Stability and differential privacy of stochastic gradient descent for pairwise learning with non-smooth loss
Yang, Zhenhuan; Lei, Yunwen; Lyu, Siwei; Ying, Yiming

License:
Creative Commons: Attribution (CC BY)

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):

Link to publication on Research at Birmingham portal
Stability and Differential Privacy of Stochastic Gradient Descent for Pairwise Learning with Non-Smooth Loss

Zhenhuan Yang\textsuperscript{1} Yunwen Lei\textsuperscript{2} Siwei Lyu\textsuperscript{3} Yiming Ying\textsuperscript{1,*}

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

In this paper, we are concerned with another important class of learning tasks called pairwise learning where the quality of a model parameter $w$ is measured by a pairwise loss $\ell(w, z, z')$ on pairs of examples $(z, z')$ as opposed to the pointwise loss $\ell(w, z)$ 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; Clémencçon 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 $W$ is assumed to be bounded. Such assumptions are very restrictive which are violated in many cases.
such as the least square loss for AUC maximization 
\((1 - w^T(x - x'))^2\) for pointwise learning 
with non-smooth losses. Our contributions can be 
summarized as follows.

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; 
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 tech-
niques are mainly motivated by the recent work 
(Bassily et al., 2020; Lei and Ying, 2020) where sta-
ilarity of SGD was established in the pointwise case. 
The main challenge here is that pairs of examples 
involving in pairwise learning are not statistically in-
dependent. To overcome this hurdle, we develop a 
novel approach for decoupling such pairwise-de-
dependent random variables in the analysis. We also 
derive the first generalization bound in high proba-
bility 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 
hance 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 affirm-
aive solution to the open question raised in Cao 
et al., 2010) that whether similarity metric learning 
with nuclear-norm constraint can yield milder de-
pendence 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 Bran-
son, 2018) have focused on pairwise learning. 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 complex-
nity or the covering number. The convergence (opti-
imization error) of SGD type algorithms for pairwise 
learning was obtained in Lin et al. (2017); Ying and 
Zhou (2016) where the algorithms there directly min-
imize 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 Lin 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 uni-
form 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 point-
wise 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 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. Specifi-
cally, 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 \(O(\max\{1/\sqrt{n}, \sqrt{d/(nc)}\})\) 
was established for \((\epsilon, \delta)\)-DP algorithms which, how-
ever, requires the loss function to be Lipschitz and 
strongly smooth, and the domain \(W\) 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 
oily SGD algorithm. In regard to the differential pri-
vate SGD in the pairwise case, the only work that we are aware of is [Huang et al. (2020)] which studied both gradient perturbation and output perturbation with Gaussian noise. They derive the rate $\hat{O}(\sqrt{d}/(\sqrt{n}e))$ for gradient perturbation and $\hat{O}(\sqrt{d}/(\sqrt{ne}))$ 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 $\ell: W \times Z \times Z \to \mathbb{R}$, we aim to minimize the following population risk

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

where $z$ and $z'$ are drawn independently from the population distribution $P$ on $Z$. The population distribution is often unknown and we only have access to a set of i.i.d. training data $S = \{z_1, z_2, \ldots, z_n\} \subset Z^n$. The task then reduces to minimizing the empirical risk

$$\min_{w \in W} R_S(w) := \frac{1}{n(n-1)} \sum_{i,j=1}^{n} \ell(w, z_i, z_j).$$

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

$$\epsilon_{\text{risk}}(A(S)) = [R(A(S)) - R_S(A(S))] + [R_S(w_*) - R(w_*)],$$

where $w_* \in \arg\min_{w \in W} R(w)$. The first term on the right hand side of (2) is called the estimation error. Since $w_*$ is fixed, the term $R_S(w_*) - R(w_*)$ can be trivially handled by the standard Hoeffding inequality. As a comparison, the estimation of the term $R(A(S)) - R_S(A(S))$, also called the generalization error, is much more challenging since $A(S)$ depends on $S$. We will develop novel stability analysis to handle this term. The last term $R_S(A(S)) - R_S(w_*)$ 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 $f: W \to \mathbb{R}$, let $\partial f(w)$ be a subgradient of $f$ at $w$. A function $f$ is said to be convex if for any $w, w' \in W$, there holds

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

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

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

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

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 $W$ 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) [Lin 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 $S = \{z_1, \ldots, z_n\}$ and $S' = \{z'_1, \ldots, z'_n\}$ be two neighborhood datasets that differ only in one single example. For any $\gamma \in (0, 1)$, $\mathcal{A}$ is called $\epsilon_{\text{stab}}$-UAS with probability $1 - \gamma$ if for any neighborhood datasets $S$ and $S'$,

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

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, $w_{t+1}$ and $w'_{t+1}$ denote the $(t+1)$-th iterate of Algorithm 1 based on samples $S$ and $S'$, respectively. And, the notation $\hat{O}(\cdot)$ indicates that the bound is up to a logarithmic term.

Algorithm 1 SGD for Pairwise Learning

**Input:** Data set $S = \{z_1, \ldots, z_n\}$, step size $\eta$, number of iterations $T$, initial point $w_1 = 0$ 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

$w_{t+1} = \Pi_{\mathcal{W}}(w_t - \frac{\eta}{n} \sum_{k=1}^{n} \partial \ell(w_t, z_{i_{t+1}}, z_{i_k}))$

**end for**

**Output:** $w_T = \frac{1}{T} \sum_{t=1}^{T} w_t$
Theorem 1. Suppose that we run Algorithm 2 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 \( A \), we have, for any \( S \) and \( S' \), that

\[
\| w_{t+1} - w_{t+1} \|^2 \leq 4en^2C^2 \left[ t + \ln^2(1t) \right] \times \left( \frac{t}{n} + \ln(1/\gamma) + \sqrt{\frac{\ln(1/\gamma)}{n}} \right)^2
\]

(3)

In particular, if \( T \geq n \), then the output of Algorithm 4 is \( \epsilon_{stab} \)-UAS with high probability where

\[
\epsilon_{stab} = \tilde{O} \left( \eta \sqrt{T} + \frac{\eta T \ln(T)}{n} \right).
\]

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 \( z_{i+t} \) but also on previous examples \( \{z_{k} : k = 1, \cdots, t\} \).

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 \( z_i \neq z_j \) 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)

\[
\mathcal{R}_t(\ell \circ W) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left[ \sup_{w \in W} \frac{1}{t} \sum_{k=1}^{t} \sigma_k \ell(w, z_i, z_{i+k}) \right].
\]

(4)

Here \( \sigma_k \) are Rademacher random variables taking values in \( \{ \pm 1 \} \) with equal probability 1/2, and the expectation is taken over \( z_i, z_{i+k} \) and \( \sigma_k \).

Theorem 2. Suppose \( W \) is bounded with diameter \( D \). Denote \( M = \sup_{z, z'} \ell(0, z, z') \). Assume we run Algorithm 4 for \( T \geq n \) 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 \( A \), we have

\[
\epsilon_{risk}(w_T) \leq \frac{4}{T} \sum_{t=1}^{T} \mathcal{R}_t(\ell \circ W) + \frac{D^2}{2T\eta} + \frac{\eta G^2}{2} + \sqrt{\frac{\ln(6T/\gamma)}{n}}
\]

\[
+ c_1 \eta [\ln(n)] \left( \sqrt{T} + \frac{\sqrt{3T \ln(eT) \ln(6/\gamma)}}{n} \right),
\]

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

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

\[
\epsilon_{risk}(w_T) = \mathcal{O} \left( \frac{\ln^2(n)}{\sqrt{n}} \right).
\]

Theorem 2 is proved in Appendix A.2. Using standard technique (Bartlett and Mendelson 2002), the Rademacher complexity estimation of \( \mathcal{R}_t(\ell \circ W) = \mathcal{O}(1/\sqrt{T}) \) holds true in many cases when \( X \) and \( W \) 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 4 even when \( W \) is unbounded. Specifically, let \( D = \| w_* \|_2 \) and \( \mathcal{W}_D = \{ w \in \mathcal{W} : \| w \|_2 \leq D \} \). The main idea is to show that the iterate \( w_t \) from Algorithm 4 has an adaptive bound, i.e. \( w_t \in \mathcal{W}_t = \{ w \in \mathcal{W} : \| w \|_2^2 \leq (G^2 + M)nt \} \).

Theorem 3. Denote \( M = \sup_{z, z'} \ell(0, z, z') \) and \( D = \| w_* \|_2 \). Suppose we run Algorithm 4 for \( T \geq n \) iterations. For any \( \gamma \in (0,1) \), with probability at least \( 1 - \gamma \) w.r.t. the sample \( S \) and the internal randomness of \( A \),
we have
\[ \epsilon_{\text{risk}}(\bar{\omega}_T) \leq 2 \sum_{t=1}^{T} \left( \mathcal{R}_t(\ell \circ \mathcal{W}_t) + \mathcal{R}_t(\ell \circ \mathcal{W}_D) \right) + \frac{D^2}{2T\eta} + \frac{\eta G^2}{2} + c_1 \sqrt{\eta \ln(6T/\gamma)} + c_2 \sqrt{\eta T \ln(n/\gamma)} + c_3 \sqrt{n \ln(6T/\gamma)} + c_4 \eta \ln(n) \left( \sqrt{T} + \frac{4T \ln(cT)}{n} \sqrt{\ln(6n/\gamma)} \right), \]
where \( c_1 = 100 \sqrt{6e^{3/2}G} \max \{1, G\} \ln(6e/\gamma), \) \( c_2 = (7 + 12\sqrt{2}e)M + 4GD + 10eG, \) \( c_4 = 3G \sqrt{G^2 + 2M} \) 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
\[ \epsilon_{\text{risk}}(\bar{\omega}_T) = \mathcal{O}\left( \frac{\ln^2(n)}{n^{1/3}} \right). \]

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 \( \mathcal{R}_t(\ell \circ \mathcal{W}_t) = \mathcal{O}(n\sqrt{T}) = \mathcal{O}(\eta \sqrt{T}) \). One can see more discussion on such estimation in Section 4. Therefore, Theorem 3 mainly differs from Theorem 2 in the additional \( \mathcal{O}(\sqrt{\eta T/n}) \) 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 \((\epsilon, \delta)\)-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 \( q : S^D \to \mathbb{R}^d \), let \( \mathcal{A}(S) = q(S) + \mathbf{u} \) where \( \mathbf{u} \sim \mathcal{N}(0, \sigma^2 I_d) \) with \( I_d \) being the identical matrix. For all neighborhood datasets \( S, S' \) that differ by one example, the \( \ell_2 \)-sensitivity \( \Delta \) of the query function \( q \) is defined as \( \Delta(q) = \sup_{S, S'} \|q(S) - q(S')\|_2 \).

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

**Theorem 4.** Given the total number of iterations \( T \), for any privacy budget \( \epsilon > 0 \) and \( \delta > 0 \), Algorithm 3 satisfies \((\epsilon, \delta)\)-differential privacy with
\[ \sigma^2 = \frac{8e\eta^2G^2 \ln(2.5/\delta)}{\epsilon^2} \left( T + \frac{3T^2 \ln^2(cT) \ln^2(2/\delta)}{n^2} \right). \]

The proof of Theorem 4 is given in Section 4.2. 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 \( \bar{\omega}_T \) can be used to quantify the high-probability sensitivity of the query function \( q(S) = \bar{\omega}_T \). Specifically, subsampling forms an event of probability measure \( 1 - \delta/2 \) under which a small sensitivity \( \mathcal{O}(\eta \sqrt{T} + \eta T \ln(T/n)) \) holds true. Hence, under this event, we only need to add noise with \( \sigma = \mathcal{O}(\eta \sqrt{T} + \eta T \ln(T)) \) 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
\[
\epsilon_{\text{risk}}(w_{\text{priv}}) = R(w_{\text{priv}}) - R(w_*) = R(w_{\text{priv}}) - R(\bar{w}_T) + R(\bar{w}_T) - R(w_*),
\]
where $R(\bar{w}_T) - R(w_*)$ measures the risk incurred by the non-private output $\bar{w}_T$ (Algorithm 1) and $R(w_{\text{priv}}) - R(\bar{w}_T)$ measures the effect of perturbation by adding random noises. The utility bound is given as follow.

**Theorem 5.** Suppose $W$ 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 $\gamma \in (\max\{4\epsilon, \exp(-d/8)\}, 1)$, with probability at least $1 - \gamma$, we have
\[
\epsilon_{\text{risk}}(w_{\text{priv}}) \leq \frac{4}{T} \sum_{t=1}^{T} R_{\ell} + D^2 \frac{27}{2\epsilon \eta} \left( \frac{n}{\sqrt{\ln(6T) / \gamma}} + 2G\sigma \sqrt{d} \ln^{1/4}(4/\gamma) \right) + c_2 \sqrt{\ln(n \ln(n)) \ln(6\epsilon / \gamma)} \left( \sqrt{\ln(3T \ln(\epsilon T / 2 \delta^2))} \right).
\]
where $c_1 = 100 \sqrt{6} e^{3/2} G \max\{1, G\} \ln(6e / \gamma)$ and $c_2 = (6 + 19\epsilon)(M + GD)$.

In particular, letting $\sigma$ satisfy [6] and choosing $T = n^2$ and $\eta = O(n^{-3/2})$, then with high probability we have
\[
\epsilon_{\text{risk}}(w_{\text{priv}}) = \tilde{O} \left( \frac{\sqrt{d}}{\sqrt{n\epsilon}} \right).
\]

Theorem 5 is proved in Appendix A.4. The difference compared to Theorem 2 is the additional $O(\sqrt{d}/\sqrt{n\epsilon})$ term caused by $R(w_{\text{priv}}) - R(\bar{w}_T)$ in [6]. The utility bound $O(\sqrt{d}/\sqrt{n\epsilon})$ 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 2

In this section, we provide technical proofs for Theorems 1 and 2. Proofs of other Theorems can be found in the Appendix. Throughout this section, we let $L_{t+1}(w_t)$ denote the accumulated loss until $z_{t+1}$ is revealed. i.e. $L_{t+1}(w_t) = \frac{1}{t} \sum_{k=1}^{t} \ell(w_t, z_{t+1}, z_{t+1})$.

#### 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 $\mu = E[X]$. Then for any $\gamma > 0$, with probability at least $1 - \exp(-\mu \gamma^2 / (2 + \gamma))$ we have $X \leq (1 + \gamma) \mu$.

**Proof of Theorem 1** Without loss of generality, assume that $S$ and $S'$ differs in $n$-th position. Denote $\delta_{t+1,k} = \partial \ell(w_t, z_{t+1}, z_{t+1}) - \partial \ell(w_t', z_{t+1}', z_{t+1}')$ and $\delta_{t+1,k} = \partial \ell(w_t, z_{t+1}, z_{t+1}) - \partial \ell(w_t', z_{t+1}', z_{t+1}')$. The following recursive inequality holds
\[
\|w_{t+1} - w'_{t+1}\|^2_2 = \|w_t - \eta \delta_{t+1, k} - w'_{t+1}\|^2 - \eta \delta_{t+1, k} - \delta_{t+1, k} \|^2_2 \leq \frac{1}{t} \sum_{k=1}^{t} \|w_t - w'_{t+1} - \eta \delta_{t+1, k}\|^2.
\]

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 $z_{i_t+1} = z'_{i_t+1}$ and $z_{i_k} = z'_{i_k}$. Then
\[
\|w_t - w'_{t+1}\|^2_2 = \|w_t - w'_{t+1}\|^2_2 + \eta^2 \|\delta_{t+1, k}\|^2_2 - 2\eta \langle w_t - w'_{t+1}, \delta_{t+1, k} \rangle \leq \|w_t - w'_{t+1}\|^2_2 + 4\eta^2 G^2,
\]
where the last inequality holds because $\ell$ is $G$-Lipschitz and convex. If $i_{t+1} = n$ or $i_k = n$, then $z_{i_t+1} \neq z'_{i_t+1}$ or $z_{i_k} \neq z'_{i_k}$. It follows from the Young’s inequality that for any $p > 0$
\[
\|w_t - w'_{t+1}\|^2_2 \leq (1 + \frac{1}{p}) \|w_t - w'_{t+1}\|^2_2 + (1 + 1/p) \eta^2 \|\delta_{t+1, k}\|^2_2 \leq (1 + \frac{1}{p}) \|w_t - w'_{t+1}\|^2_2 + 4(1 + 1/p) \eta^2 G^2.
\]
Combining the above two inequalities together and let $Y_t = \frac{1}{t} \sum_{k=1}^{t} \|w_t - w'_{t+1}\|^2_2$, we have
\[
\|w_{t+1} - w'_{t+1}\|^2_2 \leq (1 + pY_t) \|w_t - w'_{t+1}\|^2_2 + 4(1 + 1/p) \eta^2 G^2.
\]
Applying the above inequality recursively we have

\[ \|w_{t+1} - w_{t+1}'\|_2^2 \leq \sum_{j=1}^t \prod_{k=j+1}^t \left( (1+p)Y_j (4+4Y_j/p) \eta_j^2 G^2 \right) \]

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 \( \prod_{k=1}^t x_k^a \leq x_1^{l_1} \cdots x_t^{l_t} \) for \( a \geq 1 \). We note that \( Y_1, \cdots, Y_t \) are dependent variables, but the sum of \( Y_t \)'s has the following decomposition:

\[
\sum_{t=1}^\infty Y_t = \sum_{t=1}^\infty \frac{1}{t} \sum_{k=1}^t \left( \sum_{i=1}^k \Pi_{\mathbb{R}^{|X_i|\times n} \mathbb{R}^{|X_i|\times n}} \right) \leq \sum_{t=1}^\infty \left( \sum_{i=1}^k \Pi_{\mathbb{R}^{|X_i|\times n}} \right) \]

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

Lemma 3 (Post-processing). Let \( A : \mathbb{Z}^n \rightarrow \mathcal{W} \) be a (randomized) algorithm that is \((\epsilon, \delta)-DP\). Let \( f : \mathcal{W} \rightarrow \mathcal{W} \) be an arbitrary randomized mapping. Then \( f \circ A : \mathbb{Z}^n \rightarrow \mathcal{W} \) is \((\epsilon, \delta)-DP\).

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

\[
E = \{ |\Delta_T^2 + 4\eta^2 G^2 \left( T + \frac{3T^2 \ln^2 (2\epsilon T) \ln (2/\delta)}{\epsilon^2} \right) | \}
\]

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

\[
\sigma = \sqrt{2 \ln (2.5/\delta) \Delta_T}.
\]

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

\[
\mathbb{P}[w_{priv}(S) \in O] = \mathbb{P}[w_{priv}(S) \in O \cap I \in E] + \mathbb{P}[w_{priv}(S) \in O \cap I \notin E] = \mathbb{P}[w_{priv}(S) \in O | I \in E] \mathbb{P}[I \in E] + \mathbb{P}[w_{priv}(S) \in O | I \notin E] \mathbb{P}[I \notin E] \leq \left( e^\epsilon \mathbb{P}[w_{priv}(S') \in O | I \in E] + \frac{\delta}{2} \right) \mathbb{P}[I \in E] + \frac{\delta}{2} \leq e^\epsilon \mathbb{P}[w_{priv}(S') \in O \cap I \in E] + \frac{\delta}{2} + \frac{\delta}{2} \leq e^\epsilon \mathbb{P}[w_{priv}(S') \in O] + \delta
\]

where the first inequality is because when \( I \in E \), \( w_{piv} \) 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.
of pairwise learning, namely AUC maximization and similarity metric learning. According to Theorems 2 and 3, the key here is to estimate the Rademacher complexity defined by 4. 

**AUC Maximization.** AUC maximization aims to learn a ranking function \( h_w \) defined by \( h_w(x, x') = w^T(x - x') \). One expects \( h_w \) to rank positive examples higher than negative examples, i.e. \( w^T(x - x') \geq 0 \) for \( y = 1 \) and \( y' = -1 \). Using the hinge loss \( \ell(w, z, z') = (1 - h_w(x, x'))_+ + \mathbb{I}_{[y = 1, y' = -1]} \), AUC maximization can be formulated as

\[
\min_{w \in \mathcal{W}} E_{z, z'} [(1 - w^T(x - x'))_+ + \mathbb{I}_{[y = 1, y' = -1]}].
\]

(10)

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

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

Note in the case of (10), it is easy to check \( R_t(\ell \circ \mathcal{H}) \leq 4GDt/\sqrt{t} \) by Ledoux-Talagrand inequality (Ledoux and Talagrand, 2013). Combining this lemma with Theorems 2 and 3, one can derive the following excess risk and utility bound for Algorithms 1 and 2 in the context of non-smooth metric learning.

**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

\[
\epsilon_{\text{risk}}(w_T) = \tilde{O}(\sqrt{\frac{\kappa}{n}}).
\]

**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 \( \sigma \) given by 5, then, with high probability we have

\[
\epsilon_{\text{risk}}(w_{\text{priv}}) = \tilde{O}(\sqrt{\frac{kd}{n\epsilon}}).
\]

**Similarity Metric Learning.** We now turn to another notable example of pairwise learning called similarity metric learning. It aims to learn a (square) Mahalanobis distance metric which is defined by \( h_w(x, x') = (x - x')^T w (x - x') \) parametrized by a positive semi-definite matrix \( w \in \mathbb{R}^{d \times d} \). 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 \( \ell(w, z, z') = (1 + \tau(y, y')h_w(x, x'))_+ \), it can be formulated as

\[
\min_{w \in \mathcal{W}} E_{z, z'} [(1 + \tau(y, y')(x - x')^T w (x - x'))_+],
\]

(11)

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

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

\[
R_t(\mathcal{H}) = \mathcal{O}(\frac{D||E[||X||_2^2 X X^T||]}{\sqrt{t}} \sqrt{\log d} ),
\]

(12)

where \( || \cdot ||_{S_1} \) denotes the largest singular value.

The proof of Lemma 5 is postponed to Appendix A.5. As direct corollaries of Lemma 2, we can derive generalization bounds for metric learning from Theorems 2 and 3. For brevity, denote \( \chi = ||E[||X||_2^2 X X^T||]||_{S_1} \). 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 \( \eta = \mathcal{O}(n^{-3/2}) \), then, with high probability we have

\[
\epsilon_{\text{risk}}(w_T) = \tilde{O}(\frac{\chi \log(d)}{n}).
\]

**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 \( \sigma \) given by 5, then, with high probability we have

\[
\epsilon_{\text{risk}}(w_{\text{priv}}) = \tilde{O}(\frac{\chi d \log(d)}{\sqrt{\frac{n \epsilon}}}).
\]

**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

\[
R_t(\mathcal{H}) = \mathcal{O}(\frac{D \sup_{x} ||x||^2 \log d}{\sqrt{t}}).
\]

(13)

The difference between (12) and (13) is that we replace \( \sup_{x} ||x||^2 \) by the term \( ||E[||X||_2^2 X X^T||]||_{S_1} \). Notice \( ||E[||X||_2^2 X X^T||]||_{S_1} \geq \frac{1}{2} \text{tr}(E[XX^T XX^T]) = \frac{1}{2} \text{tr}(XX^T XX^T) \geq \frac{1}{2} ||X||_2^2 \). If we assume \( ||X||_2^2 \geq d^2 \), then the upper bound of (12) satisfies the relation \( \geq \sqrt{d \log d}/\sqrt{t} \) 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 \( \geq d \sqrt{\log d}/t \). That is, our algorithm 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 \( \mathcal{W} = \{w \in \mathbb{R}^{d \times d} : ||w||_F \leq D\} \), then one can show that \( R_t(\mathcal{H}) = \mathcal{O}(D^2 \sup_{x} ||x||^2/\sqrt{T}) \) (Lei and Ying, 2016).
This matches the bound in within a logarithmic factor except that there is replaced by . Since , the argument in 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 (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 nonsmooth losses and obtain optimal excess risk bounds . We extend our analysis to unbounded parameter space and achieve a rate of . We apply our stability results to study differentially private SGD algorithms in pairwise learning. Our output perturbation method achieves utility bound , which matches the previous results in 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 .

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 . 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


