Our bound in the next theorem is given for optimization bound on finite sample.

Theorem A.8. *Suppose ℓ is nonnegative, convex and G-Lipschitz. Suppose there are some non-decreasing sequence $0 \le B_t < \infty$ such that $\|\mathbf{w}_t\|_2 \le B_t$, and let $M = \sup_{\mathbf{z},\mathbf{z}'} \ell(0,\mathbf{z},\mathbf{z}')$ and $B = \|\mathbf{w}_*\|_2$. Suppose we run Algorithm 1 for T iterations, then with probability at least $1-\gamma$ with respect to the sample S and the internal randomness of Algorithm 1, we have*

$$R_S(\bar{\mathbf{w}}_T) - R_S(\mathbf{w}_*) \le \frac{2}{T}\sum_{t=1}^T \mathcal{R}_t(\ell \circ \mathcal{W}_t) + \frac{2}{T}\sum_{t=1}^T \mathcal{R}_t(\ell \circ \mathcal{W}_B) + \frac{B^2}{2T\eta} + \frac{\eta G^2}{2} + (6M + 3GB)\sqrt{\frac{\ln(2T/\gamma)}{T}} + 3GB_T\sqrt{\frac{\ln(2T/\gamma)}{T}},$$

where $\mathcal{W}_t = \{\mathbf{w} \in W | \|\mathbf{w}\|_2 \le B_t\}$ and $\mathcal{W}_B = \{\mathbf{w} \in W | \|\mathbf{w}\|_2 \le B\}$ are subspaces of $\mathcal{W}$.

In order to prove Theorem A.8, we decompose $R_S(\bar{\mathbf{w}}_T) - R_S(\mathbf{w}_*)$ as in Kar et al. (2013) and bound each part separately. In particular, recall that $\tilde{L}_{t+1}(\mathbf{w}) = \frac{1}{t}\sum_{k=1}^t \ell(\mathbf{w}, \mathbf{z}_{i_{t+1}}, \mathbf{z}_{i_k})$. We have the following lemmas.

Lemma A.7. *Assume ℓ is nonnegative, convex and G-Lipschitz. Let $\mathcal{W}_t = \big\{\mathbf{w} \in \mathcal{W} \big| \|\mathbf{w}\|_2 \le B_t\big\}$ and let $M = \sup_{\mathbf{z},\mathbf{z}'} \ell(0,\mathbf{z},\mathbf{z}')$. With probability $1-\gamma$, we have*

$$\frac{1}{T}\sum_{t=1}^T R_S(\mathbf{w}_t) - \hat{L}_{t+1}(\mathbf{w}_t) \le \frac{2}{T}\sum_{t=1}^T \mathcal{R}_t(\ell \circ \mathcal{W}_t) + 3(M + GB_T)\sqrt{\frac{\ln(T/\gamma)}{T}}.$$

Proof. For any $\mathbf{w}$, denote $\tilde{L}_{t+1}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{t}\sum_{k=1}^{t}\ell(\mathbf{w},\mathbf{z}_i,\mathbf{z}_{i_k})$. This allows us to decompose the risk as follows

$$\frac{1}{T}\sum_{t=1}^{T} R_S(\mathbf{w}_t) - \hat{L}_{t+1}(\mathbf{w}_t) = \frac{1}{T}\sum_{t=1}^{T}\underbrace{R_S(\mathbf{w}_t) - \tilde{L}_{t+1}(\mathbf{w}_t)}_{P_{t+1}} + \underbrace{\tilde{L}_{t+1}(\mathbf{w}_t) - \hat{L}_{t+1}(\mathbf{w}_t)}_{Q_{t+1}}$$

By construction, we have $\mathbb{E}_{\mathbf{z}_{i_{t+1}}}[Q_{t+1}|\mathbf{z}_{i_1},\cdots,\mathbf{z}_{i_t}] = 0$ and hence the sequence $Q_2,\cdots,Q_T$ forms a martingale difference sequence. By Lemma A.6 we have Q_{t+1} lies in $[-M-GB_t, M+GB_t] \subseteq [-M-GB_T, M+GB_T]$ as B_t's are non-decreasing. An application of the Azuma-Hoeffding inequality shows that with probability at least $1-\gamma$,

$$\frac{1}{T}\sum_{t=1}^{T} Q_t \le (M+GB_T)\sqrt{\frac{2\ln(1/\gamma)}{T}}.$$

We now analyze each term P_t individually. Let us start by introducing a ghost sample $\{\mathbf{z}'_{i_1},\cdots,\mathbf{z}'_{i_t}\}$, where each $\mathbf{z}'_{i_k}$ follows the same distribution as $\mathbf{z}_{i_k}$. By linearity of expectation, we have

$$R_S(\mathbf{w}_t) = \mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{t}\sum_{k=1}^{t}\ell(\mathbf{w}_t,\mathbf{z}_i,\mathbf{z}'_{i_k})\Big],$$

where the expectation is taken over $\{\mathbf{z}'_{i_k}\}_{k=1}^{t}$. It allows us to write P_t as follow

$$\begin{aligned}P_t =& \mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{t}\sum_{k=1}^{t}\ell(\mathbf{w}_t,\mathbf{z}_i,\mathbf{z}'_{i_k})\Big] - \tilde{L}_{t+1}(\mathbf{w}_t) \le \sup_{\mathbf{w}\in\mathcal{W}_t}\mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{t}\sum_{k=1}^{t}\ell(\mathbf{w},\mathbf{z}_i,\mathbf{z}'_{i_k})\Big] - \tilde{L}_{t+1}(\mathbf{w})\\ \triangleq& g_{t+1}(\mathbf{z}_{i_1},\cdots,\mathbf{z}_{i_t}).\end{aligned}$$

Since ℓ is bounded by A_t, the expression $g_{t+1}(\mathbf{z}_{i_1},\cdots,\mathbf{z}_{i_t})$ can have a variation of at most $(M+GB_t)/t$ when changing any of its t variables. Hence an application of McDiarmid's inequality gives us, with probability at least $1-\gamma$,

$$g_{t+1}(\mathbf{z}_{i_1},\cdots,\mathbf{z}_{i_t}) \le \mathbb{E}_{\mathbf{z}_{i_1},\cdots,\mathbf{z}_{i_t}}[g_{t+1}(\mathbf{z}_{i_1},\cdots,\mathbf{z}_{i_t})] + (M+GB_t)\sqrt{\frac{\ln(1/\gamma)}{2t}}.$$

For any $\mathbf{w}\in\mathcal{W}_t$, let $f(\mathbf{w},\mathbf{z}') = \frac{1}{t}\frac{1}{n}\sum_{i=1}^{n}\ell(\mathbf{w},\mathbf{z}_i,\mathbf{z}')$. Then we can write $\mathbb{E}_{\mathbf{z}_{i_1},\cdots,\mathbf{z}_{i_t}}[g_{t+1}(\mathbf{z}_{i_1},\cdots,\mathbf{z}_{i_t})]$ as follow

$$\begin{aligned}&\mathbb{E}_{\{\mathbf{z}_{i_k}\}}[g_{t+1}(\mathbf{z}_{i_1},\cdots,\mathbf{z}_{i_t})] = \mathbb{E}_{\{\mathbf{z}_{i_k}\}}\Big[\sup_{\mathbf{w}\in\mathcal{W}_t}\mathbb{E}_{\{\mathbf{z}'_{i_k}\}}\Big[\sum_{k=1}^{t}f(\mathbf{w},\mathbf{z}'_{i_k})\Big] - \sum_{k=1}^{t}f(\mathbf{w},\mathbf{z}_{i_k})\Big]\\ &\le \mathbb{E}_{\{\mathbf{z}_{i_k},\mathbf{z}'_{i_k}\}}\Big[\sup_{\mathbf{w}\in\mathcal{W}_t}\sum_{k=1}^{t}f(\mathbf{w},\mathbf{z}'_{i_k}) - \sum_{k=1}^{t}f(\mathbf{w},\mathbf{z}_{i_k})\Big] = \mathbb{E}_{\{\mathbf{z}_{i_k},\mathbf{z}'_{i_k},\sigma_k\}}\Big[\sup_{\mathbf{w}\in\mathcal{W}_t}\sum_{k=1}^{t}\sigma_k\big(f(\mathbf{w},\mathbf{z}'_{i_k}) - f(\mathbf{w},\mathbf{z}_{i_k})\big)\Big]\\ &\le \frac{2}{t}\mathbb{E}_{\{\mathbf{z}_{i_k},\sigma_k\}}\Big[\sup_{\mathbf{w}\in\mathcal{W}_t}\sum_{k=1}^{t}\sigma_k\frac{1}{n}\sum_{i=1}^{n}\ell(\mathbf{w},\mathbf{z}_i,\mathbf{z}_{i_k})\Big] \le \frac{2}{t}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\{\mathbf{z}_{i_k},\sigma_k\}}\Big[\sup_{\mathbf{w}\in\mathcal{W}_t}\sum_{k=1}^{t}\sigma_k\ell(\mathbf{w},\mathbf{z}_i,\mathbf{z}_{i_k})\Big] = 2\mathcal{R}_t(\ell\circ\mathcal{W}_t).\end{aligned}$$

Thus we have, with probability at least $1-\gamma$,

$$P_t \le 2\mathcal{R}_t(\ell\circ\mathcal{W}_t) + (M+GB_t)\sqrt{\frac{\ln(1/\gamma)}{2t}}.$$

The Lemma holds by applying a union bound on P_t and taking the average over t. $\square$

Lemma A.8. *Suppose ℓ is nonnegative, convex and G-Lipschitz. Let $\mathcal{W}_B = \{\mathbf{w}\in\mathcal{W}\,\big|\,\|\mathbf{w}\|_2 \le B\}$ and let $M = \sup_{\mathbf{z},\mathbf{z}'}\ell(0,\mathbf{z},\mathbf{z}')$. With probability $1-\gamma$, we have*

$$\frac{1}{T}\sum_{t=1}^{T}\hat{L}_{t+1}(\mathbf{w}_*) - R_S(\mathbf{w}_*) \le \frac{2}{T}\sum_{t=1}^{T}\mathcal{R}_t(\ell\circ\mathcal{W}_B) + 3(M+GB)\sqrt{\frac{\ln(T/\gamma)}{T}}.$$

Proof. Similar to the proof of Lemma A.7 by replacing $\mathbf{w}_t$ with $\mathbf{w}_*$. ◻

Lemma A.9. *Suppose ℓ is nonnegative, convex and G-Lipschitz. Suppose $\|\mathbf{w}_*\|_2 \leq B$. Suppose we run Algorithm 1 for T iterations, then we have*

$$\frac{1}{T}\sum_{t=1}^{T}\hat{L}_{t+1}(\mathbf{w}_t) - \hat{L}_{t+1}(\mathbf{w}_*) \leq \frac{B^2}{2T\eta} + \frac{\eta G^2}{2}$$

Proof. By the update rule of Algorithm 1, we have

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

Therefore, by the convexity of $\hat{L}_{t+1}$, we have

$$\begin{aligned}\sum_{t=1}^{T}\hat{L}_{t+1}(\mathbf{w}_t) - \hat{L}_{t+1}(\mathbf{w}_*) \leq \sum_{t=1}^{T}\big\langle\mathbf{w}_t - \mathbf{w}_*, \partial\hat{L}_{t+1}(\mathbf{w}_t)\big\rangle \leq& \sum_{t=1}^{T}\frac{\|\mathbf{w}_t - \mathbf{w}_*\|_2^2 - \|\mathbf{w}_{t+1} - \mathbf{w}_*\|_2^2}{2\eta} + \frac{T\eta G^2}{2} \\ \leq& \frac{\|\mathbf{w}_*\|_2^2}{2\eta} + \frac{\eta G^2}{2},\end{aligned}$$

the Lemma holds by dividing T over both sides. ◻

Proof of Theorem A.8. By the convexity of the empirical loss R_S, we have

$$\begin{aligned}R_S(\bar{\mathbf{w}}_T) - R_S(\mathbf{w}_*) \leq& \frac{1}{T}\sum_{t=1}^{T}R_S(\mathbf{w}_t) - R_S(\mathbf{w}_*) \\ =& \frac{1}{T}\sum_{t=1}^{T}\big(R_S(\mathbf{w}_t) - \hat{L}_{t+1}(\mathbf{w}_t) + \hat{L}_{t+1}(\mathbf{w}_*) - R_S(\mathbf{w}_*) + \hat{L}_{t+1}(\mathbf{w}_t) - \hat{L}_{t+1}(\mathbf{w}_*)\big).\end{aligned} \tag{A.15}$$

The conclusion follows from Lemma A.7, A.8 both with probability $1-\gamma/2$ and Lemma A.9. ◻

A.2 Proof of Theorem 2

Theorem A.9 (Theorem 2 restated)**.** *Suppose ℓ is nonnegative, convex and G-Lipschitz. Suppose $\mathcal{W}$ is bounded with diameter D and let $M = \sup_{\mathbf{z},\mathbf{z}'}\ell(0,\mathbf{z},\mathbf{z}')$. Assume we run Algorithm 1 for $T \geq n$ iterations under random selection with replacement rule. For any $\gamma \in (0,1)$, with probability at least $1-\gamma$, with respect to the sample S and the internal randomness of Algorithm 1, we have*

$$\begin{aligned}R(\bar{\mathbf{w}}_T) - R(\mathbf{w}_*) \leq& \frac{4}{T}\sum_{t=1}^{T}\mathcal{R}_t(\ell\circ\mathcal{W}) + \frac{D^2}{2T\eta} + \frac{\eta G^2}{2} + 6(M+GD)\sqrt{\frac{\ln(6T/\gamma)}{n}} + 19e(M+GD)\sqrt{\frac{\ln(6e/\gamma)}{n}} \\ &+ 2\sqrt{e}\eta G(4+48\sqrt{6}eG\lceil\ln(n)\rceil\ln(6e/\gamma))\Big(\sqrt{T} + \frac{\sqrt{3}T\ln(eT)\ln(6/\gamma)}{n}\Big).\end{aligned}$$

Proof of Theorem A.9. Since $\mathcal{W}$ is bounded by D, we have $B = B_t = D$. Furthermore, by Lemma A.6, we have $\sup_{\mathbf{z},\mathbf{z}'}\ell(\mathbf{w}_*,\mathbf{z},\mathbf{z}') \leq M + GD$ and $\sup_{\mathbf{z},\mathbf{z}'}\ell(\mathbf{w}_t,\mathbf{z},\mathbf{z}') \leq M + GD$. The proof is completed by recalling the error decomposition (A.14), applying Corollary A.6, Theorem A.7 and A.8 each with probability $1-\gamma/3$. ◻

A.3 Proof of Theorem 3

Theorem A.10 (Theorem 3 restated)**.** *Suppose ℓ is nonnegative, convex and G-Lipschitz. Denote $M = \sup_{\mathbf{z},\mathbf{z}'}\ell(0,\mathbf{z},\mathbf{z}')$ and $D = \|\mathbf{w}_*\|_2$. Assume we run Algorithm 1 for $T \geq n$ iterations under random selection* *with replacement rule. For any $\gamma \in (0,1)$, with probability at least $1-\gamma$ with respect to the sample S and internal randomness of Algorithm 1, we have*

$$R(\bar{\mathbf{w}}_T)-R(\mathbf{w}_*)\leq\frac{2}{T}\sum_{t=1}^{T}\big(\mathcal{R}_t(\ell\circ\mathcal{W}_t)+\mathcal{R}_t(\ell\circ\mathcal{W}_D)\big)+\frac{D^2}{2T\eta}+\frac{\eta G^2}{2}+(6M+3GD)\sqrt{\frac{\ln(6T/\gamma)}{n}}+3G\sqrt{(G^2+2M)\eta\ln(6T/\gamma)}$$
$$+2\sqrt{e}\eta G(4+48\sqrt{6}eG\lceil\ln(n)\rceil\ln(6e/\gamma))\Big(\sqrt{T}+\frac{\sqrt{3}T\ln(eT)\ln(6/\gamma)}{n}\Big)$$
$$+12\sqrt{2}e(M+G\sqrt{(G^2+2M)\eta T})\sqrt{\frac{\ln(6e/\gamma)}{n}}+(M+GD)\sqrt{\frac{\ln(3/\gamma)}{n}}.$$

Although the boundedness assumption on the parameter space $\mathcal{W}$ is removed, the next lemma characterizes the bound of the iterates $\mathbf{w}_t$ by the sum of stepsizes.

Lemma A.10. *Suppose ℓ is nonnegative, convex and G-Lipschitz. Denote $M=\sup_{\mathbf{z},\mathbf{z}'}\ell(0,\mathbf{z},\mathbf{z}')$. Let $\{\mathbf{w}_t\}$ be the sequence of iterates by Algorithm 1 with $\eta\leq 1$. Then*

$$\|\mathbf{w}_{t+1}\|_2^2\leq(G^2+2M)\eta t.$$

Proof. By the update rule of Algorithm 1, we have

$$\begin{aligned}\|\mathbf{w}_{t+1}\|_2^2 =&\big\|\mathbf{w}_t-\eta\partial\hat{L}_{t+1}(\mathbf{w}_t)\big\|_2^2=\|\mathbf{w}_t\|_2^2+\eta^2\big\|\partial\hat{L}_{t+1}(\mathbf{w}_t)\big\|_2^2-2\eta\big\langle\mathbf{w}_t,\partial\hat{L}_{t+1}(\mathbf{w}_t)\big\rangle\\ \leq&\|\mathbf{w}_t\|_2^2+\eta G^2-2\eta\big\langle\mathbf{w}_t,\partial\hat{L}_{t+1}(\mathbf{w}_t)\big\rangle\leq\|\mathbf{w}_t\|_2^2+\eta G^2+2\eta\big(\hat{L}_{t+1}(0)-\hat{L}_{t+1}(\mathbf{w}_t)\big)\\ \leq&\|\mathbf{w}_t\|_2^2+\eta(G^2+2M),\end{aligned}$$

where the first inequality holds since ℓ is G-Lipschitz and $\eta\leq1$, the second inequality is due to the convexity of ℓ and the last inequality is due to the nonnegativity of ℓ and the definition of M. □

Proof of Theorem A.10. By assumption and Lemma A.10, we have $B=D$ and $B_t=\sqrt{(G^2+2M)\eta t}$. Therefore, by Lemma A.6, we also get $\sup_{\mathbf{z},\mathbf{z}'}\ell(\mathbf{w}_*,\mathbf{z},\mathbf{z}')\leq M+GD$ and $\sup_{\mathbf{z},\mathbf{z}'}\ell(\mathbf{w}_t,\mathbf{z},\mathbf{z}')\leq M+G\sqrt{(G^2+2M)\eta t}$. The proof is completed by recalling the error decomposition (A.14), applying Corollary A.6, Theorem A.7 and A.8 with probability $1-\gamma/3$ each.

□

A.4 Proof of Theorem 5

In this section, we give utility bound of Algorithm 2. Recall the error decomposition scheme as follows

$$\begin{aligned}\epsilon_{\text{risk}}(\mathbf{w}_{\text{priv}})&=R(\mathbf{w}_{\text{priv}})-R(\mathbf{w}_*)=R(\mathbf{w}_{\text{priv}})-R(\bar{\mathbf{w}}_T)+R(\bar{\mathbf{w}}_T)-R(\mathbf{w}_*)\\&=R(\mathbf{w}_{\text{priv}})-R(\bar{\mathbf{w}}_T)+R(\bar{\mathbf{w}}_T)-R_S(\bar{\mathbf{w}}_T)+R_S(\bar{\mathbf{w}}_T)-R_S(\mathbf{w}_*)+R_S(\mathbf{w}_*)-R(\mathbf{w}_*).\end{aligned}\tag{A.16}$$

Notice that $R(\bar{\mathbf{w}}_T)-R(\mathbf{w}_*)$ yields similar excess risk as Theorem A.9. Hence the difference here is the additional term $R(\mathbf{w}_{\text{priv}})-R(\bar{\mathbf{w}}_T)$ due to the added noise $\mathbf{u}$. The next lemma is a Chernoff type bound for the ℓ_2 norm of Gaussian vectors.

Lemma A.11 (Chernoff bound for the ℓ_2 norm of Gaussian vector)**.** *Let $X_1,\cdots,X_d$ be i.i.d standard Gaussian random variables and $X=[X_1,\cdots,X_d]\in\mathbb{R}^d$. Then for any $\tilde{\gamma}\in(0,1)$, with probability at least $1-\exp(-d\tilde{\gamma}^2/8)$ there holds $\|X\|_2^2\leq(1+\tilde{\gamma})d$.*

The next lemma tells us the error incurred by $R(\mathbf{w}_{\text{priv}})-R(\bar{\mathbf{w}}_T)$ is bounded by the added noise $\mathbf{u}$.

Lemma A.12. *Suppose ℓ is nonnegative, convex and G-Lipschitz. Consider $\mathbf{w}_{priv}$ and $\bar{\mathbf{w}}_T$ from Algorithm 2. For any $\gamma>0$, and for any $\gamma\in(\exp(-d/8),1)$, with probability at least $1-\gamma$, we have*

$$R(\mathbf{w}_{priv})-R(\bar{\mathbf{w}}_T)\leq 2G\sigma\sqrt{d}\ln^{1/4}(1/\gamma).$$

Proof. By the definition of R, we have

$$\begin{aligned} R(\mathbf{w}_{\text{priv}}) - R(\bar{\mathbf{w}}_T) =& \mathbb{E}_{\mathbf{z},\mathbf{z}'}[\ell(\mathbf{w}_{\text{priv}}, \mathbf{z}, \mathbf{z}') - \ell(\bar{\mathbf{w}}_T, \mathbf{z}, \mathbf{z}')] \\ \leq& \mathbb{E}_{\mathbf{z},\mathbf{z}'}[\langle \mathbf{w}_{\text{priv}} - \bar{\mathbf{w}}_T, \partial\ell(\mathbf{w}_{\text{priv}}, \mathbf{z}, \mathbf{z}')\rangle] \\ \leq& \mathbb{E}_{\mathbf{z},\mathbf{z}'}[\|\Pi_{\mathcal{W}}(\bar{\mathbf{w}}_T + \mathbf{u}) - \bar{\mathbf{w}}_T\|_2 \|\partial\ell(\mathbf{w}_{\text{priv}}, \mathbf{z}, \mathbf{z}')\|_2] \\ \leq& \mathbb{E}_{\mathbf{z},\mathbf{z}'}[\|\mathbf{u}\|_2 \|\partial\ell(\mathbf{w}_{\text{priv}}, \mathbf{z}, \mathbf{z}')\|_2] \\ \leq& G\|\mathbf{u}\|_2 \end{aligned} \tag{A.17}$$

where the first inequality is due to the convexity of ℓ, the second inequality is by Cauchy-Schwarz inequality, the third inequality is by the non-expansiveness of projection and the last inequality is because ℓ is G-Lipschitz for any $\mathbf{w}, \mathbf{z}, \mathbf{z}'$. Now, since $\mathbf{u} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_d)$, then by Lemma A.11, for $\gamma \in (\exp(-d/8), 1)$ we have with probability $1 - \gamma$,

$$\|\mathbf{u}\|_2 \leq \sigma\sqrt{d}\Big(1 + \Big(\frac{8\ln(1/\gamma)}{d}\Big)^{1/4}\Big).$$

Plugging the above inequality back into Equation (A.17) we get the desired result. ◻

Theorem A.11 (Theorem 5 restated)**.** *Suppose ℓ is nonnegative, convex and G-Lipschitz, and $\mathcal{W}$ is bounded with diameter D. Consider Algorithm 2 for T iterations under random selection with replacement rule. For any privacy budget $\epsilon > 0$, any $\delta > 0$ and for any $\gamma \in (\max\{4\delta, \exp(-d/8)\}, 1)$, with probability at least $1-\gamma$, we have*

$$R(\mathbf{w}_{priv}) - R(\mathbf{w}_*) \leq \frac{4}{T}\sum_{t=1}^{T} \mathcal{R}_t(\ell \circ \mathcal{W}) + \frac{D^2}{2T\eta} + \frac{\eta G^2}{2} + 6(M + GD)\sqrt{\frac{\ln(8T/\gamma)}{n}} + 19e(M+GD)\sqrt{\frac{\ln(8e/\gamma)}{n}}$$
$$+ 2\sqrt{e}G\eta\big(4 + 48\sqrt{6}G\lceil\ln(n)\rceil\ln(8e/\gamma)\big)\Big(\sqrt{T} + \frac{\sqrt{3}T\ln(eT)\ln(2/\delta)}{n}\Big) + 2G\sigma\sqrt{d}\ln^{1/4}(4/\gamma).$$

Proof. For any neighborhood datasets S and S', Theorem 1 implies with probability least $1 - \delta/2$ that

$$\|\bar{\mathbf{w}}_T - \bar{\mathbf{w}}'_T\|_2 \leq 2\sqrt{e}G\eta\Big(\sqrt{T} + \frac{\sqrt{3}T\ln(eT)\ln(2/\delta)}{n}\Big). \tag{A.18}$$

Since $\gamma \geq 4\delta$, we know the (A.18) holds with probability at least $1 - \gamma/8$. Applying Corollary A.6 with (A.18) we know with probability at least $1 - \gamma/4$ we have

$$\begin{aligned} R(\bar{\mathbf{w}}_T) - R_S(\bar{\mathbf{w}}_T) \leq& 2\sqrt{e}G\eta\big(4 + 48\sqrt{6}G\lceil\ln(n)\rceil\ln(8e/\gamma)\big)\Big(\sqrt{T} + \frac{\sqrt{3}T\ln(eT)\ln(2/\delta)}{n}\Big) \\ &+ 12\sqrt{2}e(M + GD)\sqrt{\frac{\ln(8e/\gamma)}{n}}. \end{aligned} \tag{A.19}$$

Recalling the error decomposition (6) and applying Theorem A.7, Theorem A.8 and Lemma A.12 each with probability $1 - \gamma/4$ together with (A.19), we have the desired bound. ◻

A.5 Rademacher Complexity for AUC Maximization and Similarity Metric Learning

Firstly we look at the Rademacher complexity for AUC maximization.

Lemma A.13. *Given the parameter space $\mathcal{W} = \{\mathbf{w} \in \mathbb{R}^d | \|\mathbf{w}\|_2 \leq D\}$, and denote $\kappa = \sup_{\mathbf{x}}\|\mathbf{x}\|_2$. the Rademacher complexity of $\mathcal{H} = \{h_{\mathbf{w}} | \mathbf{w} \in \mathcal{W}\}$ can be upper bounded by $R_t(\mathcal{H}) \leq \frac{2D\kappa}{\sqrt{t}}$.*

Proof. Starting with the definition, the Rademacher complexity can be upper bounded by

$$\begin{aligned}R_t(\mathcal{H}) =&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\sup_{h\in\mathcal{H}}\frac{1}{t}\sum_{k=1}^{t}\sigma_k h_{\mathbf{w}}(\mathbf{x}_i,\mathbf{x}_{i_k})\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\sup_{\mathbf{w}\in\mathcal{W}_t}\frac{1}{t}\sum_{k=1}^{t}\sigma_k\langle\mathbf{w},\mathbf{x}_i-\mathbf{x}_{i_k}\rangle\right]\\ \leq&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\sup_{\mathbf{w}\in\mathcal{W}_t}\|\mathbf{w}\|_2\left\|\frac{1}{t}\sum_{k=1}^{t}\sigma_k(\mathbf{x}_i-\mathbf{x}_{i_k})\right\|_2\right] \leq \frac{D}{nt}\sum_{i=1}^{n}\left(\mathbb{E}\left[\left\|\sum_{k=1}^{t}\sigma_k(\mathbf{x}_i-\mathbf{x}_{i_k})\right\|_2^2\right]\right)^{\frac{1}{2}}\\ =&\frac{D}{nt}\sum_{i=1}^{n}\left(\sum_{k=1}^{t}\mathbb{E}\Big[\|\mathbf{x}_i-\mathbf{x}_{i_k}\|_2^2\Big]\right)^{\frac{1}{2}} \leq \frac{2D\kappa}{\sqrt{t}}\end{aligned}$$

where the first inequality is due to Cauchy-Schwarz inequality, the third identity is due to $\{\sigma_k\}_{k=1}^t$ are independent random variables with mean zero. ☐

Next we turn our focus to similarity metric learning.

Lemma A.14. *Consider the parameter space defined via the nuclear norm $\mathcal{W} = \big\{\mathbf{w}\in\mathbb{R}^{d\times d}, \|\mathbf{w}\|_{S_1}\leq D\big\}$, where $\|\mathbf{w}\|_{S_1}$ denotes the nuclear norm of a matrix $\mathbf{w}$. The complexity of $\mathcal{H}=\{h_{\mathbf{w}}:\mathbf{w}\in\mathcal{W}\}$ is bounded by*

$$R_t(\mathcal{H}) = \mathcal{O}\Big(\frac{D\big\|\mathbb{E}[\|X\|_2^2XX^\top]\big\|_{S_\infty}^{\frac{1}{2}}\sqrt{\log d}}{\sqrt{t}}\Big), \tag{A.20}$$

where $\|\cdot\|_{S_\infty}$ denotes the largest singular value.

Proof. The complexity of $\mathcal{H}=\{h_{\mathbf{w}}:\mathbf{w}\in\mathcal{W}\}$ is bounded by

$$\begin{aligned}R_t(\mathcal{H}) =&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\sup_{\mathbf{w}\in\mathcal{W}}\frac{1}{t}\sum_{k=1}^{t}\sigma_k\Big\langle\mathbf{w},(\mathbf{x}_i-\mathbf{x}_{i_k})(\mathbf{x}_i-\mathbf{x}_{i_k})^\top\Big\rangle\right] \leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\sup_{\mathbf{w}\in\mathcal{W}}\|\mathbf{w}\|_{S_1}\left\|\frac{1}{t}\sum_{k=1}^{t}\sigma_k(\mathbf{x}_i-\mathbf{x}_{i_k})(\mathbf{x}_i-\mathbf{x}_{i_k})^\top\right\|_{S_\infty}\right]\\ \leq&\frac{D}{nt}\sum_{i=1}^{n}\mathbb{E}\left\|\sum_{k=1}^{t}\sigma_k(\mathbf{x}_i-\mathbf{x}_{i_k})(\mathbf{x}_i-\mathbf{x}_{i_k})^\top\right\|_{S_\infty} = \mathcal{O}\Big(\frac{D\big\|\mathbb{E}[\|X\|_2^2XX^\top]\big\|_{S_\infty}^{\frac{1}{2}}\sqrt{\log d}}{\sqrt{t}}\Big),\end{aligned}$$

where $\|\cdot\|_{S_\infty}$ denotes the largest singular value of a matrix and we have used Lemma A.17 in the last step. ☐

For any $p\geq 1$, the Schatten-p norm of a matrix $W\in\mathbb{R}^{d\times d}$ is defined as the ℓ_p-norm of the vector of singular values $\sigma(W) := (\sigma_1(W),\ldots,\sigma_d(W))^\top$ (the singular values are assumed to be sorted in non-increasing order), i.e., $\|W\|_{S_p} := \|\sigma(W)\|_p$. Let $\Sigma = \mathbb{E}[XX^\top]$. We assume $d\geq 3$.

The following Khintchine-Kahane inequality Lust-Piquard and Pisier (1991) provides a powerful tool to control the q-th norm of the summation of Rademacher series. The following form can be found in Qiu and Wicks (2014).

Lemma A.15 (Matrix Khintchine)**.** *Let $X_1,\ldots,X_n$ be a set of symmetric matrices of the same dimension and let $\sigma_1,\ldots,\sigma_n$ be a sequence of independent Rademacher random variables. For all $q\geq 2$,*

$$\Big(\mathbb{E}_\sigma\big\|\sum_{i=1}^{n}\sigma_iX_i\big\|_{S_q}^q\Big)^{\frac{1}{q}} \leq 2^{-\frac{1}{4}}\sqrt{\frac{q\pi}{e}}\big\|\big(\sum_{i=1}^{n}X_i^2\big)^{\frac{1}{2}}\big\|_{S_q}. \tag{A.21}$$

The following inequality is the Bernstein inequality for a summation of independent matrices Tropp (2015).

Lemma A.16 (Matrix Bernstein)**.** *Let $Z_1,\ldots,Z_n$ be independent, mean-zero and symmetric random matrices in $\mathbb{R}^{d\times d}$. Assume that each one is uniformly bounded*

$$\mathbb{E}[Z_i]=0 \quad \textit{and} \quad \|Z_i\|_{S_\infty}\leq L \quad \textit{for each } i=1,\ldots,n.$$

Introduce the sum $S = \sum_{i=1}^n Z_i$ *and let* $v(S)$ *denote the matrix variance statistic of the sum*

$$v(S) = \Big\| \sum_{i=1}^n \mathbb{E}[Z_i^2] \Big\|_{S_\infty}.$$

Then

$$\mathbb{E}\big[\|S\|_{S_\infty}\big] \leq \sqrt{2v(S)\log(2d)} + \frac{L}{3}\log(2d).$$

Lemma A.17. *Let* $\sigma_1, \ldots, \sigma_n$ *be independent Rademacher variables. Then*

$$\frac{1}{n}\mathbb{E}\Big\|\sum_{i=1}^n \sigma_i \mathbf{x}_i\mathbf{x}_i^\top\Big\|_{S_\infty} \leq 2^{\frac{1}{4}}\sqrt{\pi \log d}\bigg(\frac{\sqrt{\log(2d)}\sup_{\mathbf{x}}\|\mathbf{x}\|_2^2}{n} + \frac{2\big\|\mathbb{E}[\|X\|_2^2 XX^\top]\big\|_{S_\infty}^{\frac{1}{2}}}{\sqrt{n}}\bigg). \tag{A.22}$$

Under the mild assumption $\sqrt{\log(2d)}\sup_{\mathbf{x}}\|\mathbf{x}\|_2^2 \leq \sqrt{n}\big\|\mathbb{E}[\|X\|_2^2 XX^\top]\big\|_{S_\infty}^{\frac{1}{2}}$ *we get*

$$\frac{1}{n}\mathbb{E}\Big\|\sum_{i=1}^n \sigma_i \mathbf{x}_i\mathbf{x}_i^\top\Big\|_{S_\infty} = \mathcal{O}\bigg(\frac{\sqrt{\log d}\big\|\mathbb{E}[\|X\|_2^2 XX^\top]\big\|_{S_\infty}^{\frac{1}{2}}}{\sqrt{n}}\bigg).$$

Proof. By the concavity of the square-root function, we know

$$\begin{aligned}\mathbb{E}_\sigma\big\|\sum_{i=1}^n \sigma_i\mathbf{x}_i\mathbf{x}_i^\top\big\|_{S_\infty} &\leq \Big(\mathbb{E}_\sigma\big\|\sum_{i=1}^n \sigma_i\mathbf{x}_i\mathbf{x}_i^\top\big\|_{S_q}^q\Big)^{\frac{1}{q}} \leq 2^{-\frac{1}{4}}\sqrt{\frac{q\pi}{e}}\Big\|\Big(\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top\mathbf{x}_i\mathbf{x}_i^\top\Big)^{\frac{1}{2}}\Big\|_{S_q} \\ &\leq 2^{-\frac{1}{4}}\sqrt{\frac{q\pi}{e}}d^{\frac{1}{q}}\Big\|\Big(\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top\mathbf{x}_i\mathbf{x}_i^\top\Big)^{\frac{1}{2}}\Big\|_{S_\infty},\end{aligned}$$

where we have used Lemma A.15 and $\|W\|_{S_\infty} \leq \|W\|_{S_q} \leq d^{\frac{1}{q}}\|W\|_{S_\infty}$ for all $W \in \mathbb{R}^{d\times d}$. If we choose $q = 2\log d$ ($d \geq 3$), then

$$\sqrt{q}d^{\frac{1}{q}} = \sqrt{2\log d}d^{\frac{1}{2\log d}} = \sqrt{2e\log d}$$

and therefore

$$\mathbb{E}_\sigma\Big\|\sum_{i=1}^n \sigma_i\mathbf{x}_i\mathbf{x}_i^\top\Big\|_{S_\infty} \leq 2^{\frac{1}{4}}\sqrt{\pi\log d}\Big\|\Big(\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top\mathbf{x}_i\mathbf{x}_i^\top\Big)^{\frac{1}{2}}\Big\|_{S_\infty} = 2^{\frac{1}{4}}\sqrt{\pi\log d}\Big(\Big\|\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top\mathbf{x}_i\mathbf{x}_i^\top\Big\|_{S_\infty}\Big)^{\frac{1}{2}}.$$

It then follows from the concavity of the square-root function that

$$\mathbb{E}\Big\|\sum_{i=1}^n \sigma_i\mathbf{x}_i\mathbf{x}_i^\top\Big\|_{S_\infty} \leq 2^{\frac{1}{4}}\sqrt{\pi\log d}\Big(\mathbb{E}\Big\|\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top\mathbf{x}_i\mathbf{x}_i^\top\Big\|_{S_\infty}\Big)^{\frac{1}{2}} \tag{A.23}$$

It is clear

$$\begin{aligned}\mathbb{E}\Big[\Big\|\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top\mathbf{x}_i\mathbf{x}_i^\top\Big\|_{S_\infty}\Big] &\leq \mathbb{E}\Big[\Big\|\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top\mathbf{x}_i\mathbf{x}_i^\top - \mathbb{E}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top\mathbf{x}_i\mathbf{x}_i^\top\Big\|_{S_\infty}\Big] + \Big\|\mathbb{E}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top\mathbf{x}_i\mathbf{x}_i^\top\Big\|_{S_\infty} \\ &= \mathbb{E}\Big[\Big\|\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top\mathbf{x}_i\mathbf{x}_i^\top - \mathbb{E}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top\mathbf{x}_i\mathbf{x}_i^\top\Big\|_{S_\infty}\Big] + n\Big\|\mathbb{E}\big[\|X\|_2^2 XX^\top\big]\Big\|_{S_\infty}. \end{aligned} \tag{A.24}$$

For all $i \in [n]$, denote $Z_i = \|\mathbf{x}_i\|_2^2\mathbf{x}_i\mathbf{x}_i^\top - \mathbb{E}\big[\|\mathbf{x}_i\|_2^2\mathbf{x}_i\mathbf{x}_i^\top\big]$. It is clear that

$$\begin{aligned}\mathbb{E}\big[\sum_{i=1}^n Z_i^2\big] &= \sum_{i=1}^n \mathbb{E}\Big[\|\mathbf{x}_i\|_2^6\mathbf{x}_i\mathbf{x}_i^\top\Big] - \sum_{i=1}^n\Big(\mathbb{E}[\|\mathbf{x}_i\|_2^2\mathbf{x}_i\mathbf{x}_i^\top]\Big)\Big(\mathbb{E}[\|\mathbf{x}_i\|_2^2\mathbf{x}_i\mathbf{x}_i^\top]\Big) \\ &= n\mathbb{E}\big[\|X\|_2^6 XX^\top\big] - n\mathbb{E}\big[\|X\|_2^2 XX^\top\big]\mathbb{E}\big[\|X\|_2^2 XX^\top\big] \preceq n\mathbb{E}\big[\|X\|_2^6 XX^\top\big]\end{aligned}$$

and therefore

$$\Big\|\mathbb{E}\big[\sum_{i=1}^{n} Z_i^2\big]\Big\|_{S_\infty} \le n\Big\|\mathbb{E}\big[\|X\|_2^6 XX^\top\big]\Big\|_{S_\infty}. \tag{A.25}$$

Furthermore,

$$\|Z_i\|_{S_\infty} \le \sup_{\mathbf{x}_i} \|\mathbf{x}_i\mathbf{x}_i^\top\mathbf{x}_i\mathbf{x}_i^\top\|_{S_\infty} \le \sup_{\mathbf{x}} \|\mathbf{x}\|_2^4. \tag{A.26}$$

We can apply Lemma A.16 with the above bound of variance (A.25) and magnitude (A.26), and derive

$$\mathbb{E}\Big[\|\sum_{i=1}^{n} Z_i\|_{S_\infty}\Big] \le \sqrt{2n\big\|\mathbb{E}\big[\|X\|_2^6 XX^\top\big]\big\|_{S_\infty}\log(2d)} + \frac{1}{3}\sup_{\mathbf{x}}\|\mathbf{x}\|_2^4\log(2d).$$

This together with the sub-additivity of the square-root function and (A.24) implies

$$\begin{aligned}
&\Big(\mathbb{E}\Big[\|\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top\mathbf{x}_i\mathbf{x}_i^\top\|_{S_\infty}\Big]\Big)^{\frac{1}{2}}\\
&\le \Big(\mathbb{E}\Big[\Big\|\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top\mathbf{x}_i\mathbf{x}_i^\top - \mathbb{E}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top\mathbf{x}_i\mathbf{x}_i^\top\Big\|_{S_\infty}\Big]\Big)^{\frac{1}{2}} + \Big(n\Big\|\mathbb{E}\big[\|X\|_2^2XX^\top\big]\Big\|_{S_\infty}\Big)^{\frac{1}{2}}\\
&\le \Big(2n\big\|\mathbb{E}\big[\|X\|_2^6XX^\top\big]\big\|_{S_\infty}\log(2d)\Big)^{\frac{1}{4}} + \frac{\sqrt{\log(2d)}}{\sqrt{3}}\sup_{\mathbf{x}}\|\mathbf{x}\|_2^2 + \sqrt{n\big\|\mathbb{E}[\|X\|_2^2XX^\top]\big\|_{S_\infty}}.
\end{aligned}$$

We plug the above inequality back into (A.23), and get the inequality

$$\begin{aligned}
\frac{1}{n}\mathbb{E}\Big\|\sum_{i=1}^{n}\sigma_i\mathbf{x}_i\mathbf{x}_i^\top\Big\|_{S_\infty} \le 2^{\frac{1}{4}}\sqrt{\pi\log d}\Big(&(2\log(2d))^{\frac{1}{4}}n^{-\frac{3}{4}}\big\|\mathbb{E}\big[\|X\|_2^6XX^\top\big]\big\|_{S_\infty}^{\frac{1}{4}}\\
&+\frac{\sqrt{\log(2d)}\sup_{\mathbf{x}}\|\mathbf{x}\|_2^2}{\sqrt{3}n} + \frac{\big\|\mathbb{E}[\|X\|_2^2XX^\top]\big\|_{S_\infty}^{\frac{1}{2}}}{\sqrt{n}}\Big).
\end{aligned} \tag{A.27}$$

It is clear that

$$\Big\|\mathbb{E}\big[\|X\|_2^6XX^\top\big]\Big\|_{S_\infty}^{\frac{1}{4}} \le \sup_{\mathbf{x}}\|\mathbf{x}\|_2\big\|\mathbb{E}[\|X\|_2^2]XX^\top\big\|_{S_\infty}^{\frac{1}{4}}.$$

This together with Cauchy-Schwartz inequality shows that

$$(2\log(2d))^{\frac{1}{4}}n^{-\frac{3}{4}}\big\|\mathbb{E}\big[\|X\|_2^6XX^\top\big]\big\|_{S_\infty}^{\frac{1}{4}} \le \frac{\big\|\mathbb{E}[\|X\|_2^2XX^\top]\big\|_{S_\infty}^{\frac{1}{2}}}{\sqrt{n}} + \frac{\sqrt{\log(2d)}\sup_{\mathbf{x}}\|\mathbf{x}\|_2^2}{2^{\frac{3}{2}}n}.$$

Plugging the above inequality back into (A.27) gives the stated bound (A.22) ($2^{-\frac{3}{2}} + 3^{-\frac{1}{2}} < 1$). The proof is complete. $\square$