Analysis and Applications


# STABILITY AND OPTIMIZATION ERROR OF STOCHASTIC GRADIENT DESCENT FOR PAIRWISE LEARNING

Wei Shen[†], Zhenhuan Yang[§], Yiming Ying[§] and Xiaoming Yuan[‡]

[†]*Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Kowloon, Hong Kong*
*16482530@life.hkbu.edu.hk*

[§]*Department of Mathematics and Statistics, State University of New York, Albany, New York State 12222, USA*
*{zyang6,yying}@albany.edu*

[‡]*Department of Mathematics, The University of Hong Kong, Hong Kong*
*xmyuan@hku.hk*

Received (Day Month Year)
Revised (Day Month Year)

In this paper we study the stability and its trade-off with optimization error for stochastic gradient descent (SGD) algorithms in the pairwise learning setting. Pairwise learning refers to a learning task which involves a loss function depending on pairs of instances among which notable examples are bipartite ranking, metric learning, area under ROC curve (AUC) maximization and minimum error entropy (MEE) principle. Our contribution is twofold. Firstly, we establish the stability results for SGD for pairwise learning in the convex, strongly convex and non-convex settings, from which generalization errors can be naturally derived. Secondly, we establish the trade-off between stability and optimization error of SGD algorithms for pairwise learning. This is achieved by lower-bounding the sum of stability and optimization error by the minimax statistical error over a prescribed class of pairwise loss functions. From this fundamental trade-off, we obtain lower bounds for the optimization error of SGD algorithms and the excess expected risk over a class of pairwise losses. In addition, we illustrate our stability results by giving some specific examples of AUC maximization, metric learning and MEE.

*Keywords*: Stability; Generalization; Optimization Error; Stochastic Gradient Descent; Pairwise Learning; Minimax Statistical Error

Mathematics Subject Classification 2000: 68Q32, 90C15, 90C31.

## 1. Introduction

This paper concerns with *pairwise learning* which usually involves a pairwise loss function, i.e., the loss function depending on a pair of examples which can be expressed by $\ell(f,(x,y),(x',y'))$ for a hypothesis function $f : \mathcal{X} \to \mathbb{R}$. This is in contrast to the problem of *pointwise learning* in standard classification and regression which typically involves a univariate loss function $\ell(f,x,y)$. Several important learning tasks can be viewed as pairwise learning problems. For instance, bipartite

ranking [2,10,34] and AUC maximization [15,21,45,48,49] aim to correctly predict the ordering of pairs of binary labeled samples. This involves the use of a misranking loss $\ell(f,(x,y),(x',y')) = \mathbb{I}_{\{f(x)-f(x')<0\}}\mathbb{I}_{y=1}\mathbb{I}_{y'=-1}$, where $\mathbb{I}_{\{\cdot\}}$ is the indicator function. In practice, one usually replaces the indicator function $\mathbb{I}_{\{f(x)-f(x')<0\}}$ by a smooth convex surrogate function like $(1-(f(x)-f(x')))^2$. Other important examples include metric learning [3,11,42,43,44,4] and minimum error entropy (MEE) principle [19,20,32,40].

Stochastic gradient descent (SGD) has now become the workhorse in machine learning as it scales well to big data. In particular, SGD-type algorithms for pairwise learning have been proposed and extensively studied in the recent work [15,20, 23,28,41,47,49]. The overall performance of SGD algorithms is measured by the excess expected risk which can be decomposed into two parts: the *optimization error* and *generalization error*. The optimization error is sometimes referred to as computational error which characterizes the discrepancy between an output of SGD and the empirical risk minimizer from batch learning. It portrays how fast the algorithm converges as the number of iterations grows. The generalization error describes the discrepancy between the population risk of an output of SGD and its empirical risk. One can interpret the expected and empirical risks as the test error and the training error, respectively. The analysis of optimization and generalization errors has been conducted in the existing literature using various approaches but most of them have been done separately. A natural question would be what is the trade-off between generalization and optimization errors which requires to analyze these two errors together rather than separately.

Generalization analysis has been done for SGD algorithms for pairwise learning using different techniques such as covering number [41], Rademacher complexities [22] and integral operators [20,28,47]. An alternative approach is to use the the concept of algorithmic stability [5,29]. While a large amount of work has been devoted to studying the stability for pointwise learning, there is few work on the stability for pairwise learning except the work by Agarwal and Niyogi [1] which focused on the regularized ERM formulation for bipartite ranking.

**Main Contribution.** The first contribution of our work is to establish random-uniform stability [13] of randomized SGD algorithms for pairwise learning in both convex and non-convex settings, from which generalization error bounds of SGD algorithms can be obtained very naturally. We then illustrate the stability results using concrete examples in metric learning, AUC maximization and MEE principle. Our second contribution is the trade-off framework for stability and optimization error of SGD for pairwise learning, which indicates that tight stability leads to a slow convergence rate (large optimization error), and vice versa. This is achieved by establishing minimax statistical error for the sum of stability and optimization error over a prescribed class of pairwise loss functions. To the best of our knowledge, this is the first-ever known work on the stability and its trade-off with optimization error for randomized SGD algorithms in the setting of pairwise learning.

Our work is inspired by the recent work [18] and [8] which focused on the setting of pointwise learning. Our studies differ from previous work in the following aspects. Firstly, Hardt et al. [18] established stability results for the last iterate of randomized iterative SGD algorithms for pointwise learning. Our work significantly extends the results in [18] to the setting of pairwise learning since we establish both the last iterate and the average of iterates of SDG algorithms for pairwise learning. Secondly, Chen et al. [8] studied the trade-off results between stability and optimization error for SGD in pointwise learning which employed a strong notion of stability called *uniform stability* [5] specifically tailored for deterministic algorithms. Our trade-off framework uses a weak notion called *random uniform stability* [13] which applies to the randomized iterative SGD algorithms. In addition, we established lower bounds of the average of the iterates of SGD algorithms for pairwise learning which match the upper bounds in the literature of online pairwise learning [23,41]. The results are new even for the case of pointwise learning.

**Related Work.** The stability analysis dates back to the work [12,35] where it was shown that the variance of the leave-one-out error can be upper bounded by hypothesis stability [24]. Bousequet and Elisseeff [5] used the notation of uniform stability and studied stability of regularization based algorithms. Kutin and Niyogi [25] introduced several weaker variants of stability, and showed how they are sufficient to obtain generalization bounds for certain algorithms. Rakhlin et al. [33] and Mukherjee et al. [29] studied the relation between stability and learnability. All these work considered the stability of deterministic learning algorithms such as kNN rules, ERM and regularized network and it cannot be used to study a large number of randomized learning algorithms. More recently, Chen et al. [8] employed the strong notation of uniform stability and established the trade-off between stability and convergence rates of certain iterative algorithms.

Elisseeff et al. [13] extended the work [5] and introduced a notion of random uniform stability for studying randomized algorithms such as bagging. Hardt et al. [18] first established random uniform stability for randomized iterative SGD algorithms for convex and non-convex settings in the setting of pointwise learning. The results were further improved in the work [26,31] by exploring the structures of the loss function and the data.

Concurrently, SGD algorithms for pairwise learning were originally introduced and studied in [41]. Pairwise learning involves statistically dependent pairs of instances while, in practice, the individual instances are i.i.d. according an unknown distribution. As such, standard analysis for the pointwise learning case can not be directly applied to pairwise learning. Indeed, there is a considerable efforts on developing various new techniques to study the convergence of SGD for pairwise learning. In particular, generalization bounds [7] of SGD for pairwise were established using uniform convergence approaches such as covering number [41] and Rademacher complexity [23]. The work [46] used integral operators developed in [36,37] to show the convergence of SGD for pairwise learning with a focus on the least-square loss

and the setting of reproducing kernel Hilbert spaces.

A close related concept to algorithmic stability is the statistical robustness which considers the problem of how the estimators change relatively to the perturbation of the underlying distribution generating the data. This robustness concept is more general than algorithmic stability we consider here. In the appealing work [9], it was shown that minimizers of the regularized ERM is statistically robust in the setting of reproducing kernel Hilbert spaces.

**Organization of this paper.** The rest of the paper is organized as follows. Section 2 introduces some basic notations and concepts related to stability which will be used later. In Section 3, we present stability results for SGD in the pairwise learning setting. We establish the trade-off results between stability and optimization error in Section 4. Examples are given in Section 5. We conclude the paper in Section 6.

## 2. Preliminaries

Let the sample $S = \{z_i = (x_i, y_i) : i = 1, \ldots, n\}$ be drawn i.i.d. from $D$ on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ where $\mathcal{X}$ is a domain in $\mathbb{R}^d$ and $\mathcal{Y} \subseteq \mathbb{R}$. Let $\mathbf{w} \in \mathbb{R}^d$ be the model parameter associated with the hypothesis function $f$ (e.g., the linear hypothesis function $f(x) = \mathbf{w}^T x$ ). The goal of pairwise learning is to minimize the following population risk:

$$R(\mathbf{w}) \overset{\text{def}}{=} \mathbb{E}_{(z,z')\sim D\times D}[\ell(\mathbf{w}, z, z')]. \tag{2.1}$$

The corresponding empirical risk is defined by

$$R_S(\mathbf{w}) \overset{\text{def}}{=} \frac{2}{n(n-1)} \sum_{i<j} \ell(\mathbf{w}, z_i, z_j). \tag{2.2}$$

We use the conventional notation $A$ denote the randomized SGD algorithm and $A(S)$ to denote its output based on $S$. The expected generalization error of $A(S)$ is given by

$$\epsilon_{\text{gen}} \overset{\text{def}}{=} \mathbb{E}_{S,A}[R_S(A(S)) - R(A(S))], \tag{2.3}$$

where the expectation is taken over the randomness of $A$ and $S$.

### 2.1. *SGD for Pairwise Learning*

Recall that the pairwise learning loss $\ell : \mathbb{R}^d \times \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}^+$ is defined, for any $\mathbf{w}, z, z' \in \mathcal{Z}$, by $\ell(\mathbf{w}, z, z')$. The SGD updates for pairwise learning [23,41,47,49] are given by $\mathbf{w}_1 = 0$, and for $2 \le t \le T$,

$$\mathbf{w}_t = \mathbf{w}_{t-1} - \frac{\alpha_{t-1}}{t-1} \sum_{j=1}^{t-1} \nabla \ell(\mathbf{w}_{t-1}, z_{\xi_t}, z_{\xi_j}), \tag{2.4}$$

where $\{z_{\xi_j}\}_{j=1}^T$ are examples from $S$ with the indexes $\{\xi_j\}_{j=1}^T$ chosen at random from $\{1, \cdots, n\}$, and $\nabla \ell$ denotes the gradient with respect to the first argument.

The above algorithm is an extension of the standard SGD in the pointwise learning setting to the pairwise learning setting. It was first introduced by Wang et al. [41] as online gradient descent for pairwise learning. It was further developed for AUC maximization [15,47,49] and MEE [20] for the stochastic setting (i.e. the data are assumed to be i.i.d.). For simplicity, we refer to it as SGD for pairwise learning or just SGD when it is clear from the context.

There are two schemes for choosing $\{\xi_j\}_{j=1}^T$ for the SGD update rule which are independent of the sample $S$. The first one, called the *random permutation rule*, is to choose a new random permutation over $\{1, \cdots, n\}$ at the beginning of each epoch and go through the examples in the order determined by the permutations. The other is the *random selection rule* which selects each $\xi_j$ uniformly at random in $\{1, \cdots, n\}$ at each step. In this work, our results hold true for the above two schemes.

The output of SGD algorithm (2.4) at $T$ can be the last iterate $A(S) = \mathbf{w}_T$ or the average of iterates $A(S) = \bar{\mathbf{w}}_T = \frac{1}{T}\sum_{t=1}^T \mathbf{w}_t$. We denote $A^{\text{last}}(S) = \mathbf{w}_T$ and $A^{\text{avg}}(S) = \bar{\mathbf{w}}_T$. Later on we use the conventional notation $A(S)$ to denote $A^{\text{avg}}(S)$ or $A^{\text{last}}(S)$ when it can be either of them.

### 2.2. *Algorithmic Stability and Its Relation with Generalization*

We will use a modification of $\epsilon$-uniform stability introduced by Agarwal and Niyogi [2] which considered the regularized ERM formulation for ranking problems. It can also be regarded as an extension of random uniform stability [13] to the case of pairwise learning.

**Definition 2.1.** An SGD algorithm $A$ for pairwise learning is called random uniform stable with $\varepsilon > 0$ if for all data sets $S, S' \in \mathcal{Z}^n$ according to distribution $D$ such that $S$ and $S'$ differ in at most one example, we have

$$\sup_{(z,z')\sim D\times D} \mathbb{E}_A[\ell(A(S), z, z') - \ell(A(S'), z, z')] \leq \epsilon. \tag{2.5}$$

Here, the expectation is taken only over the randomness of $A$. We denote the smallest constant $\varepsilon$ satisfies (2.5) as $\epsilon_{\text{stab}}(A, T, \ell, D, n)$.

It is worthy of noting that we always assume that the randomness for algorithm $A$ is independent of the sample $S$ which is i.i.d. generated from $D$ on $\mathcal{X} \times \mathcal{Y}$. The notation $\epsilon_{\text{stab}}(A, T, \ell, D, n)$ can be $\epsilon_{\text{stab}}(A^{\text{last}}, T, \ell, D, n)$ for the last iterate of SGD or $\epsilon_{\text{stab}}(A^{\text{avg}}, T, \ell, D, n)$ for the average of iterates.

The following theorem describes the relation between the stability and generalization for pairwise learning which is originally in the work [2,1] for bipartite ranking. We include its proof for completeness.

**Theorem 2.1.** *If the SGD algorithm $A$ is random uniform stable with $\varepsilon > 0$, then we have*

$$|\mathbb{E}_{S,A}[R_S(A(S)) - R(A(S))]| \leq 2\epsilon. \tag{2.6}$$

**Proof.** Denote by $S = (z_1, \cdots, z_n)$ and $\tilde{S} = (\tilde{z}_1, \cdots, \tilde{z}_n)$ two samples wherein the examples are i.i.d. chosen from $D$. Let $S'(i)$ be an i.i.d. copy of $S$ except the $i$th example being replaced by $\tilde{z}_i$. Let $S''(i,j) = (z_1, \cdots, \tilde{z}_i, \cdots, \tilde{z}_j, \cdots, z_n)$. Therefore,

$$\begin{aligned}
&\mathbb{E}_S\mathbb{E}_A[R_S(A(S))] = \mathbb{E}_S\mathbb{E}_A\Big[\frac{2}{n(n-1)}\sum_{i<j}\ell(A(S); z_i, z_j)\Big] \\
=&\mathbb{E}_{\tilde{S}}\mathbb{E}_S\mathbb{E}_A\Big[\frac{2}{n(n-1)}\sum_{i<j}\ell(A(S''(i,j)); \tilde{z}_i, \tilde{z}_j)\Big] \\
=&\,\mathbb{E}_{\tilde{S}}\mathbb{E}_S\mathbb{E}_A\Big[\frac{2}{n(n-1)}\sum_{i<j}\ell(A(S); \tilde{z}_i, \tilde{z}_j)\Big] + \delta = \mathbb{E}_S\mathbb{E}_A\Big[R(A(S))\Big] + \delta, \qquad (2.7)
\end{aligned}$$

where the second equality comes from the identical distribution assumption. The residual term $\delta$ in the last two equations can be expressed as

$$\begin{aligned}
\delta &= \frac{2}{n(n-1)}\sum_{i<j}\mathbb{E}_{\tilde{S}}\mathbb{E}_S\mathbb{E}_A\Big[\ell(A(S''(i,j)); \tilde{z}_i, \tilde{z}_j) - \ell(A(S); \tilde{z}_i, \tilde{z}_j)\Big] \\
&= \frac{2}{n(n-1)}\sum_{i<j}\mathbb{E}_{\tilde{S}}\mathbb{E}_S\mathbb{E}_A\Big[\ell(A(S''(i,j)); \tilde{z}_i, \tilde{z}_j) - \ell(A(S'(i)); \tilde{z}_i, \tilde{z}_j) \\
&+ \ell(A(S'(i)); \tilde{z}_i, \tilde{z}_j) - \ell(A(S); \tilde{z}_i, \tilde{z}_j)\Big] \\
&= \frac{2}{n(n-1)}\sum_{i<j}\mathbb{E}_S\mathbb{E}_A\mathbb{E}_{(\tilde{z}_i,\tilde{z}_j)\sim D\times D}\Big[\ell(A(S''(i,j)); \tilde{z}_i, \tilde{z}_j) - \ell(A(S'(i)); \tilde{z}_i, \tilde{z}_j)\Big] \\
&+ \frac{2}{n(n-1)}\sum_{i<j}\mathbb{E}_S\mathbb{E}_A\mathbb{E}_{(\tilde{z}_i,\tilde{z}_j)\sim D\times D}\Big[\ell(A(S'(i)); \tilde{z}_i, \tilde{z}_j) - \ell(A(S); \tilde{z}_i, \tilde{z}_j)\Big]. \qquad (2.8)
\end{aligned}$$

Note that $S''(i,j)$ and $S'(i)$ differ in only one example and so do $S'(i)$ and $S$. Furthermore, taking the supremum over any two data sets $S, S'$ differing in only one example, we can bound the difference as

$$|\delta| \leq 2\sup_{S,S',(z,\tilde{z})\sim D\times D}\mathbb{E}_A\left[\ell(A(S'); z, \tilde{z}) - \ell(A(S); z, \tilde{z})\right] \leq 2\epsilon, \qquad (2.9)$$

by our assumption on the random uniform stability of $A$. The claim follows. □

Theorem 2.1 bounds the expected generalization error of SGD for pairwise learning with two times of its random uniform stability bound. We will present the detailed bounds for the stability of SGD for pairwise learning in Section 3.

### 2.3. *Stability and Optimization Error Decomposition*

In this subsection, we assume $\mathbf{w} \in \Omega \subseteq \mathbb{R}^d$. Recall that $A(S)$ is the output of SGD algorithm (2.4) for pairwise learning at iteration $T$. The overall performance of the output $A(S)$ is measured in terms of the *excess risk* defined as

$$\Delta R(A(S)) \stackrel{\text{def}}{=} R(A(S)) - \inf_{\mathbf{w}\in\Omega} R(\mathbf{w}). \qquad (2.10)$$

For notional simplicity, let

$$\mathbf{w}_S^* = \text{argmin}_{\mathbf{w}\in\Omega} R_S(\mathbf{w}), \tag{2.11}$$

and

$$\mathbf{w}^* = \text{argmin}_{\mathbf{w}\in\Omega} R(\mathbf{w}). \tag{2.12}$$

Then we can obtain the following decomposition, namely,

$$\begin{aligned}\Delta R(A(S)) &= R(A(S)) - R(\mathbf{w}^*)\\ &= R(A(S)) - R_S(A(S)) + R_S(A(S)) - R_S(\mathbf{w}_S^*)\\ &\quad + R_S(\mathbf{w}_S^*) - R_S(\mathbf{w}^*) + R_S(\mathbf{w}^*) - R(\mathbf{w}^*)\\ &\le R(A(S)) - R_S(A(S)) + R_S(A(S)) - R_S(\mathbf{w}_S^*)\\ &\quad + R_S(\mathbf{w}^*) - R(\mathbf{w}^*),\end{aligned} \tag{2.13}$$

where the last inequality follows from the fact $R_S(\mathbf{w}_S^*) - R_S(\mathbf{w}^*) \le 0$ from the definition $\mathbf{w}_S^*$ (i.e. (2.11)). Taking expectation on both sides of (2.13) w.r.t. the randomness of $S$ and $A$ and noting that $\mathbb{E}_S[R_S(\mathbf{w}^*) - R(\mathbf{w}^*)] = 0$, we can decompose the expected excess risk as

$$\mathbb{E}_{S,A}[\Delta R(A(S))] \le \mathbb{E}_{S,A}[\underbrace{R(A(S)) - R_S(A(S))}_{\text{generalization error}}] + \mathbb{E}_{S,A}[\underbrace{R_S(A(S)) - R_S(\mathbf{w}_S^*)}_{\text{optimization error}}]. \tag{2.14}$$

Denote the expected generalization error and optimization error of $A(S)$ as $\epsilon_{\text{gen}}(A, T, \ell, D, n) \overset{\text{def}}{=} \mathbb{E}_{S,A}[R(A(S)) - R_S(A(S))]$ and $\epsilon_{\text{opt}}(A, T, \ell, D, n) \overset{\text{def}}{=} \mathbb{E}_{S,A}[R_S(A(S)) - R_S(\mathbf{w}_S^*)]$. Note that the above quantities are indexed by the estimator $A(S)$, loss function $\ell$, data distribution $D$ and sample size $n$. When it is clear from the context, we will omit these indexes for simplicity. As a result, we can rewrite (2.14) as

$$\mathbb{E}_{S,A}[\Delta R(A(S))] \le \epsilon_{\text{gen}}(A, T, \ell, D, n) + \epsilon_{\text{opt}}(A, T, \ell, D, n). \tag{2.15}$$

Combining the expected excess risk decomposition (2.15) and Theorem 2.1, we have, for any loss $\ell$, that

$$\mathbb{E}_{S,A}[\Delta R(A(S)] \le 2\epsilon_{\text{stab}}(A, T, \ell, D, n) + \epsilon_{\text{opt}}(A, T, \ell, D, n). \tag{2.16}$$

The above inequality means that the overall performance of SGD measured by the excess population risk $\Delta R(A(S))$ can be decomposed into stability and optimization error. This leads to a natural question that what is the trade-off between these two terms and whether SGD can achieve both the tighter stability bounds and fast convergence rate.

To answer this question, we consider the stability and optimization error for the last output of SGD (i.e. $A^{\text{last}}(S)$) over a class of convex pairwise losses $\mathcal{L}$ and $\mathcal{D}$ is the class of all probability distributions which are given by

$$\mathcal{E}_{\text{stab}}^{last}(T, \mathcal{L}, \mathcal{D}, n) \overset{\text{def}}{=} \sup_{\ell\in\mathcal{L}, D\in\mathcal{D}} \epsilon_{\text{stab}}(A^{\text{last}}(S), T, \ell, D, n),$$

and

$$\mathcal{E}_{\rm opt}^{last}\,(T,\mathcal{L},\mathcal{D},n)\stackrel{\rm def}{=}\sup_{\ell\in\mathcal{L},D\in\mathcal{D}}\epsilon_{\rm opt}(A^{\rm last}(S),T,\ell,D,n).$$

Likewise, one can define $\mathcal{E}_{\rm stab}^{avg}\,(T,\mathcal{L},\mathcal{D},n)\stackrel{\rm def}{=}\sup_{\ell\in\mathcal{L},D\in\mathcal{D}}\epsilon_{\rm stab}(A^{\rm avg}(S),T,\ell,D,n)$ and $\mathcal{E}_{\rm opt}^{avg}\,(T,\mathcal{L},\mathcal{D},n)\stackrel{\rm def}{=}\sup_{\ell\in\mathcal{L},D\in\mathcal{D}}\epsilon_{\rm opt}(A^{\rm avg}(S),T,\ell,D,n)$.

Recall that the minimax risk in nonparametric statistics [38,39] is given by $\inf_{\widetilde{\mathbf{w}}_n}\sup_{D\in\mathcal{D}}\mathbb{E}_{S\sim D^n}[\Delta R(\widetilde{\mathbf{w}}_n)]$ where the infinimum is taken with respect to all possible estimator $\widetilde{\mathbf{w}}_n:\mathcal{Z}^n\to\mathbb{R}^d$ which is a function of a random sample $S=\{z_1,\ldots,z_n\}$, i.e. $\widetilde{\mathbf{w}}_n=\widetilde{\mathbf{w}}_n(S)$. The key idea is to connect the above two errors with minimax risk in nonparametric statistics as given by the following lemma.

**Lemma 2.1.** *For any convex pairwise loss $\ell\in\mathcal{L}$, there holds*

$$2\mathcal{E}_{stab}^{last}\,(T,\mathcal{L},\mathcal{D},n)+\mathcal{E}_{opt}^{last}\,(T,\mathcal{L},\mathcal{D},n)\geq\inf_{\widetilde{\mathbf{w}}_n}\sup_{D\in\mathcal{D}}\mathbb{E}_{S\sim D^n}[\Delta R(\widetilde{\mathbf{w}}_n)],\tag{2.17}$$

*and*

$$2\mathcal{E}_{stab}^{avg}\,(T,\mathcal{L},\mathcal{D},n)+\mathcal{E}_{opt}^{avg}\,(T,\mathcal{L},\mathcal{D},n)\geq\inf_{\widetilde{\mathbf{w}}_n}\sup_{D\in\mathcal{D}}\mathbb{E}_{S\sim D^n}[\Delta R(\widetilde{\mathbf{w}}_n)].\tag{2.18}$$

**Proof.** We only prove (2.17) as the proof for (2.18) is exactly the same.

From (2.16) and definitions for $\mathcal{E}_{\rm stab}^{last}\,(T,\mathcal{L},\mathcal{D},n)$ and $\mathcal{E}_{\rm opt}^{last}\,(T,\mathcal{L},\mathcal{D},n)$, we have, for any $\ell\in\mathcal{L}$, that

$$\sup_{D\in\mathcal{D}}\mathbb{E}_{S,A}[\Delta R(A^{\rm last}(S))]\leq 2\mathcal{E}_{\rm stab}^{last}\,(T,\mathcal{L},\mathcal{D},n)+\mathcal{E}_{\rm opt}^{last}\,(T,\mathcal{L},\mathcal{D},n).\tag{2.19}$$

Notice that $\Delta R(A^{\rm last}(S))=\mathbb{E}_{(z,z')}[\ell(A^{\rm last}(S),z,z')]-\inf_{\mathbf{w}}\mathbb{E}_{(z,z')}[\ell(\mathbf{w},z,z')]$ and the randomness of the SGD algorithm $A$ is independent of $S$. Consequently,

$$\begin{aligned}&\mathbb{E}_{S,A}[\Delta R(A^{\rm last}(S))]=\mathbb{E}_S\big\{\mathbb{E}_A[\Delta R(A^{\rm last}(S))]\big\}\\&=\mathbb{E}_S\big\{\mathbb{E}_A[\mathbb{E}_{(z,z')}[\ell(A^{\rm last}(S),z,z')]\big\}-\inf_{\mathbf{w}}\mathbb{E}_{(z,z')}[\ell(\mathbf{w},z,z')].\end{aligned}\tag{2.20}$$

Since $\ell\in\mathcal{L}$ is convex with respect to the first argument, Jensen's inequality tells us that

$$\mathbb{E}_A[\mathbb{E}_{(z,z')}[\ell(A^{\rm last}(S),z,z')]\geq\mathbb{E}_{(z,z')}\big[\ell(\mathbb{E}_A[A^{\rm last}(S)],z,z')\big].\tag{2.21}$$

Putting (2.20) and (2.21) together, we have

$$\mathbb{E}_{S,A}[\Delta R(A^{\rm last}(S))]\geq\mathbb{E}_S\big[\Delta R(\,\mathbb{E}_A[A^{\rm last}(S)]\,)\big].$$

Putting this back into (2.19) yields that

$$\begin{aligned}2\mathcal{E}_{\rm stab}^{last}\,(T,\mathcal{L},\mathcal{D},n)+\mathcal{E}_{\rm opt}^{last}\,(T,\mathcal{L},\mathcal{D},n)&\geq\sup_{D\in\mathcal{D}}\mathbb{E}_S\big[\Delta R(\,\mathbb{E}_A[A^{\rm last}(S)]\,)\big]\\&\geq\inf_{\widetilde{\mathbf{w}}_n}\sup_{D\in\mathcal{D}}\mathbb{E}_{S\sim D^n}[\Delta R(\widetilde{\mathbf{w}}_n)].\end{aligned}$$

This completes the proof of the lemma. □

Using techniques from nonparametric statistics (e.g. [27,38,39]), one can estimate the minimum risk on the righthand side of (2.17) and thus derive trade-off results between stability and optimization error of SGD for pairwise learning as we will do soon in Section 4.

It is worth of mentioning that this connection (2.17) was first observed in [8] for pointwise learning which, however, focused on the deterministic algorithms. Specifically, the uniform stability in [8] is not taken with respect to the randomness of algorithm $A$ and the expectation $\mathbb{E}$ involved in Lemma 2.1 is only with respect to $S$ without the randomness of algorithm $A$. Our paper studies stability of SGD algorithm defined by (2.4) which involves the randomness of $\{\xi_j\}$, and the uniform stability defined by Definition 2.1 is taken in the sense of the expectation of $\{\xi_j\}$. In this sense, our result stated in Lemma 2.1 is a non-trivial extension of [8] to the the case of randomized SGD algorithms for pairwise learning.

## 3. Stability Analysis of SGD Algorithms

In this section we establish stability results for SGD algorithms given by (2.4). Before we present the main stability results, we introduce some definitions and background materials.

### 3.1. *Warm-up: Some Technical Preparation*

The following definitions list convexity and smoothness properties of a function $f$.

**Definition 3.1.** A function $f$ is convex if and only if $\mathrm{dom} f$ is a convex set and $f(\theta x_1+(1-\theta)x_2) \leq \theta f(x_1)+(1-\theta)f(x_2)$, for all $x_1, x_2 \in \mathrm{dom} f$ and $\theta \in [0,1]$. And a function $f$ is $\gamma$-strongly convex if and only if $g(x) = f(x) - (\gamma/2)x^\top x$ is convex.

**Definition 3.2.** A function $f$ is $L$-Lipschitz if and only if $\|f(x_2) - f(x_1)\| \leq L \cdot \|x_2 - x_1\|$, for all $x_1, x_2 \in \mathrm{dom} f$. Furthermore, a function $f$ is $\beta$-strongly smooth or $\beta$-smooth for short if and only if $f$ is differentiable and $\nabla f(x)$ is $\beta$-Lipschitz.

Let $S' = \{z_1', z_2', \cdots, z_n'\}$ be an i.i.d. copy of $S$ but differ from $S$ at precisely one location. Assume SGD for pairwise learning is run based on $S$ and $S'$ along the same path $\{\xi_1, \xi_2, \cdots, \xi_T\}$ with the same initial points $\mathbf{w}_1 = \mathbf{w}_1' = 0$. Recall, for $t = 2, \cdots, T$, the SGD updates based on $S$ are given by

$$G_t(\mathbf{w}_{t-1}) = \mathbf{w}_{t-1} - \frac{\alpha_{t-1}}{t-1} \sum_{j=1}^{t-1} \nabla \ell(\mathbf{w}_{t-1}, z_{\xi_t}, z_{\xi_j}). \tag{3.1}$$

Similarly, for $t = 2, \cdots, T$, we denote the gradient updates based on $S'$ by

$$G_t'(\mathbf{w}_{t-1}') = \mathbf{w}_{t-1}' - \frac{\alpha_{t-1}}{t-1} \sum_{j=1}^{t-1} \nabla \ell(\mathbf{w}_{t-1}', z_{\xi_t}', z_{\xi_j}').$$

We say that an operator $G_t$ is *expansive* with parameter $\eta_t > 0$ if $\|G_t(\mathbf{w}) - G_t(\mathbf{w}')\| \leq \eta_t \|\mathbf{w} - \mathbf{w}'\|$ for any $\mathbf{w}$ and $\mathbf{w}'$. The main theorems about stability rely on the following lemma which states $G_t$ is expansive.

**Lemma 3.1.** *Assume that $\ell(\cdot, z, z')$ is $\beta$-smooth for every pair $(z, z')$.*

1. *Then $G_t$ is $(1 + \alpha_{t-1}\beta)$-expansive.*
2. *Assume in addition that $\ell(\cdot, z, z')$ is convex and $\alpha_{t-1} \leq \frac{2}{\beta}$. Then $G_t$ is 1-expansive.*
3. *Assume in addition that $\ell(\cdot, z, z')$ is $\gamma$-strongly convex and $\alpha_{t-1} \leq \frac{2}{\beta+\gamma}$. Then $G_t$ is $\left(1 - \frac{\beta\gamma\alpha_{t-1}}{\beta+\gamma}\right)$-expansive.*

The proof for the above elementary results can be found in Appendix A. Note that the results of Lemma 3.1 about $G_t$ also apply to $G'_t$.

Now consider the SGD updates respectively on $S$ and $S'$ with $\mathbf{w}_t = G_t(\mathbf{w}_{t-1})$ and $\mathbf{w}'_t = G'_t(\mathbf{w}'_{t-1})$ for any $t \geq 2$ and the initial point $\mathbf{w}_1 = \mathbf{w}'_1 = 0$. The stability of SGD for pairwise learning critically depends on the following recursive property of $\delta_t = \|\mathbf{w}_t - \mathbf{w}'_t\|$.

**Theorem 3.1.** *Assume that $\ell(\cdot, z, z')$ is $L$-Lipschitz for any $z, z'$. Suppose that both $G_t$ and $G'_t$ are expansive with parameter $\eta_t$. Then for $1 < t \leq T$, under both random rules (e.g. random permutation or selection rules), the following recursive relation holds true.*

$$\mathbb{E}[\delta_t] \leq \left\{\frac{1}{n} \cdot \min(\eta_t, 1) + \left(1 - \frac{1}{n}\right) \cdot \eta_t\right\}\mathbb{E}[\delta_{t-1}] + \frac{4L}{n} \cdot \alpha_{t-1}. \tag{3.2}$$

The proof of this theorem is inspired by the work [18]. However, compared with the situation in the context of pointwise learning, the key challenge here is that at any step $t$, the computation of the new gradient direction not only depends on the current example $z_{\xi_t}$ but also on all previously used examples, i.e. $\{z_{\xi_i}\}_{i=1}^{t-1}$. We overcome this hurdle by a careful investigation into how many times SGD has encountered the different examples between $S$ and $S'$ before the $t$-th step, as illustrated below respectively for both cases of random selection and permutation rules.

We first consider the case of random selection rule.

**Lemma 3.2.** *Suppose that we run SGD based on $S$ and $S'$ under the random selection rule for $T$ steps along the same path $\{\xi_1, \xi_2, \cdots, \xi_T\}$. For a fixed $t \in (1, T]$, assume among the first $t-1$ steps, there are $m$ steps where SGD has encountered the different examples. Then we have the following properties:*

*(1) $\delta_t \leq \min(\eta_t, 1)\delta_{t-1} + 2\alpha_{t-1}L$, if $z_{\xi_t} \neq z'_{\xi_t}$;*
*(2) $\delta_t \leq \eta_t\delta_{t-1} + \frac{m}{t-1} \cdot 2\alpha_{t-1}L$, if $z_{\xi_t} = z'_{\xi_t}$,*

*wherein $\eta_t$ is the expansive parameter of the updates $G_t$ and $G'_t$.*

**Proof.** First of all, for either case, we have

$$\begin{aligned}\delta_t &= \|G_t(\mathbf{w}_{t-1}) - G'_t(\mathbf{w}'_{t-1})\| \\ &\leq \|G_t(\mathbf{w}_{t-1}) - G_t(\mathbf{w}'_{t-1})\| + \|G_t(\mathbf{w}'_{t-1}) - G'_t(\mathbf{w}'_{t-1})\| \\ &\leq \eta_t\delta_{t-1} + \frac{\alpha_{t-1}}{t-1}\sum_{j=1}^{t-1}\|\nabla_w\ell(\mathbf{w}'_{t-1}, z'_{\xi_t}, z'_{\xi_j}) - \nabla_w\ell(\mathbf{w}'_{t-1}, z_{\xi_t}, z_{\xi_j})\|. \end{aligned} \tag{3.3}$$

Then we prove the two claims in this lemma separately.

1) For the first property, if $z_{\xi_t} \neq z'_{\xi_t}$, we have $\nabla_w\ell(\mathbf{w}'_{t-1}, z'_{\xi_t}, z'_{\xi_j}) \neq \nabla_w\ell(\mathbf{w}'_{t-1}, z_{\xi_t}, z_{\xi_j})$ for all $j = 1, \cdots, t-1$. Then following the $L$-Lipschitz condition of $\ell$, we have

$$\|\nabla_w\ell(\mathbf{w}'_{t-1}, z'_{\xi_t}, z'_{\xi_j}) - \nabla_w\ell(\mathbf{w}'_{t-1}, z_{\xi_t}, z_{\xi_j})\| \leq 2L.$$

As a result, we obtain

$$\delta_t \leq \eta_t\delta_{t-1} + 2\alpha_{t-1}L. \tag{3.4}$$

Next we prove the other half of the first claim of this lemma. By the triangle inequality, we have

$$\begin{aligned}\delta_t &= \|G_t(\mathbf{w}_{t-1}) - G'_t(\mathbf{w}'_{t-1})\| \\ &\leq \|\mathbf{w}_{t-1} - \mathbf{w}'_{t-1}\| + \frac{\alpha_{t-1}}{t-1}\sum_{j=1}^{t-1}\|\nabla_w\ell(\mathbf{w}'_{t-1}; z'_{\xi_t}, z'_{\xi_j}) - \nabla_w\ell(\mathbf{w}_{t-1}; z_{\xi_t}, z_{\xi_j})\| \\ &\leq \delta_{t-1} + 2\alpha_{t-1}L. \end{aligned} \tag{3.5}$$

Thus the first property follows by combining (3.4) and (3.5).

2) We now prove the second property. Denote $U = \{1 \leq j \leq t-1 | z_{\xi_j} \neq z'_{\xi_j}\}$. From the assumption that there are $m$ steps where SGD has encountered the different examples among the first $t-1$ steps, we know there are $m$ number of elements in $\{z_{\xi_j}\}_{j=1}^{t-1}$ which are different from those in $\{z'_{\xi_j}\}_{j=1}^{t-1}$. That means $|U| = m$ where $|U|$ is the number of coordinates in the set $U$. Recall we have $z_{\xi_t} = z'_{\xi_t}$ and thus there are at most $m$ number of the pairs $\{z_{\xi_t}, z_{\xi_j}\}_{j=1}^{t-1}$ which are different from $\{z'_{\xi_t}, z'_{\xi_j}\}_{j=1}^{t-1}$. It follows that

$$\begin{aligned}&\sum_{j=1}^{t-1}\|\nabla_w\ell(\mathbf{w}'_{t-1}, z'_{\xi_t}, z'_{\xi_j}) - \nabla_w\ell(\mathbf{w}'_{t-1}, z_{\xi_t}, z_{\xi_j})\| \\ &= \sum_{j\in U}\|\nabla_w\ell(\mathbf{w}'_{t-1}, z'_{\xi_t}, z'_{\xi_j}) - \nabla_w\ell(\mathbf{w}'_{t-1}, z_{\xi_t}, z_{\xi_j})\|. \end{aligned}$$

Thus following the $L$-Lipschitz condition of $\ell$, we have

$$\sum_{j=1}^{t-1}\|\nabla_w\ell(\mathbf{w}'_{t-1}, z'_{\xi_t}, z'_{\xi_j}) - \nabla_w\ell(\mathbf{w}'_{t-1}, z_{\xi_t}, z_{\xi_j})\| \leq 2mL.$$

Plugging this into (3.3), we get the second property. □

Now we consider the permutation rule for $T$ steps. In this case, let $t_k^\star$ be the (only one) element in the set $\{t \mid z_{\xi_t} \neq z'_{\xi_t},\ (k-1)n < t \leq kn\}$ for each $k \geq 1$. In fact at the $t_k^\star$-th step, SGD encounters the different examples during the $k$-th epoch. Fix an arbitrary sequence of SGD updates $G_1, \cdots, G_T$ based on $S$ and another sequence $G'_1, \cdots, G'_T$ based on $S'$. We have the following lemma for the recursive property of the SGD updates.

**Lemma 3.3.** *Suppose that we run SGD based on $S$ and $S'$ under the random permutation rule for $T$ steps along the same path $\{\xi_1, \xi_2, \cdots, \xi_T\}$. Assume that both $G_t$ and $G'_t$ are expansive with parameter $\eta_t$. For $(k-1)n < t \leq kn$ where $k$ is the number of epochs, we have the following properties:*

*(1)* $\delta_t \leq \min(\eta_t, 1)\delta_{t-1} + 2\alpha_{t-1}L$, *if* $t = t_k^\star$,
*(2)* $\delta_t \leq \eta_t\delta_{t-1} + \frac{k-1}{t-1} \cdot 2\alpha_{t-1}L$, *if* $(k-1)n < t < t_k^\star$,
*(3)* $\delta_t \leq \eta_t\delta_{t-1} + \frac{k}{t-1} \cdot 2\alpha_{t-1}L$, *if* $t_k^\star < t \leq kn$.

**Proof.** 1) For each $(k-1)n < t \leq kn$ where $k$ is the number of epochs, we have

$$\begin{aligned}
\delta_t &= \|G_t(\mathbf{w}_{t-1}) - G'_t(\mathbf{w}'_{t-1})\| \\
&\leq \|G_t(\mathbf{w}_{t-1}) - G_t(\mathbf{w}'_{t-1})\| + \|G_t(\mathbf{w}'_{t-1}) - G'_t(\mathbf{w}'_{t-1})\| \\
&\leq \eta_t\delta_{t-1} + \frac{\alpha_{t-1}}{t-1}\sum_{j=1}^{t-1}\|\nabla_w\ell(\mathbf{w}'_{t-1}, z'_{\xi_t}, z'_{\xi_j}) - \nabla_w\ell(\mathbf{w}'_{t-1}, z_{\xi_t}, z_{\xi_j})\|. \qquad (3.6)
\end{aligned}$$

For the first property, if $t = t_k^\star$, we must have $z_{\xi_t} \neq z'_{\xi_t}$. As a result, $\nabla_w\ell(\mathbf{w}'_{t-1}, z'_{\xi_t}, z'_{\xi_j}) \neq \nabla_w\ell(\mathbf{w}'_{t-1}, z_{\xi_t}, z_{\xi_j})$ for all $j = 1, \cdots, t-1$. Then following the $L$-Lipschitz condition of $\ell$, we have

$$\delta_t \leq \eta_t\delta_{t-1} + 2\alpha_{t-1}L. \qquad (3.7)$$

Next we prove the other half. By the triangle inequality, we have

$$\begin{aligned}
\delta_t &= \|G_t(\mathbf{w}_{t-1}) - G'_t(\mathbf{w}'_{t-1})\| \\
&\leq \|\mathbf{w}_{t-1} - \mathbf{w}'_{t-1}\| + \frac{\alpha_{t-1}}{t-1}\sum_{j=1}^{t-1}\|\nabla_w\ell(\mathbf{w}'_{t-1}; z'_{\xi_t}, z'_{\xi_j}) - \nabla_w\ell(\mathbf{w}_{t-1}; z_{\xi_t}, z_{\xi_j})\| \\
&\leq \delta_{t-1} + 2\alpha_{t-1}L. \qquad (3.8)
\end{aligned}$$

Thus the first property follows by combining (3.7) and (3.8).

2) We now prove the second property. If $(k-1)n < t < t_k^\star$, we have $z_{\xi_j} \neq z'_{\xi_j}$ when $j \in U^\star := \{t_1^\star, \cdots, t_{k-1}^\star\}$, while $z_{\xi_j} = z'_{\xi_j}$ for $j$ belonging to $\{1, 2, \cdots, t\}$ but not in $U^\star$. As a result, $z_{\xi_t} = z'_{\xi_t}$ and there are at most $(k-1)$ number of the pairs $\{(z_{\xi_t}, z_{\xi_j})\}_{j\in U^\star}$ which are different from $\{(z'_{\xi_t}, z'_{\xi_j})\}_{j\in U^\star}$. Thus following the

$L$-Lipschitz condition of $\ell$, we have

$$\sum_{j=1}^{t-1} \|\nabla_w \ell(\mathbf{w}'_{t-1}, z'_{\xi_t}, z'_{\xi_j}) - \nabla_w \ell(\mathbf{w}'_{t-1}, z_{\xi_t}, z_{\xi_j})\|$$
$$= \sum_{j\in U^\star} \|\nabla_w \ell(\mathbf{w}'_{t-1}, z_{\xi_t}, z'_{\xi_j}) - \nabla_w \ell(\mathbf{w}'_{t-1}, z_{\xi_t}, z_{\xi_j})\| \leq 2(k-1)L.$$

Plugging this into (3.6), we get the second property.

3) Following the same strategy as above, if $t_k^\star < t \leq kn$, there are at most $k$ number of the pairs $\{(z_{\xi_t}, z_{\xi_j})\}_{j\in V^\star}$ which are different from $\{(z'_{\xi_t}, z'_{\xi_j})\}_{j\in V^\star}$, where $V^\star = \{t_1^\star, \cdots, t_k^\star\}$. Similarly, we have

$$\sum_{j=1}^{t-1} \|\nabla \ell(\mathbf{w}'_{t-1}; z'_{\xi_t}, z'_{\xi_j}) - \nabla \ell(\mathbf{w}'_{t-1}; z_{\xi_t}, z_{\xi_j})\| \leq 2kL.$$

Plugging this into the equation (3.6), we get the third property. □

We are now in a position to prove Theorem 3.1.

**Proof of Theorem 3.1.** Firstly, under the random selection rule, we denote $m$ as the times of SGD choosing the different examples during the first $t-1$ steps. Since the examples chosen by SGD at each step are i.i.d. under the random selection rule, $m$ follows a binomial distribution, i.e. $m \sim \mathbb{B}(t-1, 1/n)$. And we know that at step $t$, $\mathbb{P}\{z_{\xi_t} \neq z'_{\xi_t}\} = \frac{1}{n}$. Then by the independence between the $t-$th step and previous $t-1$ steps, the probability of that $z_{\xi_t} = z'_{\xi_t}$ at the $t$-th step and SGD has encountered the different examples $m$ times during the previous $t-1$ steps is $\left(1-\frac{1}{n}\right)\cdot C_{t-1}^m \left(1-\frac{1}{n}\right)^{t-1-m}\left(\frac{1}{n}\right)^m$ where $C_{t-1}^m$ is the binomial coefficient. By Lemma 3.2, for every $1 < t \leq T$, we have

$$\mathbb{E}[\delta_t] \leq \frac{1}{n}\cdot\Big(\min(\eta_t, 1)\mathbb{E}[\delta_{t-1}] + 2\alpha_{t-1}L\Big)$$
$$+ \sum_{m=0}^{t-1}\left(1-\frac{1}{n}\right)\cdot C_{t-1}^m\left(1-\frac{1}{n}\right)^{t-1-m}\left(\frac{1}{n}\right)^m \times \Big(\eta_t\mathbb{E}[\delta_{t-1}] + \frac{m}{t-1}\cdot 2\alpha_{t-1}L\Big)$$
$$\leq \Big\{\frac{1}{n}\cdot\min(\eta_t, 1) + \Big(1-\frac{1}{n}\Big)\cdot\eta_t\Big\}\mathbb{E}[\delta_{t-1}] + \frac{4L\alpha_{t-1}}{n},$$

wherein the second inequality follows from the facts

$$\sum_{m=0}^{t-1} C_{t-1}^m\left(1-\frac{1}{n}\right)^{t-1-m}\left(\frac{1}{n}\right)^m = 1$$

and

$$\sum_{m=0}^{t-1} mC_{t-1}^m\left(1-\frac{1}{n}\right)^{t-1-m}\left(\frac{1}{n}\right)^m = \frac{t-1}{n}.$$

Secondly, under the random permutation rule, $t_k^\star$ is a uniformly random number in $\{(k-1)n+1, (k-1)n+2, \cdots, kn\}$ and therefore $\forall k \geq 1$,for every $(k-1)n < t \leq kn$ we have

$$\mathbb{P}\{t_k^\star = t\} = \frac{1}{n},\ \mathbb{P}\{t_k^\star > t\} = 1 - \frac{t-(k-1)n}{n} = k - \frac{t}{n},$$

and

$$\mathbb{P}\{t_k^\star < t\} = \frac{t-1-(k-1)n}{n} = \frac{t-1}{n} - (k-1).$$

By Lemma 3.3, for every $(k-1)n < t \leq kn$ with $k \geq 1$, we have

$$\begin{aligned}
&\mathbb{E}[\delta_t] \leq \frac{1}{n} \cdot \Big( \min\big(\eta_t, 1\big)\mathbb{E}[\delta_{t-1}] + 2\alpha_{t-1}L \Big) + \left(k - \frac{t}{n}\right) \cdot \left( \eta_t \mathbb{E}[\delta_{t-1}] + \frac{k-1}{t-1} \cdot 2\alpha_{t-1}L \right) \\
&+ \Big( \frac{t-1}{n} - (k-1) \Big) \Big( \eta_t \mathbb{E}[\delta_{t-1}] + \frac{k}{t-1} \cdot 2\alpha_{t-1}L \Big) \\
&\leq \Big\{ \frac{1}{n} \cdot \min\big(\eta_t, 1\big) + \left(1 - \frac{1}{n}\right) \cdot \eta_t \Big\} \mathbb{E}[\delta_{t-1}] + \frac{4L\alpha_{t-1}}{n}.
\end{aligned}$$

Finally, combining the above two cases yields the desired result. □

Before we use Theorem 3.1 to analyze the stability of SGD for convex, strongly convex and non-convex cases respectively, we introduce the following useful lemma which reveals an important advantage of SGD: it usually takes several steps before the updates $\mathbf{w}_t$ and $\mathbf{w}_t'$ of SGD start to differ from each other.

**Lemma 3.4.** *Assume that the loss function $\ell(\cdot; z, z')$ is nonnegative and L-Lipschitz for all pairs $(z, z')$. Suppose we run SGD for $T$ steps on two samples of size $n$ namely $S$ and $S'$ which differ in at most an example. Then, for every $t_0 \in \{2, \cdots, n\}$, we have*

$$\mathbb{E}\Big[ |\ell(\mathbf{w}_T; z, z') - \ell(\mathbf{w}_T'; z, z')| \Big] \leq \frac{t_0}{n} \sup_{\mathbf{w}, z, z'} \ell(\mathbf{w}, z, z') + L\mathbb{E}[\delta_T | \delta_{t_0} = 0], \quad (3.9)$$

*where $\delta_{t_0} = \|\mathbf{w}_{t_0} - \mathbf{w}_{t_0}'\|$.*

**Proof.** Let $z, z' \in Z$ be an arbitrary pair of examples. By the conditional expectation formula and the Lipschitz assumption of $\ell$, we have

$$\begin{aligned}
&\mathbb{E}\Big[ |\ell(\mathbf{w}_T; z, z') - \ell(\mathbf{w}_T'; z, z')| \Big] \\
&= \mathbb{P}\{\delta_{t_0} \neq 0\} \mathbb{E}[|\ell(\mathbf{w}_T, z, z') - \ell(\mathbf{w}_T', z, z')| \big| \delta_{t_0} \neq 0] \\
&+ \mathbb{P}\{\delta_{t_0} = 0\} \mathbb{E}[|\ell(\mathbf{w}_T, z, z') - \ell(\mathbf{w}_T', z, z')| \big| \delta_{t_0} = 0] \\
&\leq \mathbb{P}\{\delta_{t_0} \neq 0\} \cdot \sup_{\mathbf{w}, z, z'} \ell(\mathbf{w}, z, z') + L\mathbb{E}[\delta_T \big| \delta_{t_0} = 0].
\end{aligned}$$

Now we bound $\mathbb{P}\{\delta_{t_0} \neq 0\}$ under random permutation and selection rules.

Under the random permutation rule, denote $t_1^\star = \{t \mid z_{\xi_t} \neq z'_{\xi_t}, 1 \leq t \leq n\}$. We have

$$\mathbb{P}\{\delta_{t_0} \neq 0\} \leq \mathbb{P}\{t_1^\star \leq t_0\} = \frac{t_0}{n} \quad (3.10)$$

since if $t_1^\star > t_0$, then we must have $\delta_{t_0} = 0$.

For the case of random selection rule, let $t^\star$ be the first time our algorithm encountering the different examples. For the same reason behind (3.10), we just need to bound $\mathbb{P}\{t^\star \leq t_0\}$ and we have

$$\mathbb{P}\{\delta_{t_0} \neq 0\} \leq \mathbb{P}\{t^\star \leq t_0\} \leq \sum_{t=1}^{t_0} \mathbb{P}\{t^\star = t\} = \frac{t_0}{n}.$$

Combining these two cases, we complete the proof. □

### 3.2. *Convex case*

We present below the first stability result of SGD provided that the pairwise loss $\ell(\cdot, z, z')$ is convex and strongly smooth.

**Theorem 3.2.** *Assume that the loss function $\ell(\cdot; z, z')$ is $\beta$-smooth, convex and $L$-Lipschitz for every example points $z$ and $z'$. Suppose that we run SGD with step sizes $\alpha_t \leq 2/\beta$ for $T$ steps. Then,*

$$\epsilon_{stab}(A^{last}, T, \ell, D, n) \leq \frac{4L^2}{n} \sum_{t=1}^{T-1} \alpha_t, \tag{3.11}$$

*and*

$$\epsilon_{stab}(A^{avg}, T, \ell, D, n) \leq \frac{4L^2}{Tn} \sum_{t=2}^{T} \sum_{j=1}^{t-1} \alpha_j, \tag{3.12}$$

**Proof.** We now fix a pair of examples $z$ and $z'$ and apply the Lipschitz condition on $\ell(\cdot, z, z')$ to get

$$\mathbb{E}|\ell(\mathbf{w}_T, z, z') - \ell(\mathbf{w}'_T, z, z')| \leq L\mathbb{E}[\delta_T], \tag{3.13}$$

where $\delta_T = \|\mathbf{w}_T - \mathbf{w}'_T\|$. By Lemma 3.1 and Theorem 3.1, we have $\mathbb{E}[\delta_t] \leq \mathbb{E}[\delta_{t-1}] + \frac{4L}{n} \cdot \alpha_{t-1}$. Unraveling the recursion yields

$$\mathbb{E}[\delta_T] \leq \frac{4L}{n} \sum_{t=1}^{T-1} \alpha_t. \tag{3.14}$$

Plugging this back into the equation (3.13), we obtained (3.11).

To prove (3.12), we notice that (3.14) holds true for any $T$, and therefore

$$\begin{aligned} \mathbb{E}\Big[|\ell(\bar{\mathbf{w}}_T, z, z') - \ell(\bar{\mathbf{w}}'_T, z, z')|\Big] &\leq L\mathbb{E}[\|\bar{\mathbf{w}}_T - \bar{\mathbf{w}}'_T\|] \\ \leq L \cdot \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\|\mathbf{w}_t - \mathbf{w}'_t\|] &= \frac{L}{T} \sum_{t=1}^{T} \mathbb{E}[\delta_t] \leq \frac{4L^2}{nT} \sum_{t=2}^{T} \sum_{j=1}^{t-1} \alpha_j, \end{aligned} \tag{3.15}$$

where we used the fact $\mathbf{w}_1 = \mathbf{w}'_1$. This completes the proof of the theorem. □

If we choose $\alpha_t = \frac{2}{\beta t^a}$ with $a \in (0,1)$, then Theorem 3.2 tells us that stability bounds of SGD for pairwise learning schemes are of order $O(\frac{T^{1-a}}{n})$. If the iteration of SGD is linear with respect to the size of the training data, e.g. $T = n$, SGD for pairwise learning will achieve stability and generalization error of order $O(\frac{1}{T^a})$. In this sense, faster training SGD will lead to reasonably good generalization.

### 3.3. *Strongly convex case*

If, furthermore, the function $\ell$ is strongly convex, we can establish stronger results.

**Theorem 3.3.** *Assume that the loss function $\ell(\cdot, z, z')$ is $\gamma$-strongly convex, $\beta$-smooth and $L$-Lipschitz for every example points $z$ and $z'$. Suppose that we run SGD with the constant step size $\alpha \leq \frac{2}{\beta+\gamma}$ for $T$ steps. Then, SGD satisfies uniform stability with*

$$\epsilon_{stab}(A^{last}, T, \ell, D, n) \leq \frac{8L^2}{\gamma n}\left[1 - (1 - \frac{\alpha\gamma}{2})^{T-1}\right]. \tag{3.16}$$

*and*

$$\epsilon_{stab}(A^{avg}, T, \ell, D, n) \leq \frac{8L^2}{\gamma T n}\sum_{t=2}^{T}\left[1 - (1 - \frac{\alpha\gamma}{2})^{t-1}\right]. \tag{3.17}$$

**Proof.** Fix a pair of examples $z$ and $z'$ and apply the boundedness of the gradient of $\ell(\cdot, z, z')$ to get

$$\mathbb{E}\Big[|\ell(\mathbf{w}_T; z, z') - \ell(\mathbf{w}'_T; z, z')|\Big] \leq L\mathbb{E}[\delta_T], \tag{3.18}$$

where $\delta_T = \|\mathbf{w}_T - \mathbf{w}'_T\|$. We then use the recursive relation between $\delta_t$ and $\delta_{t-1}$ as established in Theorem 3.1 to bound $\delta_T$. Since $\alpha \leq \frac{2}{\beta+\gamma}$ by assumption, we have $G_t$ is $\left(1 - \frac{\beta\gamma\alpha}{\beta+\gamma}\right)$-expansive by Lemma 3.1. Moreover we have $1 - \frac{\beta\gamma\alpha}{\beta+\gamma} \leq 1 - \frac{\alpha\gamma}{2}$ following from $\beta \geq \gamma$ by the definitions. As a result we have $G_t$ is $\left(1 - \frac{\alpha\gamma}{2}\right)$-expansive. Hence $\eta = 1 - \frac{\alpha\gamma}{2} \in (0,1)$. Then by Theorem 3.1, we have $\mathbb{E}[\delta_t] \leq \eta\mathbb{E}[\delta_{t-1}] + \frac{4L}{n} \cdot \alpha$. Unravel the recursion and we have

$$\mathbb{E}[\delta_T] \leq \frac{4L\alpha}{n}\sum_{j=0}^{T-2}\eta^j \leq \frac{8L}{\gamma n}(1 - \eta^{T-1}). \tag{3.19}$$

Plugging this back into the equation (3.18) yields (3.16).

To prove (3.17), notice that (3.19) holds true for any $T$. Consequently,

$$\begin{aligned}
&\mathbb{E}\Big[|\ell(\bar{\mathbf{w}}_T, z, z') - \ell(\bar{\mathbf{w}}'_T, z, z')|\Big] \leq L\mathbb{E}[\|\bar{\mathbf{w}}_T - \bar{\mathbf{w}}'_T\|] \\
&\leq L \cdot \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}'_t\|] = \frac{L}{T}\sum_{t=1}^{T}\mathbb{E}[\delta_t] \leq \frac{8L^2}{\gamma T n}\sum_{t=2}^{T}(1 - \eta^{t-1}),
\end{aligned}$$

where we used the fact that $\delta_1 = 0$. This completes the proof of the theorem. □

Theorem 3.3 indicates that, in the strongly convex case, although the uniform stability is also increasing w.r.t. $T$, it is upper bounded by a finite bound, i.e. $\frac{8L^2}{\gamma n}$ which is independent of the running time $T$.

Note that Theorem 3.3 only analyzes the uniform stability of SGD with constant step size which is not commonly used in practice. With the help of Lemma 3.4, we can establish the following theorem on the stability of a more popular form of SGD in which "staircase" decaying step sizes are chosen as in the machine learning and stochastic optimization fields [23,30].

**Theorem 3.4.** *Assume that the loss function $\ell(\cdot, z, z')$ is $\gamma$-strongly convex, $\beta$-smooth and $L$-Lipschitz for every example points $z$ and $z'$ and $\rho = \sup_{\mathbf{w},z,z'} \ell(\mathbf{w}, z, z')$. Let $\lceil \beta/\gamma \rceil$ be the smallest positive integer which is larger than or equals to $\beta/\gamma$. Suppose that we run SGD with the varying step sizes $\alpha_t = \frac{2}{\gamma t}$ for $t = 1, \ldots, T$ and $T \geq \lceil \beta/\gamma \rceil + 1$. Then,*

$$\epsilon_{stab}(A^{last}, \ell, D, n) \leq \frac{8L^2}{\gamma n}\left(1 - \frac{\lceil \beta/\gamma \rceil}{T-1}\right) + \frac{\rho}{n}\left(1 + \lceil \beta/\gamma \rceil\right).$$

**Proof.** It is easy to check that $\alpha_t \leq \frac{2}{\beta+\gamma}$ when $t \geq 1 + \frac{\beta}{\gamma}$. Thus if $t \geq t_0 := 1 + \lceil \frac{\beta}{\gamma} \rceil$, we have $G_t$ is $\eta_t$-expansive with $\eta_t = 1 - \frac{1}{t-1}$ by Lemma 3.1 and the fact $1 - \frac{\beta\gamma}{\beta+\gamma} \cdot \frac{2}{\gamma(t-1)} \leq 1 - \frac{1}{t-1}$. To this end, recalling Lemma 3.4, we have

$$\mathbb{E}\Big[|\ell(\mathbf{w}_T; z, z') - \ell(\mathbf{w}'_T; z, z')|\Big] \leq \frac{\rho}{n}\left(1 + \left\lceil \frac{\beta}{\gamma} \right\rceil\right) + L\mathbb{E}[\delta_T | \delta_{t_0} = 0]. \quad (3.20)$$

Next we will bound $\Delta_T := \mathbb{E}[\delta_T | \delta_{t_0} = 0]$. By Theorem 3.1, we have $\Delta_t \leq (1 - \frac{1}{t-1})\Delta_{t-1} + \frac{4L}{n} \cdot \alpha_{t-1}$ for $t_0 \leq t \leq T$. Unravel the recursion from $t = T$ to $t = t_0$ and we have $\Delta_T \leq \frac{8L}{n\gamma} \cdot \frac{T-t_0}{T-1}$. Plugging this back into the equation (3.20) yields the desired result. □

For the "staircase" decaying step sizes, it remains a question to us on how to get similar stability results when the output of SGD is the average of iterates, i.e. $A^{\text{avg}}(S)$.

### 3.4. *Non-convex case*

If $\ell(\cdot, z, z')$ is not convex such as in the case of MEE principle [19,20,32], we have the following result.

**Theorem 3.5.** *Assume that the loss function $\ell(\cdot, z, z') \in [0, 1]$ is $\beta$-smooth and $L$-Lipschitz for any $z$ and $z'$. Suppose that we run SGD for $T$ steps with the step sizes $\alpha_t \leq \frac{c}{t}$ where $c > 0$ is a scale parameter determined by the users in practice. Then, we have*

$$\epsilon_{stab}(A^{last}, \ell, D, n) \leq \frac{1 + 1/(\beta c)}{n-1}(4cL^2)^{\frac{1}{1+\beta c}}(T-1)^{\frac{\beta c}{1+\beta c}} + \frac{4cL^2}{n(T-1)} + \frac{1}{n}.$$

**Proof.** Firstly, by Lemma 3.4, we have for every $t_0 \in \{2, \cdots, n\}$,

$$\mathbb{E}\Big[|\ell(\mathbf{w}_T; z, z') - \ell(\mathbf{w}'_T; z, z')|\Big] \leq \frac{t_0}{n} + L\mathbb{E}[\delta_T | \delta_{t_0} = 0]. \tag{3.21}$$

Next, we will bound $\Delta_T := \mathbb{E}[\delta_T | \delta_{t_0} = 0]$ as a function of $t_0$ and then minimize for $t_0$. By Lemma 3.1 and a variant of Theorem 3.1 modified for conditional expectation, we have $\Delta_t \leq \Big(1 + (1-1/n)\frac{\beta c}{t-1}\Big)\Delta_{t-1} + \frac{4cL}{n(t-1)} \leq \exp\Big\{(1-1/n)\frac{\beta c}{t-1}\Big\}\Delta_{t-1} + \frac{4cL}{n(t-1)}$. Unwind this recurrence relation from $T$ down to $t_0 + 1$. This gives

$$\Delta_T \leq \frac{4cL}{n(T-1)} + \sum_{t=t_0}^{T-2}\prod_{s=t+1}^{T-1} \exp\Big\{(1-1/n)\frac{\beta c}{s}\Big\}\frac{4cL}{nt},$$

wherein the second term

$$\begin{aligned}
&\sum_{t=t_0}^{T-2}\prod_{s=t+1}^{T-1} \exp\Big\{(1-1/n)\frac{\beta c}{s}\Big\}\frac{4cL}{nt} \\
&= \frac{4cL}{n}\sum_{t=t_0}^{T-2}\Big\{\exp\Big[(1-1/n)c\beta\sum_{s=t+1}^{T-1}\frac{1}{s}\Big]\Big\}\frac{1}{t} \\
&\leq \frac{4cL}{n}\sum_{t=t_0}^{T-2}\Big\{\exp\Big[(1-1/n)c\beta\ln\Big(\frac{T-1}{t}\Big)\Big]\Big\}\frac{1}{t} \\
&\leq \frac{4cL}{n}(T-1)^{(1-1/n)c\beta}\sum_{t=t_0}^{T-2} t^{-(1-1/n)c\beta-1} \leq \frac{4L}{(n-1)\beta}\Big(\frac{T-1}{t_0-1}\Big)^{c\beta}.
\end{aligned}$$

Thus we have $\Delta_T \leq \frac{4cL}{n(T-1)} + \frac{4L}{(n-1)\beta}\left(\frac{T-1}{t_0-1}\right)^{c\beta}$. Plugging this bound into (3.21), we have

$$\mathbb{E}\Big[|\ell(\mathbf{w}_T; z, z') - \ell(\mathbf{w}'_T; z, z')|\Big] \leq \frac{t_0}{n} + \frac{4L^2}{(n-1)\beta}\left(\frac{T-1}{t_0-1}\right)^{c\beta} + \frac{4cL^2}{n(T-1)}.$$

The right hand side is approximately minimized when $t_0 = (4cL^2)^{\frac{1}{1+\beta c}}(T-1)^{\frac{\beta c}{1+\beta c}} + 1$. Thus we obtain

$$\mathbb{E}\Big[|\ell(\mathbf{w}_T; z, z') - \ell(\mathbf{w}'_T; z, z')|\Big] \leq \frac{1+1/(\beta c)}{n-1}(4cL^2)^{\frac{1}{1+\beta c}}(T-1)^{\frac{\beta c}{1+\beta c}} + \frac{4cL^2}{n(T-1)} + \frac{1}{n}$$

and we complete the proof. □

For the non-convex case, it also remains a question to us on how to get similar stability results when the output of SGD is the average of iterates, i.e., $A^{\text{avg}}(S)$. Note that Lemma 3.4 plays a key role in the stability analysis of SGD in the general non-convex case, where the gradient updates $G_t$ are no longer non-expansive operations in contrast to the convex case using Lemma 3.1.

We end Section 3 with a useful remark. The stability results above also hold true for the projected SGD algorithm defined by

$$\mathbf{w}_t = \Pi_\Omega\Big\{\mathbf{w}_{t-1} - \frac{\alpha_{t-1}}{t-1}\sum_{j=1}^{t-1}\nabla\ell(\mathbf{w}_{t-1}, z_{\xi_t}, z_{\xi_j})\Big\}, \tag{3.22}$$

where $\Omega$ is a bounded convex domain in $\mathbb{R}^d$, and $\Pi_\Omega$ is the projection operator defined by $\Pi_\Omega(\mathbf{u}) = \text{argmin}_{\mathbf{w}\in\Omega}\|\mathbf{u}-\mathbf{w}\|$. Typically, one can choose $\Omega$ to be a bounded ball with center zero, i.e. $\Omega = \{\mathbf{w} : \|\mathbf{w}\| \leq r_0\}$ for which the projection operator can be computed analytically. In this case, the projection onto a convex set is a non-expansive operation, i.e.

$$\delta_t = \|\Pi_\Omega(G_t(\mathbf{w}_{t-1}) - G'_t(\mathbf{w}'_{t-1}))\| \leq \|G_t(\mathbf{w}_{t-1}) - G'_t(\mathbf{w}'_{t-1})\|.$$

As a result, our previous proof techniques in the case of the original (non-projected) SGD algorithm defined by (2.4) can still be applied to this situation. Consequently, the stability results stated in the above theorems hold true for the projected SGD.

## 4. Trade-off between Stability and Optimization Error

In this section, we will start from the trade-off connection in Lemma 2.1 to establish the minimax lower bound for the excess expected risk. Then, we will combine this with the stability results in Section 3 to derive the lower bounds for the optimization error of SGD algorithms in the setting of pairwise learning.

### 4.1. *Minimax Lower Bounds*

In particular, let $\Omega$ be a bounded convex domain with finite diameter, i.e. $|\Omega| < \infty$. We consider the class $\mathcal{L}_c$ of convex and strongly smooth pairwise losses which is defined by

$$\mathcal{L}_{\text{c}} = \{\ell : \Omega \times \mathcal{Z} \times \mathcal{Z} \to \mathbb{R} \mid \ell \text{ is convex}, \beta - \text{smooth}; |\Omega| < \infty\},$$

and the class of strongly convex and smooth pairwise losses which is given by

$$\mathcal{L}_{\text{sc}} = \{\ell : \Omega \times \mathcal{Z} \times \mathcal{Z} \to \mathbb{R} \mid \ell \text{ is } \gamma - \text{strongly convex}, \beta - \text{smooth}; |\Omega| < \infty\}.$$

For the class $\mathcal{L}_c$ of pairwise loss functions, we have the following lower bound for the minimax risk.

**Theorem 4.1.** *There exists a pairwise loss $\ell \in \mathcal{L}_c$ such that*

$$\inf_{\widetilde{\mathbf{w}}_n}\sup_{D\in\mathcal{D}}\mathbb{E}_{S\sim D^n}[\Delta R(\widetilde{\mathbf{w}}_n(S))] \geq \frac{3\beta|\Omega|^2}{128\sqrt{6n}}. \tag{4.1}$$

The proof of Theorem 4.1 can be found in Appendix B which involves the Le Cam's method [27].

An immediate by-product result from the above theorem is the following corollary which states the lower bound for the excess expected risk when $\ell \in \mathcal{L}_c$.

**Corollary 4.1.** *There holds*

$$\inf_{\widetilde{\mathbf{w}}_n} \sup_{\ell\in\mathcal{L}_c, D\in\mathcal{D}} \mathbb{E}_S[\Delta R(\widetilde{\mathbf{w}}_n)] \geq \frac{3\beta|\Omega|^2}{128\sqrt{6n}}. \tag{4.2}$$

**Proof.** The result follows directly from Theorem 4.1 and the elementary inequality:

$$\inf_{\widetilde{\mathbf{w}}_n} \sup_{\ell\in\mathcal{L}_c, D\in\mathcal{D}} \mathbb{E}_S[\Delta R(\widetilde{\mathbf{w}}_n)] \geq \sup_{\ell\in\mathcal{L}_c} \inf_{\widetilde{\mathbf{w}}_n} \sup_{D\in\mathcal{D}} \mathbb{E}_S[\Delta R(\widetilde{\mathbf{w}}_n).$$ □

We now present the lower bound for the minimax risk for the class $\mathcal{L}_{\rm sc}$ of pairwise losses.

**Theorem 4.2.** *There exists a pairwise loss $\ell \in \mathcal{L}_{sc}$ such that*

$$\inf_{\widetilde{\mathbf{w}}_n} \sup_{D\in\mathcal{D}} \mathbb{E}_S[\Delta R(\widetilde{\mathbf{w}}_n)] \geq \frac{\beta|\Omega|^2}{32n}. \tag{4.3}$$

We postpone the proof of Theorem 4.2 to Appendix C. An immediate result is the following lower bound for the excess expected risk for $\ell \in \mathcal{L}_{\rm sc}$.

**Corollary 4.2.** *There holds*

$$\inf_{\widetilde{\mathbf{w}}_n} \sup_{\ell\in\mathcal{L}_{sc}, D\in\mathcal{D}} \mathbb{E}_S[\Delta R(\widetilde{\mathbf{w}}_n)] \geq \frac{\beta|\Omega|^2}{32n}. \tag{4.4}$$

**Proof.** The result follows directly from Theorem 4.2 and the elementary inequality:

$$\inf_{\widetilde{\mathbf{w}}_n} \sup_{\ell\in\mathcal{L}_{sc}, D\in\mathcal{D}} \mathbb{E}_S[\Delta R(\widetilde{\mathbf{w}}_n)] \geq \sup_{\ell\in\mathcal{L}_{sc}} \inf_{\widetilde{\mathbf{w}}_n} \sup_{D\in\mathcal{D}} \mathbb{E}_S[\Delta R(\widetilde{\mathbf{w}}_n)].$$ □

### 4.2. *Optimization Lower Bounds for SGD of Pairwise Learning*

In this subsection, we assume now that there exists an absolute constant $b > 0$ such that, for any loss $\ell \in \mathcal{L}$ where $\mathcal{L}$ can be $\mathcal{L}_{\rm c}$ or $\mathcal{L}_{\rm sc}$ for different settings in our consideration, there holds

$$\sup_{z,z'\in\mathcal{Z}} \min_{\mathbf{w}\in\Omega} \|\nabla\ell(\mathbf{w}, z, z')\| \leq b.$$

Under this condition, we can see that $\ell$ is $(|\Omega|\beta+b)$-Lipschitz. Indeed, for any fixed $z, z' \in \mathcal{Z}$, assume $\mathbf{w}_0 = \arg\min_{\mathbf{w}\in\Omega} \|\nabla\ell(\mathbf{w}, z, z')\|$. Then, by the $\beta$-smoothness of $\ell$, we have, for any $\mathbf{w} \in \Omega$, that $\|\nabla\ell(\mathbf{w}, z, z') - \nabla\ell(\mathbf{w}_0, z, z')\| \leq \beta\|\mathbf{w} - \mathbf{w}_0\| \leq \beta|\Omega|$. This indicates that $\|\nabla\ell(\mathbf{w}, z, z')\| \leq \|\nabla\ell(\mathbf{w}, z, z') - \nabla\ell(\mathbf{w}_0, z, z')\| + \|\nabla\ell(\mathbf{w}_0, z, z')\| \leq \beta|\Omega| + b$. Since $z, z'$ and $\mathbf{w}$ are arbitrary, it follows that $\ell$ is $(|\Omega|\beta + b)$-Lipschitz, i.e., $L = |\Omega|\beta + b$.

Combining the minimum lower bound in Theorem 4.1 with Lemma 2.1, one can derive the following lower bound for SGD of pairwise learning with smooth convex loss functions.

**Theorem 4.3.** *Consider the output $A^{avg}(S)$ of the projected SGD with step sizes $\alpha_t$ at iteration $T$ based on a pairwise loss $\ell \in \mathcal{L}_c$, and the following cases:*

*(1)* ***Constant step size:*** $\alpha_t \equiv \alpha = \frac{c}{T^a} \leq \frac{2}{\beta}$ *with* $a \in [0,1)$*;*
*(2)* ***Staircase decaying step sizes:*** $\alpha_t = \frac{c}{t^a}$ *with* $a \in (0,1)$ *and* $c \leq \frac{2}{\beta}$.

*Then, for either of the above cases, there exists a universal constant* $\widetilde{C}_1$*, and* $T_0$ *such that, for any* $T \geq T_0$*, there holds* $\mathcal{E}_{opt}^{avg}(T, \mathcal{L}_c, \mathcal{D}, n) \geq \frac{\widetilde{C}_1}{T^{1-a}}$.

**Proof.** 1) Putting $\alpha_t \equiv \frac{c}{T^a}$ back into (3.12) in Theorem 3.2 implies that

$$\mathcal{E}_{\text{stab}}^{avg}(T, \mathcal{L}_c, \mathcal{D}, n) \leq \frac{4L^2}{nT} \sum_{t=2}^{T} \sum_{j=1}^{t-1} \alpha_j \leq \frac{4cL^2}{nT^{1+a}} \sum_{t=2}^{T} (t-1) \leq \frac{4cL^2}{n} \cdot T^{1-a}. \tag{4.5}$$

Noting the relation (2.17) and applying Theorem 4.1, we have

$$\frac{8cL^2}{n} \cdot T^{1-a} + \mathcal{E}_{\text{opt}}^{avg}(T, \mathcal{L}_c, \mathcal{D}, n) \geq \frac{3\beta|\Omega|^2}{128\sqrt{6n}}.$$

It follows that

$$\mathcal{E}_{\text{opt}}^{avg}(T, \mathcal{L}_c, \mathcal{D}, n) \geq \frac{3\beta|\Omega|^2}{128\sqrt{6n}} - \frac{8cL^2}{n} \cdot T^{1-a} := Q(n).$$

Note that it is well known that the optimization error of the projected SGD is independent of the sample size $n$ (see the results in [23] for example). That means $\mathcal{E}_{\text{opt}}^{avg}(T, \mathcal{L}_c, \mathcal{D}, n)$ is actually not a function of $n$ although we include $n$ in its construction. As a result, we can take maximum of $Q(n)$ over $n$ so that the resulting lower bound is "best". To this end, letting $\tau_0 = \left[\frac{3\beta|\Omega|^2}{2048\sqrt{6}cL^2}\right]^{1/(1-a)}$ and $C_0 = \frac{3\beta^2|\Omega|^4}{1048576cL^2}$, we can rewrite $Q(n)$ as

$$Q(n) = \frac{C_0}{T^{1-a}} - 8cL^2T^{1-a}\left[\frac{1}{\sqrt{n}} - (\frac{\tau_0}{T})^{1-a}\right]^2.$$

Thus for sufficiently large $T \geq \tau_0$, we can always find an integer $n_0$ such that $\frac{2}{3}\left(\frac{T}{\tau_0}\right)^{1-a} \leq \sqrt{n_0} \leq 2\left(\frac{T}{\tau_0}\right)^{1-a}$. Let $C_1 = \frac{9\beta^2|\Omega|^4}{4194304cL^2}$. As a result, we have

$$Q(n_0) \geq \frac{C_0}{T^{1-a}} - 2cL^2T^{1-a}\left(\frac{\tau_0}{T}\right)^{2(1-a)} = \frac{C_1}{T^{1-a}}.$$

Thus we obtain, for any $T \geq \tau_0$

$$\mathcal{E}_{\text{opt}}^{\text{avg}}(T, \mathcal{L}_c, \mathcal{D}, n_0) \geq \frac{C_1}{T^{1-a}}.$$

Since $\mathcal{E}_{\text{opt}}^{\text{avg}}(T, \mathcal{L}_c, \mathcal{D}, n)$ is independent of $n$, we establish the desired result.

2) Plug $\alpha_t = \frac{c}{t^a}$ into (3.12) and let $c' = c/(1-a)$. We have

$$\mathcal{E}_{\text{stab}}^{avg}(T, \mathcal{L}_c, \mathcal{D}, n) \leq \frac{4L^2}{Tn} \sum_{t=2}^{T} \sum_{j=1}^{t-1} \frac{c}{j^a} \leq \frac{4cL^2}{(1-a)n} \sum_{t=2}^{T} \frac{t^{1-a}}{T} \leq \frac{4c'L^2}{n} \cdot T^{1-a}. \tag{4.6}$$

Recall the first equation namely (4.5) in the proof of the first case. We can find that the only difference between these two stability results, namely (4.6) and (4.5),

comes from we replacing $c$ by $c'$. Likewise denote $\tau_0' = \left[\frac{3\beta|\Omega|^2}{2048\sqrt{6}c'L^2}\right]^{1/(1-a)}$ and $C_0' = \frac{3\beta^2|\Omega|^4}{1048576c'L^2}$. Thus for sufficiently large $T \geq \tau_0'$, we can always find an integer $n_0'$ s.t. $\frac{2}{3}\left(\frac{T}{\tau_0'}\right)^{1-a} \leq \sqrt{n_0'} \leq 2\left(\frac{T}{\tau_0'}\right)^{1-a}$. Let $C_1' = \frac{9\beta^2|\Omega|^4}{4194304c'L^2}$. It is natural to use the same strategy to obtain almost the same lower bound for the optimization error as the case of constant step size, viz.,

$$\mathcal{E}_{\text{opt}}^{avg}\left(T, \mathcal{L}_c, \mathcal{D}, n_0'\right) \geq \frac{C_1'}{T^{1-a}}.$$

Again, as $\mathcal{E}_{\text{opt}}^{\text{avg}}(T, \mathcal{L}_c, \mathcal{D}, n)$ is independent of $n$, the desired result is proved. □

The work of Wang et al. [41] considered the regret bound for projected online gradient descent algorithm in pairwise learning which is exactly SGD algorithm we consider here in the stochastic setting. Specifically, in [41, Theorem 13], they gave the regret rate of $O(\sqrt{T})$ of the projected online gradient descent algorithm with varying step sizes $\alpha_t = O\left(\frac{1}{\sqrt{t}}\right)$ for pairwise learning. While in [23, Theorem 3], they obtained the online to batch conversion bound. Combining the above results, we can obtain an upper bound of the convergence rate, i.e., $O\left(\frac{1}{\sqrt{T}}\right)$ (up to a $\log T$ factor). This result meets the lower bound we have established in Theorem 4.3 which says this algorithm can not have better worst-case convergence rate than $O\left(\frac{1}{\sqrt{T}}\right)$ with step sizes $\alpha_t = O\left(\frac{1}{\sqrt{t}}\right)$ in the general convex smooth case. Thus our results confirm its optimality up to a logarithmic factor.

From Theorem 4.2, we can get the following two theorems about the lower bounds for the optimization error of SGD with fixed step size and varying step sizes respectively in the setting of smooth strongly convex loss functions.

**Theorem 4.4.** *Let the projected SGD with fixed step size $\alpha_t \equiv \alpha \leq \frac{2}{\beta+\gamma}$ for $T$ iterations to get an output $A^{avg}(S)$ based on a pairwise loss $\ell \in \mathcal{L}_{sc}$. Then we can get the following results, viz.,*

$$\mathcal{E}_{opt}^{avg}(T, \mathcal{L}_{sc}, \mathcal{D}, n) \geq \frac{16(|\Omega|\beta + b)^2}{\gamma n}\left(1 - \frac{\alpha\gamma}{2}\right)^{T-1} - C,$$

*wherein the offset $C = \frac{1}{n}\left(\frac{16(|\Omega|\beta+b)^2}{\gamma} - \frac{\beta|\Omega|^2}{32}\right) > 0$.*

**Proof.** Recall (3.17). We have

$$\begin{aligned}
&\mathbb{E}|\ell(\bar{\mathbf{w}}_T, z, z') - \ell(\bar{\mathbf{w}}_T', z, z')| \leq L\mathbb{E}[\|\bar{\mathbf{w}}_T - \bar{\mathbf{w}}_T'\|] \\
&\leq \frac{8L^2}{\gamma n}\frac{1}{T}\sum_{t=2}^{T}(1 - \eta^{t-1}) \leq \frac{8L^2}{\gamma n}\left[1 - \eta^{T-1}\right]. \qquad (4.7)
\end{aligned}$$

Thus we have

$$\mathcal{E}_{\text{stab}}^{\text{avg}}(T, \mathcal{L}_{sc}, \mathcal{D}, n) \leq \frac{8L^2}{\gamma n}\left[1 - (1 - \frac{\alpha\gamma}{2})^{T-1}\right].$$

Noting the relation (2.18) and applying Theorem 4.2, we have

$$\frac{16L^2}{\gamma n}\left[1-(1-\frac{\alpha\gamma}{2})^{T-1}\right]+\mathcal{E}^{\text{avg}}_{\text{opt}}(T,\mathcal{L}_{sc},\mathcal{D},n)\geq\frac{\beta|\Omega|^2}{32n}.$$

It follows that

$$\mathcal{E}^{\text{avg}}_{\text{opt}}(T,\mathcal{L}_{sc},\mathcal{D},n)\geq\frac{\beta|\Omega|^2}{32n}-\frac{16L^2}{\gamma n}\left[1-(1-\frac{\alpha\gamma}{2})^{T-1}\right].$$

Recall $L=|\Omega|\beta+b$ and we have finished the proof. □

**Theorem 4.5.** *Let the projected SGD with step sizes $\alpha_t$ for $T$ iterations to get an averaged output $A^{avg}(S)$ based on a pairwise loss $\ell\in\mathcal{L}_{sc}$. Let $\alpha_t=\frac{2}{\gamma t}$. Denote $C:=\frac{2(\beta|\Omega|^2+b|\Omega|)}{n}\cdot\left(\frac{\beta}{\gamma}+3\right)-\frac{\beta|\Omega|^2}{32n}+\frac{16(|\Omega|\beta+b)^2}{n\gamma}\cdot\ln\left(\frac{\beta}{\gamma}+3\right)$. Then,*

$$\mathcal{E}^{avg}_{opt}(T,\mathcal{L}_{sc},\mathcal{D},n)\geq\frac{16L^2(\beta+\gamma)}{\gamma^2 n}\cdot\frac{\ln T}{T}-C.$$

**Proof.** It is easy to check that $\alpha_t\leq\frac{2}{\beta+\gamma}$ when $t\geq 1+\frac{\beta}{\gamma}$. Let $\lceil\beta/\gamma\rceil$ be the smallest positive integer which is larger than or equals to $\beta/\gamma$. Thus if $t\geq t_0:=2+\lceil\frac{\beta}{\gamma}\rceil$, we have $G_t$ is $\eta_t$-expansive with $\eta_t=1-\frac{1}{t-1}$ due to Lemma 3.1 and the fact $1-\frac{\beta\gamma}{\beta+\gamma}\cdot\frac{2}{\gamma(t-1)}\leq 1-\frac{1}{t-1}$.

Let $\delta_t=\|\mathbf{w}_t-\mathbf{w}'_t\|$ and $t_1^\star$ be the first time that the SGD algorithms encounter the different examples. By the conditional expectation formula, we have

$$\begin{aligned}\mathbb{E}[\delta_t]&=\mathbb{P}\{t_1^\star\leq t_0\}\mathbb{E}[\delta_t\big|t_1^\star\leq t_0]+\mathbb{P}\{t_1^\star>t_0\}\mathbb{E}[\delta_t\big|t_1^\star>t_0]\\&=\frac{t_0}{n}\cdot\mathbb{E}[\delta_t\big|t_1^\star\leq t_0]+\left(1-\frac{t_0}{n}\right)\cdot\mathbb{E}[\delta_t\big|t_1^\star>t_0].\end{aligned}$$

If $t<t_0$, we have $\mathbb{E}[\delta_t\big|t_1^\star>t_0]=0$ as the SGD algorithms have not encountered the different examples during the first $t$ steps. Thus when $t<t_0$ we have

$$\mathbb{E}[\delta_t]=\frac{t_0}{n}\cdot\mathbb{E}[\delta_t\big|t_1^\star\leq t_0]\leq\frac{t_0}{n}\cdot|\Omega|. \tag{4.8}$$

If $t\geq t_0$, we have

$$\mathbb{E}[\delta_t]\leq\frac{t_0}{n}\cdot|\Omega|+\mathbb{E}[\delta_t\big|t_1^\star>t_0].$$

Denote $\Delta_t:=\mathbb{E}[\delta_t|t_1^\star>t_0]$. Recall $G_t$ is $\eta_t$-expansive with $\eta_t=1-\frac{1}{t-1}$. By Theorem 3.1, we have $\Delta_t\leq(1-\frac{1}{t-1})\Delta_{t-1}+\frac{4L}{n}\cdot\alpha_{t-1}$ for $t\geq t_0$. Unravel the recursion from $t$ to $t_0$ and we have $\Delta_t\leq\frac{8L}{n\gamma}\cdot\frac{t-t_0}{t-1}$. Thus when $t\geq t_0$, we have

$$\mathbb{E}[\delta_t]\leq\frac{t_0}{n}\cdot|\Omega|+\frac{8L}{n\gamma}\cdot\frac{t-t_0}{t-1}. \tag{4.9}$$

Combining (4.8) and (4.9), for $t\geq 2$, we have

$$\mathbb{E}[\delta_t]\leq\frac{t_0}{n}\cdot|\Omega|+\frac{8L}{n\gamma}\cdot\frac{(t-t_0)_+}{t-1}, \tag{4.10}$$

where $(t-t_0)_+ = \max(0, t-t_0)$.

Let $\bar{\mathbf{w}}_T = \frac{1}{T}\sum_{t=1}^{T}\mathbf{w}_t$. Using the Lipschitz condition of $\ell(\cdot; z, z')$, we further have

$$\begin{aligned}
&\mathbb{E}\Big[|\ell(\bar{\mathbf{w}}_T, z, z') - \ell(\bar{\mathbf{w}}'_T, z, z')|\Big] \leq L\mathbb{E}[\|\bar{\mathbf{w}}_T - \bar{\mathbf{w}}'_T\|] \\
&\leq L \cdot \frac{1}{T}\sum_{t=2}^{T}\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}'_t\|] = L \cdot \frac{1}{T}\sum_{t=2}^{T}\mathbb{E}[\delta_t] \\
&\leq \frac{t_0}{n} \cdot L|\Omega| + \frac{8L^2}{n\gamma} \cdot \frac{1}{T}\sum_{t=2}^{T}\frac{(t-t_0)_+}{t-1}, \qquad (4.11)
\end{aligned}$$

wherein the last inequality comes from (4.10). Next we will bound $\sum_{t=2}^{T}\frac{(t-t_0)_+}{t-1}$. Actually we can write

$$\begin{aligned}
&\sum_{t=2}^{T}\frac{(t-t_0)_+}{t-1} = \sum_{t=t_0+1}^{T}\frac{t-t_0}{t-1} = \sum_{t=t_0+1}^{T}\frac{t-1+1-t_0}{t-1} = T - t_0 - \sum_{t=t_0+1}^{T}\frac{t_0-1}{t-1} \\
&\leq T - t_0 - \int_{t=t_0+1}^{T+1}\frac{t_0-1}{t-1}dt = T - t_0 - (t_0-1)(\ln T - \ln t_0) \\
&= T - t_0 + (t_0-1)\cdot\ln t_0 - (t_0-1)\cdot\ln T \\
&\leq (T-1)\cdot\ln t_0 - (t_0-1)\cdot\ln T, \qquad (4.12)
\end{aligned}$$

where the last inequality comes from the fact $\ln t_0 \geq 1$ as $t_0 := 2 + \lceil\frac{\beta}{\gamma}\rceil \geq 3$. Substituting (4.12) into (4.11), we have

$$\mathbb{E}[|\ell(\bar{\mathbf{w}}_T, z, z') - \ell(\bar{\mathbf{w}}'_T, z, z')|] \leq \frac{t_0}{n}\cdot L|\Omega| + \frac{8L^2}{n\gamma}\cdot\ln t_0 - \frac{8L^2}{n\gamma}\cdot(t_0-1)\cdot\frac{\ln T}{T}.$$

Recall $t_0 = 2 + \lceil\frac{\beta}{\gamma}\rceil$. Thus $2 + \frac{\beta}{\gamma} \leq t_0 \leq 3 + \frac{\beta}{\gamma}$. As a result, we have

$$\mathcal{E}^{\text{avg}}_{\text{stab}}(T, \mathcal{L}_{sc}, \mathcal{D}, n) \leq \frac{L|\Omega|}{n}\cdot\left(\frac{\beta}{\gamma}+3\right) + \frac{8L^2}{n\gamma}\cdot\ln\left(\frac{\beta}{\gamma}+3\right) - \frac{8L^2}{n\gamma}\cdot\left(\frac{\beta}{\gamma}+1\right)\cdot\frac{\ln T}{T}.$$

Noting the relation (2.18) and applying Theorem 4.2, we have

$$2\mathcal{E}^{\text{avg}}_{\text{stab}}(T, \mathcal{L}_{sc}, \mathcal{D}, n) + \mathcal{E}^{\text{avg}}_{\text{opt}}(T, \mathcal{L}_{sc}, \mathcal{D}, n) \geq \frac{\beta|\Omega|^2}{32n}.$$

It follows that

$$\begin{aligned}
&\mathcal{E}^{\text{avg}}_{\text{opt}}(T, \mathcal{L}_{sc}, \mathcal{D}, n) \\
&\geq \frac{\beta|\Omega|^2}{32n} - \frac{2L|\Omega|}{n}\cdot\left(\frac{\beta}{\gamma}+3\right) - \frac{16L^2}{n\gamma}\cdot\ln\left(\frac{\beta}{\gamma}+3\right) + \frac{16L^2}{n\gamma}\cdot\left(\frac{\beta}{\gamma}+1\right)\cdot\frac{\ln T}{T} \\
&\geq \frac{16L^2}{n\gamma}\cdot\left(\frac{\beta}{\gamma}+1\right)\cdot\frac{\ln T}{T} - \left\{\frac{2L|\Omega|}{n}\cdot\left(\frac{\beta}{\gamma}+3\right) - \frac{\beta|\Omega|^2}{32n} + \frac{16L^2}{n\gamma}\cdot\ln\left(\frac{\beta}{\gamma}+3\right)\right\}.
\end{aligned}$$

Recall that $L = |\Omega|\beta + b$ and $C = \frac{2L|\Omega|}{n}\cdot\left(\frac{\beta}{\gamma}+3\right) - \frac{\beta|\Omega|^2}{32n} + \frac{16L^2}{n\gamma}\cdot\ln\left(\frac{\beta}{\gamma}+3\right)$. We have obtained the desired lower bound. □

To illustrate the practical value of Theorem 4.5, we recall the work of [23]. In [23, Theorem 5], they established the first fast convergence rate for averaged outputs of online gradient descent algorithm for strongly convex loss functions. Following a variant of [50, Theorem 1] in which we choose the step sizes $\alpha_t = O\Big(\frac{1}{t}\Big)$, we can get a regret bound of $\log(T)$ for the projected online gradient descent algorithm. Combine these two results and we obtain an upper bound of the optimization error, i.e., $O\Big(\frac{\log T}{T}\Big)$. However, our theory can only obtain a matching lower bound with an undesirable offset $C$.

## 5. Examples

In this section, we illustrate the stability results obtained in Section 3 using three specific examples, namely, AUC maximization, metric learning and MEE. In the following examples, the model parameter $\mathbf{w}$ is assumed to be in $\Omega = \{\mathbf{w} : \|\mathbf{w}\| \leq r_0\}$. In addition, we assume $\|x\| \leq B_1$ and $|y| \leq B_2$.

In the following, $\epsilon_{stab}(A, T, \ell, D, n)$ means the stability parameter for both the last output of SGD and the average of its iterates.

### 5.1. *AUC Maximization*

Area under ROC (AUC) is a metric which is widely used for measuring the classification performance for imbalanced data [6,14,17]. The AUC score of a scoring funciton is the probability of a random positive example ranked higher than a random negative example [17,10]. Here we consider a population version of the regularization framework for AUC maximization in [45]:

$$\min_{\mathbf{w}} R(\mathbf{w}) := \mathbb{E}[\ell(\mathbf{w}, z, z')], \tag{5.1}$$

where $\ell(\mathbf{w}, z, z') = (1-(x-x')^\top \mathbf{w})^2 \mathbb{I}_{\{y=1 \wedge y'=-1\}} + (\mu/2)\|\mathbf{w}\|^2$. Note that an optimal solution $\mathbf{w}^\star$ for $R(\mathbf{w})$ must lie in a ball about 0 with the radius $r_0 = \sqrt{2/\mu}$ since $(\mu/2)\|\mathbf{w}^\star\|^2 \leq R(\mathbf{w}^\star) \leq R(0) \leq 1$. Hence one can let $\mathbf{w}$ in (5.1) satisfying $\|\mathbf{w}\| \leq r_0$. As an application of Theorem 3.3, we have

**Corollary 5.1.** *For the AUC maximization problem (5.1), the loss function $\ell(\cdot; z, z')$ is $\mu$-strongly convex, $(4B_1 + 8B_1^2\sqrt{2/\mu} + \sqrt{2\mu})$-Lipschitz and $(8B_1^2 + \mu)$-smooth for every example points $z$ and $z'$. The projected SGD with the constant step size $\alpha \leq 1/(4B_1^2 + \mu)$ has the stability*

$$\epsilon_{stab}(A, T, \ell, D, n) \leq \frac{8(4B_1 + 8B_1^2\sqrt{2/\mu} + \sqrt{2\mu})^2}{n\mu} \Big[1 - (1 - \frac{\alpha\mu}{2})^{T-1}\Big].$$

**Proof.** Since $\ell(\mathbf{w}; z, z') = (1 - (x - x')^\top \mathbf{w})^2 \mathbb{I}_{\{y=1 \wedge y'=-1\}} + (\mu/2)\|\mathbf{w}\|^2$, it is easy to check that $\ell(\mathbf{w}; z, z')$ is $(4B_1 + 8B_1^2 r_0 + \mu r_0)$-Lipschitz and $(8B_1^2 + \mu)$-smooth for every example points $z$ and $z'$. Note that $r_0 = \sqrt{2/\mu}$. Then we finish the proof by substituting these constants into Theorem 3.3. □

For the case of varying step sizes, applying Theorem 3.4, we have

**Corollary 5.2.** *For the AUC maximization problem (5.1), the loss function $\ell(\cdot; z, z')$ is $\mu$-strongly convex, $(4B_1 + 8B_1^2\sqrt{2/\mu} + \sqrt{2\mu})$-Lipschitz and $(8B_1^2 + \mu)$-smooth for every example points $z$ and $z'$. The projected SGD with the varying step size $\alpha_t = \frac{2}{\gamma t}$ has the stability*

$$\epsilon_{stab}(A^{last}, T, \ell, D, n) \leq \frac{8(4B_1 + 8B_1^2\sqrt{2/\mu} + \sqrt{2\mu})^2}{n\mu}\left(1 - \frac{1 + \lceil 8B_1^2/\mu \rceil}{T-1}\right) + \frac{\rho}{n}\left(2 + \lceil 8B_1^2/\mu \rceil\right),$$

*wherein $\rho = 1 + (1 + 2B_1\sqrt{2/\mu})^2$.*

**Proof.** We just need to show that $\sup_{\mathbf{w},z,z'} \ell(\mathbf{w}, z, z') = 1 + (1 + 2B_1\sqrt{2/\mu})^2$. Since $\ell(\mathbf{w}; z, z') = (1 - (x - x')^\top \mathbf{w})^2 \mathbb{I}_{\{y=1 \wedge y'=-1\}} + (\mu/2)\|\mathbf{w}\|^2$ and $\|x\| \leq B_1$, $\|\mathbf{w}\| \leq r_0$ by assumption, it is direct to find that $\sup_{\mathbf{w},z,z'} \ell(\mathbf{w}, z, z') \leq \frac{\mu r_0^2}{2} + (1 + 2B_1 r_0)^2$. Recall $r_0 = \sqrt{\frac{2}{\mu}}$. Thus we obtain $\rho = \sup_{\mathbf{w},z,z'} \ell(\mathbf{w}, z, z') = 1 + (1 + 2B_1\sqrt{2/\mu})^2$.□

### 5.2. *Metric Learning*

In supervised metric learning, the distance between two examples $x$ and $x'$ w.r.t $M \in \mathbb{S}_+^d$ is defined by $\|x - x'\|_M^2 = (x - x')^\top M(x - x')$, where $\mathbb{S}_+^d$ denotes the cone of all $d \times d$ p.s.d. matrices. For every pair of examples with labels $(x, y)$ and $(x', y')$, denote $\mathcal{I}_{yy'} = 1$ if $y = y'$, otherwise $\mathcal{I}_{yy'} = -1$. Using the following logistic loss (e.g. [16]), the ERM formulation for metric learning can be written as

$$\min_{M \in \Omega} \sum_{i=1}^n \sum_{j=1}^n \log\left[1 + \exp\left(\mathcal{I}_{y_i y_j}(\|x_i - x_j\|_M^2)\right)\right], \tag{5.2}$$

where $\Omega := \{M \in \mathbb{S}_+^d : \|M\|_F \leq r_0\}$ with $\|\cdot\|_F$ denoting the Frobenius norm of matrix. Its population risk can be expressed as $\mathbb{E}\left[\log\left(1 + \exp\left(\mathcal{I}_{yy'}(\|x - x'\|_M^2)\right)\right)\right]$.

For any matrices $A$ and $B$, let $\langle A, B\rangle_{\text{tr}} = \text{trace}(A^\top B)$. In this case, the model parameter $\mathbf{w} = M$ and

$$\ell(\mathbf{w}, z, z') = \log\left[1 + \exp\left\{\mathcal{I}_{yy'}\langle \mathbf{w}, (x - x')(x - x')^\top\rangle_{\text{tr}}\right\}\right].$$

By Theorem 3.2, we have the following result.

**Corollary 5.3.** *For the metric learning problem (5.2), the loss function $\ell(\cdot; z, z')$ is $(4B_1^4)$-smooth, convex and $(4B_1^2)$-Lipschitz for every example points $z$ and $z'$. The projected SGD with the step sizes $\alpha_t \leq 1/(2B_1^4)$ has the stability*

$$\epsilon_{stab}(A, T, \ell, D, n) \leq \frac{64B_1^4}{n}\sum_{t=1}^{T-1} \alpha_t,$$

*where $T$ is the number of updates.*

**Proof.** We first give one claim which is easy to be verified. Rewrite $\ell(\mathbf{w}; z, z') = g_1(g_2(\mathbf{w}))$, where $g_1$ is $L_1-$Lipschitz, $\beta_1-$smooth and $g_2$ is $L_2-$Lipschitz, $\beta_2-$smooth. Then, $\ell(\mathbf{w}; z, z')$ is $(L_1L_2)-$Lipschitz and $(L_1\beta_2 + L_2^2\beta_1)-$smooth.

Rewrite $\ell(\mathbf{w}; z, z') = g_1(g_2(\mathbf{w}))$, where

$$\begin{cases} g_1(u) &= \log\{1 + \exp(u)\}, \\ u &= g_2(\mathbf{w}), \\ g_2(\mathbf{w}) &= \mathcal{I}_{yy'}\langle \mathbf{w}, (x - x')(x - x')^\top\rangle_{\mathrm{tr}}. \end{cases}$$

We have $g_1$ is $1-$Lipschitz, $1/4-$smooth and $g_2$ is $(4B_1^2)$-Lipschitz, $0-$smooth as $\nabla g_2(\mathbf{w}) = \mathcal{I}_{yy'}(x - x')(x - x')^\top$. Thus we have $L = 4B_1^2$ and $\beta = 4B_1^4$. Substituting these constants into Theorem 3.2 we have proved the Corollary 5.3. □

### 5.3. *Minimum Error Entropy Principle*

For simplification, we concentrate on a simple linear regression case of the general framework of MEE principle in [20,19,32], i.e.,

$$\min_{\|\mathbf{w}\| \le r_0} R(\mathbf{w}) := \mathbb{E}[\ell(\mathbf{w}, z, z')], \tag{5.3}$$

where the loss

$$\ell(\mathbf{w}, z, z') = 1 - \exp\left(-\frac{((y - y') - (x - x')^\top \mathbf{w})^2}{2h^2}\right)$$

with a scaling parameter $h > 0$. It is obvious that the loss function is non-convex.

Notice that $\ell(\mathbf{w}, z, z')$ is negative and bounded with $\sup_{\mathbf{w},z,z'} \ell(\mathbf{w}, z, z') = 1$. Then we can use Theorem 3.5 to give a uniform stability of the projected SGD for MEE in the following corollary.

**Corollary 5.4.** *For MEE problem (5.3), the loss function $\ell(\cdot; z, z') \in [0, 1)$ is $L$-Lipschitz and $\beta$-smooth for every example points $z$ and $z'$ with the following constants*

$$\begin{cases} L = \frac{4}{h^2} \cdot (B_1^2 r_0 + B_1 B_2), \\ \beta = \frac{4}{h^2} \cdot B_1^2 + \frac{16}{h^4} \cdot (B_1^2 r_0) + B_1 B_2)^2. \end{cases}$$

*The projected SGD with step sizes $\alpha_t \le \frac{c}{t}$ for somec $> 0$ satisfies the uniform stability with*

$$\epsilon_{stab}(A^{last}, T, \ell, D, n) \le \frac{1 + 1/(\beta c)}{n - 1} \cdot (4cL^2)^{\frac{1}{1+\beta c}} (T - 1)^{\frac{\beta c}{1+\beta c}} + \frac{4cL^2}{n(T - 1)} + \frac{1}{n},$$

*where $T$ is the number of updates.*

**Proof.** We now calculate $L$ and $\beta$ of the loss $\ell$ in the MEE problem (5.3). Rewrite $\ell(\mathbf{w}; z, z') = g_1(g_2(\mathbf{w}))$, where $g_1(u) = 1 - \exp\{-\frac{u^2}{2h^2}\}$, $u = g_2(\mathbf{w})$ and $g_2(\mathbf{w}) = (x - x')^\top \mathbf{w} - (y - y')$. Assume $g_1$ is $L_1-$Lipschitz, $\beta_1-$smooth and $g_2$ is $L_2-$Lipschitz, $\beta_2-$smooth. Thus we have

$$\begin{cases} L_1 = \frac{2}{h^2} \cdot (B_1 r_0 + B_2),\ L_2 = 2B_1, \\ \beta_1 = \frac{1}{h^2} + \frac{4}{h^4} \cdot (B_1 r_0 + B_2)^2,\ \beta_2 = 0. \end{cases}$$

And recalling the simple claim at the beginning of the proof of Corollary 5.3, we have

$$\begin{cases} L = \frac{4}{h^2} \cdot (B_1^2 r_0 + B_1 B_2), \\ \beta = \frac{4}{h^2} \cdot B_1^2 + \frac{16}{h^4} \cdot (B_1^2 r_0 + B_1 B_2)^2. \end{cases}$$ □

## 6. Conclusion

In this paper we establish the stability and its trade-off with optimization error of SGD algorithms for pairwise learning. Stability results of SGD hold true for both convex and non-convex cases. The trade-off results are established by deriving the lower bound for the minimax statistical error from which lower bounds for the convergence rate of SGD can be obtained for the cases of smooth convex and strongly convex losses. Examples are given to illustrate our main results in specific pairwise learning tasks such as AUC maximization, metric learning and MEE principle.

There are several directions for future work. Firstly, the stability results we established are not data-dependent. It would be nice to obtain data-dependent bounds related to the curvature of the loss function and the geometry of the training data. Secondly, the lower bounds for optimization error of SGD have an undesired bias term in Theorems 4.4 and 4.5. We do not know how to get rid of this term. Thirdly, the stability and generalization bounds here can not explain why SGD iterates converge to a good local minimum for the non-convex case of MEE. It was shown in [20] that the iterates of SGD for pairwise learning converge to the target function for large enough $h$. However, it remains an open question how to establish similar results for a general scaling parameter $h$. Finally, generalization bounds and stability results are obtained in expectation. It is unclear to us how to derive the bounds with high probability.

## Acknowledgments

This work was done when Wei Shen was a visiting student at SUNY Albany. The corresponding author is Yiming Ying, whose work is supported by the National Science Foundation (NSF) under Grant No #1816227. The work of Xiaoming Yuan is supported by the General Research Fund from the Hong Kong Research Grants Council, 12302318.

## Appendix A. Proof of Lemma 3.1

Let $\ell_{M_t}(\mathbf{w}) = \frac{1}{t-1}\sum_{j=1}^{t-1} \ell(\mathbf{w}, z_{\xi_t}, z_{\xi_j})$ wherein $M_t = \{z_{\xi_1}, \cdots, z_{\xi_t}\}$. We can simplify the equation of $G_t$ as $G_t(\mathbf{w}_{t-1}) = \mathbf{w}_{t-1} - \alpha_{t-1}\nabla_{\mathbf{w}}\ell_{M_t}(\mathbf{w}_{t-1})$. It is obvious that $\ell_{M_t}(\mathbf{w})$ has the same properties of convexity and smoothness with $\ell(\mathbf{w}; z_{\xi_t}, z_{\xi_j})$. Then we prove the three claims in Lemma 3.1.

1. If $\ell$ is $\beta$-smooth, then $\ell_{M_t}(\mathbf{w})$ is also $\beta$-smooth. By the triangle inequality and

the $\beta$-smoothness of $\ell_{M_t}$,

$$\|G_t(\mathbf{w}') - G_t(\mathbf{w})\| \leq \|\mathbf{w}' - \mathbf{w}\| + \alpha_{t-1}\|\nabla_{\mathbf{w}}\ell_{M_t}(\mathbf{w}') - \nabla_{\mathbf{w}}\ell_{M_t}(\mathbf{w})\|$$
$$\leq \|\mathbf{w}' - \mathbf{w}\| + \alpha_{t-1}\beta\|\mathbf{w}' - \mathbf{w}\| = (1 + \alpha_{t-1}\beta)\|\mathbf{w}' - \mathbf{w}\|.$$

2. We have

$$\|G_t(\mathbf{w}') - G_t(\mathbf{w})\|^2 = \|(\mathbf{w}' - \mathbf{w}) - \alpha_{t-1}(\nabla_{\mathbf{w}}\ell_{M_t}(\mathbf{w}') - \nabla_{\mathbf{w}}\ell_{M_t}(\mathbf{w}))\|^2$$
$$= \|\mathbf{w}' - \mathbf{w}\|^2 + \alpha_{t-1}^2\|\nabla_{\mathbf{w}}\ell_{M_t}(\mathbf{w}') - \nabla_{\mathbf{w}}\ell_{M_t}(\mathbf{w})\|^2$$
$$-2\alpha_{t-1}\langle\nabla_{\mathbf{w}}\ell_{M_t}(\mathbf{w}') - \nabla_{\mathbf{w}}\ell_{M_t}(\mathbf{w}), \mathbf{w}' - \mathbf{w}\rangle$$
$$\leq \|\mathbf{w}' - \mathbf{w}\|^2 - \Big(\frac{2\alpha_{t-1}}{\beta} - \alpha_{t-1}^2\Big)\|\nabla_{\mathbf{w}}\ell_{M_t}(\mathbf{w}') - \nabla_{\mathbf{w}}\ell_{M_t}(\mathbf{w})\|^2$$
$$\leq \|\mathbf{w}' - \mathbf{w}\|^2, \tag{A.1}$$

wherein the first inequality follows from the $\frac{1}{\beta}$-co-coerciveness of $\nabla_{\mathbf{w}}\ell_{M_t}(\cdot)$, namely

$$\langle\nabla_{\mathbf{w}}\ell_{M_t}(\mathbf{w}') - \nabla_{\mathbf{w}}\ell_{M_t}(\mathbf{w}), \mathbf{w}' - \mathbf{w}\rangle \geq \frac{1}{\beta}\|\nabla_{\mathbf{w}}\ell_{M_t}(\mathbf{w}') - \nabla_{\mathbf{w}}\ell_{M_t}(\mathbf{w})\|^2,$$

since $\ell_{M_t}$ is both convex and $\beta$-smooth from our assumptions of $\ell$. The last inequality in (A.1) holds because we assume $\alpha_{t-1} \leq \frac{2}{\beta}$.

3. We have $\phi(\mathbf{w}) = \ell_{M_t}(\mathbf{w}) - \frac{\gamma}{2}\|\mathbf{w}\|^2$ is convex and $(\beta - \gamma)$-smooth, which implies the gradient of $\phi$ is $\Big(\frac{1}{\beta-\gamma}\Big)$-co-coercive. Thus

$$\langle\nabla_{\mathbf{w}}\ell_{M_t}(\mathbf{w}') - \nabla_{\mathbf{w}}\ell_{M_t}(\mathbf{w}), \mathbf{w}' - \mathbf{w}\rangle \geq \frac{\beta\gamma}{\beta+\gamma}\|\mathbf{w}' - \mathbf{w}\|^2$$
$$+\frac{1}{\beta+\gamma}\|\nabla_{\mathbf{w}}\ell_{M_t}(\mathbf{w}') - \nabla_{\mathbf{w}}\ell_{M_t}(\mathbf{w})\|^2.$$

With this inequality in mind we have

$$\|G_t(\mathbf{w}') - G_t(\mathbf{w})\|^2 = \|\mathbf{w}' - \mathbf{w}\|^2 + \alpha_{t-1}^2\|\nabla_{\mathbf{w}}\ell_{M_t}(\mathbf{w}') - \nabla_{\mathbf{w}}\ell_{M_t}(\mathbf{w})\|^2$$
$$-2\alpha_{t-1}\langle\nabla_{\mathbf{w}}\ell_{M_t}(\mathbf{w}') - \nabla_{\mathbf{w}}\ell_{M_t}(\mathbf{w}), \mathbf{w}' - \mathbf{w}\rangle$$
$$\leq \left(1 - 2\frac{\beta\gamma\alpha_{t-1}}{\beta+\gamma}\right)\|\mathbf{w}' - \mathbf{w}\|^2 - \left(\frac{2\alpha_{t-1}}{\beta+\gamma} - \alpha_{t-1}^2\right)\|\nabla_{\mathbf{w}}\ell_{M_t}(\mathbf{w}') - \nabla_{\mathbf{w}}\ell_{M_t}(\mathbf{w})\|^2$$
$$\leq \left(1 - \frac{\beta\gamma\alpha_{t-1}}{\beta+\gamma}\right)^2\|\mathbf{w}' - \mathbf{w}\|^2,$$

wherein the last inequality follows from our assumption $\alpha_{t-1} \leq \frac{2}{\beta+\gamma}$ and the inequality $\sqrt{1-x} \leq 1 - \frac{x}{2}$ which holds for $x \in [0, 1]$.

□

## Appendix B. Proof of Theorem 4.1

As we are considering the worst case over the data distribution family $\mathcal{D}$ and the loss function family $\mathcal{L}_c$, we just need to find some special distributions from $\mathcal{D}$ and a specific loss from $\mathcal{L}_c$ and under these specific cases to derive the desired lower bound.

Specifically, we consider a particular classification problem. Recall the sample space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ where $\mathcal{X}$ is a domain in $\mathbb{R}^d$ and $\mathcal{Y} = \{-1, +1\}$. Naturally, $\mathcal{Z}$ can be divided into two parts, viz., $\mathcal{Z}_+ = \mathcal{X} \times \{+1\}$ and $\mathcal{Z}_- = \mathcal{X} \times \{-1\}$. Also we divide $\mathcal{X}$ into two disjoint parts, namely, $\mathcal{X}_1$ and $\mathcal{X}_2$.

We first consider a special distribution $P_1$ on the sample space $\mathcal{Z}$. Denote the marginal distribution of $P_1$ on $\mathcal{X}$ by $P_{1\mathcal{X}}$. We assume $P_{1\mathcal{X}}(x \in \mathcal{X}_1) = P_{1\mathcal{X}}(x \in \mathcal{X}_2) = \frac{1}{2}$. Accordingly, we can write $\mathcal{Z}_+ = (\mathcal{X}_1 \times \{+1\}) \sqcup (\mathcal{X}_2 \times \{+1\})$ and $\mathcal{Z}_- = (\mathcal{X}_1 \times \{-1\}) \sqcup (\mathcal{X}_2 \times \{-1\})$ using $\sqcup$ to denote the disjoint union. Then, define corresponding conditional probabilities as follows:

$$P_{1,y|\mathcal{X}}(y = 1|x \in \mathcal{X}_1) = \frac{1}{2} + \frac{\nu}{\sqrt{6n}}, \quad P_{1,y|\mathcal{X}}(y = -1|x \in \mathcal{X}_1) = \frac{1}{2} - \frac{\nu}{\sqrt{6n}},$$
$$P_{1,y|\mathcal{X}}(y = 1|x \in \mathcal{X}_2) = \frac{1}{2} - \frac{\nu - 1}{\sqrt{6n}}, \quad P_{1,y|\mathcal{X}}(y = -1|x \in \mathcal{X}_2) = \frac{1}{2} + \frac{\nu - 1}{\sqrt{6n}},$$

wherein the constant $\nu \in (1, \frac{\sqrt{6}}{2})$ to ensure that the above four probabilities are all in $(0, 1)$. Using the law of total probability, we have

$$P_1(z \in \mathcal{Z}_+) = \frac{1}{2} \cdot \Big(\frac{1}{2} + \frac{\nu}{\sqrt{6n}} + \frac{1}{2} - \frac{\nu - 1}{\sqrt{6n}}\Big) = \frac{1}{2} + \frac{1}{2\sqrt{6n}},$$
$$P_1(z \in \mathcal{Z}_-) = \frac{1}{2} \cdot \Big(\frac{1}{2} - \frac{\nu}{\sqrt{6n}} + \frac{1}{2} + \frac{\nu - 1}{\sqrt{6n}}\Big) = \frac{1}{2} - \frac{1}{2\sqrt{6n}}.$$

Similarly, we can define another distribution $P_2$ on the same splitting of $\mathcal{Z}$. Assume $P_{2\mathcal{X}}(x \in \mathcal{X}_1) = P_{2\mathcal{X}}(x \in \mathcal{X}_2) = \frac{1}{2}$. Its conditional probabilities are given by

$$P_{2,y|\mathcal{X}}(y = 1|x \in \mathcal{X}_1) = \frac{1}{2} - \frac{\nu}{\sqrt{6n}}, P_{2,y|\mathcal{X}}(y = -1|x \in \mathcal{X}_1) = \frac{1}{2} + \frac{\nu}{\sqrt{6n}},$$
$$P_{2,y|\mathcal{X}}(y = 1|x \in \mathcal{X}_2) = \frac{1}{2} + \frac{\nu - 1}{\sqrt{6n}}, P_{2,y|\mathcal{X}}(y = -1|x \in \mathcal{X}_2) = \frac{1}{2} - \frac{\nu - 1}{\sqrt{6n}}.$$

Then, we have

$$P_2(z \in \mathcal{Z}_+) = \frac{1}{2} - \frac{1}{2\sqrt{6n}}, \quad P_2(z \in \mathcal{Z}_-) = \frac{1}{2} + \frac{1}{2\sqrt{6n}}.$$

Let the sample $S_1$ and $S_2$ are i.i.d. drawn from $P_1$ and $P_2$, respectively.

Next, we define a specific convex and $\beta$-smooth loss function from the loss function family $\mathcal{L}_c$. Denote $\mathbf{w} \in \Omega$ as the parameter of the hypothesis function $h$, where $\Omega$ is the parameter space. Recall that we have assumed $\Omega$ has a finite diameter i.e. $|\Omega| < \infty$ and for simplicity, we also assume $\Omega$ is centered by 0 without

loss of generality. Let $\mathbf{w}[1]$ be the first coordinate of $\mathbf{w}$ and denote

$$f_1(\mathbf{w}) = \begin{cases} \frac{\beta}{2}(\mathbf{w}[1]-r)^2 & \text{for } |\mathbf{w}[1]-r| \leq \frac{r}{2}, \\ \frac{\beta r}{2}|\mathbf{w}[1]-r| - \frac{\beta r^2}{8} & \text{otherwise;} \end{cases}$$
$$f_2(\mathbf{w}) = \begin{cases} \frac{\beta}{2}(\mathbf{w}[1]+r)^2 & \text{for } |\mathbf{w}[1]+r| \leq \frac{r}{2}, \\ \frac{\beta r}{2}|\mathbf{w}[1]+r| - \frac{\beta r^2}{8} & \text{otherwise.} \end{cases}$$

The pairwise loss function $\ell(\mathbf{w}; z, z') : \Omega \times \mathcal{Z} \times \mathcal{Z} \longrightarrow \mathbb{R}$ in our purpose is defined as

$$\ell(\mathbf{w}; z, z') = \begin{cases} f_1(\mathbf{w}) & \text{for } z \in \mathcal{Z}_+, z' \in \mathcal{Z}_+, \\ \frac{1}{2}(f_1(\mathbf{w}) + f_2(\mathbf{w})) & \text{for } z \in \mathcal{Z}_+, z' \in \mathcal{Z}_- \text{ or } z \in \mathcal{Z}_-, z' \in \mathcal{Z}_+, \\ f_2(\mathbf{w}) & \text{for } z \in \mathcal{Z}_-, z' \in \mathcal{Z}_-. \end{cases}$$

It is easy to see that that $\ell$ is convex and $\beta$-smooth with respect to the first argument.

Now we consider the excess risks of the above specific loss $\ell$ under these two distributions which is given by

$$\begin{aligned}
&R_1(\mathbf{w}) = \mathbb{E}_{(z,z')\sim P_1\times P_1}[\ell(\mathbf{w}; z, z')] \\
&= \mathbb{P}(z \in \mathcal{Z}_+, z' \in \mathcal{Z}_+) \cdot f_1(\mathbf{w}) + \mathbb{P}(z \in \mathcal{Z}_-, z' \in \mathcal{Z}_-) \cdot f_2(\mathbf{w}) \\
&+ \mathbb{P}(z \in \mathcal{Z}_+, z' \in \mathcal{Z}_-) \cdot \frac{1}{2}(f_1(\mathbf{w}) + f_2(\mathbf{w})) + \mathbb{P}(z \in \mathcal{Z}_-, z' \in \mathcal{Z}_+) \cdot \frac{1}{2}(f_1(\mathbf{w}) + f_2(\mathbf{w})) \\
&= P_1(z \in \mathcal{Z}_+)P_1(z' \in \mathcal{Z}_+) \cdot f_1(\mathbf{w}) + P_1(z \in \mathcal{Z}_-)P_1(z' \in \mathcal{Z}_-) \cdot f_2(\mathbf{w}) \\
&+ P_1(z \in \mathcal{Z}_+)P_1(z' \in \mathcal{Z}_-) \cdot \frac{1}{2}(f_1(\mathbf{w}) + f_2(\mathbf{w})) + P_1(z \in \mathcal{Z}_-)P_1(z' \in \mathcal{Z}_+) \cdot \frac{1}{2}(f_1(\mathbf{w}) + f_2(\mathbf{w})) \\
&= \left(\frac{1}{2} + \frac{1}{2\sqrt{6n}}\right)^2 \cdot f_1(\mathbf{w}) + \left(\frac{1}{2} - \frac{1}{2\sqrt{6n}}\right)^2 \cdot f_2(\mathbf{w}) \\
&+ \left(\frac{1}{2} + \frac{1}{2\sqrt{6n}}\right)\left(\frac{1}{2} - \frac{1}{2\sqrt{6n}}\right) \cdot (f_1(\mathbf{w}) + f_2(\mathbf{w})) \\
&= \left(\frac{1}{2} + \frac{1}{2\sqrt{6n}}\right) \cdot f_1(\mathbf{w}) + \left(\frac{1}{2} - \frac{1}{2\sqrt{6n}}\right) \cdot f_2(\mathbf{w}).
\end{aligned}$$

Similarly, we have that

$$R_2(\mathbf{w}) = \mathbb{E}_{(z,z')\sim P_2\times P_2}[\ell(\mathbf{w}; z, z')] = \left(\frac{1}{2} - \frac{1}{2\sqrt{6n}}\right) \cdot f_1(\mathbf{w}) + \left(\frac{1}{2} + \frac{1}{2\sqrt{6n}}\right) \cdot f_2(\mathbf{w}).$$

Denote the excess risks as $\Delta R_1(\mathbf{w}) := R_1(\mathbf{w}) - \inf_{\mathbf{w}\in\Omega} R_1(\mathbf{w})$ and $\Delta R_2(\mathbf{w}) := R_2(\mathbf{w}) - \inf_{\mathbf{w}\in\Omega} R_2(\mathbf{w})$.

With the above preparations, we are now in the position to use the Le Cam's method ([27,38,39]) to estimate the minimax statistical error, i.e.,

$\inf_{\widetilde{\mathbf{w}}_n} \max_{i\in\{1,2\}} \mathbb{E}_{S_i\sim P_i^n}[\Delta R_i(\widetilde{\mathbf{w}}_n(S_i))]$. To this end, we write $R_1(\mathbf{w})$ in details as

$$R_1(\mathbf{w}) = \begin{cases} \left(\frac{1}{2}+\frac{1}{2\sqrt{6n}}\right)\frac{\beta r}{2}\left(\frac{3r}{4}-\mathbf{w}[1]\right)+\left(\frac{1}{2}-\frac{1}{2\sqrt{6n}}\right)\frac{\beta r}{2}\left(-\frac{5r}{4}-\mathbf{w}[1]\right), & \text{if } \mathbf{w}[1]\le\frac{-3r}{2}, \\ \left(\frac{1}{2}+\frac{1}{2\sqrt{6n}}\right)\frac{\beta r}{2}\left(\frac{3r}{4}-\mathbf{w}[1]\right)+\left(\frac{1}{2}-\frac{1}{2\sqrt{6n}}\right)\frac{\beta}{2}(r+\mathbf{w}[1])^2, & \text{if } |\mathbf{w}[1]+r|\le\frac{r}{2}, \\ \left(\frac{1}{2}+\frac{1}{2\sqrt{6n}}\right)\frac{\beta r}{2}\left(\frac{3r}{4}-\mathbf{w}[1]\right)+\left(\frac{1}{2}-\frac{1}{2\sqrt{6n}}\right)\frac{\beta r}{2}\left(\frac{3r}{4}+\mathbf{w}[1]\right), & \text{if } |\mathbf{w}[1]|\le\frac{r}{2}, \\ \left(\frac{1}{2}+\frac{1}{2\sqrt{6n}}\right)\frac{\beta}{2}(\mathbf{w}[1]-r)^2+\left(\frac{1}{2}-\frac{1}{2\sqrt{6n}}\right)\frac{\beta r}{2}\left(\frac{3r}{4}+\mathbf{w}[1]\right), & \text{if } |\mathbf{w}[1]-r|\le\frac{r}{2} \\ \left(\frac{1}{2}+\frac{1}{2\sqrt{6n}}\right)\frac{\beta r}{2}\left(\mathbf{w}[1]-\frac{5r}{4}\right)+\left(\frac{1}{2}-\frac{1}{2\sqrt{6n}}\right)\frac{\beta r}{2}\left(\frac{3r}{4}+\mathbf{w}[1]\right), & \text{if } \mathbf{w}[1]>\frac{3r}{2}. \end{cases}$$

Thus, we have

$$\nabla_{\mathbf{w}} R_1(\mathbf{w}) = \begin{cases} \left(-\frac{\beta r}{2},0,\cdots,0\right)^\top, & \text{if } \mathbf{w}[1]\le\frac{-3r}{2}, \\ \left((\frac{1}{2}-\frac{1}{2\sqrt{6n}})\beta\cdot\mathbf{w}[1]+(\frac{1}{4}-\frac{3}{4\sqrt{6n}})\beta r,0,\cdots,0\right)^\top, & \text{if } |\mathbf{w}[1]+r|\le\frac{r}{2}, \\ \left(-\frac{\beta r}{2\sqrt{6n}},0,\cdots,0\right)^\top, & \text{if } |\mathbf{w}[1]|\le\frac{r}{2}, \\ \left((\frac{1}{2}+\frac{1}{2\sqrt{6n}})\beta\cdot\mathbf{w}[1]+(-\frac{1}{4}-\frac{3}{4\sqrt{6n}})\beta r,0,\cdots,0\right)^\top, & \text{if } |\mathbf{w}[1]-r|\le\frac{r}{2} \\ \left(\frac{\beta r}{2},0,\cdots,0\right)^\top, & \text{if } \mathbf{w}[1]>\frac{3r}{2}. \end{cases}$$

Let $\mathbf{w}_1^*$ be (any) one of the minimum points of $R_1(\mathbf{w})$, i.e. $R_1(\mathbf{w}_1^*)=\inf_{\mathbf{w}\in\Omega} R_1(\mathbf{w})$. From the explicit form of $\nabla_{\mathbf{w}} R_1(\mathbf{w})$, it is direct to find that $(\frac{1}{2}+\frac{1}{2\sqrt{6n}})\beta\cdot\mathbf{w}_1^*[1]+(-\frac{1}{4}-\frac{3}{4\sqrt{6n}})\beta r=0$. As a result, we have $\mathbf{w}_1^*[1]=\frac{r}{2}+\frac{r}{1+\sqrt{6n}}:=\delta$. To be more specific, we can further assume that the other coordinates of $\mathbf{w}_1^*$ except $\mathbf{w}_1^*[1]$ all equal to zero. Thus $R_1(\mathbf{w}_1^*)=\inf_{\mathbf{w}\in\Omega} R_1(\mathbf{w})=\frac{3(\sqrt{6n}-1)\beta r^2}{8\sqrt{6n}}$. Denote $\mathbf{w}_{1,\text{right}}:=(2\delta,0,\ldots,0)$. So we have for any estimator $\widetilde{\mathbf{w}}_n$ s.t. $|\widetilde{\mathbf{w}}_n[1]-\mathbf{w}_1^*[1]|\ge\delta$, we have $\Delta R_1(\widetilde{\mathbf{w}}_n)=R_1(\widetilde{\mathbf{w}}_n)-R_1(\mathbf{w}_1^*)\ge\min\{R_1(0),R_1(\mathbf{w}_{1,\text{right}})\}-\frac{3(\sqrt{6n}-1)\beta r^2}{8\sqrt{6n}}=\frac{3\beta r^2}{8}-\frac{3(\sqrt{6n}-1)\beta r^2}{8\sqrt{6n}}=\frac{3\beta r^2}{8\sqrt{6n}}$.

Similarly, we write $R_2(\mathbf{w})$ in details as

$$R_2(\mathbf{w}) = \begin{cases} \left(\frac{1}{2}+\frac{1}{2\sqrt{6n}}\right)\frac{\beta r}{2}\left(-\mathbf{w}[1]-\frac{5r}{8}\right)+\left(\frac{1}{2}-\frac{1}{2\sqrt{6n}}\right)\frac{\beta r}{2}\left(\frac{3r}{4}-\mathbf{w}[1]\right), & \text{if } \mathbf{w}[1]\le\frac{-3r}{2}, \\ \left(\frac{1}{2}+\frac{1}{2\sqrt{6n}}\right)\frac{\beta}{2}(r+\mathbf{w}[1])^2+\left(\frac{1}{2}-\frac{1}{2\sqrt{6n}}\right)\frac{\beta r}{2}\left(\frac{3r}{4}-\mathbf{w}[1]\right), & \text{if } |\mathbf{w}[1]+r|\le\frac{r}{2}, \\ \left(\frac{1}{2}+\frac{1}{2\sqrt{6n}}\right)\frac{\beta r}{2}\left(\mathbf{w}[1]+\frac{3r}{4}\right)+\left(\frac{1}{2}-\frac{1}{2\sqrt{6n}}\right)\frac{\beta r}{2}\left(\frac{3r}{4}-\mathbf{w}[1]\right), & \text{if } |\mathbf{w}[1]|\le\frac{r}{2}, \\ \left(\frac{1}{2}+\frac{1}{2\sqrt{6n}}\right)\frac{\beta r}{2}\left(\mathbf{w}[1]+\frac{3r}{4}\right)+\left(\frac{1}{2}-\frac{1}{2\sqrt{6n}}\right)\frac{\beta}{2}(\mathbf{w}[1]-r)^2, & \text{if } |\mathbf{w}[1]-r|\le\frac{r}{2} \\ \left(\frac{1}{2}+\frac{1}{2\sqrt{6n}}\right)\frac{\beta r}{2}\left(\mathbf{w}[1]+\frac{3r}{4}\right)+\left(\frac{1}{2}-\frac{1}{2\sqrt{6n}}\right)\frac{\beta r}{2}\left(\mathbf{w}[1]-\frac{5r}{4}\right), & \text{if } \mathbf{w}[1]>\frac{3r}{2}. \end{cases}$$

Thus, we have

$$\nabla_{\mathbf{w}} R_2(\mathbf{w}) = \begin{cases} \left(-\frac{\beta r}{2}, 0, \cdots, 0\right)^\top, & \text{if } \mathbf{w}[1] \leq \frac{-3r}{2}, \\ \left((\frac{1}{2} + \frac{1}{2\sqrt{6n}})\beta \cdot \mathbf{w}[1] + (\frac{1}{4} + \frac{3}{4\sqrt{6n}})\beta r, 0, \cdots, 0\right)^\top, & \text{if } |\mathbf{w}[1] + r| \leq \frac{r}{2}, \\ \left(\frac{\beta r}{2\sqrt{6n}}, 0, \cdots, 0\right)^\top, & \text{if } |\mathbf{w}[1]| \leq \frac{r}{2}, \\ \left((\frac{1}{2} - \frac{1}{2\sqrt{6n}})\beta \cdot \mathbf{w}[1] + (-\frac{1}{4} + \frac{3}{4\sqrt{6n}})\beta r, 0, \cdots, 0\right)^\top, & \text{if } |\mathbf{w}[1] - r| \leq \frac{r}{2} \\ \left(\frac{\beta r}{2}, 0, \cdots, 0\right)^\top, & \text{if } \mathbf{w}[1] > \frac{3r}{2}. \end{cases}$$

Let $\mathbf{w}_2^*$ be (any) one of the minimum points of $R_2(\mathbf{w})$, i.e. $R_2(\mathbf{w}_2^*) = \inf_{\mathbf{w}\in\Omega} R_2(\mathbf{w})$. From the explicit form of $\nabla_{\mathbf{w}} R_2(\mathbf{w})$, it is direct to find that $(\frac{1}{2} - \frac{1}{2\sqrt{6n}})\beta \cdot \mathbf{w}_2^*[1] + (-\frac{1}{4} + \frac{3}{4\sqrt{6n}})\beta r = 0$. So we have $\mathbf{w}_2^*[1] = -\frac{r}{2} - \frac{r}{1+\sqrt{6n}} = -\delta$. For simplicity, we can further assume that the other coordinates of $\mathbf{w}_2^*$ except $\mathbf{w}_2^*[1]$ all equal to zero. Thus $R_2(\mathbf{w}_2^*) = \inf_{\mathbf{w}\in\Omega} R_2(\mathbf{w}) = \frac{3(\sqrt{6n}-1)\beta r^2}{8\sqrt{6n}}$. Let $\mathbf{w}_{2,\text{left}} = (-2\delta, 0, \ldots, 0)$. So we have for any $\widetilde{\mathbf{w}}_n$ s.t. $|\widetilde{\mathbf{w}}_n[1] - \mathbf{w}_2^*[1]| \geq \delta$, we have $\Delta R_2(\widetilde{\mathbf{w}}_n) = R_2(\widetilde{\mathbf{w}}_n) - R_2(\mathbf{w}_2^*) \geq \min\{R_2(0), R_2(\mathbf{w}_{2,\text{left}})\} - \frac{3(\sqrt{6n}-1)\beta r^2}{8\sqrt{6n}} = \frac{3\beta r^2}{8} - \frac{3(\sqrt{6n}-1)\beta r^2}{8\sqrt{6n}} = \frac{3\beta r^2}{8\sqrt{6n}}$.

Combining the above two situations, we have that for any output $\widetilde{\mathbf{w}}_n$, and $\forall i \in \{1, 2\}$, if $|\widetilde{\mathbf{w}}_n[1] - \mathbf{w}_i^*[1]| \geq \delta$, then $\Delta R_i(\widetilde{\mathbf{w}}_n) \geq \frac{3\beta r^2}{8\sqrt{6n}}$.

Then, for any $i = 1, 2$ there holds

$$\mathbb{E}_{S_i}[\Delta R_i(\widetilde{\mathbf{w}}_n(S_i))] \geq P_i^n(|\widetilde{\mathbf{w}}_n[1] - \mathbf{w}_i^*[1]| \geq \delta) \cdot \frac{3\beta r^2}{8\sqrt{6n}}.$$

Consequently,

$$\inf_{\widetilde{\mathbf{w}}_n} \max_{i\in\{1,2\}} \mathbb{E}_{S_i}[\Delta R_i(\widetilde{\mathbf{w}}_n(S_i))] \geq \frac{3\beta r^2}{8\sqrt{6n}} \inf_{\widetilde{\mathbf{w}}_n} \max_{i\in\{1,2\}} P_i^n(|\widetilde{\mathbf{w}}_n[1] - \mathbf{w}_i^*[1]| \geq \delta). \quad \text{(B.1)}$$

By Le Cam's method ([27,38,39]), when $|\mathbf{w}_1^*[1] - \mathbf{w}_2^*[1]| = 2\delta$, we can further reduce the estimation of the lower bound of the right hand side of (B.1) to a binary hypothesis testing problem:

$$\inf_{\hat{\mathbf{w}}\in\Omega} \max_{i\in\{1,2\}} P_i^n(|\hat{\mathbf{w}}[1] - \mathbf{w}_i^*[1]| \geq \delta) \geq \inf_{\Phi} \max_{i\in\{1,2\}} P_i^n(\Phi(\mathcal{Z}_i^n) \neq i), \quad \text{(B.2)}$$

where the infimum is taken over all binary testing functions $\Phi : \mathcal{Z}^n \to \{1, 2\}$. Thus by the standard analysis of Le Cam's method, we can further obtain

$$\inf_{\Phi} \max_{i\in\{1,2\}} P_i^n(\Phi(\mathcal{Z}_i^n) \neq i) \geq \frac{1}{2} \cdot (1 - \sqrt{\text{KL}(P_1^n \| P_2^n)/2}), \quad \text{(B.3)}$$

where $\text{KL}(P_1^n \| P_2^n)$ is the KL divergence. By the assumption of sampling independence, we have $\text{KL}(P_1^n \| P_2^n) = n\text{KL}(P_1 \| P_2)$. Furthermore, using the formulation of the distributions $P_1$ and $P_2$, we have $\text{KL}(P_1 \| P_2) = \frac{1}{\sqrt{6n}} \log\left(\frac{1+\frac{1}{\sqrt{6n}}}{1-\frac{1}{\sqrt{6n}}}\right)$. Note that $\log\left(\frac{1+x}{1-x}\right) \leq 3x$ for $x \in [0, 0.5]$. Thus $\text{KL}(P_1 \| P_2) \leq \frac{3}{6n} = \frac{1}{2n}$. Plugging the above

results into (B.3) gives

$$\inf_{\Phi} \max_{i\in\{1,2\}} P_i^n(\Phi(\mathcal{Z}_i^n) \neq i) \geq \frac{1}{2}\left(1 - \sqrt{\frac{1}{4}}\right) = \frac{1}{4}. \tag{B.4}$$

Combining the results (B.1), (B.2) and (B.4), we have

$$\inf_{\widetilde{\mathbf{w}}_n} \max_{i\in\{1,2\}} \mathbb{E}_{S_i}[\Delta R_i(\widetilde{\mathbf{w}}_n(S_i))] \geq \frac{3\beta r^2}{8\sqrt{6n}} \cdot \frac{1}{4} = \frac{3\beta r^2}{32\sqrt{6n}}. \tag{B.5}$$

To ensure both $\mathbf{w}_1^*$ and $\mathbf{w}_2^*$ are included in $\Omega$, it must hold that $\|\mathbf{w}_1^*\|_2 = \|\mathbf{w}_2^*\|_2 = \delta \leq \frac{|\Omega|}{2}$. Recall that $\delta = \frac{r}{2} + \frac{r}{1+\sqrt{6n}} < r$. Thus it is sufficient to assume $r \leq \frac{|\Omega|}{2}$. This means that we can take $r$ as large as $\frac{|\Omega|}{2}$. Take this into account and there exists $\ell$ such that

$$\inf_{\widetilde{\mathbf{w}}_n} \sup_{\mathcal{D}\in\mathcal{D}} \mathbb{E}_{S\sim D^n}[\Delta R(\widetilde{\mathbf{w}}_n(S))] \geq \frac{3\beta|\Omega|^2}{128\sqrt{6n}}. \tag{B.6}$$

The completes the proof of the theorem. □

## Appendix C. Proof of Theorem 4.2

We will follow the same procedure as the proof for Theorem 4.1. Specifically, we first define two distributions $P_1$ and $P_2$ which are exactly the same as the definitions in the proof of Theorem 4.1.

Then we define a specific strongly convex and strongly smooth loss function. Let $\Omega$ be the parameter space. with a finite diameter i.e. $|\Omega| < \infty$ and without loss of generality, we also assume $\Omega$ is centered by 0. Denote

$$f_1(\mathbf{w}) = \frac{\beta}{2}(\mathbf{w}[1] - r)^2 + \frac{\beta}{2}\left(\mathbf{w}[2]^2 + \cdots + \mathbf{w}[d]^2\right),$$
$$f_2(\mathbf{w}) = \frac{\beta}{2}(\mathbf{w}[1] + r)^2 + \frac{\beta}{2}\left(\mathbf{w}[2]^2 + \cdots + \mathbf{w}[d]^2\right).$$

We define the pairwise loss function $\ell(\mathbf{w}; z, z') : \Omega \times \mathcal{Z} \times \mathcal{Z} \longrightarrow \mathbb{R}$ as

$$\ell(\mathbf{w}; z, z') = \begin{cases} f_1(\mathbf{w}) & \text{for } z \in \mathcal{Z}_+, z' \in \mathcal{Z}_+, \\ \frac{1}{2}(f_1(\mathbf{w}) + f_2(\mathbf{w})) & \text{for } z \in \mathcal{Z}_+, z' \in \mathcal{Z}_- \text{ or } z \in \mathcal{Z}_-, z' \in \mathcal{Z}_+, \\ f_2(\mathbf{w}) & \text{for } z \in \mathcal{Z}_-, z' \in \mathcal{Z}_-. \end{cases}$$

It is easy to see that the above loss function $\ell(\mathbf{w}; z, z')$ is strongly convex and $\beta$-smooth w.r.t $\mathbf{w}$. It is sufficient to show that $f_1(\mathbf{w})$ and $f_2(\mathbf{w})$ are both strongly convex and $\beta$-smooth w.r.t $\mathbf{w}$. Firstly the Hessian matrices of both $f_1(\mathbf{w})$ and $f_2(\mathbf{w})$ have eigenvalues lower bounded by $\beta > 0$. So both $f_1(\mathbf{w})$ and $f_2(\mathbf{w})$ are strongly convex. To show they are $\beta$-smooth, we calculate the gradients of $f_1(\mathbf{w})$ and $f_2(\mathbf{w})$. We have $\nabla f_1(\mathbf{w}) = (\beta(\mathbf{w}[1] - r), \beta \cdot \mathbf{w}[2], \ldots, \beta \cdot \mathbf{w}[d])^\top$ It is easy to check that $\|\nabla f_1(\mathbf{w}_1) - \nabla f_1(\mathbf{w}_2)\| \leq \beta\|\mathbf{w}_1 - \mathbf{w}_2\|$. Similarly we can show $\nabla f_2(\mathbf{w})$ is $\beta$-Lipschitz.

Let distributions $P_1$ and $P_2$ be defined as in the proof of Theorem 4.1. Then,

$$R_1(\mathbf{w}) = \mathbb{E}_{(z,z')\sim P_1\times P_1}[\ell(\mathbf{w};z,z')] = \left(\frac{1}{2}+\frac{1}{2\sqrt{6n}}\right)\cdot f_1(\mathbf{w}) + \left(\frac{1}{2}-\frac{1}{2\sqrt{6n}}\right)\cdot f_2(\mathbf{w}).$$

We denote the excess risk under the distribution $P_1$ as $\Delta R_1(\mathbf{w}) := R_1(\mathbf{w}) - \inf_{\mathbf{w}\in\Omega} R_1(\mathbf{w})$. Similarly,

$$R_2(\mathbf{w}) = \mathbb{E}_{(z,z')\sim P_2\times P_2}[\ell(\mathbf{w};z,z')] = \left(\frac{1}{2}-\frac{1}{2\sqrt{6n}}\right)\cdot f_1(\mathbf{w}) + \left(\frac{1}{2}+\frac{1}{2\sqrt{6n}}\right)\cdot f_2(\mathbf{w}).$$

We denote the excess risk under the distribution $P_2$ as $\Delta R_2(\mathbf{w}) := R_2(\mathbf{w}) - \inf_{\mathbf{w}\in\Omega} R_2(\mathbf{w})$. Consequently,

$$\inf_{\widetilde{\mathbf{w}}_n}\sup_{\ell\in\mathcal{L}_{\rm sc},D\in\mathcal{D}} \mathbb{E}_{S\sim D^n}[\Delta R(\widetilde{\mathbf{w}}_n(S)] \geq \inf_{\widetilde{\mathbf{w}}_n}\max_{i\in\{1,2\}} \mathbb{E}_{S_i\sim P_i^n}[\Delta R_i(\widetilde{\mathbf{w}}_n(S_i))]. \tag{C.1}$$

Thus it is sufficient to lower bound the right hand side of (C.1) using the Le Cam's method ([27,38,39]).

To this end, we write $R_1(\mathbf{w})$ as

$$\begin{aligned} R_1(\mathbf{w}) &= \big(\frac{1}{2}+\frac{1}{2\sqrt{6n}}\big)\frac{\beta}{2}(r-\mathbf{w}[1])^2 + \big(\frac{1}{2}-\frac{1}{2\sqrt{6n}}\big)\frac{\beta}{2}(r+\mathbf{w}[1])^2 \\ &+ \frac{\beta}{2}\left(\mathbf{w}[2]^2+\cdots+\mathbf{w}[d]^2\right) \\ &= \frac{\beta}{2}\big(\mathbf{w}[1]-\frac{r}{\sqrt{6n}}\big)^2 + \frac{\beta r^2}{2}\big(1-\frac{1}{6n}\big) + \frac{\beta}{2}\big(\mathbf{w}[2]^2+\cdots+\mathbf{w}[d]^2\big). \end{aligned}$$

Let $\mathbf{w}_1^* = \arg\min_{\mathbf{w}\in\Omega} R_1(\mathbf{w})$. It is easy to see that $\mathbf{w}_1^*[1] = \frac{r}{\sqrt{6n}} := \delta$ and $\mathbf{w}_1^*[2] = \cdots = \mathbf{w}_1^*[d] = 0$. Thus $R_1(\mathbf{w}_1^*) = \inf_{\mathbf{w}\in\Omega} R_1(\mathbf{w}) = \frac{\beta r^2}{2}\big(1-\frac{1}{6n}\big)$. Also, for any $\widetilde{\mathbf{w}}_n$ s.t. $|\widetilde{\mathbf{w}}_n[1]-\mathbf{w}_1^*[1]| \geq \delta$, we have $\Delta R_1(\widetilde{\mathbf{w}}_n) = R_1(\widetilde{\mathbf{w}}_n) - R_1(\mathbf{w}_1^*) \geq R_1(0) - \frac{\beta r^2}{2}\big(1-\frac{1}{6n}\big) = \frac{\beta r^2}{2} - \frac{\beta r^2}{2}\big(1-\frac{1}{6n}\big) = \frac{\beta r^2}{12n}$.

Likewise,

$$R_2(\mathbf{w}) == \frac{\beta}{2}\left(\mathbf{w}[1]+\frac{r}{\sqrt{6n}}\right)^2 + \frac{\beta r^2}{2}\left(1-\frac{1}{6n}\right) + \frac{\beta}{2}\left(\mathbf{w}[2]^2+\cdots+\mathbf{w}[d]^2\right).$$

It is easy to see that $\mathbf{w}_2^* = \arg\min_{\mathbf{w}\in\Omega} R_2(\mathbf{w})$ is given by $\mathbf{w}_2^*[1] = -\frac{r}{\sqrt{6n}} = -\delta$ and $\mathbf{w}_2^*[2] = \cdots = \mathbf{w}_2^*[d] = 0$. Thus $R_2(\mathbf{w}_1^*) = \inf_{\mathbf{w}\in\Omega} R_2(\mathbf{w}) = \frac{\beta r^2}{2}\big(1-\frac{1}{6n}\big)$. For any estimator $\widetilde{\mathbf{w}}_n$ such that $|\widetilde{\mathbf{w}}_n[1]-\mathbf{w}_2^*[1]| \geq \delta$, we have $\Delta R_2(\widetilde{\mathbf{w}}_n) = R_2(\widetilde{\mathbf{w}}_n) - R_2(\mathbf{w}_2^*) \geq R_2(0) - \frac{\beta r^2}{2}\big(1-\frac{1}{6n}\big) = \frac{\beta r^2}{2} - \frac{\beta r^2}{2}\big(1-\frac{1}{6n}\big) = \frac{\beta r^2}{12n}$.

Combining the above estimation implies the following: for any output $\widetilde{\mathbf{w}}_n$, and $\forall i \in \{1,2\}$, if $|\widetilde{\mathbf{w}}_n[1]-\mathbf{w}_i^*[1]| \geq \delta$, then $\Delta R_i(\widetilde{\mathbf{w}}_n) \geq \frac{\beta r^2}{12n}$. Consequently, for any $i = 1, 2$, we obtain

$$\mathbb{E}_{S_i\sim P_i^n}[\Delta R_i(\widetilde{\mathbf{w}}_n(S_i))] \geq P_i^n(|\widetilde{\mathbf{w}}_n[1]-\mathbf{w}_i^*[1]| \geq \delta)\cdot\frac{\beta r^2}{12n},$$

which implies that

$$\inf_{\widetilde{\mathbf{w}}_n} \max_{i\in\{1,2\}} \mathbb{E}_{S_i\sim P_i^n}[\Delta R_i(\widetilde{\mathbf{w}}_n(S_i))] \geq \frac{\beta r^2}{12n} \cdot \inf_{\widetilde{\mathbf{w}}_n} \max_{i\in\{1,2\}} P_i^n(|\widetilde{\mathbf{w}}_n[1] - \mathbf{w}_i^*[1]| \geq \delta). \tag{C.2}$$

By exactly the same analysis as (B.2), (B.3) and (B.4) in the proof of Theorem 4.1, we further have

$$\inf_{\widetilde{\mathbf{w}}_n} \max_{i\in\{1,2\}} P_i^n(|\widetilde{\mathbf{w}}_n[1] - \mathbf{w}_i^*[1]| \geq \delta) \geq \frac{1}{2}\left(1 - \sqrt{\frac{1}{4}}\right) = \frac{1}{4}. \tag{C.3}$$

Combining the results (C.2), (C.2) and (C.3), we have

$$\inf_{\widetilde{\mathbf{w}}_n} \max_{D\in\mathcal{D}} \mathbb{E}_{S\sim D^n}[\Delta R(\widetilde{\mathbf{w}}_n(S))] \geq \inf_{\widetilde{\mathbf{w}}_n} \max_{i\in\{1,2\}} \mathbb{E}_{S_i\sim P_i^n}[\Delta R_i(\widetilde{\mathbf{w}}_n(S_i))] \tag{C.4}$$

$$\geq \frac{\beta r^2}{12n} \cdot \frac{1}{4} = \frac{\beta r^2}{48n}. \tag{C.5}$$

This completes the proof the theorem. □

## References

[1] S. Agarwal and P. Niyogi. Stability and generalization of bipartite ranking algorithms. In *International Conference on Computational Learning Theory*, pages 32–47. Springer, 2005.

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

[3] A. Bellet and A. Habrard. Robustness and generalization for metric learning. *Neurocomputing*, 151:259–267, 2015.

[4] C. L. Blake and C. J. Merz. Uci repository of machine learning databases [http://www. ics. uci. edu/˜ mlearn/mlrepository. html]. irvine, ca: University of california. *Department of Information and Computer Science*, 55, 1998.

[5] O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002.

[6] A. P. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.

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

[8] Yuansi Chen, Chi Jin, and Bin Yu. Stability and convergence trade-off of iterative optimization algorithms. *arXiv preprint arXiv:1804.01619*, 2018.

[9] Andreas Christmann and Ding-Xuan Zhou. On the robustness of regularized pairwise learning methods based on kernels. *Journal of Complexity*, 37:1–33, 2016.

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

[11] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S Dhillon. Information-theoretic metric learning. In *Proceedings of the 24th international conference on Machine learning*, pages 209–216. ACM, 2007.

[12] L. Devroye and T. Wagner. Distribution-free performance bounds for potential function rules. *IEEE Transactions on Information Theory*, 25(5):601–604, 1979.

[13] A. Elisseeff, T. Evgeniou, and M. Pontil. Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6(Jan):55–79, 2005.
[14] T. Fawcett. Prie: a system for generating rulelists to maximize roc performance. *Data Mining and Knowledge Discovery*, 17(2):207–224, 2008.
[15] W. Gao, R. Jin, S. Zhu, and Z. H. Zhou. One-pass auc optimization. In *International Conference on Machine Learning*, pages 906–914, 2013.
[16] M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? metric learning approaches for face identification. In *ICCV 2009-International Conference on Computer Vision*, pages 498–505. IEEE, 2009.
[17] J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.
[18] M. Hardt, B. Recht, and Y. Singer. Train faster, generalize better: Stability of stochastic gradient descent. *arXiv preprint arXiv:1509.01240*, 2015.
[19] T. Hu, J. Fan, Q. Wu, and D. X. Zhou. Learning theory approach to minimum error entropy criterion. *Journal of Machine Learning Research*, 14(Feb):377–397, 2013.
[20] T. Hu, Q. Wu, and D. X. Zhou. Convergence of gradient descent for minimum error entropy principle in linear regression. *IEEE Transactions on Signal Processing*, 64(24):6571–6579, 2016.
[21] T. Joachims. A support vector method for multivariate performance measures. In *Proceedings of the 22nd international conference on Machine learning*, pages 377–384. ACM, 2005.
[22] P. Kar and H. Karnick. Random feature maps for dot product kernels. In *International conference on artificial intelligence and statistics*, pages 583–591, 2012.
[23] P. Kar, B. Sriperumbudur, P. Jain, and H. Karnick. On the generalization ability of online learning algorithms for pairwise loss functions. In *International Conference on Machine Learning*, pages 441–449, 2013.
[24] Michael Kearns and Dana Ron. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural computation*, 11(6):1427–1453, 1999.
[25] Samuel Kutin and Partha Niyogi. Almost-everywhere algorithmic stability and generalization error. *arXiv preprint arXiv:1301.0579*, 2012.
[26] I. Kuzborskij and C. H. Lampert. Data-dependent stability of stochastic gradient descent. *arXiv preprint arXiv:1703.01678*, 2017.
[27] Lucien Le Cam. *Asymptotic methods in statistical decision theory*. Springer Science & Business Media, 2012.
[28] J. Lin, Y. Lei, B. Zhang, and D. X. Zhou. Online pairwise learning algorithms with convex loss functions. *Information Sciences*, 406:57–70, 2017.
[29] S. Mukherjee, P. Niyogi, T. Poggio, and R. Rifkin. Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, 25(1-3):161–193, 2006.
[30] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.
[31] A. Pensia, V. Jog, and P. L. Loh. Generalization error bounds for noisy, iterative algorithms. *arXiv preprint arXiv:1801.04295*, 2018.
[32] J. C. Principe. *Information theoretic learning: Renyi's entropy and kernel perspectives*. Springer Science & Business Media, 2010.
[33] A. Rakhlin, O. Shamir, and K. Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Conference on Machine Learning*, pages 449–456, 2012.
[34] W. Rejchel. On ranking and generalization bounds. *Journal of Machine Learning*

*Research*, 13:1373–1392, 2012.

[35] William H Rogers and Terry J Wagner. A finite sample distribution-free performance bound for local discrimination rules. *The Annals of Statistics*, pages 506–514, 1978.

[36] L. Rosasco, M. Belkin, and E. Vito. On learning with integral operators. *Journal of Machine Learning Research*, 11(Feb):905–934, 2010.

[37] S. Smale and D. X. Zhou. Learning theory estimates via integral operators and their approximations. *Constructive approximation*, 26(2):153–172, 2007.

[38] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Science & Business Media, 2009.

[39] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[40] C. Wang and T. Hu. Online minimum error entropy algorithm with unbounded sampling. *To appear in Analysis and Applications*, 2018.

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

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

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

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

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

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

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

[48] X. Zhang, A. Saha, and S.V.N. Vishwanathan. Smoothing multivariate performance measures. *Journal of Machine Learning Research*, 13:3623–3680, 2012.

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

[50] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 928–936, 2003.