Generalization Guarantee of SGD for Pairwise Learning

Yunwen Lei[1] **Mingrui Liu**[2] **Yiming Ying**[3]
[1]School of Computer Science, University of Birmingham, Birmingham B15 2TT, United Kingdom
[2]Department of Computer Science, George Mason University, Fairfax, VA 22030, USA
[3]Department of Mathematics and Statistics, State University of New York at Albany, USA
y.lei@bham.ac.uk mingruil@gmu.edu yying@albany.edu

Abstract

Recently, there is a growing interest in studying pairwise learning since it includes many important machine learning tasks as specific examples, e.g., metric learning, AUC maximization and ranking. While stochastic gradient descent (SGD) is an efficient method, there is a lacking study on its generalization behavior for pairwise learning. In this paper, we present a systematic study on the generalization analysis of SGD for pairwise learning to understand the balance between generalization and optimization. We develop a novel high-probability generalization bound for uniformly-stable algorithms to incorporate the variance information for better generalization, based on which we establish the first *nonsmooth* learning algorithm to achieve almost optimal *high-probability* and dimension-independent excess risk bounds with $O(n)$ gradient computations. We consider both convex and nonconvex pairwise learning problems. Our stability analysis for convex problems shows how the interpolation can help generalization. We establish a uniform convergence of gradients, and apply it to derive the first excess risk bounds on population gradients for nonconvex pairwise learning. Finally, we extend our stability analysis to pairwise learning with gradient-dominated problems.

1 Introduction

Many machine learning problems can be formulated as learning with pairwise loss functions, where the performance of the associated models needs to be quantified on a pair of training examples. Representative problems include AUC maximization [14, 25, 42, 63, 66], metric learning [8, 31], ranking [1, 13] and learning with minimum error entropy loss functions [29]. For example, in supervised metric learning we wish to find a distance function between pairs of examples so that examples within the same class are relatively close while examples from different classes are far apart from each other. In ranking, we aim to find a function to predict the ordering of examples. This motivates the recent growing interest in a unifying study of these problems, under the framework of pairwise learning [29, 32, 40, 58].

Stochastic gradient descent (SGD) is a workhorse for machine learning due to its cheap computation complexity, simplicity and efficiency [7, 20, 49, 53, 62, 65]. SGD iteratively updates the model based on stochastic gradients computed on one or several randomly selected training examples, which can achieve sample-size independent iteration complexity for a prescribed optimization accuracy. This is especially attractive for pairwise learning as the objective function involves $O(n^2)$ terms for problems with n training examples. An important problem on SGD is to understand its generalization performance, i.e., how the models trained by SGD would behave on testing examples. While there are some interesting work on the generalization analysis of SGD for pointwise learning [9, 10, 28, 36, 39], there is much less work on SGD for pairwise learning. A notable difference between pairwise learning

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

and pointwise learning is that the objective function in pairwise learning involves $O(n^2)$ *dependent* terms, which introduces a difficulty in handling this dependency. For example, one needs to decouple this dependency to apply concentration inequalities established for independent data. Furthermore, for algorithmic stability analysis, a perturbation of the training dataset by a single example can change $O(n)$ terms in the objective function, which is more challenging than stability analysis of pointwise learning. To our best knowledge, the only work on generalization analysis of SGD for pairwise learning are [40, 55, 59]. However, their analysis requires restrictive assumptions on convexity, smoothness and Lipschitz continuity of loss functions. Furthermore, they fail to incorporate the interpolation (low noise) assumption into their generalization guarantee.

In this paper, we initialize a systematic generalization analysis of SGD for pairwise learning under general assumptions. Our contributions are listed as follows.

1. We develop a novel high-probability generalization bound for uniformly-stable algorithms, which incorporates the variance information to improve the learning performance. We apply this result to develop the first dimension-independent high-probability bound $O(1/\sqrt{n})$ (up to a log factor) for an algorithm with $O(n)$ gradient computations to solve nonsmooth learning problems.

2. We study the stability and generalization guarantee of SGD for pairwise learning with convex loss functions, covering both smooth and nonsmooth problems. Our analysis suggests an early-stopping strategy for getting excess population risk bounds of the order $O(1/\sqrt{n})$ and $O(1/(n\sigma))$ for convex and σ-strongly convex problems, respectively. Under an interpolation or a low noise assumption, we improve our excess risk bounds to $O(1/n)$ by exploiting the smoothness assumption.

3. We provide the first generalization analysis of SGD for pairwise learning with nonconvex loss functions. We establish a uniform convergence of empirical gradients to population gradients by showing its connection to Rademacher chaos complexities. We then apply this uniform convergence to develop high-probability generalization guarantees for general nonconvex pairwise learning. Under a gradient dominance assumption, our stability analysis gives dimension-independent bounds.

The paper is organized as follows. We survey the related work in Section 2 and give the problem formulation in Section 3. We study convex and nonconvex pairwise learning in Section 4 and Section 5, respectively. Conclusion is given in Section 6. In the appendix, we present all the proofs, specific examples of pairwise learning and preliminary experimental results.

2 Related Work

We first review the related work on algorithmic stability. Algorithmic stability is an important concept in statistical learning theory (SLT) with close connection to learnability [52, 54]. The modern framework of stability analysis was established in a seminal paper [5], where the important uniform stability was introduced. This stability measure was extended to study randomized algorithms in [18] and motivates several concepts including argument stability [39, 43], Bayes stability [41] and on-average stability [39, 54]. Algorithmic stability has shown its remarkable effectiveness in deriving dimension-independent generalization bounds for various domains including stochastic optimization [10, 28, 36, 47, 51], structured prediction [44], transfer learning [37] and differential privacy [3, 48]. Recent progress shows a tradeoff between optimization and stability [11, 55], and its applications to yield almost optimal high-probability generalization bounds [6, 22, 34].

We now review the related work on pairwise learning. There are two popular approaches to studying the generalization performance of pairwise learning: the uniform convergence approach and the algorithmic stability approach. The idea of uniform convergence is to control the uniform deviation between training errors and testing errors over a hypothesis space. The complexity measures of function spaces play an important role in this approach, including VC dimension, covering numbers [17, 58, 60] and Rademacher complexities [8, 32]. Furthermore, one needs to use concentration inequalities to handle the associated U-statistics and U-processes [13, 16, 50]. Algorithmic stability of ranking [1] and metric learning [31, 57] was studied for strongly convex objectives [30]. High-probability bounds of the order $O(\epsilon \log n + 1/\sqrt{n})$ were recently developed for ϵ-uniformly stable pairwise learning algorithms [40]. A nice property of stability analysis is its ability to yield dimension-independent bounds, while a square-root dependency on the dimension is inevitable for uniform convergence analysis when considering general problems [21]. However, algorithmic stability generally requires a convexity assumption which is not required for the uniform convergence approach. Other than the above two approaches, several researchers have studied the generalization behavior of pairwise learning using the algorithmic robustness [4, 12] and integral operators [19, 27].

Before we move on, we add more discussions with a very related work on generalization analysis of pairwise learning [13, 50]. The work [50] considers a very general problem setting for SGD with K-sample U-statistic of degrees $(d_1, \ldots, d_K)$, which includes our algorithm as a special case with $K = 1$ and $d_1 = 2$. It shows the advantage of reducing variances using gradient estimates through incomplete U-statistics over that through complete U-statistics based on subsamples. We sketch the difference as follows. First, the generalization analysis in Papa et al. [50] requires smoothness and strong convexity assumptions. As a comparison, we also consider nonsmooth problems (Section 4.2) and nonconvex problems (Section 5). Second, the work [50] studies generalization via the uniform convergence approach and requires a complexity assumption. As a comparison, we also study generalization via a fundamentally different algorithmic stability approach. The classical work [13] focuses on the exact solution of empirical risk minimizer for pairwise learning from the perspective of uniform convergence. As a comparison, we study the excess risk of SGD via both an algorithmic stability approach and an uniform convergence approach, for which we also consider the tradeoff between optimization and generalization.

3 Problem Formulation

3.1 Pairwise Learning and Stochastic Gradient Descent

Let ρ be a probability measure defined on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ with an input space $\mathcal{X}$ and an output space $\mathcal{Y}$. Let $S = \{z_1, \ldots, z_n\}$ be drawn independently according to ρ, from which we aim to learn a prediction function $h : \mathcal{X} \mapsto \mathbb{R}$ or $h : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$. We consider parametric models where the prediction function $h_{\mathbf{w}}$ can be indexed by an element $\mathbf{w} \in \mathcal{W}$, where $\mathcal{W}$ is a d-dimensional Hilbert space. We consider pairwise learning problems where the performance of a model $h_{\mathbf{w}}$ on an example pair (z, z') can be measured by a nonnegative loss function $f(\mathbf{w}; z, z')$. This is in contrast to standard pointwise learning (e.g., classification and regression) where we can measure the quality of a model via its behavior on an individual point. Two notable examples of pairwise learning include ranking and supervised metric learning. For ranking, we build a function $h_{\mathbf{w}} : \mathcal{X} \mapsto \mathcal{Y}$ to rank instances in a way consistent with the outputs, i.e., $h_{\mathbf{w}}(x) < h_{\mathbf{w}}(x')$ if $y < y'$ for two example pairs $z = (x, y), z' = (x', y')$. Then we can formulate ranking as a pairwise learning problem with $f(\mathbf{w}; z, z') = \psi(\mathrm{sgn}(y - y')(h_{\mathbf{w}}(x) - h_{\mathbf{w}}(x')))$, where sgn is the sign function and ψ can be either the hinge loss $\psi(t) = \max\{1 - t, 0\}$ or the logistic loss $\psi(t) = \log(1 + \exp(-t))$. For supervised metric learning with $\mathcal{Y} = \{-1, +1\}$, we find a distance function under which examples with the same label are similar while examples with different labels are apart from each other. A popular distance function takes the form $h_{\mathbf{w}}(x, x') = \langle \mathbf{w}, (x - x')(x - x')^\top \rangle$, where $\mathbf{w} \in \mathbb{R}^{d \times d}$ is positive definite. We can formulate supervised metric learning as pairwise learning with $f(\mathbf{w}; z, z') = \psi(\tau(y, y')(1 - h_{\mathbf{w}}(x, x')))$, where $\tau(y, y') = 1$ if $y = y'$ and -1 otherwise.

The population risk of $\mathbf{w}$ in pairwise learning is $F(\mathbf{w}) = \mathbb{E}_{Z,Z'}[f(\mathbf{w}; Z, Z')]$, where $\mathbb{E}_{Z,Z'}$ denotes the expectation with respect to (w.r.t.) $Z, Z' \sim \rho$. The empirical risk of $\mathbf{w}$ is

$$F_S(\mathbf{w}) = \frac{1}{n(n-1)} \sum_{i,j \in [n]: i \neq j} f(\mathbf{w}; z_i, z_j),$$

where $[n] := \{1, \ldots, n\}$. Let $\mathbf{w}_S^* = \arg\min_{\mathbf{w} \in \mathcal{W}} F_S(\mathbf{w})$ and $\mathbf{w}^* = \arg\min_{\mathbf{w} \in \mathcal{W}} F(\mathbf{w})$. For a randomized algorithm A, we use $A(S)$ to denote the output model produced by applying A to the dataset S. We are interested in the excess risk $F(A(S)) - F(\mathbf{w}^*)$, which measures the relative behavior of $A(S)$ as compared to the best model. A standard approach to handle $F(A(S)) - F(\mathbf{w}^*)$ is to use the following error decomposition

$$\mathbb{E}_{S,A}\big[F(A(S)) - F(\mathbf{w}^*)\big] = \mathbb{E}_{S,A}\big[F(A(S)) - F_S(A(S))\big] + \mathbb{E}_{S,A}\big[F_S(A(S)) - F_S(\mathbf{w}^*)\big], \quad (3.1)$$

where the first term $F(A(S)) - F_S(A(S))$ is called the generalization error and the second term $F_S(A(S)) - F_S(\mathbf{w}^*)$ is the optimization error. These two errors can be handled by tools in SLT and optimization theory, respectively. We are interested in the specific SGD for pairwise learning.

Definition 1 (SGD for Pairwise Learning)**.** Let $\mathbf{w}_1 = 0 \in \mathbb{R}^d$ and $\{\eta_t\}_t$ be a stepsize sequence. Let $\nabla f(\mathbf{w}_t; z_{i_t}, z_{j_t})$ denote the gradient of f w.r.t. the first argument. At the t-th iteration, we first draw $\{(i_t, j_t)\}$ from the uniform distribution over all pairs $\{(i, j) : i, j \in [n], i \neq j\}$ and then update

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla f(\mathbf{w}_t; z_{i_t}, z_{j_t}). \quad (3.2)$$

Note our problem setting is totally different from some online learning setting where the streaming examples are assumed to be drawn from the true probability measure ρ [15]. Indeed, we consider the offline setting where the dataset is given beforehand and during the optimization process we actually randomly draw an example from the empirical measure. This necessitates the consideration of the generalization gap which is not touched in the online learning setting [15]. An advantage of SGD is that its computational complexity to achieve an accuracy is independent of the number of pairs, which is particularly attractive for pairwise learning (gradient descent requires $O(n^2)$ gradient computations per iteration). We describe SGD($S, T, f, \{\eta_t\}$) in Algorithm 1 of SGD with dataset S, iteration number T, loss function f and stepsize $\{\eta_t\}$. Algorithm 1 was also studied in [40], which however requires restrictive assumptions on convexity, smoothness and Lipschitz continuity. We significantly extend their discussions by considering either nonconvex, nonsmooth or non-Lipschitz loss. Moreover, our analysis can clarify the effect of interpolation on generalization.

Algorithm 1: $\mathrm{SGD}(S, T, f, \{\eta_t\})$

Input: initial point $\mathbf{w}_1 = 0$, learning rates $\{\eta_t\}_t$, and dataset $S = \{z_1, \ldots, z_n\}$
for $t = 1, 2, \ldots, T$ **do**
  draw (i_t, j_t) uniformly over all pairs $\{(i,j) : i, j \in [n], i \neq j\}$
  update $\mathbf{w}_{t+1}$ according to Eq. (3.2)
end

Below we introduce necessary definitions and assumptions. Let $\|\cdot\|_2$ be the Euclidean norm and $\langle \cdot, \cdot \rangle$ be the associated inner product. Let $b = \sup_{z,z' \in \mathcal{Z}} f(0; z, z')$ and $b' = \sup_{z,z' \in \mathcal{Z}} \|\nabla f(0; z, z')\|_2$. We denote $B \asymp \widetilde{B}$ if there are absolute constants c_1 and c_2 such that $c_1 B \leq \widetilde{B} \leq c_2 B$. We collect the notations of this paper in Table A.1.

Definition 2. Let $g : \mathcal{W} \mapsto \mathbb{R}, L, G > 0, \sigma \geq 0$.

1. We say g is L-smooth if $\|\nabla g(\mathbf{w}) - \nabla g(\mathbf{w}')\|_2 \leq L \|\mathbf{w} - \mathbf{w}'\|_2$ for all $\mathbf{w}, \mathbf{w}' \in \mathcal{W}$.
2. We say g is G-Lipschitz continuous if $|g(\mathbf{w}) - g(\mathbf{w}')| \leq G \|\mathbf{w} - \mathbf{w}'\|_2$ for all $\mathbf{w}, \mathbf{w}' \in \mathcal{W}$.
3. We say g is σ-strongly convex w.r.t. $\|\cdot\|_2$ if $g(\mathbf{w}) - \big(g(\mathbf{w}') + \langle \mathbf{w} - \mathbf{w}', \nabla g(\mathbf{w}') \rangle\big) \geq \sigma \|\mathbf{w} - \mathbf{w}'\|_2^2 / 2$ for all $\mathbf{w}, \mathbf{w}' \in \mathcal{W}$. We say g is convex if g is σ-strongly convex with $\sigma = 0$.

Assumption 1 (Convexity)**.** Assume for all $z, z' \in \mathcal{Z}$, the function $\mathbf{w} \mapsto f(\mathbf{w}; z, z')$ is convex.

Assumption 2 (Boundedness of Gradients)**.** Assume for all z, z' and $\mathbf{w} \in \mathcal{W}$, $\|\nabla f(\mathbf{w}; z, z')\|_2 \leq G$.

Assumption 3 (Smoothness)**.** Assume for all z, z', $\mathbf{w} \mapsto f(\mathbf{w}; z, z')$ is nonnegative and L-smooth.

3.2 Algorithmic Stability and Generalization

A fundamental concept in SLT is the algorithmic stability, which measures the sensitivity of an algorithm w.r.t. the perturbation of the training dataset. Various stability measures have been introduced in the literature, including uniform stability [5], hypothesis stability [5, 18], argument stability [43] and on-average stability [54]. We focus on uniform stability and on-average stability here. The following on-average loss stability was introduced in [40], while the on-average argument stability was motivated by a similar concept in pointwise learning [39]. Let $S = \{z_1, \ldots, z_n\}, S' = \{z'_1, \ldots, z'_n\}$ be independently drawn from ρ. We denote

$$S_i = \big\{z_1, \ldots, z_{i-1}, z'_i, z_{i+1}, \ldots, z_n\big\}, \quad \forall i \in [n], \tag{3.3}$$

$$S_{i,j} = \big\{z_1, \ldots, z_{i-1}, z'_i, z_{i+1}, \ldots, z_{j-1}, z'_j, z_{j+1}, \ldots, z_n\big\}, \quad \forall i < j \in [n]. \tag{3.4}$$

Definition 3 (Algorithmic Stability)**.** Let $S = \{z_1, \ldots, z_n\}$ and $S' = \{z'_1, \ldots, z'_n\}$ be drawn independently from ρ. For any $i, j \in [n]$, denote S_i as (3.3) and $S_{i,j}$ as (3.4).

1. We say a deterministic algorithm $A : \mathcal{Z}^n \mapsto \mathcal{W}$ is ϵ-uniformly stable if for any datasets $S, \widetilde{S} \in \mathcal{Z}^n$ that differ by at most a single example we have $\sup_{z, \tilde{z} \in \mathcal{Z}} \big| f(A(S); z, \tilde{z}) - f(A(\widetilde{S}); z, \tilde{z}) \big| \leq \epsilon$.
2. We say a randomized algorithm A is on-average (loss) ϵ-stable if

$$\frac{1}{n(n-1)} \sum_{i,j \in [n]: i \neq j} \mathbb{E}_{S,S',A} \Big[f\big(A(S_{i,j}); z_i, z_j\big) - f\big(A(S); z_i, z_j\big) \Big] \leq \epsilon.$$

3. We say A is on-average argument ϵ-stable if $\mathbb{E}_{S,\widetilde{S},A}\Big[\frac{1}{n}\sum_{i=1}^n \|A(S)-A(S_i)\|_2^2\Big] \leq \epsilon^2$.

The following theorem establishes the connection between on-average stability and generalization in expectation for pairwise learning. Part (a) was due to [40], while Part (b) was motivated by a similar result in pointwise learning [39]. The proof is given in Appendix B.

Theorem 1. *(a) If A is on-average (loss) ϵ-stable, then $\mathbb{E}[F(A(S)) - F_S(A(S))] \leq \epsilon$.*

(b) If A is on-average argument ϵ-stable and Assumption 3 holds, then for any $\gamma > 0$ we have

$$\mathbb{E}[F(A(S)) - F_S(A(S))] \leq 2(L+\gamma)\epsilon^2 + L\gamma^{-1}\mathbb{E}[F_S(A(S))].$$

Remark 1. We can choose $\gamma \asymp \sqrt{\mathbb{E}[F_S(A(S))]}/\epsilon$ in Part (b) and get the bound $\mathbb{E}[F(A(S)) - F_S(A(S))] = O(\epsilon^2 + \epsilon\sqrt{\mathbb{E}[F_S(A(S))]})$. Therefore, if $\mathbb{E}[F_S(A(S))]$ is small, Part (b) can imply bounds better than $O(\epsilon)$. In particular, if $\mathbb{E}[F_S(A(S))] = O(\epsilon^2)$ the bound in Part (b) becomes $O(\epsilon^2)$, which is much faster than $O(\epsilon)$ in Part (a).

Theorem 2 establishes the connection between uniform stability and generalization with high probability for pairwise learning. The proof is given in Section C, whose novelty is to use decoupling techniques to address the coupling among $O(n^2)$ terms in the objective function of pairwise learning.

Theorem 2. *Let A be an ϵ-uniformly stable and deterministic algorithm. Let $B := \sup_{z,z'}\big|\mathbb{E}_S[f(A(S);z,z')] - f(\mathbf{w}^*;z,z')\big|$ and $\sigma_0^2 := \mathbb{E}_{Z,Z',S}\big[\big(f(A(S);Z,Z') - f(\mathbf{w}^*;Z,Z')\big)^2\big]$. For any $\delta \in (0,1)$, the following inequality holds with probability at least $1-\delta$*

$$F(A(S)) - F_S(A(S)) - F(\mathbf{w}^*) + F_S(\mathbf{w}^*) \leq 98\sqrt{2}e\epsilon\log n\log(2e/\delta) + \frac{2B\log(2/\delta)}{3\lfloor n/2\rfloor} + \sqrt{\frac{2\sigma_0^2\log(2/\delta)}{\lfloor n/2\rfloor}}.$$

Remark 2. Note we only impose a bounded loss assumption on $A(S)$, which can be achieved by truncating the value of the output function. Theorem 2 was motivated by the recent work [34, 40]. For pointwise learning, high-probability generalization bounds of the order $O(\epsilon\log n)$ were developed for ϵ-uniformly stable algorithms under a further Bernstein condition on the variance-expectation relationship [34]. High-probability bounds $O(\epsilon\log n + 1/\sqrt{n})$ were also developed for ϵ-uniformly stable algorithms in pairwise learning [40]. We refine these results by developing generalization bounds $O(\epsilon\log n + \sqrt{\sigma_0^2/n})$, where σ_0^2 is the variance of the excess loss at the output model. If this variance is small, then our bounds can be much better than that in [40]. For example, if F is σ-strongly convex, then one can show that this variance can be bounded by $O(\mathbb{E}[F(A(S)) - F(\mathbf{w}^*)]/\sigma)$, and in this case the term $\sqrt{\sigma_0^2/n}$ in our bound would be $o(n^{-\frac{1}{2}})$ instead of $O(n^{-\frac{1}{2}})$ in [40]. As we will show, Theorem 2 can imply almost optimal excess risk bounds for an algorithm with $O(n)$ gradient computations to solve nonsmooth problems (Theorem 8), for which the existing high-probability analysis can only imply bounds of the order $O(n^{-\frac{1}{4}})$ [40].

4 Pairwise Learning with Convex Loss Functions

In this section, we study the generalization performance of SGD for pairwise learning with convex loss functions. We consider convex/strongly-convex and smooth/nonsmooth problems.

4.1 Convex and Smooth Problems

We first consider stability and risk bounds for convex and smooth pairwise learning problems. The proofs of results in this subsection can be found in Section E. Theorem 3 gives the bounds for on-average argument stability of SGD. Note we do not require the loss functions to be Lipschitz continuous. A nice property is that the upper bound involves the empirical risk of $\mathbf{w}_j$. Since we are minimizing the empirical risk by SGD, it is reasonable to assume that $F_S(\mathbf{w}_j)$ would become smaller and smaller along the learning process. It should be mentioned that a similar result was derived for pointwise learning [39]. A key difference in the stability analysis for pairwise learning is that a change of z_i would influence $2(n-1)$ pairs $(z_j,z_i),(z_i,z_j)$ for $j \neq i$. We need to use the U-structure of the empirical risk to prove this result.

Theorem 3 (Stability bound)**.** *Let Assumptions 1, 3 hold. Let A be SGD (Algorithm 1) with $\eta_j \leq 2/L$. Then A with t iterations is on-average argument ϵ-stable with $\epsilon^2 \leq \frac{16L(1+2t/n)e}{n}\sum_{j=1}^t \eta_j^2 \mathbb{E}[F_S(\mathbf{w}_j)]$.*

Based on Theorem 3, we get error bounds for pairwise learning with convex and smooth functions.

Theorem 4 (Excess risk bound)**.** *Let Assumptions 1, 3 hold. Let $\{\mathbf{w}_t\}$ be the sequence produced by SGD (Algorithm 1) on a dataset of size n with $\eta_t = \eta \leq 2/L$. Then for $\bar{\mathbf{w}}_T = \frac{1}{T}\sum_{t=1}^T \mathbf{w}_t$ and any $\gamma \geq 1$ we have the following inequality for all $\mathbf{w}$ independent of A (can depend on S)*

$$\mathbb{E}[F(\bar{\mathbf{w}}_T) - F_S(\mathbf{w})] = O\Big(\Big(\eta + \frac{1}{\gamma} + \frac{\gamma(T+T^2/n)\eta^2}{n}\Big)\mathbb{E}[F_S(\mathbf{w})]\Big) + O\Big(\frac{\mathbb{E}[\|\mathbf{w}\|_2^2]}{T\eta} + \frac{\gamma\eta(1+T/n)\mathbb{E}[\|\mathbf{w}\|_2^2]}{n}\Big). \quad (4.1)$$

A notable property of the above bound is that it holds for any $\mathbf{w}$ independent of A. If $\mathbb{E}[\|\mathbf{w}_S^*\|_2^2]$ is finite, we can choose $\mathbf{w} = \mathbf{w}_S^*$ in Eq. (4.1) and get a bound involving $\mathbb{E}[F_S(\mathbf{w}_S^*)]$. Furthermore, if we are in an interpolation or overparameterized setting [45] then $\mathbb{E}[F_S(\mathbf{w}_S^*)] = o(1/\sqrt{n})$ and the generalization will improve according to (4.1). Therefore, our stability analysis provides an explanation on how interpolation/overparameterization can help in generalization. Note SGD has an implicit bias to choose a model with a small norm and therefore it is reasonable to assume $\mathbb{E}[\|\mathbf{w}_S^*\|_2^2] < \infty$. We can also choose $\mathbf{w} = \mathbf{w}^*$ in Eq. (4.1) to get optimistic bounds in the sense of involving $F(\mathbf{w}^*)$, which decay fast if $F(\mathbf{w}^*)$ is small [56, 60]. Indeed, the following corollary gives the bound $O(1/\sqrt{n})$ in the general case and improves it to $O(1/n)$ if $F(\mathbf{w}^*) = O(1/n)$. Note here we use the assumption $F(\mathbf{w}^*) = O(1/n)$ just to show that we can get improved bound under low noise conditions. The term $F(\mathbf{w}^*)$ should be independent of n.

Corollary 5. *Let Assumptions in Theorem 4 hold.*

(a) We can choose $\eta_t = \eta \asymp 1/\sqrt{T}$ and $T \asymp n$ to get $\mathbb{E}[F(\bar{\mathbf{w}}_T)] - F(\mathbf{w}^*) = O(1/\sqrt{n})$.

(b) If $F(\mathbf{w}^*) = O(1/n)$, choosing $\eta_t = \eta \leq 2/L$ and $T \asymp n$ yields $\mathbb{E}[F(\bar{\mathbf{w}}_T)] - F(\mathbf{w}^*) = O(1/n)$.

Remark 3. Stability and excess risk bounds of the order $O(1/\sqrt{n})$ were studied for SGD applied to pairwise learning [40, 55]. However, these discussions require the loss functions to be smooth, Lipschitz continuous and convex. As a comparison, we remove the Lipschitz continuity assumption. Furthermore, their discussion can only imply non-optimistic bounds of the order $O(1/\sqrt{n})$. As a comparison, our discussions can fully exploit the property of $F(\mathbf{w}^*)$ to imply fast bounds $O(1/n)$.

4.2 Convex and Nonsmooth Problems

We now turn to pairwise learning with convex and nonsmooth functions. Theorem 6 gives the argument stability bounds based on which we develop excess risk bounds in Theorem 7. The proofs of results in this subsection are given in Section F.

Theorem 6 (Stability bounds)**.** *Let Assumptions 1, 2 hold. Let $S = \{z_1, \ldots, z_n\}$ and $S' = \{z'_1, \ldots, z'_n\}$ be two datasets that differ by a single point. Let $\{\mathbf{w}_t\}, \{\mathbf{w}'_t\}$ be the sequence produced by SGD (Algorithm 1) w.r.t. S and S' with $\eta_t = \eta$, respectively. Then*

$$\mathbb{E}\big[\|\mathbf{w}_{t+1} - \mathbf{w}'_{t+1}\|_2^2\big] \leq 4G^2 et\big(1 + 4t/n^2\big)\eta^2. \quad (4.2)$$

For any $\delta \in (0,1)$ with probability at least $1-\delta$ we have

$$\|\mathbf{w}_{t+1} - \mathbf{w}'_{t+1}\|_2^2 \leq 4G^2\eta^2 e\Big(t + \big(2t/n + \log(1/\delta) + \sqrt{4tn^{-1}\log(1/\delta)}\big)^2\Big).$$

Theorem 7 (Excess risk bounds)**.** *Let Assumptions 1, 2 hold. Let $\{\mathbf{w}_t\}$ be the sequence produced by SGD with $\eta_t = \eta \asymp T^{-\frac{3}{4}}$. If $T \asymp n^2$, then $\mathbb{E}[F(\bar{\mathbf{w}}_T)] - F(\mathbf{w}^*) = O(n^{-\frac{1}{2}})$.*

Remark 4. As compared to the stability bounds in the smooth case (Theorem 3), the stability bounds in (4.2) are worse in the sense that we do not have a factor of $1/n$ in (4.2). Therefore, one needs to choose very small stepsizes to let the right-hand side of (4.2) vanish to 0. Indeed, Theorem 7 suggests $\eta_t \asymp T^{-\frac{3}{4}}$, which are much smaller than the $\eta_t \asymp T^{-\frac{1}{2}}$ in the smooth case. As a result, we require to run SGD with $T \asymp n^2$ to get the optimal excess risk bounds $O(1/\sqrt{n})$. Note we also give high-probability bounds on the argument stability in Theorem 6. It should be mentioned that the stability of a variant of SGD as $\mathbf{w}_{t+1} = \mathbf{w}_t - \frac{\eta_t}{t-1}\sum_{k=1}^{t-1}\nabla f(\mathbf{w}_t; z_{i_t}, z_{i_k})$ was recently studied for pairwise learning with convex and nonsmooth loss functions [59]. Note this update requires $O(t)$ gradient computations at the t-th iteration, while (3.2) only requires a single gradient computation.

Algorithm 2: Iterative Localized Algorithm for Pairwise Learning

Input: initial point $\mathbf{w}_0 = 0$, parameter $\gamma > 0, k = \lceil \frac{1}{2}\log_2 n \rceil$

for $i = 1, 2, \ldots, k$ **do**

set $T_i \asymp n_i = \lceil \frac{n}{2^i} \rceil, \gamma_i = \frac{\gamma}{2^i}, \eta_t = \frac{\gamma_i n_i}{t+1}, t \in \mathbb{N}, \tilde{f}(\mathbf{w}; z, z') = f(\mathbf{w}; z, z') + \frac{1}{\gamma_i n_i}\|\mathbf{w} - \mathbf{w}_{i-1}\|_2^2$

draw a sample S_i of size n_i independently from ρ

apply $\mathrm{SGD}(S_i, T_i, \tilde{f}, \{\eta_t\})$ to minimize the following problem and get $\mathbf{w}_i$

$$\widetilde{F}_{S_i}(\mathbf{w}) := \frac{1}{n_i(n_i - 1)} \sum_{z, z' \in S_i : z \neq z'} f(\mathbf{w}; z, z') + \frac{1}{\gamma_i n_i}\|\mathbf{w} - \mathbf{w}_{i-1}\|_2^2. \tag{4.3}$$

end

Note SGD requires the undesirable $O(n^2)$ gradient computations to achieve the bound $O(1/\sqrt{n})$ for nonsmooth problems. The following theorem shows one can also achieve the bound $O(1/\sqrt{n})$ with $O(n)$ gradient computations by considering Algorithm 2. Algorithm 2 is motivated from the iterative localization approach established in pointwise learning [23], which was also used to develop efficient differentially private algorithms [2, 35]. Note the choice $\gamma \asymp n^{-\frac{1}{2}}\|\mathbf{w}^*\|_2$ depends on the unknown $\|\mathbf{w}^*\|_2$. However, one can get the bound $O(D/\sqrt{n})$ by choosing $\gamma \asymp D/\sqrt{n}$ if $\|\mathbf{w}^*\|_2 \leq D$.

Theorem 8. *Let Assumptions 1, 2 hold and $\delta \in (0, 1)$. Let $\mathbf{w}_k$ be produced by Algorithm 2 and assume $\sup_{z,z'} |\mathbb{E}[f(\mathbf{w}_i, z, z')]| \leq B$ for some $B > 0$. If we choose $\gamma \asymp n^{-\frac{1}{2}}\|\mathbf{w}^*\|_2$, then with probability at least $1 - \delta$ we have $F(\mathbf{w}_k) - F(\mathbf{w}^*) = O(\log(\log n/\delta)(\log n)\|\mathbf{w}^*\|_2/\sqrt{n})$. Moreover, Algorithm 2 requires only $O(n)$ gradient computations to achieve this excess risk bound.*

Remark 5. Iterative localization approach was developed in [23, 35] to develop novel algorithms with $O(n)$ gradient computations and $O(1/\sqrt{n})$ excess risk bounds for nonsmooth pointwise learning problems. We extend this technique to the pairwise learning setting. Furthermore, the excess risk bounds in [23, 35] are stated in expectation. As a comparison, we use the novel high-probability bounds for uniformly stable algorithms established in Theorem 2 to develop high-probability bounds of the order $(\log^2 n)/\sqrt{n}$, which has not been developed even for pointwise learning (the best *high-probability* excess risk bound for SGD with nonsmooth loss functions requires $O(n^2)$ gradient computations [3]). Note that the sample size $n_k \asymp \sqrt{n}$ in the k-th epoch, and therefore the high-probability bounds in [40] can only yield bounds $O(n_k^{-\frac{1}{2}}) = O(n^{-\frac{1}{4}})$ for Algorithm 2. As a comparison, Theorem 2 applied to the ERM of $\widetilde{F}_{S_k}$ yields the bounds $O(n_k^{-1} + \gamma_k) = O(n^{-\frac{1}{2}})$ (ERM of $\widetilde{F}_{S_k}$ is γ_k-uniformly stable and the last term of the bound in Theorem 2 is dominated). This demonstrates the advantage of our new high-probability bounds in Theorem 2 for developing almost optimal bounds with $O(n)$ gradient computations. Another difference is that we apply iterative localization framework with $k = \lceil \frac{1}{2}\log_2 n \rceil$ instead of $k' = \lceil \log_2 n \rceil$ epochs in [23, 35]. The underlying reason is that the high-probability bounds for ERM of $\widetilde{F}_{S_k}$ involve $n_k^{-1} + \gamma_k^{-1}$, while the bounds in expectation only involve γ_k^{-1} [40]. Since $n_{k'} \asymp 1$, our high-probability analysis only implies vacuous bounds $O(1)$ if we use $k' = \lceil \log_2 n \rceil$ epochs.

4.3 Strongly Convex Problems

We now turn to strongly convex cases. Theorem 9 gives bounds for smooth problems, while Theorem 10 gives excess risk bounds for nonsmooth problems. Note Theorems 9 and 10 apply to any algorithm, which give a general relationship between excess risks and optimization errors. One can plug the optimization error bounds for any algorithm to immediately derive the corresponding excess risk bounds. The proofs of results in this subsection are given in Section G.

Theorem 9 (Strongly Convex and Smooth Problems)**.** *Let Assumption 3 hold. Assume for all $S \in \mathcal{Z}^n$, F_S is σ-strongly convex w.r.t. $\|\cdot\|_2$. Let A be a randomized algorithm and $\sigma n \geq 8L$. Then*

$$\mathbb{E}[F(A(S))] - F(\mathbf{w}^*) \leq 128L\Big(\frac{16L}{n^2\sigma^2} + \frac{1}{n\sigma}\Big)\mathbb{E}\big[F_S(\mathbf{w}_S^*)\big] + \frac{2L}{\sigma}\mathbb{E}\big[F_S(A(S)) - F_S(\mathbf{w}_S^*)\big].$$

Remark 6. The above bound involves two components. The first component $O(1/(n\sigma))$ depends only on the landscape of the learning problem. The second component involves optimization errors. It shows that optimization is always beneficial to improve generalization for strongly convex problems. Furthermore, it also shows that one can stop the algorithm once we achieve the optimization error bounds $O(1/n)$ since further optimization would not improve essentially the generalization. Theorem 9 also implies bounds of the order $o(1/(n\sigma))$ if $F_S(\mathbf{w}^*)$ is small.

Theorem 10 (Strongly Convex and Nonsmooth Problems)**.** *Assume f takes a structure as $f(\mathbf{w}; z, z') = \ell(\mathbf{w}; z, z') + r(\mathbf{w})$. Assume for all z, z', the map $\mathbf{w} \mapsto \ell(\mathbf{w}; z, z')$ is G-Lipschitz. Assume for all $S \in \mathcal{Z}^n$, F_S is σ-strongly convex w.r.t. $\|\cdot\|_2$. For any algorithm A we have*

$$\mathbb{E}[F(A(S))] - F(\mathbf{w}^*) \leq \frac{8G^2}{n\sigma} + G\sqrt{\frac{2\mathbb{E}\big[F_S(A(S)) - F_S(\mathbf{w}_S^*)\big]}{\sigma}}.$$

We present the specific applications of the above results to SGD in Corollary G.2 (Section G).

5 Pairwise Learning with Nonconvex Loss Functions

In this section, we consider excess risk bounds for pairwise learning in a nonconvex setting. In this case, the excess population risk is not a reasonable measure since we cannot guarantee that the algorithm can find a global minimizer. We therefore use the norm of gradients at $A(S)$ to measure the performance of A [26, 68]. In a general nonconvex setting, SGD requires to choose $\eta_t = O(1/t)$ for a meaningful stability bound [28], for which the optimization errors would decay logarithmically w.r.t the number of iterations [26]. Then, stability analysis fails to trade-off the stability and optimization for a model with good generalization performance in a general nonconvex problem. Therefore, we turn to a different uniform convergence approach in a general nonconvex setting [38]. After that, we study stability of SGD for nonconvex problems under a further PL condition.

5.1 Uniform Convergence of Gradients for Pairwise Learning

Our first result for nonconvex pairwise learning is a uniform convergence of gradients. Specifically, we show that the uniform deviation between population gradients and empirical gradients over a space can be bounded by the associated Rademacher chaos complexity. Let $\mathcal{W}_R = \{\mathbf{w} \in \mathcal{W} : \|\mathbf{w}\|_2 \leq R\}$ for $R > 0$. The proofs of results in this subsection are given in Section H.

Definition 4. Let $\mathcal{F} := \{f : \mathcal{Z}^4 \mapsto \mathbb{R}\}$ be a function class and $S = \{z_i\}_{i=1}^n \subset \mathcal{Z}$. Let $\{\epsilon_i\}_{i=1}^{\lfloor \frac{n}{2} \rfloor}$ be independent Rademacher variables with $\Pr\{\epsilon_i = 1\} = \Pr\{\epsilon_i = -1\} = 1/2$. The empirical Rademacher chaos complexity for $\mathcal{F}$ w.r.t. S is defined as

$$\mathcal{U}_S(\mathcal{F}) = \frac{1}{\lfloor \frac{n}{2} \rfloor}\mathbb{E}_\epsilon\Big[\sup_{f \in \mathcal{F}} \sum_{1 \leq i < j \leq \lfloor \frac{n}{2} \rfloor} \epsilon_i \epsilon_j f(z_i, z_{i+\lfloor \frac{n}{2} \rfloor}, z_j, z_{j+\lfloor \frac{n}{2} \rfloor})\Big].$$

Theorem 11 (Uniform Convergence of Gradients)**.** *Let $\delta \in (0,1)$, $R > 0$ and $S = \{z_1, \ldots, z_n\}$ be drawn independently from ρ. If Assumption 3 holds, then with probability at least $1 - \delta$ we have*

$$\sup_{\mathbf{w} \in \mathcal{W}_R} \big\|\nabla F(\mathbf{w}) - \nabla F_S(\mathbf{w})\big\|_2 \leq \frac{2\sqrt{2}(LR + b')\big(2 + \sqrt{\log(1/\delta)}\big)}{\sqrt{n}} + 4\sqrt{\frac{\mathcal{U}_S(\mathcal{F}_R)}{n}},$$

where

$$\mathcal{F}_R = \big\{(z_1, z_2, z_3, z_4) \mapsto \langle \nabla f(\mathbf{w}; z_1, z_2), \nabla f(\mathbf{w}; z_3, z_4) \rangle : \mathbf{w} \in \mathcal{W}_R\big\}.$$

We can apply the entropy integral to control the above Rademacher chaos complexity [16, 61], and get the following result. Note d is the dimension of the space $\mathcal{W}$.

Corollary 12. *Under Assumptions of Theorem 11, with probability at least* $1-\delta$ *we have*

$$\sup_{\mathbf{w}\in\mathcal{W}_R}\big\|\nabla F(\mathbf{w})-\nabla F_S(\mathbf{w})\big\|_2\leq\frac{2\sqrt{2}(LR+b')}{\sqrt{n}}\Big(2+\sqrt{96e\big(\log 2+d\log(3e)\big)}+\sqrt{\log(1/\delta)}\Big).$$

The above uniform convergence rate involves a square-root dependency on d. We show that this dependency can be avoided if we consider a special class of functions with a specific structure

$$f(\mathbf{w};z,z')=\psi(\langle\mathbf{w},\phi(x,x')\rangle,\tau(y,y')),\tag{5.1}$$

where $\phi:\mathcal{X}\times\mathcal{X}\mapsto\mathcal{W}$ is a feature map, $\psi:\mathbb{R}\times\mathbb{R}\mapsto\mathbb{R}$ is L_ψ-smooth w.r.t. the first argument and $\tau:\mathcal{Y}\times\mathcal{Y}\mapsto\mathbb{R}$. Loss functions of the structure (5.1) have wide applications in robust optimization and generalized linear models [24, 46]. We assume $\kappa=\sup_{x,x'\in\mathcal{X}}\|\phi(x,x')\|_2$.

Corollary 13. *Let* $\delta\in(0,1)$, $R>0$ *and* $S=\{z_1,\ldots,z_n\}$ *be examples drawn independently from* ρ. *Suppose* $f:\mathcal{W}\times\mathcal{Z}^2\mapsto\mathbb{R}$ *takes the form* (5.1) *with* ψ *being* L_ψ*-smooth w.r.t. the first argument. Then the following inequality holds with probability at least* $1-\delta$

$$\sup_{\mathbf{w}\in\mathcal{W}_R}\big\|\nabla F(\mathbf{w})-\nabla F_S(\mathbf{w})\big\|_2\leq\frac{4\kappa\big(2L_\psi R\kappa+b'\big)}{\sqrt{n}}+\sqrt{\frac{8\big(L_\psi\kappa^2R+b'\big)^2\log(1/\delta)}{n}}.$$

Remark 7. Uniform convergence of gradients was studied for pointwise learning based on covering numbers [46, 64] and Rademacher complexities [24]. We extend these results to the pairwise learning setting. A key difference between pointwise learning and pairwise learning is that the empirical gradients for pairwise learning can be no longer written as a summation of i.i.d. terms. Indeed, the $n(n-1)$ terms in $\nabla F_S(\mathbf{w})$ are correlated, which introduces difficulties in applying concentration inequalities. We need to apply decoupling techniques in U-process to handle this correlation.

5.2 Smooth Problems

We now study the generalization performance of SGD for pairwise learning based on the uniform convergence of gradients developed in the previous subsection. We first introduce necessary assumptions. Since η_t is always small (a typical choice is $\eta_t\asymp 1/\sqrt{T}$), Eq. (5.2) is milder than a bounded gradient assumption. Eq. (5.3) imposes a bounded variance assumption on stochastic gradients, which is a standard assumption for the analysis of SGD [26, 36, 67].

Assumption 4. Assume the existence of $G>0$ and $\sigma_1>0$ such that

$$\sqrt{\eta_t}\|\nabla f(\mathbf{w}_t;z,z')\|_2\leq G,\forall t\in\mathbb{N},z,z'\in\mathcal{Z},\tag{5.2}$$

$$\mathbb{E}_{i_t,j_t}\big[\|\nabla f(\mathbf{w}_t;z_{i_t},z_{j_t})-\nabla F_S(\mathbf{w}_t)\|_2^2\big]\leq\sigma_1^2,\quad\forall t\in\mathbb{N}.\tag{5.3}$$

Theorem 14 gives high-probability bounds on the norm of population gradients. Our basic idea is to use the following error decomposition

$$\|\nabla F(\mathbf{w}_t)\|_2^2\leq 2\|\nabla F(\mathbf{w}_t)-\nabla F_S(\mathbf{w}_t)\|_2^2+2\|\nabla F_S(\mathbf{w}_t)\|_2^2.$$

We refer to the first term on the right-hand side as the generalization error for nonconvex pairwise learning, which can be bounded by the uniform convergence of gradients established in Theorem 11. The second term is the optimization error and can be addressed by techniques in optimization theory. The proof is given in Section I.

Theorem 14 (Smooth Problems)**.** *Let Assumptions 3 and 4 hold. Let* $\{\mathbf{w}_t\}_t$ *be the sequence produced by* (3.2) *with* $\eta_t=\eta/\sqrt{T}$ *and* $\eta\leq\sqrt{T}/(2L)$. *Then for any* $\delta\in(0,1)$, *we can choose* $T\asymp nd^{-1}$ *to derive the following inequality with probability at least* $1-\delta$

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\mathbf{w}_t)\|_2^2=O\Big(n^{-\frac{1}{2}}\log^2(1/\delta)\big(d+\log(1/\delta)\big)^{\frac{1}{2}}\Big).\tag{5.4}$$

Furthermore, if f *takes the specific structure* (5.1) *we can choose* $T\asymp n$ *to derive the following inequality with probability at least* $1-\delta$

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\mathbf{w}_t)\|_2^2=O\Big(n^{-\frac{1}{2}}\log^{\frac{5}{2}}(1/\delta)\Big).\tag{5.5}$$

Remark 8. It is clear that the bound in (5.4) is dimension-dependent, which is due to the use of the uniform convergence approach. Eq. (5.5) further shows that this dependency on the dimension can be avoided for problems of a specific structure. Note the optimization errors of SGD with T iterations for nonconvex problems satisfy $\frac{1}{T}\sum_{t=1}^{T}\|\nabla F_S(\mathbf{w}_t)\|_2^2 = O(T^{-\frac{1}{2}})$ [26]. This is consistent with the error bounds $\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\mathbf{w}_t)\|_2^2 = O\big(n^{-\frac{1}{2}}\log^{\frac{5}{2}}(1/\delta)\big)$ in (5.5) by noting $T \asymp n$. Therefore, our analysis shows that the extension from optimization to generalization comes for free.

5.3 Gradient Dominated Problems

We now study the stability and generalization of SGD for pairwise learning with gradient-dominated objectives (or PL condition). PL condition is widely used in nonconvex learning [24, 33], and was shown to hold true for deep (linear) and shallow neural networks [10]. Intuitively speaking, PL condition means that the suboptimality measured by function values can be bounded by gradients.

Assumption 5 (Polyak-Lojasiewicz Condition)**.** Denote $\hat{F}_S = \inf_{\mathbf{w}'\in\mathcal{W}} F_S(\mathbf{w}')$. We assume F_S satisfies PL or gradient-dominated condition (in expectation) with parameter $\beta > 0$, i.e.,

$$\mathbb{E}_S\big[F_S(\mathbf{w}) - \hat{F}_S\big] \leq \frac{1}{2\beta}\mathbb{E}_S\big[\|\nabla F_S(\mathbf{w})\|_2^2\big], \quad \forall \mathbf{w} \in \mathcal{W}. \tag{5.6}$$

Under the PL condition, we can get excess population risk bounds based on the stability analysis. The proof of Theorem 15 is given in Section J.

Theorem 15 (Gradient Dominated Problems)**.** *Let Assumptions 2, 3, 5 hold. Assume* $|f(\mathbf{w}; z, z')| \leq B$ *for all* $\mathbf{w} \in \mathcal{W}, z, z' \in \mathcal{Z}$. *Let* $\{\mathbf{w}_t\}_t$ *be the sequence produced by* (3.2) *with* $\eta_t = \frac{2t+1}{2\beta(t+1)^2}$. *Then*

$$\mathbb{E}\big[F(\mathbf{w}_T)\big] - F(\mathbf{w}^*) = O\Big(\frac{T^{\frac{L}{L+\beta}}}{n}\Big) + O\big(1/(T\beta^2)\big). \tag{5.7}$$

We can choose $T \asymp n^{\frac{1+L/\beta}{1+2L/\beta}}\beta^{-\frac{2+2L/\beta}{1+2L/\beta}}$ *to get* $\mathbb{E}\big[F(\mathbf{w}_T)\big] - F(\mathbf{w}^*) = O\big(n^{-\frac{1+L/\beta}{1+2L/\beta}}\beta^{-\frac{2L/\beta}{1+2L/\beta}}\big)$.

Remark 9. Note the above bounds depend on the condition number $\mathrm{cond} := L/\beta$. If $\mathrm{cond} \approx 1$, then we get $\mathbb{E}\big[F(\mathbf{w}_T)\big] - F(\mathbf{w}^*) \approx O\big(n^{-\frac{2}{3}}\big)$. As cond increases, the bound increases to $O(n^{-\frac{1}{2}})$.

6 Conclusion

In this paper, we present a systematic study on the generalization performance for pairwise learning. We develop novel high-probability bounds for uniformly stable algorithms, and apply them to develop algorithms with optimal bounds with $O(n)$ gradient computations for nonsmooth problems. We conduct the stability analysis for various convex problems including smooth, nonsmooth and strongly convex objectives, and get optimal excess population risk bounds of the order $O(1/\sqrt{n})$ for convex problems and $O(1/(n\sigma))$ for σ-strongly convex problems, respectively. We conduct the uniform convergence analysis for general nonconvex problems, which imply the bounds of the order $O(1/\sqrt{n})$ for population gradients. We further study the stability and generalization for nonconvex pairwise learning with gradient dominated objectives. Our discussions can clarify the effect of interpolation on generalization. In Section L we present preliminary experimental results to verify our stability bounds.

It would be interesting to study the stability of SGD in a general nonconvex case for getting dimension-independent bounds. It would also be very interesting to study other stochastic optimization methods for pairwise learning, including variance reduction variants and momentum technique.

Acknowledgments

We are also grateful to the anonymous reviewers and area chairs for their insightful and constructive comments. Yiming's work is supported by National Science Foundation (NSF) under grants DMS-2110836, IIS-1816227, IIS-2110546, IIS-2103450.

References

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

[2] H. Asi, V. Feldman, T. Koren, and K. Talwar. Private stochastic convex optimization: Optimal rates in ℓ_1 geometry. *arXiv preprint arXiv:2103.01516*, 2021.

[3] R. Bassily, V. Feldman, C. Guzmán, and K. Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. *Advances in Neural Information Processing Systems*, 33, 2020.

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

[5] O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

[6] O. Bousquet, Y. Klochkov, and N. Zhivotovskiy. Sharper bounds for uniformly stable algorithms. In *Conference on Learning Theory*, pages 610–626, 2020.

[7] R. Camoriano, T. Angles, A. Rudi, and L. Rosasco. Nytro: When subsampling meets early stopping. In *Artificial Intelligence and Statistics*, pages 1403–1411. PMLR, 2016.

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

[9] L. Carratino, A. Rudi, and L. Rosasco. Learning with sgd and random features. In *Advances in Neural Information Processing Systems*, pages 10213–10224, 2018.

[10] Z. Charles and D. Papailiopoulos. Stability and generalization of learning algorithms that converge to global optima. In *International Conference on Machine Learning*, pages 744–753, 2018.

[11] Y. Chen, C. Jin, and B. Yu. Stability and convergence trade-off of iterative optimization algorithms. *arXiv preprint arXiv:1804.01619*, 2018.

[12] A. Christmann and D.-X. Zhou. On the robustness of regularized pairwise learning methods based on kernels. *Journal of Complexity*, 37:1–33, 2016.

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

[14] C. Cortes and M. Mohri. AUC optimization vs. error rate minimization. In *Advances in Neural Information Processing Systems*, pages 313–320, 2004.

[15] A. Cutkosky. Anytime online-to-batch, optimism and acceleration. In *International Conference on Machine Learning*, pages 1446–1454. PMLR, 2019.

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

[17] M. Dong, X. Yang, R. Zhu, Y. Wang, and J. Xue. Generalization bound of gradient descent for non-convex metric learning. In *Advances in Neural Information Processing Systems*, 2020.

[18] A. Elisseeff, T. Evgeniou, and M. Pontil. Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6(Jan):55–79, 2005.

[19] J. Fan, T. Hu, Q. Wu, and D.-X. Zhou. Consistency analysis of an empirical minimum error entropy algorithm. *Applied and Computational Harmonic Analysis*, 41(1):164–189, 2016.

[20] B. Fehrman, B. Gess, and A. Jentzen. Convergence rates for the stochastic gradient descent method for non-convex objective functions. *Journal of Machine Learning Research*, 21, 2020.

[21] V. Feldman. Generalization of erm in stochastic convex optimization: The dimension strikes back. In *Advances in Neural Information Processing Systems*, pages 3576–3584, 2016.

[22] V. Feldman and J. Vondrak. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Conference on Learning Theory*, pages 1270–1279, 2019.

[23] V. Feldman, T. Koren, and K. Talwar. Private stochastic convex optimization: optimal rates in linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 439–449, 2020.

[24] D. J. Foster, A. Sekhari, and K. Sridharan. Uniform convergence of gradients for non-convex learning and optimization. In *Advances in Neural Information Processing Systems*, pages 8759–8770, 2018.

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

[26] S. Ghadimi and G. Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[27] Z.-C. Guo, D.-H. Xiang, X. Guo, and D.-X. Zhou. Thresholded spectral algorithms for sparse approximations. *Analysis and Applications*, 15(03):433–455, 2017.

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

[29] T. Hu, J. Fan, Q. Wu, and D.-X. Zhou. Regularization schemes for minimum error entropy principle. *Analysis and Applications*, 13(04):437–455, 2015.

[30] M. Huai, D. Wang, C. Miao, J. Xu, and A. Zhang. Pairwise learning with differential privacy guarantees. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 694–701, 2020.

[31] R. Jin, S. Wang, and Y. Zhou. Regularized distance metric learning: Theory and algorithm. In *Advances in Neural Information Processing Systems*, pages 862–870, 2009.

[32] 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.

[33] H. Karimi, J. Nutini, and M. Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *European Conference on Machine Learning*, pages 795–811, 2016.

[34] Y. Klochkov and N. Zhivotovskiy. Stability and deviation optimal risk bounds with convergence rate O(1/n). *arXiv preprint arXiv:2103.12024*, 2021.

[35] J. Kulkarni, Y. T. Lee, and D. Liu. Private non-smooth empirical risk minimization and stochastic convex optimization in subquadratic steps. *arXiv preprint arXiv:2103.15352*, 2021.

[36] I. Kuzborskij and C. Lampert. Data-dependent stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 2820–2829, 2018.

[37] I. Kuzborskij and F. Orabona. Stability and hypothesis transfer learning. In *International Conference on Machine Learning*, pages 942–950. PMLR, 2013.

[38] Y. Lei and K. Tang. Learning rates for stochastic gradient descent with nonconvex objectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. doi: 10.1109/TPAMI.2021.3068154.

[39] Y. Lei and Y. Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *International Conference on Machine Learning*, pages 5809–5819, 2020.

[40] Y. Lei, A. Ledent, and M. Kloft. Sharper generalization bounds for pairwise learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 21236–21246, 2020.

[41] J. Li, X. Luo, and M. Qiao. On generalization error bounds of noisy gradient methods for non-convex learning. In *International Conference on Learning Representations*, 2020.

[42] M. Liu, X. Zhang, Z. Chen, X. Wang, and T. Yang. Fast stochastic AUC maximization with O(1/n)-convergence rate. In *International Conference on Machine Learning*, pages 3195–3203, 2018.

[43] T. Liu, G. Lugosi, G. Neu, and D. Tao. Algorithmic stability and hypothesis complexity. In *International Conference on Machine Learning*, pages 2159–2167, 2017.

[44] B. London, B. Huang, and L. Getoor. Stability and generalization in structured prediction. *The Journal of Machine Learning Research*, 17(1):7808–7859, 2016.

[45] S. Ma, R. Bassily, and M. Belkin. The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning. In *International Conference on Machine Learning*, pages 3325–3334, 2018.

[46] S. Mei, Y. Bai, and A. Montanari. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A):2747–2774, 2018.

[47] W. Mou, L. Wang, X. Zhai, and K. Zheng. Generalization bounds of sgld for non-convex learning: Two theoretical viewpoints. In *Conference on Learning Theory*, pages 605–638, 2018.

[48] L. Oneto, M. Donini, M. Pontil, and J. Shawe-Taylor. Randomized learning and generalization of fair and private classifiers: From pac-bayes to stability and differential privacy. *Neurocomputing*, 416:231–243, 2020.

[49] F. Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

[50] G. Papa, S. Clémençon, and A. Bellet. SGD algorithms based on incomplete U-statistics: large-scale minimization of empirical risk. In *Neural Information Processing Systems*, pages 1027–1035, 2015.

[51] V. Patel, B. Tian, and S. Zhang. Global convergence and stability of stochastic gradient descent. *arXiv preprint arXiv:2110.01663*, 2021.

[52] A. Rakhlin, S. Mukherjee, and T. Poggio. Stability results in learning theory. *Analysis and Applications*, 3(04):397–417, 2005.

[53] N. L. Roux, M. Schmidt, and F. Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. In *Advances in Neural Information Processing Systems*, pages 2663–2671, 2012.

[54] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11:2635–2670, 2010.

[55] W. Shen, Z. Yang, Y. Ying, and X. Yuan. Stability and optimization error of stochastic gradient descent for pairwise learning. *Analysis and Applications*, pages 1–41, 2019.

[56] N. Srebro, K. Sridharan, and A. Tewari. Smoothness, low noise and fast rates. In *Advances in Neural Information Processing Systems*, pages 2199–2207, 2010.

[57] B. Wang, H. Zhang, P. Liu, Z. Shen, and J. Pineau. Multitask metric learning: Theory and algorithm. In *International Conference on Artificial Intelligence and Statistics*, pages 3362–3371, 2019.

[58] Y. Wang, R. Khardon, D. Pechyony, and R. Jones. Generalization bounds for online learning algorithms with pairwise loss functions. In *Conference on Learning Theory*, volume 23, pages 13–1, 2012.

[59] Z. Yang, Y. Lei, S. Lyu, and Y. Ying. Stability and differential privacy of stochastic gradient descent for pairwise learning with non-smooth loss. In *International Conference on Artificial Intelligence and Statistics*, pages 2026–2034. PMLR, 2021.

[60] H.-J. Ye, D.-C. Zhan, and Y. Jiang. Fast generalization rates for distance metric learning. *Machine Learning*, 108(2):267–295, 2019.

[61] Y. Ying and C. Campbell. Rademacher chaos complexities for learning the kernel problem. *Neural computation*, 22(11):2858–2886, 2010.

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

[63] Y. Ying, L. Wen, and S. Lyu. Stochastic online AUC maximization. In *Advances in Neural Information Processing Systems*, pages 451–459, 2016.

[64] L. Zhang, T. Yang, and R. Jin. Empirical risk minimization for stochastic convex optimization: $\mathrm{O}(1/n)$-and $\mathrm{O}(1/n^2)$-type of risk bounds. In *Conference on Learning Theory*, pages 1954–1979, 2017.

[65] T. Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *International Conference on Machine Learning*, pages 919–926, 2004.

[66] P. Zhao, S. C. Hoi, R. Jin, and T. Yang. Online AUC maximization. In *International Conference on Machine Learning*, pages 233–240, 2011.

[67] Y. Zhou, Y. Liang, and H. Zhang. Generalization error bounds with probabilistic guarantee for SGD in nonconvex optimization. *arXiv preprint arXiv:1802.06903*, 2018.

[68] Y. Zhou, B. Karimi, J. Yu, Z. Xu, and P. Li. Towards better generalization of adaptive gradient methods. *Advances in Neural Information Processing Systems*, 33, 2020.

Appendix for "Generalization Guarantee of SGD for Pairwise Learning"

Yunwen Lei[1] **Mingrui Liu**[2] **Yiming Ying**[3]
[1]School of Computer Science, University of Birmingham, Birmingham B15 2TT, United Kingdom
[2]Department of Computer Science, George Mason University, Fairfax, VA 22030, USA
[3]Department of Mathematics and Statistics, State University of New York at Albany, USA
y.lei@bham.ac.uk mingruil@gmu.edu yying@albany.edu

A Table of Notations

We collect in Table A.1 the notations of performance measures used in this paper.

$\mathcal{X}$	input space	$\mathcal{Y}$	output space	$\mathcal{Z}$	sample space
S	training dataset	n	sample size	z_i	i-th training example
$f(\mathbf{w}; z, z')$	loss function	F_S	training risk	F	population risk
$\mathbf{w}_S^*$	$\arg\min_{\mathbf{w}} F_S(\mathbf{w})$	$\mathbf{w}^*$	$\arg\min_{\mathbf{w}} F(\mathbf{w})$	$A(S)$	output of algorithm A to S
L	smoothness parameter	G	Lipschitz parameter	σ	strong convexity parameter
η_t	step size	T	largest iteration number	i_t	randomly selected index
b	$\sup_{z,z'\in\mathcal{Z}} f(0; z, z')$	b'	$\sup_{z,z'\in\mathcal{Z}} \|\nabla f(0; z, z')\|_2$	ρ	probability measure

Table A.1: Table of Notations.

B Proof of Theorem 1

In this section, we prove Theorem 1 on the connection between on-average stability and generalization bounds, following the arguments in [15]. To this aim, we require the following lemma on the self-bounding property of smooth loss functions.

Lemma B.1 ([20]). *Assume for all z, z', the function $\mathbf{w} \mapsto f(\mathbf{w}; z, z')$ is nonnegative and L-smooth. Then $\|\nabla f(\mathbf{w}; z, z')\|_2^2 \leq 2Lf(\mathbf{w}; z, z')$.*

Proof of Theorem 1. Part (a) was established in [16]. We only consider Part (b). According to the symmetry between z_i, z_j and z_i', z_j', we know

$$\mathbb{E}[F(A(S)) - F_S(A(S))] = \frac{1}{n(n-1)} \sum_{i,j\in[n]:i\neq j} \mathbb{E}\big[F(A(S_{i,j})) - F_S(A(S))\big]$$
$$= \frac{1}{n(n-1)} \sum_{i,j\in[n]:i\neq j} \mathbb{E}\big[f(A(S_{i,j}); z_i, z_j) - f(A(S); z_i, z_j)\big], \quad \text{(B.1)}$$

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

where we have used $\mathbb{E}_{z_i,z_j}\big[f(A(S_{i,j});z_i,z_j)\big] = F(A(S_{i,j}))$ since z_i, z_j are independent of $A(S_{i,j})$. By the L-smoothness of f, we know

$$
\begin{aligned}
&f(A(S_{i,j});z_i,z_j) - f(A(S);z_i,z_j) \le \langle A(S_{i,j}) - A(S), \nabla f(A(S);z_i,z_j)\rangle + \frac{L}{2}\|A(S_{i,j}) - A(S)\|_2^2 \\
&\le \|A(S_{i,j}) - A(S)\|_2 \|\nabla f(A(S);z_i,z_j)\|_2 + \frac{L}{2}\|A(S_{i,j}) - A(S)\|_2^2 \\
&\le \frac{\gamma}{2}\|A(S_{i,j}) - A(S)\|_2^2 + \frac{1}{2\gamma}\|\nabla f(A(S);z_i,z_j)\|_2^2 + \frac{L}{2}\|A(S_{i,j}) - A(S)\|_2^2 \\
&\le \frac{L+\gamma}{2}\|A(S_{i,j}) - A(S)\|_2^2 + \frac{L}{\gamma} f(A(S);z_i,z_j) \\
&\le (L+\gamma)\|A(S_{i,j}) - A(S_i)\|_2^2 + (L+\gamma)\|A(S_i) - A(S)\|_2^2 + \frac{L}{\gamma} f(A(S);z_i,z_j),
\end{aligned}
$$

where we have used Lemma B.1 in the last second inequality and the following inequality in the last step

$$\|A(S_{i,j}) - A(S)\|_2^2 \le 2\|A(S_{i,j}) - A(S_i)\|_2^2 + 2\|A(S_i) - A(S)\|_2^2.$$

Since $\mathbb{E}[\|A(S_{i,j}) - A(S_i)\|_2^2] = \mathbb{E}[\|A(S_j) - A(S)\|_2^2]$, we know

$$
\begin{aligned}
\mathbb{E}\big[f(A(S_{i,j});z_i,z_j) - f(A(S);z_i,z_j)\big] &\le (L+\gamma)\mathbb{E}[\|A(S_i) - A(S)\|_2^2] \\
&\quad + (L+\gamma)\mathbb{E}[\|A(S_j) - A(S)\|_2^2] + \frac{L}{\gamma}\mathbb{E}[f(A(S);z_i,z_j)].
\end{aligned}
$$

We can plug the above inequality back into (B.1), and get

$$
\begin{aligned}
\mathbb{E}[F(A(S)) - F_S(A(S))] &\le \frac{1}{n(n-1)} \sum_{i,j\in[n]:i\ne j} \Big(2(L+\gamma)\mathbb{E}[\|A(S_i) - A(S)\|_2^2] + \frac{L}{\gamma}\mathbb{E}[f(A(S);z_i,z_j)]\Big) \\
&= \frac{2(L+\gamma)}{n}\sum_{i=1}^n \mathbb{E}[\|A(S_i) - A(S)\|_2^2] + \frac{L}{\gamma}\mathbb{E}[F_S(A(S))].
\end{aligned}
$$

The proof is complete. □

C Proof of Theorem 2

In this section, we prove Theorem 2. To this aim, we first introduce some lemmas. The following lemma provides moment bounds for a summation of weakly dependent and mean-zero random functions with bounded increments under a change of any single coordinate [1, 10]. We denote by $S\backslash\{z_i\}$ the set $\{z_1,\ldots,z_{i-1},z_{i+1},\ldots,z_n\}$. The L_p-norm of a real-valued random variable Z is denoted by $\|Z\|_p := \big(\mathbb{E}[|Z|^p]\big)^{\frac{1}{p}}, p \ge 1$.

Lemma C.1 ([1]). *Let $S = \{z_1,\ldots,z_n\}$ be a set of independent random variables each taking values in $\mathcal{Z}$ and $M \ge 0$. Let $h_1,\ldots,h_n$ be some functions $h_i : \mathcal{Z}^n \mapsto \mathbb{R}$ such that the following holds for any $i \in [n]$*

1. *$\big|\mathbb{E}_{S\backslash\{z_i\}}[h_i(S)]\big| \le M$ almost surely (a.s.),*
2. *$\mathbb{E}_{z_i}\big[h_i(S)\big] = 0$ a.s.,*
3. *for any $j \in [n]$ with $j \ne i$, and $z_j'' \in \mathcal{Z}$*

$$\big|h_i(S) - h_i(z_1,\ldots,z_{j-1},z_j'',z_{j+1},\ldots,z_n)\big| \le \beta. \tag{C.1}$$

Then, for any $p \ge 2$

$$\Big\|\sum_{i=1}^n h_i(S)\Big\|_p \le 12\sqrt{2}pn\beta\lceil\log_2 n\rceil + 4M\sqrt{pn}.$$

The bounds on moments of random variables can be used to establish concentration inequalities, as shown in the following lemma [1, 10].

Lemma C.2. *Let $a, b \in \mathbb{R}_+$ and $\delta \in (0, 1/e)$. Let Z be a random variable with $\|Z\|_p \le \sqrt{p}a + pb$ for any $p \ge 2$. Then with probability at least $1 - \delta$*

$$|Z| \le e\Big(a\sqrt{\log(e/\delta)} + b\log(e/\delta)\Big).$$

The following lemma relates $F(A(S) - F_S(A(S)) - \mathbb{E}[F(A(S))]$ to $\mathbb{E}_{S'}\big[f(A(S'); z_i, z_j)\big]$. A notable property is that $A(S')$ is independent of S and therefore can be considered as a fixed point, which simplifies the application of concentration inequalities. Lemma C.3 is motivated by a recent work in pointwise learning [10].

Lemma C.3. *Let A be an ϵ-uniformly stable deterministic algorithm. Let $S = \{z_1, \ldots, z_n\}, S' = \{z'_1, \ldots, z'_n\}$ be independent datasets. Then for any $p \ge 2$ there holds*

$$\Big\|F(A(S)-F_S(A(S))-\mathbb{E}[F(A(S))]+\frac{1}{n(n-1)}\sum_{i\neq j}\mathbb{E}_{S'}\left[f(A(S');z_i,z_j)\right]\Big\|_p \le 4\epsilon+96\sqrt{2}p\epsilon\lceil\log_2 n/2\rceil.$$

Proof. Let $p \ge 2$ be any number. It was shown that [16]

$$\Big|\mathbb{E}_{Z,\tilde{Z}}\big[f(A(S); Z, \tilde{Z})\big] - \frac{1}{n(n-1)}\sum_{i\neq j} f(A(S); z_i, z_j) - \frac{1}{n(n-1)}\sum_{i\neq j} g_{i,j}(S)\Big| \le 4\epsilon, \quad \text{(C.2)}$$

where we introduce

$$g_{i,j}(S) = \mathbb{E}_{z'_i,z'_j}\Big[\mathbb{E}_{Z,\tilde{Z}}\big[f(A(S_{i,j}); Z, \tilde{Z})\big] - f(A(S_{i,j}); z_i, z_j)\Big], \quad \forall i, j \in [n]$$

and $S_{i,j}$ is defined in Eq. (3.4). For any $i \neq j \in [n]$, define

$$h_{i,j}(S) := g_{i,j}(S) - \mathbb{E}_{S\backslash\{z_i \cup z_j\}} g_{i,j}(S),$$

from which and (C.2) we get the following inequality for any $p \ge 1$

$$\Big\|F(A(S) - F_S(A(S)) - \frac{1}{n(n-1)}\sum_{i\neq j}\mathbb{E}_{S\backslash\{z_i\cup z_j\}} g_{i,j}(S)\Big\|_p \le 4\epsilon + \frac{1}{n(n-1)}\Big\|\sum_{i\neq j} h_{i,j}(S)\Big\|_p. \quad \text{(C.3)}$$

We have the following representation of U-statistic [3]

$$\frac{1}{n(n-1)}\sum_{i\neq j} h_{i,j}(S) = \frac{1}{n!}\sum_{\pi}\frac{1}{\lfloor\frac{n}{2}\rfloor}\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor} h_{\pi(i),\pi(i+\lfloor\frac{n}{2}\rfloor)}(S),$$

where the sum is taken over all permutations π of $\{1, \ldots, n\}$. It then follows from Jensen's inequality that

$$\frac{1}{n(n-1)}\big\|\sum_{i\neq j} h_{i,j}(S)\big\|_p \le \frac{1}{n!}\sum_{\pi}\frac{1}{\lfloor\frac{n}{2}\rfloor}\Big\|\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor} h_{\pi(i),\pi(i+\lfloor\frac{n}{2}\rfloor)}(S)\Big\|_p = \frac{1}{\lfloor\frac{n}{2}\rfloor}\Big\|\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor} h_{i,i+\lfloor\frac{n}{2}\rfloor}(S)\Big\|_p, \quad \text{(C.4)}$$

where the last identity is due to the symmetry of permutations (note $\|\cdot\|_p$ involves an expectation). It is clear that

$$\mathbb{E}_{S\backslash\{z_i\cup z_{i+\lfloor\frac{n}{2}\rfloor}\}} h_{i,i+\lfloor\frac{n}{2}\rfloor}(S) = \mathbb{E}_{S\backslash\{z_i\cup z_{i+\lfloor\frac{n}{2}\rfloor}\}}\Big[g_{i,i+\lfloor\frac{n}{2}\rfloor}(S) - \mathbb{E}_{S\backslash\{z_i\cup z_{i+\lfloor\frac{n}{2}\rfloor}\}} g_{i,i+\lfloor\frac{n}{2}\rfloor}(S)\Big] = 0, \quad \text{(C.5)}$$

where $\mathbb{E}_{S\backslash\{z_i\cup z_{i+\lfloor\frac{n}{2}\rfloor}\}}$ denotes the expectation w.r.t. $S\backslash\{z_i \cup z_{i+\lfloor\frac{n}{2}\rfloor}\}$. Furthermore, there holds

$$\begin{aligned}&\mathbb{E}_{z_i\cup z_{i+\lfloor\frac{n}{2}\rfloor}}[g_{i,i+\lfloor\frac{n}{2}\rfloor}(S)]\\ &= \mathbb{E}_{z_i\cup z_{i+\lfloor\frac{n}{2}\rfloor}}\mathbb{E}_{z'_i,z'_{i+\lfloor\frac{n}{2}\rfloor}}\Big[\mathbb{E}_{Z,\tilde{Z}}\big[f(A(S_{i,i+\lfloor\frac{n}{2}\rfloor}); Z, \tilde{Z})\big] - f(A(S_{i,i+\lfloor\frac{n}{2}\rfloor}); z_i, z_{i+\lfloor\frac{n}{2}\rfloor})\Big] = 0.\end{aligned} \quad \text{(C.6)}$$

For any $k \in [\lfloor\frac{n}{2}\rfloor]$ with $k \neq i$ and $z''_k, z''_{k+\lfloor\frac{n}{2}\rfloor} \in \mathcal{Z}$, it is clear from the uniform stability of A that

$$\Big|\mathbb{E}_{z'_i,z'_{i+\lfloor\frac{n}{2}\rfloor}}\mathbb{E}_{Z,\tilde{Z}}\big[f(A(S_{i,i+\lfloor\frac{n}{2}\rfloor}); Z, \tilde{Z})\big] - \mathbb{E}_{z'_i,z'_{i+\lfloor\frac{n}{2}\rfloor}}\mathbb{E}_{Z,\tilde{Z}}\big[f(A(S^{(k,k+\lfloor\frac{n}{2}\rfloor)}_{i,i+\lfloor\frac{n}{2}\rfloor}); Z, \tilde{Z})\big]\Big| \le 2\epsilon,$$

where $S_{i,i+\lfloor\frac{n}{2}\rfloor}^{(k,k+\lfloor\frac{n}{2}\rfloor)}$ is the set derived by replacing the k-th element of $S_{i,i+\lfloor\frac{n}{2}\rfloor}$ with z''_k and $k+\lfloor\frac{n}{2}\rfloor$-th element with $z''_{k+\lfloor\frac{n}{2}\rfloor}$. In a similar way, one can show

$$\Big|\mathbb{E}_{z'_i,z'_{i+\lfloor\frac{n}{2}\rfloor}}\big[f(A(S_{i,i+\lfloor\frac{n}{2}\rfloor});z_i,z_{i+\lfloor\frac{n}{2}\rfloor})\big]-\mathbb{E}_{z'_i,z'_{i+\lfloor\frac{n}{2}\rfloor}}\big[f(A(S_{i,i+\lfloor\frac{n}{2}\rfloor}^{(k,k+\lfloor\frac{n}{2}\rfloor)});z_i,z_{i+\lfloor\frac{n}{2}\rfloor})\big]\Big|\le 2\epsilon.$$

We can combine the above two inequalities together and get

$$\Big|g_{i,i+\lfloor\frac{n}{2}\rfloor}(S)-g_{i,i+\lfloor\frac{n}{2}\rfloor}(S^{(k,k+\lfloor\frac{n}{2}\rfloor)})\Big|\le 4\epsilon,$$

where $S^{(k,k+\lfloor\frac{n}{2}\rfloor)}$ is the set derived by replacing the k-th element of S with z''_k and $k+\lfloor\frac{n}{2}\rfloor$-th element with $z''_{k+\lfloor\frac{n}{2}\rfloor}$ Similarly, one can show

$$\Big|\mathbb{E}_{S\backslash\{z_i\cup z_{i+\lfloor\frac{n}{2}\rfloor}\}}\big[g_{i,i+\lfloor\frac{n}{2}\rfloor}(S)\big]-\mathbb{E}_{S\backslash\{z_i\cup z_{i+\lfloor\frac{n}{2}\rfloor}\}}\big[g_{i,i+\lfloor\frac{n}{2}\rfloor}(S^{(k,k+\lfloor\frac{n}{2}\rfloor)})\big]\Big|\le 4\epsilon.$$

We can combine the above two inequalities together and get

$$\Big|h_{i,i+\lfloor\frac{n}{2}\rfloor}(S)-h_{i,i+\lfloor\frac{n}{2}\rfloor}(S^{(k,k+\lfloor\frac{n}{2}\rfloor)})\Big|\le 8\epsilon. \tag{C.7}$$

According to (C.5), (C.6) and (C.7), we know that the conditions of Lemma C.1 hold with $M=0,n=\lfloor\frac{n}{2}\rfloor,\beta=8\epsilon,z_i=z_i\cup z_{i+\lfloor\frac{n}{2}\rfloor}$ and $h_i(S)=h_{i,i+\lfloor\frac{n}{2}\rfloor}(S)$. Therefore, one can apply Lemma C.1 to show that

$$\frac{1}{\lfloor\frac{n}{2}\rfloor}\Big\|\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor}h_{i,i+\lfloor\frac{n}{2}\rfloor}(S)\Big\|_p\le 96\sqrt{2}p\epsilon\lceil\log_2 n/2\rceil.$$

We can plug the above inequality and (C.4) back into (C.3) and get the following inequality for any $p\ge 2$

$$\Big\|F(A(S)-F_S(A(S))-\frac{1}{n(n-1)}\sum_{i\neq j}\mathbb{E}_{S\backslash\{z_i\cup z_j\}}g_{i,j}(S)\Big\|_p\le 4\epsilon+96\sqrt{2}p\epsilon\lceil\log_2 n/2\rceil. \tag{C.8}$$

Furthermore, the symmetry between S and S' implies (note $\mathbb{E}_{S'}[A(S');z_i,z_j]=\mathbb{E}_{S\backslash\{z_i\cup z_j\}}\mathbb{E}_{z'_i,z'_j}[f(A(S_{i,j});z_i,z_j)]$)

$$\begin{aligned}\mathbb{E}_{S\backslash\{z_i\cup z_j\}}[g_{i,j}(S)]&=\mathbb{E}_{S\backslash\{z_i\cup z_j\}}\mathbb{E}_{z'_i,z'_j}\Big[\mathbb{E}_{Z,\tilde{Z}}\big[f(A(S_{i,j});Z,\tilde{Z})\big]-f(A(S_{i,j});z_i,z_j)\Big]\\&=\mathbb{E}[F(A(S))]-\mathbb{E}_{S'}\big[f(A(S');z_i,z_j)\big].\end{aligned}$$

The stated bound then follows by combining the above two inequalities together. The proof is complete. □

We require a Bernstein inequality for U-Statistic [3] (inequality A.1 on page 868) to prove Theorem 2.

Lemma C.4 (Bernstein inequality for U-Statistic)**.** *Let $Z_1,\ldots,Z_n$ be independent variables taking values in $\mathcal{Z}$ and $q:\mathcal{Z}\times\mathcal{Z}\mapsto\mathbb{R}$. Let $B=\sup_{z,\tilde{z}}|q(z,\tilde{z})|$ and σ_0^2 be the variance of $q(Z,\tilde{Z})$. Then for any $\delta\in(0,1)$ with probability at least $1-\delta$*

$$\Big|\frac{1}{n(n-1)}\sum_{i,j\in[n]:i\neq j}q(Z_i,Z_j)-\mathbb{E}_{Z,\tilde{Z}}[q(Z,\tilde{Z})]\Big|\le\frac{2B\log(1/\delta)}{3\lfloor n/2\rfloor}+\sqrt{\frac{2\sigma_0^2\log(1/\delta)}{\lfloor n/2\rfloor}}. \tag{C.9}$$

Proof of Theorem 2. Let $S=\{z_1,\ldots,z_n\},S'=\{z'_1,\ldots,z'_n\}$ be independent datasets. We have the following error decomposition

$$\begin{aligned}F(A(S))-F_S(A(S))-F(\mathbf{w}^*)+F_S(\mathbf{w}^*)=\xi+&\\ \mathbb{E}_S[F(A(S))]-F(\mathbf{w}^*)-\frac{1}{n(n-1)}\sum_{i\neq j}&\mathbb{E}_{S'}\big[f(A(S');z_i,z_j)\big]+F_S(\mathbf{w}^*),\end{aligned}$$

where

$$\xi = F(A(S) - F_S(A(S)) - \mathbb{E}_S[F(A(S))] + \frac{1}{n(n-1)} \sum_{i \neq j} \mathbb{E}_{S'}\big[f(A(S'); z_i, z_j)\big].$$

Due to the symmetry between S and S' we further get

$$F(A(S)) - F_S(A(S)) - F(\mathbf{w}^*) + F_S(\mathbf{w}^*) = \xi + \mathbb{E}_{S'}[F(A(S'))] - F(\mathbf{w}^*) - \mathbb{E}_{S'}[F_S(A(S'))] + F_S(\mathbf{w}^*)$$

and therefore

$$F(A(S)) - F_S(A(S)) - F(\mathbf{w}^*) + F_S(\mathbf{w}^*) = \xi + \mathbb{E}_{S'}\Big[F(A(S')) - F(\mathbf{w}^*) - F_S(A(S')) + F_S(\mathbf{w}^*)\Big]. \tag{C.10}$$

Note $A(S')$ is independent of S and can be considered as a fixed model if we only consider the randomness induced from S. We now apply a concentration inequality to study the behavior of $\mathbb{E}_{S'}\big[F(A(S')) - F(\mathbf{w}^*) - F_S(A(S')) + F_S(\mathbf{w}^*)\big]$. For any z, z', define

$$q(z, z') = \mathbb{E}_{S'}[f(A(S'); z, z')] - f(\mathbf{w}^*; z, z').$$

Then it is clear

$$\mathbb{E}_{S'}\Big[F(A(S')) - F(\mathbf{w}^*) - F_S(A(S')) + F_S(\mathbf{w}^*)\Big] = \mathbb{E}_{Z,Z'}[q(Z, Z')] - \frac{1}{n(n-1)} \sum_{i \neq j} q(z_i, z_j).$$

The variance of q can be bounded by

$$\begin{aligned} \mathbb{E}_{Z,Z'}[q^2(Z, Z')] &\leq \mathbb{E}_{Z,Z'}\mathbb{E}_{S'}\big[\big(f(A(S'); Z, Z') - f(\mathbf{w}^*; Z, Z')\big)^2\big] \\ &= \mathbb{E}_{Z,Z',S}\big[\big(f(A(S); Z, Z') - f(\mathbf{w}^*; Z, Z')\big)^2\big], \end{aligned}$$

where we have used the symmetry between S and S' as well as the Jensen's inequality. We can apply Lemma C.4 with the above q to show the following inequality with probability at least $1 - \delta/2$

$$\mathbb{E}_{S'}\Big[F(A(S')) - F(\mathbf{w}^*) - F_S(A(S')) + F_S(\mathbf{w}^*)\Big] \leq \frac{2B \log(2/\delta)}{3\lfloor n/2 \rfloor} + \sqrt{\frac{2\sigma_0^2 \log(2/\delta)}{\lfloor n/2 \rfloor}}.$$

Furthermore, Lemma C.3 implies that $\|\xi\|_p \leq 2p\epsilon\big(1 + 48\sqrt{2}\lceil \log_2 n/2 \rceil\big)$ for any $p \geq 2$, from which and Lemma C.2 we derive the following inequality with probability at least $1 - \delta/2$

$$\xi \leq 2e\epsilon\big(1 + 48\sqrt{2}\lceil \log_2 n/2 \rceil\big) \log(2e/\delta).$$

We can combine the above two inequalities and Eq. (C.10) together and derive the stated inequality with probability at least $1 - \delta$. The proof is complete. □

D Optimization Errors

The following lemma provides the optimization error bounds of SGD for convex, strongly convex and nonconvex problems. The optimization error analysis of SGD (Algorithm 1) for pairwise learning is the same as that for pointwise learning. The underlying reason is that both algorithms build an unbiased estimator (stochastic gradient) of the true gradient, and perform the update along the negative direction of the stochastic gradient. Part (a) is standard, see, e.g., [17]. Part (b) was given in [15]. Part (c) can be found in [7, 12]. Part (d) was given in [14]. Part (e) can be found in [9].

Lemma D.1. *Let $\{\mathbf{w}_t\}_t$ be produced by* (3.2) *and $\mathbf{w} \in \mathcal{W}$ be independent of SGD.*

(a) Let $\mathbf{w}_t^{(1)} = \big(\sum_{j=1}^t \eta_j \mathbf{w}_j\big) / \sum_{j=1}^t \eta_j$. If F_S is convex and Assumption 2 holds, then for all $t \in \mathbb{N}$

$$\mathbb{E}_A[F_S(\mathbf{w}_t^{(1)})] - F_S(\mathbf{w}) \leq \frac{G^2 \sum_{j=1}^t \eta_j^2 + \|\mathbf{w}\|_2^2}{2 \sum_{j=1}^t \eta_j}. \tag{D.1}$$

(b) Let Assumptions 1, 3 hold. If $\eta_t \le 1/(2L)$ and is nonincreasing, then for all $t \in \mathbb{N}$

$$\sum_{j=1}^{t} \eta_j \mathbb{E}_A[F_S(\mathbf{w}_j) - F_S(\mathbf{w})] \le (1/2 + L\eta_1)\|\mathbf{w}\|_2^2 + 2L\sum_{j=1}^{t} \eta_j^2 F_S(\mathbf{w}) \tag{D.2}$$

and

$$\sum_{j=1}^{t} \eta_j^2 \mathbb{E}_A[F_S(\mathbf{w}_j)] \le \eta_1\|\mathbf{w}\|_2^2 + 2\sum_{j=1}^{t} \eta_j^2 \mathbb{E}_A[F_S(\mathbf{w})]. \tag{D.3}$$

(c) Let F_S be σ-strongly convex and $\eta_t = 2/(\sigma(t+1))$. Let $\bar{\mathbf{w}}_t' = \big(\sum_{j=1}^t j\mathbf{w}_j\big)/\sum_{j=1}^t j$. If either Assumption 3 or Assumption 2 holds, then

$$\mathbb{E}_A[F_S(\bar{\mathbf{w}}_t')] - F_S(\mathbf{w}) = O\big(1/(t\sigma) + \|\mathbf{w}\|_2^2/t^2\big). \tag{D.4}$$

If Assumption 3 holds, then with probability at least $1-\delta$

$$F_S(\bar{\mathbf{w}}_t') - F_S(\mathbf{w}) = O\Big(\log(1/\delta)/(t\sigma)\Big). \tag{D.5}$$

(d) Let Assumptions 3, 4 hold and $\eta_j \le 1/(2L)$. For any $\delta \in (0,1)$, the following inequality holds with probability at least $1-\delta$

$$\sum_{j=1}^{t} \eta_j \|\nabla F_S(\mathbf{w}_j)\|_2^2 = O\Big(\sum_{j=1}^{t} \eta_j^2 + \log(1/\delta)\Big). \tag{D.6}$$

Furthermore, the following inequality holds with probability at least $1-\delta$ simultaneously for all $t = 1, \ldots, T$

$$\|\mathbf{w}_{t+1}\|_2 = O\Big(\Big(1+\sum_{k=1}^{T}\eta_k^2\Big)^{\frac{1}{2}}\Big(1+\sum_{k=1}^{t}\eta_k\Big)^{\frac{1}{2}} \log(1/\delta)\Big). \tag{D.7}$$

(e) Let Assumptions 3, 5 hold. If $\eta_t = \frac{2t+1}{2\beta(t+1)^2}$, then

$$\mathbb{E}_A[F_S(\mathbf{w}_t)] - \inf_{\mathbf{w}}[F_S(\mathbf{w})] = O\big(1/(t\beta^2)\big). \tag{D.8}$$

E Proofs on Smooth and Convex Problems

In this section, we present the proof related to stability and generalization for pairwise learning with convex and smooth loss functions. The following lemma shows the gradient map $\mathbf{w} \mapsto \mathbf{w} - \eta\nabla f(\mathbf{w}; z, z')$ is nonexpansive, which is very useful to study the stability bounds.

Lemma E.1 ([6]). *Assume for all $z \in \mathcal{Z}$, the function $\mathbf{w} \mapsto f(\mathbf{w}; z, z')$ is convex and L-smooth. Then for all $\eta \le 2/L$ and $z, z' \in \mathcal{Z}$ there holds*

$$\|\mathbf{w} - \eta\nabla f(\mathbf{w}; z, z') - \mathbf{w}' + \eta\nabla f(\mathbf{w}'; z, z')\|_2 \le \|\mathbf{w} - \mathbf{w}'\|_2.$$

Based on Lemma E.1, we can prove Theorem 3 on stability bounds.

Proof of Theorem 3. For any $i \in [n]$, define S_i as (3.3). Let $\{\mathbf{w}_t\}, \{\mathbf{w}_t^{(i)}\}$ be produced by SGD (Algorithm 1) w.r.t. S and S_i , respectively. For any S and $i \in [n]$, we denote

$$L_{S,i}(\mathbf{w}) = \sum_{j\in[n]:j\neq i} \big(f(\mathbf{w}; z_i, z_j) + f(\mathbf{w}; z_j, z_i)\big),\ L_{S_i,i}(\mathbf{w}) = \sum_{j\in[n]:j\neq i} \big(f(\mathbf{w}; z_i', z_j) + f(\mathbf{w}; z_j, z_i')\big). \tag{E.1}$$

If $i_t \neq i$ and $j_t \neq i$, it follows from (3.2) that

$$\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2^2 = \big\|\mathbf{w}_t - \mathbf{w}_t^{(i)} - \eta_t\nabla f(\mathbf{w}_t; z_{i_t}, z_{j_t}) + \eta_t\nabla f(\mathbf{w}_t^{(i)}; z_{i_t}, z_{j_t})\big\|_2^2 \le \|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2,$$

where we have used Lemma E.1 in the last inequality. If $i_t = i$, it follows from (3.2) that

$$\begin{aligned}\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2^2 &= \big\|\mathbf{w}_t - \mathbf{w}_t^{(i)} - \eta_t \nabla f(\mathbf{w}_t; z_i, z_{j_t}) + \eta_t \nabla f(\mathbf{w}_t^{(i)}; z_i', z_{j_t})\big\|_2^2 \\ &\le (1+p)\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2 + (1+1/p)\eta_t^2 \|\nabla f(\mathbf{w}_t; z_i, z_{j_t}) - \nabla f(\mathbf{w}_t^{(i)}; z_i', z_{j_t})\|_2^2 \\ &\le (1+p)\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2 + 2(1+1/p)\eta_t^2 \Big(\|\nabla f(\mathbf{w}_t; z_i, z_{j_t})\|_2^2 + \|\nabla f(\mathbf{w}_t^{(i)}; z_i', z_{j_t})\|_2^2\Big) \\ &\le (1+p)\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2 + 4L(1+1/p)\eta_t^2 \Big(f(\mathbf{w}_t; z_i, z_{j_t}) + f(\mathbf{w}_t^{(i)}; z_i', z_{j_t})\Big),\end{aligned}$$

where we have used the elementary inequality

$$(a+b)^2 \le (1+p)a^2 + (1+1/p)b^2, \quad \forall p > 0$$

and the self-bounding property (Lemma B.1). If $j_t = i$, we can similarly show that

$$\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2^2 \le (1+p)\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2 + 4L(1+1/p)\eta_t^2 \Big(f(\mathbf{w}_t; z_{i_t}, z_i) + f(\mathbf{w}_t^{(i)}; z_{i_t}, z_i')\Big).$$

Note the event $i_t \neq i$ and $j_t \neq i$ happens with the probability $\frac{(n-1)(n-2)}{n(n-1)}$, and $i_t = i, j_t = j$ for $i \neq j$ happens with probability $1/(n(n-1))$. We can combine the above three cases together and derive

$$\begin{aligned}\mathbb{E}_{k_t}[\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2^2] \le{} & \frac{(n-1)(n-2)}{n(n-1)}\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2 \\ &+ \frac{1}{n(n-1)} \sum_{j\in[n]:j\neq i} \Big((1+p)\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2 + 4L(1+1/p)\eta_t^2\big(f(\mathbf{w}_t; z_i, z_j) + f(\mathbf{w}_t^{(i)}; z_i', z_j)\big)\Big) \\ &+ \frac{1}{n(n-1)} \sum_{j\in[n]:j\neq i} \Big((1+p)\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2 + 4L(1+1/p)\eta_t^2\big(f(\mathbf{w}_t; z_j, z_i) + f(\mathbf{w}_t^{(i)}; z_j, z_i')\big)\Big) \\ &= \big(1 + 2p/n\big)\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2 + \frac{4L(1+1/p)\eta_t^2}{n(n-1)}\big(L_{S,i}(\mathbf{w}_t) + L_{S_i,i}(\mathbf{w}_t^{(i)})\big),\end{aligned}$$

where $\mathbb{E}_{k_t}$ means the conditional expectation w.r.t. $k_t := (i_t, j_t)$. It then follows that

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{k_t}[\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2^2] \le \frac{1}{n}\Big(1 + \frac{2p}{n}\Big)\sum_{i=1}^n \|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2 + \frac{4L(1+1/p)\eta_t^2}{n^2(n-1)}\sum_{i=1}^n \big(L_{S,i}(\mathbf{w}_t) + L_{S_i,i}(\mathbf{w}_t^{(i)})\big).$$

We can take expectation over both sides and get

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2^2] \le \frac{1}{n}\Big(1 + \frac{2p}{n}\Big)\sum_{i=1}^n \mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2] + \frac{8L(1+1/p)\eta_t^2}{n^2(n-1)}\sum_{i=1}^n \mathbb{E}\big[L_{S,i}(\mathbf{w}_t)\big],$$

where we have used the following identity due to the symmetry between z_i and z_i'

$$\mathbb{E}[L_{S_i,i}(\mathbf{w}_t^{(i)})] = \mathbb{E}\big[L_{S,i}(\mathbf{w}_t)\big].$$

According to the definition of $L_{S,i}$ we know

$$\sum_{i=1}^n L_{S,i}(\mathbf{w}) = \sum_{i=1}^n \sum_{j\in[n]:j\neq i} \big(f(\mathbf{w}; z_i, z_j) + f(\mathbf{w}; z_j, z_i)\big) = 2n(n-1)F_S(\mathbf{w}).$$

We can combine the above two equations together and get

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2^2] \le \frac{1}{n}\Big(1 + \frac{2p}{n}\Big)\sum_{i=1}^n \mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2] + \frac{16L(1+1/p)\eta_t^2}{n}\mathbb{E}\big[F_S(\mathbf{w}_t)\big].$$

We can apply the above inequality recursively and get

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2^2] \le \frac{16L(1+1/p)}{n}\sum_{j=1}^t \Big(1 + \frac{2p}{n}\Big)^{t-j}\eta_j^2 \mathbb{E}[F_S(\mathbf{w}_j)].$$

We can choose $p = n/(2t)$ in the above inequality and note

$$\Big(1 + \frac{2p}{n}\Big)^{t-j} \leq (1 + 1/t)^t \leq e.$$

It then follows that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2^2] \leq \frac{16L(1 + 2t/n)e}{n}\sum_{j=1}^{t}\eta_j^2\mathbb{E}[F_S(\mathbf{w}_j)].$$

The proof is complete. ☐

We now use the above stability bounds to prove generalization bounds in Theorem 4.

Proof of Theorem 4. We can plug the on-average argument stability bounds in Theorem 3 into Theorem 1 with $A(S) = \mathbf{w}_t$ and get

$$\mathbb{E}[F(\mathbf{w}_t)] \leq \frac{32L(L+\gamma)(1+2t/n)e}{n}\sum_{j=1}^{t-1}\eta_j^2\mathbb{E}[F_S(\mathbf{w}_j)] + \big(1 + L/\gamma\big)\mathbb{E}[F_S(\mathbf{w}_t)].$$

Multiplying both sides by η_t and taking a summation then gives

$$\sum_{t=1}^{T}\eta_t\mathbb{E}[F(\mathbf{w}_t)] \leq \big(1+L/\gamma\big)\sum_{t=1}^{T}\eta_t\mathbb{E}[F_S(\mathbf{w}_t)] + \frac{32L(L+\gamma)(1+2T/n)e}{n}\sum_{t=1}^{T}\eta_t\sum_{j=1}^{t-1}\eta_j^2\mathbb{E}[F_S(\mathbf{w}_j)].$$

It then follows that

$$\sum_{t=1}^{T}\eta_t\mathbb{E}[F(\mathbf{w}_t) - F_S(\mathbf{w})] \leq \big(1 + L/\gamma\big)\sum_{t=1}^{T}\eta_t\mathbb{E}[F_S(\mathbf{w}_t) - F_S(\mathbf{w})] + \\ L/\gamma\sum_{t=1}^{T}\eta_t\mathbb{E}[F_S(\mathbf{w})] + \frac{32L(L+\gamma)(1+2T/n)e}{n}\sum_{t=1}^{T}\eta_t\sum_{j=1}^{t-1}\eta_j^2\mathbb{E}[F_S(\mathbf{w}_j)].$$

According to (D.3) and $\eta_t = \eta$, the above inequality implies further

$$\sum_{t=1}^{T}\eta\mathbb{E}[F(\mathbf{w}_t) - F_S(\mathbf{w})] \leq \big(1 + L/\gamma\big)\sum_{t=1}^{T}\eta\mathbb{E}[F_S(\mathbf{w}_t) - F_S(\mathbf{w})] + \\ L/\gamma T\eta\mathbb{E}[F_S(\mathbf{w})] + \frac{32L(L+\gamma)(1+2T/n)e}{n}\sum_{t=1}^{T}\eta\big(\eta\mathbb{E}[\|\mathbf{w}\|_2^2] + 2t\eta^2\mathbb{E}[F_S(\mathbf{w})]\big).$$

We can plug (D.2) into the above inequality and get

$$\sum_{t=1}^{T}\eta\mathbb{E}[F(\mathbf{w}_t) - F_S(\mathbf{w})] \leq \big(1 + L/\gamma\big)\Big((1/2 + L\eta)\mathbb{E}[\|\mathbf{w}\|_2^2] + 2L\sum_{t=1}^{T}\eta^2\mathbb{E}[F_S(\mathbf{w})]\Big) + \\ L/\gamma T\eta\mathbb{E}[F_S(\mathbf{w})] + \frac{32L(L+\gamma)(1+2T/n)e}{n}\sum_{t=1}^{T}\eta\big(\eta\mathbb{E}[\|\mathbf{w}\|_2^2] + 2t\eta^2\mathbb{E}[F_S(\mathbf{w})]\big).$$

It then follows from the Jensen's inequality that

$$\mathbb{E}[F(\bar{\mathbf{w}}_T)] - \mathbb{E}[F_S(\mathbf{w})] = O\Big(\big(T\eta\big)^{-1}\big(\mathbb{E}[\|\mathbf{w}\|_2^2] + T\eta^2\mathbb{E}[F_S(\mathbf{w})]\big)\Big) + \frac{\mathbb{E}[F_S(\mathbf{w})]}{\gamma} + \\ O\Big(\frac{\gamma(1+T/n)}{n}\Big(\eta\mathbb{E}[\|\mathbf{w}\|_2^2] + T\eta^2\mathbb{E}[F_S(\mathbf{w})]\Big)\Big).$$

The stated bound then follows directly. The proof is complete. ☐

Finally, we present the proof of Corollary 5.

Proof of Corollary 5. We choose $\mathbf{w} = \mathbf{w}^*$ in Theorem 4 and get

$$\mathbb{E}[F(\bar{\mathbf{w}}_T) - F_S(\mathbf{w}^*)] = O\Big(\Big(\eta + \frac{1}{\gamma} + \frac{\gamma(T + T^2/n)\eta^2}{n}\Big)\mathbb{E}[F_S(\mathbf{w}^*)]\Big) + O\Big(\frac{1}{T\eta} + \frac{\gamma\eta(1 + T/n)}{n}\Big). \tag{E.2}$$

Note $\mathbb{E}[F_S(\mathbf{w}^*)] = F(\mathbf{w}^*)$.

We first prove Part (a). Since $\eta \asymp 1/\sqrt{T}$ and $T \asymp n$, the inequality (E.2) becomes

$$\mathbb{E}[F(\bar{\mathbf{w}}_T)] - F(\mathbf{w}^*) = O\Big(T^{-\frac{1}{2}} + \frac{1}{\gamma} + \frac{\gamma}{T} + \frac{\gamma(1 + T/n)}{n\sqrt{T}}\Big).$$

We can choose $\gamma = \sqrt{n}$ to get that $\mathbb{E}[F(\bar{\mathbf{w}}_T)] - F(\mathbf{w}^*) = O(1/\sqrt{n})$.

We now turn to Part (b). In this case, the inequality (E.2) becomes

$$\mathbb{E}[F(\bar{\mathbf{w}}_T)] - F(\mathbf{w}^*) = O\Big(\frac{1}{n} + \frac{1}{n\gamma} + \frac{\gamma}{n}\Big).$$

We can choose $\gamma = 1$ to get $\mathbb{E}[F(\bar{\mathbf{w}}_T)] - F(\mathbf{w}^*) = O(1/n)$. The proof is complete. □

F Proofs on Convex and Nonsmooth Problems

In this section, we present the proof related to stability and generalization for pairwise learning with convex and nonsmooth loss functions. We first prove stability (Theorem 6) and excess risk bounds (Theorem 7) for Algorithm 1. Then we move to excess risk bounds for Algorithm 2 (Theorem 8).

F.1 Proofs of Theorem 6 and Theorem 7

We need to introduce a concentration inequality [19] which is useful for developing high-probability bounds.

Lemma F.1 (Chernoff's Bound)**.** *Let $X_1, \ldots, X_t$ be independent random variables taking values in $\{0, 1\}$. Let $X = \sum_{j=1}^t X_j$ and $\mu = \mathbb{E}[X]$. Then for any $\tilde{\delta} > 0$ with probability at least $1 - \exp\big(-\mu\tilde{\delta}^2/(2 + \tilde{\delta})\big)$ we have $X \le (1 + \tilde{\delta})\mu$. Furthermore, for any $\delta \in (0, 1)$ with probability at least $1 - \delta$ we have*

$$X \le \mu + \log(1/\delta) + \sqrt{2\mu\log(1/\delta)}.$$

Proof of Theorem 6. Suppose S and S' differ by the first example. If $i_t \ne 1$ and $j_t \ne 1$, then

$$\begin{aligned}
\|\mathbf{w}_{t+1} - \mathbf{w}'_{t+1}\|_2^2 &= \big\|\mathbf{w}_t - \eta_t\nabla f(\mathbf{w}_t; z_{i_t}, z_{j_t}) - \mathbf{w}'_t + \eta_t\nabla f(\mathbf{w}'_t; z'_{i_t}, z'_{j_t})\big\|_2^2 \\
&= \big\|\mathbf{w}_t - \eta_t\nabla f(\mathbf{w}_t; z_{i_t}, z_{j_t}) - \mathbf{w}'_t + \eta_t\nabla f(\mathbf{w}'_t; z_{i_t}, z_{j_t})\big\|_2^2 \\
&= \|\mathbf{w}_t - \mathbf{w}'_t\|_2^2 - \langle \mathbf{w}_t - \mathbf{w}'_t, \eta_t\nabla f(\mathbf{w}_t; z_{i_t}, z_{j_t}) - \eta_t\nabla f(\mathbf{w}'_t; z_{i_t}, z_{j_t})\rangle + 4\eta_t^2G^2 \\
&\le \|\mathbf{w}_t - \mathbf{w}'_t\|_2^2 + 4\eta_t^2G^2,
\end{aligned}$$

where we have used the fact $\langle \mathbf{w}_t - \mathbf{w}'_t, \nabla f(\mathbf{w}_t; z_{i_t}, z_{j_t}) - \nabla f(\mathbf{w}'_t; z_{i_t}, z_{j_t})\rangle \ge 0$ due to the convexity of f. Otherwise,

$$\begin{aligned}
\|\mathbf{w}_{t+1} - \mathbf{w}'_{t+1}\|_2^2 &= \big\|\mathbf{w}_t - \eta_t\nabla f(\mathbf{w}_t; z_{i_t}, z_{j_t}) - \mathbf{w}'_t + \eta_t\nabla f(\mathbf{w}'_t; z'_{i_t}, z'_{j_t})\big\|_2^2 \\
&\le (1 + p)\|\mathbf{w}_t - \mathbf{w}'_t\|_2^2 + (1 + 1/p)\eta_t^2\big\|\nabla f(\mathbf{w}_t; z_{i_t}, z_{j_t}) - \nabla f(\mathbf{w}'_t; z'_{i_t}, z'_{j_t})\big\|_2^2 \\
&\le (1 + p)\|\mathbf{w}_t - \mathbf{w}'_t\|_2^2 + 4(1 + 1/p)\eta_t^2G^2,
\end{aligned}$$

where we have used Assumption 2. Combining the above two cases, we derive

$$\begin{aligned}
\|\mathbf{w}_{t+1} - \mathbf{w}'_{t+1}\|_2^2 &\le \big(1 + p\mathbb{I}_{[i_t=1 \text{ or } j_t=1]}\big)\|\mathbf{w}_t - \mathbf{w}'_t\|_2^2 + 4G^2\eta_t^2\big(1 + p^{-1}\mathbb{I}_{[i_t=1 \text{ or } j_t=1]}\big) &\text{(F.1)} \\
&= \big(1 + p\big)^{\mathbb{I}_{[i_t=1 \text{ or } j_t=1]}}\|\mathbf{w}_t - \mathbf{w}'_t\|_2^2 + 4G^2\eta_t^2\big(1 + p^{-1}\mathbb{I}_{[i_t=1 \text{ or } j_t=1]}\big), &\text{(F.2)}
\end{aligned}$$

where $\mathbb{I}_{[\cdot]}$ denotes the indicator function. Taking expectations over both sides of (F.1), we get

$$\mathbb{E}\big[\|\mathbf{w}_{t+1} - \mathbf{w}'_{t+1}\|_2^2\big] \le \big(1 + 2p/n\big)\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}'_t\|_2^2] + 4G^2\eta_t^2\big(1 + 2/(pn)\big),$$

where we have used $\mathbb{E}[\mathbb{I}_{[i_t=1 \text{ or } j_t=1]}] \leq 2/n$. We apply the above inequality recursively and get

$$\mathbb{E}\big[\|\mathbf{w}_{t+1}-\mathbf{w}'_{t+1}\|_2^2\big] \leq 4G^2\big(1+2/(pn)\big)\sum_{j=1}^{t}\eta_j^2\big(1+2p/n\big)^{t-j} \leq 4G^2\big(1+2/(pn)\big)\eta^2 t\big(1+2p/n\big)^t.$$

We can choose $p = n/(2t)$ and use the standard inequality $(1+1/t)^t \leq e$ to get

$$\mathbb{E}\big[\|\mathbf{w}_{t+1}-\mathbf{w}'_{t+1}\|_2^2\big] \leq 4G^2 et\big(1+4t/n^2\big)\eta^2.$$

This proves the stability bound in expectation. We now turn to high-probability bounds. It follows from (F.2) that

$$\begin{aligned}\|\mathbf{w}_{t+1}-\mathbf{w}'_{t+1}\|_2^2 &\leq 4G^2\sum_{k=1}^{t}\eta_k^2\big(1+p^{-1}\mathbb{I}_{[i_k=1 \text{ or } j_k=1]}\big)\prod_{k'=k+1}^{t}(1+p)^{\mathbb{I}_{[i_{k'}=1 \text{ or } j_{k'}=1]}}\\ &\leq 4G^2\eta^2\prod_{k=1}^{t}(1+p)^{\mathbb{I}_{[i_k=1 \text{ or } j_k=1]}}\sum_{k=1}^{t}\big(1+p^{-1}\mathbb{I}_{[i_k=1 \text{ or } j_k=1]}\big)\\ &= 4G^2\eta^2(1+p)^{\sum_{k=1}^{t}\mathbb{I}_{[i_k=1 \text{ or } j_k=1]}}\Big(t+p^{-1}\sum_{k=1}^{t}\mathbb{I}_{[i_k=1 \text{ or } j_k=1]}\Big).\end{aligned}$$

We can apply Lemma F.1 with $X_k = \mathbb{I}_{[i_k=1 \text{ or } j_k=1]}, \mu \leq 2t/n$ to get the following inequality with probability at least $1-\delta$

$$\sum_{k=1}^{t}\mathbb{I}_{[i_k=1 \text{ or } j_k=1]} \leq 2t/n + \log(1/\delta) + \sqrt{4tn^{-1}\log(1/\delta)}.$$

Therefore, with probability at least $1-\delta$ there holds

$$\begin{aligned}&\|\mathbf{w}_{t+1}-\mathbf{w}'_{t+1}\|_2^2 \leq\\ &4G^2\eta^2(1+p)^{2t/n+\log(1/\delta)+\sqrt{4tn^{-1}\log(1/\delta)}}\Big(t+p^{-1}\big(2t/n+\log(1/\delta)+\sqrt{4tn^{-1}\log(1/\delta)}\big)\Big).\end{aligned}$$

We can choose

$$p = \frac{1}{2t/n+\log(1/\delta)+\sqrt{4tn^{-1}\log(1/\delta)}}$$

in the above inequality and derive the following inequality with probability at least $1-\delta$ ($(1+1/x)^x \leq e$)

$$\|\mathbf{w}_{t+1}-\mathbf{w}'_{t+1}\|_2^2 \leq 4G^2\eta^2 e\Big(t+\big(2t/n+\log(1/\delta)+\sqrt{4tn^{-1}\log(1/\delta)}\big)^2\Big).$$

The proof is complete. □

We can use the above stability bounds to develop excess risk bounds in Theorem 7 for SGD with nonsmooth problems.

Proof of Theorem 7. Let $\{\mathbf{w}_t\}, \{\mathbf{w}'_t\}$ be defined in Theorem 6. According to (4.2) and Jensen's inequality, we know

$$\mathbb{E}[\|\mathbf{w}_{t+1}-\mathbf{w}'_{t+1}\|_2] \leq 2G\sqrt{2et}\big(1+2\sqrt{t}/n\big)\eta.$$

It then follows that SGD with t-iterations for nonsmooth problems is on-average loss ϵ-stable with

$$\epsilon \leq 4G^2\sqrt{2et}\big(1+2\sqrt{t}/n\big)\eta.$$

This together with the relationship between on-average stability and generalization shows

$$\mathbb{E}[F(\mathbf{w}_t)-F_S(\mathbf{w}_t)] \leq 4G^2\sqrt{2et}\big(1+2\sqrt{t}/n\big)\eta.$$

We can take an average of the above inequalities to get

$$\frac{1}{n}\sum_{t=1}^{T}\mathbb{E}[F(\mathbf{w}_t)-F_S(\mathbf{w}_t)] \leq 4G^2\sqrt{2eT}\big(1+2\sqrt{T}/n\big)\eta.$$

It then follows from (D.1) that

$$\mathbb{E}[F(\bar{\mathbf{w}}_T)] - F(\mathbf{w}^*) = \mathbb{E}[F(\bar{\mathbf{w}}_T) - F_S(\bar{\mathbf{w}}_T)] + \mathbb{E}\big[F_S(\bar{\mathbf{w}}_T) - F_S(\mathbf{w}^*)\big]$$
$$\leq 4G^2\sqrt{2eT}\big(1 + 2\sqrt{T}/n\big)\eta + \frac{G^2T\eta^2 + \|\mathbf{w}^*\|_2^2}{2T\eta},$$

where we have used the Jensen's inequality and (D.1). The stated bound then follows from the choice $T \asymp n^2$ and $\eta = T^{-\frac{3}{4}}$. The proof is complete. □

F.2 Proof of Theorem 8

We now turn to Theorem 8 on excess risk bounds of Algorithm 2 based on the iterative localization technique [5, 11]. We need to introduce some definitions. For any i, let

$$\hat{\mathbf{w}}_i = \arg\min_{\mathbf{w}} \widetilde{F}_{S_i}(\mathbf{w}). \tag{F.3}$$

Note $\mathbf{w}_i$ is derived by applying SGD with $\eta_t = \gamma_i n_i/(t+1)$ to minimize $\widetilde{F}_{S_i}(\mathbf{w})$, with the iterates weighted according to Part (c) of Lemma D.1. We need the following lemmas.

Lemma F.2. *Let Assumptions 1, 2 hold. For any* $\delta \in (0,1)$*, the following inequality holds with probability at least* $1-\delta/(2k)$*:* $\|\hat{\mathbf{w}}_i - \mathbf{w}_i\|_2 = O\big(\sqrt{n_i}\gamma_i \log^{\frac{1}{2}}(2k/\delta)\big)$.

Proof. It is clear that $\widetilde{F}_{S_i}$ is $\lambda_i := 2/(\gamma_i n_i)$-strongly convex. According to (D.5), the following inequality holds with probability at least $1-\delta/(2k)$

$$\widetilde{F}_{S_i}(\mathbf{w}_i) - \widetilde{F}_{S_i}(\hat{\mathbf{w}}_i) = O(\log(2k/\delta)/(T_i\lambda_i)) = O(\log(2k/\delta)/(n_i\lambda_i)).$$

It then follows from the definition of $\hat{\mathbf{w}}_i$ and the strong convexity that

$$\frac{\lambda_i}{2}\|\hat{\mathbf{w}}_i - \mathbf{w}_i\|_2^2 \leq \widetilde{F}_{S_i}(\mathbf{w}_i) - \widetilde{F}_{S_i}(\hat{\mathbf{w}}_i) = O\big(\log(2k/\delta)/(n_i\lambda_i)\big) \tag{F.4}$$

and therefore

$$\|\hat{\mathbf{w}}_i - \mathbf{w}_i\|_2^2 = O\big(\log(2k/\delta)/(n_i\lambda_i^2)\big) = O\big(n_i\gamma_i^2\log(2k/\delta)\big).$$

The proof is complete. □

The following lemma establishes the uniform stability of pairwise learning with strongly convex objectives.

Lemma F.3 ([16])**.** *Suppose* $f : \mathcal{W}\times\mathcal{Z}\times\mathcal{Z} \mapsto \mathbb{R}$ *takes a structure* $f = \ell + r$*, where* $\ell : \mathcal{W}\times\mathcal{Z}\times\mathcal{Z}\mapsto\mathbb{R}$ *and* $r : \mathcal{W}\mapsto\mathbb{R}$*. Assume for all* z, z'*, we have* $\|\nabla\ell(\mathbf{w}; z, z')\|_2 \leq G$*. Suppose* F_S *is* σ*-strongly convex and define* A *as* $A(S) = \arg\min_{\mathbf{w}\in\mathcal{W}} F_S(\mathbf{w})$*. Then* A *is* $\frac{8G^2}{n\sigma}$*-uniformly stable.*

The following lemma establishes the excess risk bounds for the empirical risk minimizer defined in (F.3).

Lemma F.4. *Let Assumptions 1, 2 hold. Let* $\hat{\mathbf{w}}_i$ *be defined in* (F.3)*. With probability at least* $1-\delta/(2k)$ *the following inequality holds for any* $\mathbf{w}\in\mathcal{W}$

$$F(\hat{\mathbf{w}}_i) - F(\mathbf{w}) = O\Big(\gamma_i \log n_i \log(k/\delta) + n_i^{-1}\log(k/\delta)\Big) + \frac{1}{\gamma_i n_i}\|\mathbf{w} - \mathbf{w}_{i-1}\|_2^2.$$

Proof. For any i, define

$$\widetilde{F}_i(\mathbf{w}) = \mathbb{E}_{Z,Z'}\big[f(\mathbf{w}; Z, Z')\big] + \frac{1}{\gamma_i n_i}\|\mathbf{w} - \mathbf{w}_{i-1}\|_2^2$$

and $\mathbf{w}_i^* = \arg\min_{\mathbf{w}} \widetilde{F}_i(\mathbf{w})$. Denote by A_i' the deterministic algorithm outputting the minimizer of $\widetilde{F}_{S_i}$. Since $\widetilde{F}_{S_i}$ is $\lambda_i = 2/(\gamma_i n_i)$-strongly convex and f is Lipschitz continuous, it follows from Lemma F.3 that A'_i is $4G^2\gamma_i$-uniformly stable. Furthermore, we have the following bound on variances

$$\begin{aligned}
&\mathbb{E}_{Z,Z',S_i}\big[\big(f(A'_i(S_i);Z,Z')-f(\mathbf{w}_i^*;Z,Z')\big)^2\big] \leq G^2\mathbb{E}_{S_i}\big[\|A'_i(S_i)-\mathbf{w}_i^*\|_2^2\big] \\
&\leq \frac{2G^2}{\lambda_i}\mathbb{E}_{S_i}\big[\widetilde{F}_i(A'_i(S_i))-\widetilde{F}_i(\mathbf{w}_i^*)\big] = G^2\gamma_i n_i\mathbb{E}_{S_i}\big[\widetilde{F}_i(A'_i(S_i))-\widetilde{F}_i(\mathbf{w}_i^*)\big].
\end{aligned}$$

It follows from the definition of $\hat{\mathbf{w}}_i$ and Theorem 2 ($A=A'_i, F=\widetilde{F}_i$) that with probability at least $1-\delta/(2k)$

$$\begin{aligned}
&\widetilde{F}_i(\hat{\mathbf{w}}_i)-\widetilde{F}_i(\mathbf{w}_i^*) \leq \widetilde{F}_i(\hat{\mathbf{w}}_i)-\widetilde{F}_{S_i}(\hat{\mathbf{w}}_i)-\widetilde{F}_i(\mathbf{w}_i^*)+\widetilde{F}_{S_i}(\mathbf{w}_i^*) = \\
&O\Big(\gamma_i\log n_i\log(k/\delta)+n_i^{-1}\log(k/\delta)+n_i^{-\frac{1}{2}}\big(\gamma_i n_i\log(k/\delta)\mathbb{E}_{S_i}\big[\widetilde{F}_i(A'_i(S_i))-\widetilde{F}_i(\mathbf{w}_i^*)\big]\big)^{\frac{1}{2}}\Big).
\end{aligned} \tag{F.5}$$

On the other hand, the uniform stability of A'_i and Part (a) of Theorem 1 implies that

$$\mathbb{E}_{S_i}\big[\widetilde{F}_i(\hat{\mathbf{w}}_i)-\widetilde{F}_{S_i}(\hat{\mathbf{w}}_i)\big] = O(\gamma_i).$$

It then follows that (note $\mathbb{E}_{S_i}[\widetilde{F}_{S_i}(\mathbf{w}_i^*)]=\widetilde{F}_i(\mathbf{w}_i^*)$ since $\mathbf{w}_i^*$ is independent of S_i)

$$\begin{aligned}
\mathbb{E}_{S_i}\big[\widetilde{F}_i(\hat{\mathbf{w}}_i)-\widetilde{F}_i(\mathbf{w}_i^*)\big] &= \mathbb{E}_{S_i}\big[\widetilde{F}_i(\hat{\mathbf{w}}_i)-\widetilde{F}_{S_i}(\mathbf{w}_i^*)\big] \\
&\leq \mathbb{E}_{S_i}\big[\widetilde{F}_i(\hat{\mathbf{w}}_i)-\widetilde{F}_{S_i}(\hat{\mathbf{w}}_i)\big] = O(\gamma_i).
\end{aligned}$$

We can plug the above inequality back into (F.5) and get the following inequality with probability at least $1-\delta/(2k)$

$$\widetilde{F}_i(\hat{\mathbf{w}}_i)-\widetilde{F}_i(\mathbf{w}) \leq \widetilde{F}_i(\hat{\mathbf{w}}_i)-\widetilde{F}_i(\mathbf{w}_i^*) = O\Big(\gamma_i\log n_i\log(k/\delta)+n_i^{-1}\log(k/\delta)\Big).$$

Then the following inequality holds with probability at least $1-\delta/(2k)$

$$\begin{aligned}
&F(\hat{\mathbf{w}}_i)-F(\mathbf{w}) = \widetilde{F}_i(\hat{\mathbf{w}}_i)-\widetilde{F}_i(\mathbf{w})-\frac{1}{\gamma_i n_i}\|\hat{\mathbf{w}}_i-\mathbf{w}_{i-1}\|_2^2+\frac{1}{\gamma_i n_i}\|\mathbf{w}-\mathbf{w}_{i-1}\|_2^2 \\
&= O\Big(\gamma_i\log n_i\log(k/\delta)+n_i^{-1}\log(k/\delta)\Big)-\frac{1}{\gamma_i n_i}\|\hat{\mathbf{w}}_i-\mathbf{w}_{i-1}\|_2^2+\frac{1}{\gamma_i n_i}\|\mathbf{w}-\mathbf{w}_{i-1}\|_2^2.
\end{aligned}$$

The stated bound then follows directly. The proof is complete. □

Based on the above lemmas, we are now ready to prove Theorem 8.

Proof of Theorem 8. Let $\hat{\mathbf{w}}_i$ be defined by (F.3). Let $\hat{\mathbf{w}}_0=\mathbf{w}^*$. We have the following error decomposition

$$F(\mathbf{w}_k)-F(\mathbf{w}^*) = \sum_{i=1}^k\big(F(\hat{\mathbf{w}}_i)-F(\hat{\mathbf{w}}_{i-1})\big)+F(\mathbf{w}_k)-F(\hat{\mathbf{w}}_k). \tag{F.6}$$

According to Lemma F.2, we know the following inequality with probability at least $1-\delta/(2k)$

$$F(\mathbf{w}_k)-F(\hat{\mathbf{w}}_k) \leq G\|\mathbf{w}_k-\hat{\mathbf{w}}_k\|_2 = O(\sqrt{n_k}\gamma_k\log^{\frac{1}{2}}(2k/\delta)). \tag{F.7}$$

Furthermore, we can apply Lemma F.4 with $\mathbf{w}=\hat{\mathbf{w}}_{i-1}$ for different i to get the following inequality with probability $1-\delta$

$$\begin{aligned}
&\sum_{i=1}^k\big(F(\hat{\mathbf{w}}_i)-F(\hat{\mathbf{w}}_{i-1})\big) = \sum_{i=1}^k\Big(O(\gamma_i\log n_i\log(k/\delta)+n_i^{-1}\log(k/\delta))+\frac{\|\hat{\mathbf{w}}_{i-1}-\mathbf{w}_{i-1}\|_2^2}{\gamma_i n_i}\Big) \\
&= O(\gamma_1\log n_1+n_1^{-1})\log(k/\delta)+\frac{\|\hat{\mathbf{w}}_0-\mathbf{w}_0\|_2^2}{\gamma_1 n_1}+\sum_{i=2}^k\Big(O(\gamma_i\log n_i+n_i^{-1})\log(k/\delta)+\frac{\|\hat{\mathbf{w}}_{i-1}-\mathbf{w}_{i-1}\|_2^2}{\gamma_i n_i}\Big) \\
&= O(\gamma_1\log n_1+n_1^{-1})\log(k/\delta)+\frac{\|\mathbf{w}^*\|_2^2}{\gamma_1 n_1}+\sum_{i=2}^k O\Big(\gamma_i\log n_i+n_i^{-1}+\frac{n_{i-1}\gamma_{i-1}^2}{\gamma_i n_i}\Big)\log(k/\delta),
\end{aligned}$$

where we have used Lemma F.2 in the last step. We can combine the above three inequalities together and get the following inequality with probability at least $1-\delta$

$$
\begin{aligned}
&F(\mathbf{w}_k) - F(\mathbf{w}^*) = O(\sqrt{n_k}\gamma_k \log^{1/2}(k/\delta)) \\
&+ O(\gamma_1 \log n_1 + n_1^{-1}) \log(k/\delta) + \frac{\|\mathbf{w}^*\|_2^2}{\gamma_1 n_1} + \sum_{i=2}^{k} O\Big(\gamma_i \log n_i + n_i^{-1} + \frac{n_{i-1}\gamma_{i-1}^2}{\gamma_i n_i}\Big) \log(k/\delta) \\
&= O\Big(\sqrt{n}\gamma 2^{-k-k/2} + \gamma \log n + \frac{1}{\gamma n}\|\mathbf{w}^*\|_2^2 + \sum_{i=2}^{k} \big(2^{-i}\gamma \log n + 2^i n^{-1} + \frac{2^{1-i} n 2^{2(1-i)}\gamma^2}{2^{-i}\gamma 2^{-i} n}\big)\Big) \log(k/\delta) \\
&= O\Big(\sqrt{n}\gamma n^{-\frac{3}{4}} + \gamma \log n + \frac{1}{\gamma n}\|\mathbf{w}^*\|_2^2 + n^{-\frac{1}{2}} + \gamma\Big) \log(\log n/\delta),
\end{aligned}
$$

where we have used $2^k \asymp \sqrt{n}$ and

$$
k = \frac{1}{2}\lceil \log_2 n \rceil, \quad \gamma_i = \gamma/2^i, \quad n_i = \lceil n/2^i \rceil.
$$

We can take $\gamma \asymp n^{-\frac{1}{2}}\|\mathbf{w}^*\|_2$ to get $F(\mathbf{w}_k) - F(\mathbf{w}^*) = O(\log(\log n/\delta)\log n \|\mathbf{w}^*\|/\sqrt{n})$ with probability at least $1-\delta$.

Furthermore, it is clear that the total number of gradient computations is of the order of

$$
\sum_{i=1}^{k} T_i \asymp \sum_{i=1}^{k} n/2^i \asymp n.
$$

The proof is complete. □

G Proofs on Strongly Convex Problems

In this section, we present the proofs related to excess risk bounds for pairwise learning with strongly convex objectives (Theorem 9 and Theorem 10). We first prove generalization bounds for smooth problems. To this aim, we introduce a lemma.

Lemma G.1 ([16]). *Assume for all $S \in \mathcal{Z}^n$, F_S is σ-strongly convex w.r.t. $\|\cdot\|$. Let $A(S) = \arg\min_{\mathbf{w}\in\mathcal{W}} F_S(\mathbf{w})$ and Assumption 3 hold. If $\sigma n \geq 8L$, then*

$$
\mathbb{E}\big[F(A(S))\big] - F_S(A(S)) \leq \Big(\frac{1024L^2}{n^2\sigma^2} + \frac{64L}{n\sigma}\Big)\mathbb{E}\big[F_S(A(S))\big]. \tag{G.1}
$$

Proof of Theorem 9. According to (G.1), we know the following generalization bound for ERM applied to strongly convex and smooth problems

$$
\mathbb{E}[F(\mathbf{w}_S^*) - F_S(\mathbf{w}_S^*)] \leq 64L\Big(\frac{16L}{n^2\sigma^2} + \frac{1}{n\sigma}\Big)\mathbb{E}\big[F_S(\mathbf{w}_S^*)\big]. \tag{G.2}
$$

The L-smoothness of f implies the L-smoothness of F, which implies

$$
\begin{aligned}
F(A(S)) - F(\mathbf{w}_S^*) &\leq \langle A(S) - \mathbf{w}_S^*, \nabla F(\mathbf{w}_S^*)\rangle + \frac{L}{2}\|A(S) - \mathbf{w}_S^*\|_2^2 \\
&\leq \|A(S) - \mathbf{w}_S^*\|_2 \|\nabla F(\mathbf{w}_S^*)\|_2 + \frac{L}{2}\|A(S) - \mathbf{w}_S^*\|_2^2 \\
&\leq \frac{1}{2L}\|\nabla F(\mathbf{w}_S^*)\|_2^2 + L\|A(S) - \mathbf{w}_S^*\|_2^2,
\end{aligned}
$$

where we have used the Cauchy-Schwartz inequality in the last step. According to Lemma B.1 and the inequality $F_S(\mathbf{w}_S^*) \leq F_S(\mathbf{w}^*)$, we know

$$
\begin{aligned}
\mathbb{E}\big[\|\nabla F(\mathbf{w}_S^*)\|_2^2\big] &= \mathbb{E}\big[\|\nabla F(\mathbf{w}_S^*) - \nabla F(\mathbf{w}^*)\|_2^2\big] \leq 2L\mathbb{E}[F(\mathbf{w}_S^*) - F(\mathbf{w}^*)] = 2L\mathbb{E}[F(\mathbf{w}_S^*) - F_S(\mathbf{w}^*)] \\
&\leq 2L\mathbb{E}[F(\mathbf{w}_S^*) - F_S(\mathbf{w}_S^*)] \leq 128L^2\Big(\frac{16L}{n^2\sigma^2} + \frac{1}{n\sigma}\Big)\mathbb{E}\big[F_S(\mathbf{w}_S^*)\big],
\end{aligned}
$$

where we have used (G.2). Furthermore, the σ-strong convexity of F_S implies

$$\|A(S)-\mathbf{w}_S^*\|_2^2 \le \frac{2}{\sigma}\big(F_S(A(S))-F_S(\mathbf{w}_S^*)\big).$$

We can combine the above three inequalities together and derive

$$\mathbb{E}[F(A(S))-F(\mathbf{w}_S^*)] \le 64L\Big(\frac{16L}{n^2\sigma^2}+\frac{1}{n\sigma}\Big)\mathbb{E}\big[F_S(\mathbf{w}_S^*)\big]+\frac{2L}{\sigma}\mathbb{E}\big[F_S(A(S))-F_S(\mathbf{w}_S^*)\big].$$

We can combine the above inequality and (G.2) together and get

$$\begin{aligned}\mathbb{E}[F(A(S))-F_S(\mathbf{w}_S^*)] &= \mathbb{E}[F(A(S))-F(\mathbf{w}_S^*)]+\mathbb{E}[F(\mathbf{w}_S^*)-F_S(\mathbf{w}_S^*)] \\ &\le 128L\Big(\frac{16L}{n^2\sigma^2}+\frac{1}{n\sigma}\Big)\mathbb{E}\big[F_S(\mathbf{w}_S^*)\big]+\frac{2L}{\sigma}\mathbb{E}\big[F_S(A(S))-F_S(\mathbf{w}_S^*)\big].\end{aligned} \tag{G.3}$$

The stated bound then follows since

$$\mathbb{E}[F_S(\mathbf{w}_S^*)] \le \mathbb{E}[F_S(\mathbf{w}^*)] = F(\mathbf{w}^*). \tag{G.4}$$

The proof is complete. □

We now turn to Theorem 10 on nonsmooth problems.

Proof of Theorem 10. According to Lemma F.3 and Part (a) of Theorem 1, we know the following generalization bound for ERM applied to strongly convex and Lipschitz continuous problems

$$\mathbb{E}[F(\mathbf{w}_S^*)-F_S(\mathbf{w}_S^*)] \le \frac{8G^2}{n\sigma}.$$

The Lipschitz continuity of f implies the Lipschitz continuity of F. Therefore, it follows from the strong convexity of F_S that

$$F(A(S))-F(\mathbf{w}_S^*) \le G\|A(S)-\mathbf{w}_S^*\|_2 \le G\sqrt{\frac{2\big(F_S(A(S))-F_S(\mathbf{w}_S^*)\big)}{\sigma}}.$$

We can combine the above two inequalities together and use (G.3) to derive

$$\mathbb{E}[F(A(S))-F_S(\mathbf{w}_S^*)] \le \frac{8G^2}{n\sigma}+G\sqrt{\frac{2\mathbb{E}\big[F_S(A(S)-F_S(\mathbf{w}_S^*))\big]}{\sigma}}.$$

The stated bound then follows from (G.4). The proof is complete. □

We now consider the application to the specific SGD. It shows how we should early-stop the algorithm to get the optimal bound $O(1/(n\sigma))$. Part (a) and Part (b) are for smooth and nonsmooth cases, respectively.

Corollary G.2 (SGD)**.** *Let $\{\mathbf{w}_t\}$ be the sequence produced by SGD with $\eta_t = 2/(\sigma(t+1))$. Let F_S be σ-strongly convex and $\bar{\mathbf{w}}_t' = \big(\sum_{j=1}^t j\mathbf{w}_j\big)/\sum_{j=1}^t j$.*

(a) If Assumption 3 holds and $\sigma n \ge 8L$, then

$$\mathbb{E}[F(\bar{\mathbf{w}}_T')-F(\mathbf{w}^*)] = O\Big(\frac{\mathbb{E}\big[F_S(\mathbf{w}_S^*)\big]}{n\sigma}+1/(T\sigma^2)+\mathbb{E}[\|\mathbf{w}_S^*\|_2^2]/(T^2\sigma)\Big). \tag{G.5}$$

In particular, one can choose $T \asymp n/\sigma$ to get the excess population risk bound $O(1/(n\sigma))$.

(b) If Assumption 2 holds, then

$$\mathbb{E}[F(A(S))-F_S(\mathbf{w}_S^*)] = \frac{8G^2}{n\sigma}+O\Big(\sqrt{\frac{1/(T\sigma)+\mathbb{E}[\|\mathbf{w}_S^*\|_2^2]/T^2}{\sigma}}\Big). \tag{G.6}$$

In particular, one can choose $T \asymp n^2$ to get the excess population risk bound $O(1/(n\sigma))$.

Proof. According to Eq. (D.4), we know

$$\mathbb{E}_A[F_S(\bar{\mathbf{w}}'_T)] - F_S(\mathbf{w}^*_S) = O\big(1/(T\sigma) + \|\mathbf{w}^*_S\|_2^2/T^2\big). \tag{G.7}$$

We first prove Part (a). We can plug the above optimization error bounds into Theorem 9 and derive

$$\mathbb{E}[F(\bar{\mathbf{w}}'_T) - F(\mathbf{w}^*)] = 128L\Big(\frac{16L}{n^2\sigma^2} + \frac{1}{n\sigma}\Big)\mathbb{E}\big[F_S(\mathbf{w}^*_S)\big] + O\Big(1/(T\sigma^2) + \mathbb{E}[\|\mathbf{w}^*_S\|_2^2]/(T^2\sigma)\Big).$$

This gives (G.5).

We now consider Part (b). We plug (G.7) into Theorem 10 and derive

$$\mathbb{E}[F(A(S)) - F_S(\mathbf{w}^*_S)] \le \frac{8G^2}{n\sigma} + O\Big(\sqrt{\frac{1/(T\sigma) + \mathbb{E}[\|\mathbf{w}^*_S\|_2^2]/T^2}{\sigma}}\Big).$$

This gives (G.6). The proof is complete. □

H Proofs on Uniform Convergence of Gradients for Pairwise Learning

In this section, we present the proofs on the uniform convergence of gradients (Theorem 11, Corollary 12 and Corollary 13).

H.1 Proof of Theorem 11

To prove Theorem 11, we first introduce a useful lemma called the McDiarmid's inequality [19] for handling the concentration of functions with bounded increments.

Lemma H.1. *Let $c_1, \ldots, c_n \in \mathbb{R}_+$. Let $Z_1, \ldots, Z_n$ be independent random variables taking values in a set $\mathcal{Z}$, and assume that $g : \mathcal{Z}^n \mapsto \mathbb{R}$ satisfies*

$$\sup_{z_1,\ldots,z_n,\bar{z}_i\in\mathcal{Z}} |g(z_1,\cdots,z_n) - g(\cdots, z_{i-1}, \bar{z}_i, z_{i+1}, \cdots)| \le c_i \tag{H.1}$$

for $i = 1, \ldots, n$. Then, for any $0 < \delta < 1$, with probability at least $1-\delta$ we have

$$g(Z_1,\ldots,Z_n) \le \mathbb{E}\big[g(Z_1,\ldots,Z_n)\big] + \Big(\frac{1}{2}\sum_{i=1}^n c_i^2 \log(1/\delta)\Big)^{\frac{1}{2}}.$$

The following lemma gives a high-probability bound on the uniform deviation between population gradients and empirical gradients.

Lemma H.2. *Let $\delta \in (0,1)$ and $S = \{z_1,\ldots,z_n\}$ be examples drawn independently from ρ. Suppose Assumption 3 holds. Then with probability at least $1-\delta$ we have*

$$\sup_{\mathbf{w}\in\mathcal{W}_R} \big\|\nabla F(\mathbf{w}) - \nabla F_S(\mathbf{w})\big\|_2 \le \frac{2}{\lfloor\frac{n}{2}\rfloor}\mathbb{E}_S\mathbb{E}_\epsilon \sup_{\mathbf{w}\in\mathcal{W}_R}\Big\|\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor}\epsilon_i\nabla f(\mathbf{w}; z_i, z_{i+\lfloor\frac{n}{2}\rfloor})\Big\|_2 + \sqrt{\frac{8\big(LR+b'\big)^2\log(1/\delta)}{n}},$$

where ϵ_i are independent Rademacher variables.

Proof. By the L-Lipschitz continuity of ∇f, the following inequality holds for all $\mathbf{w}\in\mathcal{W}_R$

$$\|\nabla f(\mathbf{w}; z_i, z_j)\|_2 \le \|\nabla f(0; z_i, z_j)\|_2 + L\|\mathbf{w}\|_2 \le LR + b'. \tag{H.2}$$

Let $S' = \{z'_1,\ldots,z'_n\}$ be independent examples drawn independently from ρ and $S_i = \{z_1,\ldots,z_{i-1},z'_i,z_{i+1},\ldots,z_n\}$. Then, we have

$$\Big|\sup_{\mathbf{w}\in\mathcal{W}_R}\big\|\nabla F(\mathbf{w}) - \nabla F_S(\mathbf{w})\big\|_2 - \sup_{\mathbf{w}\in\mathcal{W}_R}\big\|\nabla F(\mathbf{w}) - \nabla F_{S_i}(\mathbf{w})\big\|_2\Big|$$

$$\le \sup_{\mathbf{w}\in\mathcal{W}_R}\Big|\big\|\nabla F(\mathbf{w}) - \nabla F_S(\mathbf{w})\big\|_2 - \big\|\nabla F(\mathbf{w}) - \nabla F_{S_i}(\mathbf{w})\big\|_2\Big| \le \sup_{\mathbf{w}\in\mathcal{W}_R}\big\|\nabla F_S(\mathbf{w}) - \nabla F_{S_i}(\mathbf{w})\big\|_2$$

$$= \frac{1}{n(n-1)}\sup_{\mathbf{w}\in\mathcal{W}_R}\Big\|\sum_{j\in[n]:j\ne i}\big(\nabla f(\mathbf{w}; z_i, z_j) + f(\mathbf{w}; z_j, z_i) - \nabla f(\mathbf{w}; z'_i, z_j) - \nabla f(\mathbf{w}; z_j, z'_i)\big)\Big\|$$

$$\le \frac{4(n-1)}{n(n-1)}\big(LR+b'\big) = \frac{4(LR+b')}{n},$$

where we have used (H.2) for all $\mathbf{w} \in \mathcal{W}_R$. Therefore, (H.1) holds with

$$g(z_1, \ldots, z_n) := \sup_{\mathbf{w}\in\mathcal{W}_R} \big[\|\nabla F(\mathbf{w}) - \nabla F_S(\mathbf{w})\|_2\big]$$

and $c_i = 4(LR + b')/n$. We can apply Lemma H.1 to derive the following inequality with probability $1-\delta$

$$\sup_{\mathbf{w}\in\mathcal{W}_R} \big\|\nabla F(\mathbf{w}) - \nabla F_S(\mathbf{w})\big\|_2 \le \mathbb{E}_S\Big[\sup_{\mathbf{w}\in\mathcal{W}_R} \big\|\nabla F(\mathbf{w}) - \nabla F_S(\mathbf{w})\big\|_2\Big] + \sqrt{\frac{8\big(LR + b'\big)^2 \log(1/\delta)}{n}}. \tag{H.3}$$

For any $\mathbf{w} \in \mathcal{W}_R$, define $q_{\mathbf{w}} : \mathcal{Z} \times \mathcal{Z} \mapsto \mathbb{R}$ as

$$q_{\mathbf{w}}(z, z') = \mathbb{E}_{Z,Z'}\big[\nabla f(\mathbf{w}; Z, Z')\big] - \nabla f(\mathbf{w}; z, z').$$

Then it is clear that

$$\nabla F(\mathbf{w}) - \nabla F_S(\mathbf{w}) = \frac{1}{n(n-1)} \sum_{i,j\in[n]:i\neq j} q_{\mathbf{w}}(z_i, z_j).$$

Analogous to Eq. (C.4), we have

$$\mathbb{E}_S\Big[\sup_{\mathbf{w}\in\mathcal{W}_R} \big\|\nabla F(\mathbf{w}) - \nabla F_S(\mathbf{w})\big\|_2\Big] \le \mathbb{E}_S\Big[\sup_{\mathbf{w}\in\mathcal{W}_R} \Big\|\frac{1}{\lfloor\frac{n}{2}\rfloor} \sum_{i=1}^{\lfloor\frac{n}{2}\rfloor} q_{\mathbf{w}}\big(z_i, z_{i+\lfloor\frac{n}{2}\rfloor}\big)\Big\|_2\Big].$$

By the standard symmetrization trick, we get

$$\begin{aligned}
&\mathbb{E}_S\Big[\sup_{\mathbf{w}\in\mathcal{W}_R} \Big\|\frac{1}{\lfloor\frac{n}{2}\rfloor} \sum_{i=1}^{\lfloor\frac{n}{2}\rfloor} \big(\mathbb{E}_{Z,Z'}\big[\nabla f(\mathbf{w}; Z, Z')\big] - \nabla f(\mathbf{w}; z_i, z_{i+\lfloor\frac{n}{2}\rfloor})\big)\Big\|_2\Big] \\
&= \mathbb{E}_S\Big[\sup_{\mathbf{w}\in\mathcal{W}_R} \Big\|\frac{1}{\lfloor\frac{n}{2}\rfloor} \sum_{i=1}^{\lfloor\frac{n}{2}\rfloor} \mathbb{E}_{S'}\big[\nabla f(\mathbf{w}; z_i', z'_{i+\lfloor\frac{n}{2}\rfloor}) - \nabla f(\mathbf{w}; z_i, z_{i+\lfloor\frac{n}{2}\rfloor})\big]\Big\|_2\Big] \\
&\le \mathbb{E}_{S,S'}\Big[\sup_{\mathbf{w}\in\mathcal{W}_R} \Big\|\frac{1}{\lfloor\frac{n}{2}\rfloor} \sum_{i=1}^{\lfloor\frac{n}{2}\rfloor} \big(\nabla f(\mathbf{w}; z_i', z'_{i+\lfloor\frac{n}{2}\rfloor}) - \nabla f(\mathbf{w}; z_i, z_{i+\lfloor\frac{n}{2}\rfloor})\big)\Big\|_2\Big] \\
&= \frac{1}{\lfloor\frac{n}{2}\rfloor} \mathbb{E}_{S,S',\boldsymbol{\epsilon}}\Big[\sup_{\mathbf{w}\in\mathcal{W}_R} \Big\|\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor} \epsilon_i\big(\nabla f(\mathbf{w}; z_i', z'_{i+\lfloor\frac{n}{2}\rfloor}) - \nabla f(\mathbf{w}; z_i, z_{i+\lfloor\frac{n}{2}\rfloor})\big)\Big\|_2\Big] \\
&\le \frac{2}{\lfloor\frac{n}{2}\rfloor} \mathbb{E}_S \mathbb{E}_{\boldsymbol{\epsilon}} \sup_{\mathbf{w}\in\mathcal{W}_R} \Big\|\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor} \epsilon_i \nabla f(\mathbf{w}; z_i, z_{i+\lfloor\frac{n}{2}\rfloor})\Big\|_2.
\end{aligned}$$

We can plug the above two inequalities back into (H.3) to derive the stated inequality with probability $1-\delta$. The proof is complete. □

We now use Lemma H.2 to prove Theorem 11.

Proof of Theorem 11. According to Jensen's inequality, we know

$$\begin{aligned}
\Big(\mathbb{E}_\epsilon \sup_{\mathbf{w}\in\mathcal{W}_R} \Big[\Big\|\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor} \epsilon_i \nabla f(\mathbf{w}; z_i, z_{i+\lfloor\frac{n}{2}\rfloor})\Big\|_2\Big]\Big)^2 &\le \mathbb{E}_\epsilon\Big[\sup_{\mathbf{w}\in\mathcal{W}_R} \Big\|\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor} \epsilon_i \nabla f(\mathbf{w}; z_i, z_{i+\lfloor\frac{n}{2}\rfloor})\Big\|_2^2\Big] \\
&= \mathbb{E}_\epsilon\Big[\sup_{\mathbf{w}\in\mathcal{W}_R} \Big\langle \sum_{i=1}^{\lfloor\frac{n}{2}\rfloor} \epsilon_i \nabla f(\mathbf{w}; z_i, z_{i+\lfloor\frac{n}{2}\rfloor}), \sum_{i=1}^{n} \epsilon_i \nabla f(\mathbf{w}; z_i, z_{i+\lfloor\frac{n}{2}\rfloor})\Big\rangle\Big] \\
&\le \sup_{\mathbf{w}\in\mathcal{W}_R} \sum_{i=1}^{\lfloor\frac{n}{2}\rfloor} \big\langle \nabla f(\mathbf{w}; z_i, z_{i+\lfloor\frac{n}{2}\rfloor}), \nabla f(\mathbf{w}; z_i, z_{i+\lfloor\frac{n}{2}\rfloor})\big\rangle \\
&\quad + 2\mathbb{E}_\epsilon\Big[\sup_{\mathbf{w}\in\mathcal{W}_R} \sum_{1\le i<j\le\lfloor\frac{n}{2}\rfloor} \epsilon_i\epsilon_j \big\langle \nabla f(\mathbf{w}; z_i, z_{i+\lfloor\frac{n}{2}\rfloor}), \nabla f(\mathbf{w}; z_j, z_{j+\lfloor\frac{n}{2}\rfloor})\big\rangle\Big] \\
&\le \lfloor\frac{n}{2}\rfloor (LR + b')^2 + n\mathcal{U}_S(\mathcal{F}_R),
\end{aligned} \tag{H.4}$$

where we have used (H.2) and the definition of Rademacher chaos complexities. It then follows that

$$\mathbb{E}_\epsilon \sup_{\mathbf{w}\in\mathcal{W}_R} \Big[\Big\| \sum_{i=1}^n \epsilon_i \nabla f(\mathbf{w}; z_i, z_{i+\lfloor \frac{n}{2}\rfloor})\Big\|_2\Big] \leq \sqrt{\lfloor \frac{n}{2}\rfloor}(LR+b') + \sqrt{n\mathcal{U}_S(\mathcal{F}_R)}.$$

We can plug the above bound into Lemma H.2 to derive the stated bound with high probability. □

H.2 Proof of Corollary 12

To prove Corollary 12, it suffices to estimate the involved Rademacher chaos complexity [13, 21]. We handle this term by applying the entropy integral (Lemma H.3) in terms of covering numbers.

Definition 1 (Covering number)**.** Let $(\mathcal{G}, d)$ be a metric space and set $\mathcal{F} \subseteq \mathcal{G}$. For any $\epsilon > 0$, a set $\mathcal{F}^\triangle \subset \mathcal{F}$ is called an ϵ-cover of $\mathcal{F}$ if for every $f \in \mathcal{F}$ we can find an element $g \in \mathcal{F}^\triangle$ satisfying $d(f, g) \leq \epsilon$. The covering number $\mathcal{N}(\epsilon, \mathcal{F}, d)$ is the cardinality of the minimal ϵ-cover of $\mathcal{F}$:

$$\mathcal{N}(\epsilon, \mathcal{F}, d) := \min\Big\{|\mathcal{F}^\triangle| : \mathcal{F}^\triangle \text{ is an } \epsilon\text{-cover of } \mathcal{F}\Big\}.$$

Lemma H.3 ([21])**.** *Let $\mathcal{F} : \widetilde{\mathcal{Z}} \times \widetilde{\mathcal{Z}} \mapsto \mathbb{R}$ be a function class with $\sup_{f\in\mathcal{F}} d_S(f, 0) \leq D$ and $S = \{\tilde{z}_1, \ldots, \tilde{z}_n\} \subset \widetilde{\mathcal{Z}}$, where d_S is a pseudometric on $\mathcal{F}$ defined as follows*

$$d_S(f, g) := \Big(\frac{1}{n^2} \sum_{1\leq i<j\leq n} |f(\tilde{z}_i, \tilde{z}_j) - g(\tilde{z}_i, \tilde{z}_j)|^2\Big)^{1/2}. \tag{H.5}$$

Then

$$\frac{1}{n}\mathbb{E}_\epsilon\Big[\sup_{f\in\mathcal{F}} \sum_{1\leq i<j\leq n} \epsilon_i\epsilon_j f(\tilde{z}_i, \tilde{z}_j)\Big] \leq 24e \int_0^D \log\big(\mathcal{N}(r, \mathcal{F}, d_S) + 1\big)\mathrm{d}r.$$

Proof of Corollary 12. For any $i \in [\lfloor \frac{n}{2}\rfloor]$, we define $\tilde{z}_i = (z_i, z_{i+\lfloor \frac{n}{2}\rfloor})$ and $\tilde{f}(\mathbf{w}; \tilde{z}_i) = f(\mathbf{w}; z_i, z_{i+\lfloor \frac{n}{2}\rfloor})$. Then the Rademacher chaos complexity $\mathcal{U}_n(\mathcal{F}_R)$ can be written as

$$\mathcal{U}_{\widetilde{S}}(\mathcal{F}_R) = \frac{1}{\lfloor \frac{n}{2}\rfloor}\mathbb{E}_\epsilon\Big[\sup_{\mathbf{w}\in\mathcal{W}_R} \sum_{1\leq i<j\leq \lfloor \frac{n}{2}\rfloor} \epsilon_i\epsilon_j\big\langle \nabla\tilde{f}(\mathbf{w}; \tilde{z}_i), \nabla\tilde{f}(\mathbf{w}; \tilde{z}_j)\big\rangle\Big], \tag{H.6}$$

where $\widetilde{S} = \{\tilde{z}_1, \ldots, \tilde{z}_{\lfloor \frac{n}{2}\rfloor}\}$. We define a metric $d_{\widetilde{S}}$ over $\mathcal{F}_R$ by

$$d_{\widetilde{S}}(\mathbf{w}, \mathbf{w}') = \Big(\frac{1}{\lfloor \frac{n}{2}\rfloor^2} \sum_{1\leq i<j\leq \lfloor \frac{n}{2}\rfloor} \big|\langle \nabla\tilde{f}(\mathbf{w}; \tilde{z}_i), \nabla\tilde{f}(\mathbf{w}; \tilde{z}_j)\rangle - \langle \nabla\tilde{f}(\mathbf{w}'; \tilde{z}_i), \nabla\tilde{f}(\mathbf{w}'; \tilde{z}_j)\rangle\big|^2\Big)^{1/2}.$$

For any $\mathbf{w}$ and $\mathbf{w}'$ in $\mathcal{W}_R$, there holds

$$\begin{aligned}
\lfloor \frac{n}{2}\rfloor^2 d^2_{\widetilde{S}}(\mathbf{w}, \mathbf{w}') &= \sum_{1\leq i<j\leq \lfloor \frac{n}{2}\rfloor} \big|\langle \nabla\tilde{f}(\mathbf{w}; \tilde{z}_i), \nabla\tilde{f}(\mathbf{w}; \tilde{z}_j)\rangle - \langle \nabla\tilde{f}(\mathbf{w}'; \tilde{z}_i), \nabla\tilde{f}(\mathbf{w}'; \tilde{z}_j)\rangle\big|^2 \\
&\leq 2 \sum_{1\leq i<j\leq \lfloor \frac{n}{2}\rfloor} \big\langle \nabla\tilde{f}(\mathbf{w}; \tilde{z}_i) - \nabla\tilde{f}(\mathbf{w}'; \tilde{z}_i), \nabla\tilde{f}(\mathbf{w}; \tilde{z}_j)\big\rangle^2 + 2 \sum_{1\leq i<j\leq \lfloor \frac{n}{2}\rfloor} \big\langle \nabla\tilde{f}(\mathbf{w}'; \tilde{z}_i), \nabla\tilde{f}(\mathbf{w}; \tilde{z}_j) - \nabla\tilde{f}(\mathbf{w}'; \tilde{z}_j)\big\rangle^2 \\
&\leq 2 \sum_{1\leq i<j\leq \lfloor \frac{n}{2}\rfloor} \big\|\nabla\tilde{f}(\mathbf{w}; \tilde{z}_i) - \nabla\tilde{f}(\mathbf{w}'; \tilde{z}_i)\big\|_2^2\big\|\nabla\tilde{f}(\mathbf{w}; \tilde{z}_j)\big\|_2^2 + 2 \sum_{1\leq i<j\leq \lfloor \frac{n}{2}\rfloor} \|\nabla\tilde{f}(\mathbf{w}'; \tilde{z}_i)\|_2^2\big\|\nabla\tilde{f}(\mathbf{w}; \tilde{z}_j) - \nabla\tilde{f}(\mathbf{w}'; \tilde{z}_j)\big\|_2^2 \\
&\leq 2L^2 \sum_{1\leq i<j\leq \lfloor \frac{n}{2}\rfloor} \Big[\big\|\nabla\tilde{f}(\mathbf{w}; \tilde{z}_j)\big\|_2^2 + \|\nabla\tilde{f}(\mathbf{w}'; \tilde{z}_i)\|_2^2\Big]\|\mathbf{w} - \mathbf{w}'\|_2^2 \\
&\leq L^2(LR+b')^2 n(n-1)\|\mathbf{w} - \mathbf{w}'\|_2^2,
\end{aligned} \tag{H.7}$$

where we have used the elementary inequality $(a_1 + a_2)^2 \leq 2(a_1^2 + a_2^2)$ and the decomposition

$$\langle \nabla\tilde{f}(\mathbf{w}; \tilde{z}_i), \nabla\tilde{f}(\mathbf{w}; \tilde{z}_j)\rangle - \langle \nabla\tilde{f}(\mathbf{w}'; \tilde{z}_i), \nabla\tilde{f}(\mathbf{w}'; \tilde{z}_j)\rangle = \langle \nabla\tilde{f}(\mathbf{w}; \tilde{z}_i) - \nabla\tilde{f}(\mathbf{w}'; \tilde{z}_i), \nabla\tilde{f}(\mathbf{w}; \tilde{z}_j)\rangle + \langle \nabla\tilde{f}(\mathbf{w}'; \tilde{z}_i), \nabla\tilde{f}(\mathbf{w}; \tilde{z}_j) - \nabla\tilde{f}(\mathbf{w}'; \tilde{z}_j)\rangle$$

in the first inequality, the L-smoothness of f in the third inequality and (H.2) in the last inequality. It then follows that

$$\log\mathcal{N}(r,\mathcal{F}_R,d_{\widetilde{S}})\le\log\mathcal{N}\Big(r/\big(2L(LR+b')\big),\mathcal{W}_R,d_2\Big)$$
$$\le d\log\Big(6LR(LR+b')r^{-1}\Big),$$

where we have used the classical result $\log\mathcal{N}(r,\mathcal{W}_R,d_2)\le d\log(3R/r)$ [18] and d_2 is the metric over $\mathcal{W}_R$ defined by $d_2(\mathbf{w},\tilde{\mathbf{w}})=\|\mathbf{w}-\tilde{\mathbf{w}}\|_2$. Furthermore, (H.7) also implies $d_{\widetilde{S}}(\mathbf{w},0)\le 2LR(LR+b')$ for $\mathbf{w}\in\mathcal{W}_R$. We can now apply Lemma H.3 to show that

$$\mathcal{U}_{\widetilde{S}}(\mathcal{F}_R)\le 24e\int_0^{2(LR+b')LR}\log\big(1+\mathcal{N}(r,\mathcal{F}_R,d_{\widetilde{S}})\big)dr$$
$$\le 24e\int_0^{2(LR+b')LR}\Big(\log 2+d\log\Big(6LR(LR+b')r^{-1}\Big)\Big)\mathrm{d}r$$
$$\le 48e(LR+b')LR\big(\log 2+d\log(3e)\big),$$

where we have used

$$\int_0^{2(LR+b')LR}\log\Big(6LR(LR+b')r^{-1}\Big)\mathrm{d}r=2LR(LR+b')\int_0^1\log(3/r)\mathrm{d}r=2LR(LR+b')\log(3e).$$

The stated bound then follows by plugging the above bound on Rademacher chaos complexities into Theorem 11. The proof is complete. □

H.3 Proof of Corollary 13

Our scheme to prove Corollary 13 is to directly control the term $\Big\|\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor}\epsilon_i\nabla f(\mathbf{w};z_i,z_{i+\lfloor\frac{n}{2}\rfloor})\Big\|_2$ in Lemma H.2. In more details, we show this term is related to two Gaussian processes which are more easy to handle. Our analysis requires the following classical lemma on comparison between two Gaussian processes (Slepian's lemma).

Lemma H.4. *Let $\{\mathfrak{X}_\theta:\theta\in\Theta\}$ and $\{\mathfrak{Y}_\theta:\theta\in\Theta\}$ be two mean-zero separable Gaussian processes indexed by the same set Θ and suppose that*

$$\mathbb{E}[(\mathfrak{X}_\theta-\mathfrak{X}_{\bar\theta})^2]\le\mathbb{E}[(\mathfrak{Y}_\theta-\mathfrak{Y}_{\bar\theta})^2],\quad\forall\theta,\bar\theta\in\Theta.\tag{H.8}$$

Then $\mathbb{E}[\sup_{\theta\in\Theta}\mathfrak{X}_\theta]\le\mathbb{E}[\sup_{\theta\in\Theta}\mathfrak{Y}_\theta]$.

Lemma H.5. *Suppose $f:\mathcal{W}\times\mathcal{Z}^2\mapsto\mathbb{R}$ takes the form* (5.1)*. Then*

$$\mathbb{E}_\epsilon\sup_{\mathbf{w}\in\mathcal{W}_R}\Big\|\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor}\epsilon_i\nabla f(\mathbf{w};z_i,z_j)\Big\|_2\le\sqrt{2}\big(2L_\psi R\kappa+b'\big)\Big(\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor}\|\phi(x_i,x_{i+\lfloor\frac{n}{2}\rfloor})\|_2^2\Big)^{\frac{1}{2}}.$$

Proof. By the structure of f, we know

$$\mathbb{E}_\epsilon\sup_{\mathbf{w}\in\mathcal{W}_R}\Big\|\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor}\epsilon_i\nabla f(\mathbf{w};z_i,z_{i+\lfloor\frac{n}{2}\rfloor})\Big\|_2$$
$$=\mathbb{E}_\epsilon\sup_{\mathbf{w}\in\mathcal{W}_R}\Big\|\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor}\epsilon_i\psi'\big(\langle\mathbf{w},\phi(x_i,x_{i+\lfloor\frac{n}{2}\rfloor})\rangle,\tau(y_i,y_{i+\lfloor\frac{n}{2}\rfloor})\big)\phi(x_i,x_{i+\lfloor\frac{n}{2}\rfloor})\Big\|_2$$
$$=\mathbb{E}_\epsilon\sup_{\mathbf{w}\in\mathcal{W}_R,\mathbf{v}\in\mathcal{W}_1}\Big\langle\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor}\epsilon_i\psi'\big(\langle\mathbf{w},\phi(x_i,x_{i+\lfloor\frac{n}{2}\rfloor})\rangle,\tau(y_i,y_{i+\lfloor\frac{n}{2}\rfloor})\big)\phi(x_i,x_{i+\lfloor\frac{n}{2}\rfloor}),\mathbf{v}\Big\rangle$$
$$\le\mathbb{E}_g\sup_{\mathbf{w}\in\mathcal{W}_R,\mathbf{v}\in\mathcal{W}_1}\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor}g_i\psi'\big(\langle\mathbf{w},\phi(x_i,x_{i+\lfloor\frac{n}{2}\rfloor})\rangle,\tau(y_i,y_{i+\lfloor\frac{n}{2}\rfloor})\big)\langle\phi(x_i,x_{i+\lfloor\frac{n}{2}\rfloor}),\mathbf{v}\rangle,\tag{H.9}$$

where ψ' denotes the derivative of ψ w.r.t. the first argument, $g_1,\ldots,g_n$ are independent $N(0,1)$ random variables. Note the last step follows from the following inequality on Rademacher and Gaussian complexities

$$\mathbb{E}_\epsilon \sup_f \sum_{i=1}^{\lfloor \frac{n}{2}\rfloor} \epsilon_i f(z_i) \le \mathbb{E}_g \sup_f \sum_{i=1}^{\lfloor \frac{n}{2}\rfloor} g_i f(z_i).$$

Define two mean-zero separable Gaussian processes indexed by $\mathcal{W}_R \times \mathcal{W}_1$

$$\mathfrak{X}_{\mathbf{w},\mathbf{v}} = \sum_{i=1}^{\lfloor \frac{n}{2}\rfloor} g_i \psi'\big(\langle \mathbf{w}, \phi(x_i, x_{i+\lfloor \frac{n}{2}\rfloor})\rangle, \tau(y_i, y_{i+\lfloor \frac{n}{2}\rfloor})\big)\langle \phi(x_i, x_{i+\lfloor \frac{n}{2}\rfloor}), \mathbf{v}\rangle$$

$$\mathfrak{Y}_{\mathbf{w},\mathbf{v}} = \sqrt{2}\kappa \sum_{i=1}^{\lfloor \frac{n}{2}\rfloor} g_i \psi'\big(\langle \mathbf{w}, \phi(x_i, x_{i+\lfloor \frac{n}{2}\rfloor})\rangle, \tau(y_i, y_{i+\lfloor \frac{n}{2}\rfloor})\big) + \sqrt{2}\big(b' + L_\psi R\kappa\big) \sum_{i=1}^{\lfloor \frac{n}{2}\rfloor} \tilde{g}_i \langle \phi(x_i, x_{i+\lfloor \frac{n}{2}\rfloor}), \mathbf{v}\rangle,$$

where $\tilde{g}_1,\ldots,\tilde{g}_n$ are independent $N(0,1)$ random variables. For any $\mathbf{w},\mathbf{w}' \in \mathcal{W}_R$ and $\mathbf{v},\mathbf{v}' \in \mathcal{W}_1$, it follows from the independence among g_i and $\mathbb{E}g_i^2 = 1, \forall i = 1,\ldots,n$ that

$$\begin{aligned}
\mathbb{E}_g\big[\big(\mathfrak{X}_{\mathbf{w},\mathbf{v}} - \mathfrak{X}_{\mathbf{w}',\mathbf{v}'}\big)^2\big] &= \sum_{i=1}^{\lfloor \frac{n}{2}\rfloor} \Big(\psi'\big(\langle \mathbf{w}, \phi(x_i, x_{i+\lfloor \frac{n}{2}\rfloor})\rangle, \tau(y_i, y_{i+\lfloor \frac{n}{2}\rfloor})\big)\langle \phi(x_i, x_{i+\lfloor \frac{n}{2}\rfloor}), \mathbf{v}\rangle \\
&\qquad - \psi'\big(\langle \mathbf{w}', \phi(x_i, x_{i+\lfloor \frac{n}{2}\rfloor})\rangle, \tau(y_i, y_{i+\lfloor \frac{n}{2}\rfloor})\big)\langle \phi(x_i, x_{i+\lfloor \frac{n}{2}\rfloor}), \mathbf{v}'\rangle\Big)^2 \\
&\le 2\sum_{i=1}^{\lfloor \frac{n}{2}\rfloor} \Big(\psi'\big(\langle \mathbf{w}, \phi(x_i, x_{i+\lfloor \frac{n}{2}\rfloor})\rangle, \tau(y_i, y_{i+\lfloor \frac{n}{2}\rfloor})\big) - \psi'\big(\langle \mathbf{w}', \phi(x_i, x_{i+\lfloor \frac{n}{2}\rfloor})\rangle, \tau(y_i, y_{i+\lfloor \frac{n}{2}\rfloor})\big)\Big)^2 \big(\langle \phi(x_i, x_{i+\lfloor \frac{n}{2}\rfloor}), \mathbf{v}\rangle\big)^2 \\
&+ 2\sum_{i=1}^{\lfloor \frac{n}{2}\rfloor} \Big(\psi'\big(\langle \mathbf{w}', \phi(x_i, x_{i+\lfloor \frac{n}{2}\rfloor})\rangle, \tau(y_i, y_{i+\lfloor \frac{n}{2}\rfloor})\big)\Big)^2 \big(\langle \phi(x_i, x_{i+\lfloor \frac{n}{2}\rfloor}), \mathbf{v}\rangle - \langle \phi(x_i, x_{i+\lfloor \frac{n}{2}\rfloor}), \mathbf{v}'\rangle\big)^2 \\
&\le 2\kappa^2 \sum_{i=1}^{\lfloor \frac{n}{2}\rfloor} \Big(\psi'\big(\langle \mathbf{w}, \phi(x_i, x_{i+\lfloor \frac{n}{2}\rfloor})\rangle, \tau(y_i, y_{i+\lfloor \frac{n}{2}\rfloor})\big) - \psi'\big(\langle \mathbf{w}', \phi(x_i, x_{i+\lfloor \frac{n}{2}\rfloor})\rangle, \tau(y_i, y_{i+\lfloor \frac{n}{2}\rfloor})\big)\Big)^2 \\
&+ 2\big(b' + L_\psi R\kappa\big)^2 \sum_{i=1}^{\lfloor \frac{n}{2}\rfloor} \big(\langle \phi(x_i, x_{i+\lfloor \frac{n}{2}\rfloor}), \mathbf{v}\rangle - \langle \phi(x_i, x_{i+\lfloor \frac{n}{2}\rfloor}), \mathbf{v}'\rangle\big)^2 \\
&= \mathbb{E}_g\big[\big(\mathfrak{Y}_{\mathbf{w},\mathbf{v}} - \mathfrak{Y}_{\mathbf{w}',\mathbf{v}'}\big)^2\big],
\end{aligned}$$

where we have used $(a+b)^2 \le 2a^2 + 2b^2$, the decomposition

$$\begin{aligned}
&\psi'\big(\langle \mathbf{w}, \phi(x_i, x_{i+\lfloor \frac{n}{2}\rfloor})\rangle, \tau(y_i, y_{i+\lfloor \frac{n}{2}\rfloor})\big)\langle \phi(x_i, x_{i+\lfloor \frac{n}{2}\rfloor}), \mathbf{v}\rangle - \psi'\big(\langle \mathbf{w}', \phi(x_i, x_{i+\lfloor \frac{n}{2}\rfloor})\rangle, \tau(y_i, y_{i+\lfloor \frac{n}{2}\rfloor})\big)\langle \phi(x_i, x_{i+\lfloor \frac{n}{2}\rfloor}), \mathbf{v}'\rangle \\
&= \Big(\psi'\big(\langle \mathbf{w}, \phi(x_i, x_{i+\lfloor \frac{n}{2}\rfloor})\rangle, \tau(y_i, y_{i+\lfloor \frac{n}{2}\rfloor})\big) - \psi'\big(\langle \mathbf{w}', \phi(x_i, x_{i+\lfloor \frac{n}{2}\rfloor})\rangle, \tau(y_i, y_{i+\lfloor \frac{n}{2}\rfloor})\big)\Big)\langle \phi(x_i, x_{i+\lfloor \frac{n}{2}\rfloor}), \mathbf{v}\rangle \\
&\qquad + \psi'\big(\langle \mathbf{w}', \phi(x_i, x_{i+\lfloor \frac{n}{2}\rfloor})\rangle, \tau(y_i, y_{i+\lfloor \frac{n}{2}\rfloor})\big)\big(\langle \phi(x_i, x_{i+\lfloor \frac{n}{2}\rfloor}), \mathbf{v}\rangle - \langle \phi(x_i, x_{i+\lfloor \frac{n}{2}\rfloor}), \mathbf{v}'\rangle\big)
\end{aligned}$$

and the following inequality due to the L_ψ-smoothness of ϕ

$$\big|\psi'\big(\langle \mathbf{w}', \phi(x_i, x_{i+\lfloor \frac{n}{2}\rfloor})\rangle, \tau(y_i, y_{i+\lfloor \frac{n}{2}\rfloor})\big)\big| \le b' + L_\psi |\langle \mathbf{w}', \phi(x_i, x_{i+\lfloor \frac{n}{2}\rfloor})\rangle - 0| \le b' + L_\psi R\kappa.$$

Therefore, we can apply Lemma H.4 to show

$$\begin{aligned}
&\mathbb{E}_g \sup_{\mathbf{w}\in\mathcal{W}_R, \mathbf{v}\in\mathcal{W}_1} \mathfrak{X}_{\mathbf{w},\mathbf{v}} \le \mathbb{E}_g \sup_{\mathbf{w}\in\mathcal{W}_R, \mathbf{v}\in\mathcal{W}_1} \mathfrak{Y}_{\mathbf{w},\mathbf{v}} \\
&\le \sqrt{2}\kappa \mathbb{E}_g \sup_{\mathbf{w}\in\mathcal{W}_R} \sum_{i=1}^{\lfloor \frac{n}{2}\rfloor} g_i \psi'\big(\langle \mathbf{w}, \phi(x_i, x_{i+\lfloor \frac{n}{2}\rfloor})\rangle, \tau(y_i, y_{i+\lfloor \frac{n}{2}\rfloor})\big) + \sqrt{2}\big(b' + L_\psi R\kappa\big)\mathbb{E}_g \sup_{\mathbf{v}\in\mathcal{W}_1} \sum_{i=1}^{\lfloor \frac{n}{2}\rfloor} g_i \langle \phi(x_i, x_{i+\lfloor \frac{n}{2}\rfloor}), \mathbf{v}\rangle \\
&\le \sqrt{2}L_\psi\kappa \mathbb{E}_g \sup_{\mathbf{w}\in\mathcal{W}_R} \sum_{i=1}^{\lfloor \frac{n}{2}\rfloor} g_i \langle \mathbf{w}, \phi(x_i, x_{i+\lfloor \frac{n}{2}\rfloor})\rangle + \sqrt{2}\big(b' + L_\psi R\kappa\big)\mathbb{E}_g \sup_{\mathbf{v}\in\mathcal{W}_1} \sum_{i=1}^{\lfloor \frac{n}{2}\rfloor} g_i \langle \phi(x_i, x_{i+\lfloor \frac{n}{2}\rfloor}), \mathbf{v}\rangle,
\end{aligned}$$

where we have used the L_ψ-Lipschitz continuity of ψ' and the contraction lemma of Gaussian complexities in the last step. Furthermore, it follows from the Jensen's inequality that

$$\begin{aligned}
&\mathbb{E}_g \sup_{\mathbf{w}\in\mathcal{W}_R} \sum_{i=1}^{\lfloor \frac{n}{2}\rfloor} g_i\langle \mathbf{w}, \phi(x_i, x_{i+\lfloor \frac{n}{2}\rfloor})\rangle = \mathbb{E}_g \sup_{\mathbf{w}\in\mathcal{W}_R} \Big\langle \mathbf{w}, \sum_{i=1}^{\lfloor \frac{n}{2}\rfloor} g_i\phi(x_i, x_{i+\lfloor \frac{n}{2}\rfloor})\Big\rangle \\
&\leq R\mathbb{E}_g\Big\|\sum_{i=1}^{\lfloor \frac{n}{2}\rfloor} g_i\phi(x_i, x_{i+\lfloor \frac{n}{2}\rfloor})\Big\|_2 \leq R\sqrt{\mathbb{E}_g\Big[\Big\langle \sum_{i=1}^{\lfloor \frac{n}{2}\rfloor} g_i\phi(x_i, x_{i+\lfloor \frac{n}{2}\rfloor}), \sum_{i=1}^{\lfloor \frac{n}{2}\rfloor} g_i\phi(x_i, x_{i+\lfloor \frac{n}{2}\rfloor})\Big\rangle\Big]} \\
&= R\Big(\sum_{i=1}^{\lfloor \frac{n}{2}\rfloor}\|\phi(x_i, x_{i+\lfloor \frac{n}{2}\rfloor})\|_2^2\Big)^{\frac{1}{2}}.
\end{aligned}$$

In a similar way, we can show

$$\mathbb{E}_g \sup_{\mathbf{v}\in\mathcal{W}_1} \sum_{i=1}^{\lfloor \frac{n}{2}\rfloor} g_i\langle \phi(x_i, x_{i+\lfloor \frac{n}{2}\rfloor}), \mathbf{v}\rangle \leq \Big(\sum_{i=1}^{\lfloor \frac{n}{2}\rfloor}\|\phi(x_i, x_{i+\lfloor \frac{n}{2}\rfloor})\|_2^2\Big)^{\frac{1}{2}}.$$

Therefore,

$$\mathbb{E}_g \sup_{\mathbf{w}\in\mathcal{W}_R, \mathbf{v}\in\mathcal{W}_1} \mathfrak{X}_{\mathbf{w},\mathbf{v}} \leq \big(2L_\psi R\kappa + b'\big)\sqrt{2}\Big(\sum_{i=1}^{\lfloor \frac{n}{2}\rfloor}\|\phi(x_i, x_{i+\lfloor \frac{n}{2}\rfloor})\|_2^2\Big)^{\frac{1}{2}}.$$

Plugging the above inequality into (H.9) then gives the stated bound. The proof is complete. □

We now apply Lemma H.5 to prove Corollary 13.

Proof of Corollary 13. By Lemma H.5 and the definition of κ, we know

$$\mathbb{E}_\epsilon \sup_{\mathbf{w}\in\mathcal{W}_R} \Big\|\sum_{i=1}^{\lfloor \frac{n}{2}\rfloor} \epsilon_i \nabla f(\mathbf{w}; z_i, z_j)\Big\|_2 \leq \sqrt{n}\kappa\big(2L_\psi R\kappa + b'\big). \tag{H.10}$$

According to the L_ψ-smoothness of ψ, the function f is $(L_\psi\kappa^2)$-smooth

$$\begin{aligned}
&\big\|\nabla f(\mathbf{w}; z, z') - \nabla f(\tilde{\mathbf{w}}; z, z')\big\|_2 \\
&= \big|\psi'(\langle \mathbf{w}, \phi(x, x')\rangle, \tau(y, y')) - \psi'(\langle \tilde{\mathbf{w}}, \phi(x, x')\rangle, \tau(y, y'))\big|\|\phi(x, x')\|_2 \\
&\leq L_\psi|\langle \mathbf{w} - \tilde{\mathbf{w}}, \phi(x, x')\rangle|\|\phi(x, x')\|_2 \leq L_\psi\kappa^2\|\mathbf{w} - \tilde{\mathbf{w}}\|_2.
\end{aligned}$$

Therefore, Lemma H.2 holds with $L = L_\psi\kappa^2$. We can plug (H.10) into Lemma H.2 and get the stated bound. The proof is complete. □

I Proofs on Nonconvex Problems

In this section, we apply the uniform convergence of gradients to prove Theorem 14.

Proof of Theorem 14. By the elementary inequality $(a+b)^2 \leq 2(a^2+b^2)$ and (D.6), we derive the following inequality with probability $1 - \delta/3$

$$\begin{aligned}
\sum_{t=1}^{T}\eta_t\|\nabla F(\mathbf{w}_t)\|_2^2 &= \sum_{t=1}^{T}\eta_t\big\|\nabla F(\mathbf{w}_t) - \nabla F_S(\mathbf{w}_t) + \nabla F_S(\mathbf{w}_t)\big\|_2^2 \\
&\leq 2\sum_{t=1}^{T}\eta_t\big\|\nabla F(\mathbf{w}_t) - \nabla F_S(\mathbf{w}_t)\big\|_2^2 + 2\sum_{t=1}^{T}\eta_t\big\|\nabla F_S(\mathbf{w}_t)\big\|_2^2 \\
&\leq 2\sum_{t=1}^{T}\eta_t \max_{t=1,\ldots,T}\big\|\nabla F(\mathbf{w}_t) - \nabla F_S(\mathbf{w}_t)\big\|_2^2 + O\big(\sum_{t=1}^{T}\eta_t^2 + \log(1/\delta)\big).
\end{aligned}$$

It then follows that

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\mathbf{w}_t)\|_2^2 \leq 2\max_{t=1,\ldots,T}\big\|\nabla F(\mathbf{w}_t)-\nabla F_S(\mathbf{w}_t)\big\|_2^2 + O(1)\Big(\sum_{t=1}^{T}\eta_t\Big)^{-1}\big(\sum_{t=1}^{T}\eta_t^2+\log(1/\delta)\big)$$

$$= 2\max_{t=1,\ldots,T}\big\|\nabla F(\mathbf{w}_t)-\nabla F_S(\mathbf{w}_t)\big\|_2^2 + O\Big(T^{-\frac{1}{2}}\log(1/\delta)\Big). \tag{I.1}$$

According to (D.7), with probability $1-\delta/3$ we have the following inequality uniformly for all $t=1,\ldots,T$

$$\|\mathbf{w}_t\|_2 \leq R_T := O\Big(T^{\frac{1}{4}}\log(1/\delta)\Big). \tag{I.2}$$

According to Corollary 12, the following inequality holds with probability $1-\delta/3$ (we assume $R_T\geq 1$)

$$\sup_{\mathbf{w}\in\mathcal{W}_{R_T}}\|\nabla F(\mathbf{w})-\nabla F_S(\mathbf{w})\|_2 = O\Big(R_T\sqrt{d+\log(1/\delta)}n^{-\frac{1}{2}}\Big). \tag{I.3}$$

Combining (I.1), (I.2) and (I.3) together, with probability $1-\delta$ we derive the following inequality

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\mathbf{w}_t)\|_2^2 &= 2\max_{t=1,\ldots,T}\big\|\nabla F(\mathbf{w}_t)-\nabla F_S(\mathbf{w}_t)\big\|_2^2 + O\Big(T^{-\frac{1}{2}}\log(1/\delta)\Big)\\ &= O\Big(R_T^2\big(d+\log(1/\delta)\big)n^{-1}\Big) + O\Big(T^{-\frac{1}{2}}\log(1/\delta)\Big)\\ &= O\Big(\sqrt{T}\log^2(1/\delta)\big(d+\log(1/\delta)\big)n^{-1}\Big) + O\Big(T^{-\frac{1}{2}}\log(1/\delta)\Big).\end{aligned}$$

Therefore, we can choose $T\asymp nd^{-1}$ to derive the following inequality with probability $1-\delta$

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\mathbf{w}_t)\|_2^2 = O\Big(n^{-\frac{1}{2}}\log^2(1/\delta)\big(d+\log(1/\delta)\big)^{\frac{1}{2}}\Big).$$

This gives the bound (5.4).

The proof of (5.5) is the same except using the uniform convergence of gradients established in Corollary 13 instead of Corollary 12. We omit the proof for simplicity. The proof is complete. □

J Proofs on Gradient Dominated Problems

In this section, we prove Theorem 15 on excess risk bounds for learning with gradient dominated problems. The following lemma is a simple extension of a similar result in [6].

Lemma J.1. *Assume for all z,z', the function $\mathbf{w}\mapsto f(\mathbf{w};z,z')$ is nonnegative and G-Lipschitz. Let $S=\{z_1,\ldots,z_n\}$ and $S'=\{z'_1,\ldots,z'_n\}$ be two datasets that differ by the first point. Let $\{\mathbf{w}_t\},\{\mathbf{w}'_t\}$ be the sequence produced by SGD (Algorithm 1) w.r.t. S and S', respectively. Then for every $z,z'\in\mathcal{Z}$ and every $t_0\in[n]$ we have*

$$\mathbb{E}\big[|f(\mathbf{w}_T;z,z')-f(\mathbf{w}'_T;z,z')|\big] \leq \frac{2Bt_0}{n}\sup_{\mathbf{w};z,z'}f(\mathbf{w};z,z')+G\mathbb{E}\big[\|\mathbf{w}_T-\mathbf{w}'_T\||1\notin I_{t_0}(A)\big]\Pr\{1\notin I_{t_0}(A)\},$$

where $I_t(A):=\{i_1,j_1,\ldots,i_t,j_t\}$ is the set of indices selected by A in the first t iterations.

Proof. According to the law of total expectation, we know

$$\begin{aligned}\mathbb{E}\big[|f(\mathbf{w}_T;z,z')-f(\mathbf{w}'_T;z,z')|\big] &= \mathbb{E}\big[|f(\mathbf{w}_T;z,z')-f(\mathbf{w}'_T;z,z')||1\in I_{t_0}(A)\big]\Pr\{1\in I_{t_0}(A)\}\\ &+\mathbb{E}\big[|f(\mathbf{w}_T;z,z')-f(\mathbf{w}'_T;z,z')||1\notin I_{t_0}(A)\big]\Pr\{1\notin I_{t_0}(A)\}.\end{aligned}$$

According to the update rule, we know

$$\Pr\{1\in I_{t_0}(A)\}\leq\sum_{t=1}^{t_0}\Pr\{i_t=1 \text{ or } j_t=1\}=\sum_{t=1}^{t_0}\frac{2(n-1)}{n(n-1)}=\frac{2t_0}{n}.$$

The stated bound then follows from the Lipschitz continuity of f. The proof is complete. □

We follow the arguments in [6] to prove Theorem 15.

Proof of Theorem 15. We first give the stability bounds. Suppose S and S' differ by the first example. If $i_t \neq 1$ and $j_t \neq 1$, then

$$\begin{aligned}\|\mathbf{w}_{t+1} - \mathbf{w}'_{t+1}\|_2 &= \big\|\mathbf{w}_t - \eta_t \nabla f(\mathbf{w}_t; z_{i_t}, z_{j_t}) - \mathbf{w}'_t + \eta_t \nabla f(\mathbf{w}'_t; z'_{i_t}, z'_{j_t})\big\|_2 \\ &\leq \big\|\mathbf{w}_t - \mathbf{w}'_t\big\|_2 + \big\|\eta_t \nabla f(\mathbf{w}_t; z_{i_t}, z_{j_t}) - \eta_t \nabla f(\mathbf{w}'_t; z_{i_t}, z_{j_t})\big\|_2 \\ &\leq \big(1 + L\eta_t\big)\big\|\mathbf{w}_t - \mathbf{w}'_t\big\|_2.\end{aligned}$$

Otherwise, we have

$$\|\mathbf{w}_{t+1} - \mathbf{w}'_{t+1}\|_2 \leq \|\mathbf{w}_t - \mathbf{w}'_t\|_2 + 2G\eta_t.$$

It then follows that

$$\begin{aligned}&\mathbb{E}_{(i_t,j_t)}\big[\|\mathbf{w}_{t+1} - \mathbf{w}'_{t+1}\|_2\big] \\ &\leq \big(1 + L\eta_t\big)\big\|\mathbf{w}_t - \mathbf{w}'_t\big\|_2 \Pr\{i_t \neq 1 \text{ and } j_t \neq 1\} + \big(\|\mathbf{w}_t - \mathbf{w}'_t\|_2 + 2G\eta_t\big)\Pr\{i_t = 1 \text{ or } j_t = 1\} \\ &= \frac{(n-2)\big(1 + L\eta_t\big)}{n}\big\|\mathbf{w}_t - \mathbf{w}'_t\big\|_2 + \frac{2}{n}\big(\|\mathbf{w}_t - \mathbf{w}'_t\|_2 + 2G\eta_t\big). \qquad (J.1)\end{aligned}$$

Let $\triangle_t = \mathbb{E}[\|\mathbf{w}_t - \mathbf{w}'_t\| | 1 \notin I_{t_0}(A)]$, where $I_{t_0}(A)$ is defined in Lemma J.1. Then it follows from (J.1) that

$$\begin{aligned}\triangle_{t+1} &\leq \frac{(n-2)\big(1 + L\eta_t\big)}{n}\triangle_t + \frac{2}{n}\big(\triangle_t + 2G\eta_t\big) \leq \big(1 + L(1 - 2/n)\eta_t\big)\triangle_t + \frac{4G\eta_t}{n} \\ &\leq \exp\big(L(1 - 2/n)\eta_t\big)\triangle_t + \frac{4G\eta_t}{n}.\end{aligned}$$

Since $\triangle_{t_0+1} = 0$, we can apply the above inequality repeatedly and get

$$\begin{aligned}\triangle_T &\leq \sum_{t=t_0+1}^{T} \prod_{k=t+1}^{T} \exp\big(L(1 - 2/n)\eta_k\big)\frac{4G\eta_t}{n} \leq \sum_{t=t_0+1}^{T} \prod_{k=t+1}^{T} \exp\big(Lc(1 - 2/n)/k\big)\frac{4Gc}{nt} \\ &\leq \sum_{t=t_0+1}^{T} \exp\Big(Lc(1 - 2/n)\sum_{k=t+1}^{T}\frac{1}{k}\Big)\frac{4Gc}{nt} \leq \sum_{t=t_0+1}^{T} \exp\Big(Lc(1 - 2/n)\log(T/t)\Big)\frac{4Gc}{nt} \\ &\leq \sum_{t=t_0+1}^{T}\Big(\frac{T}{t}\Big)^{Lc(1-2/n)}\frac{4Gc}{nt} = \frac{4Gc}{n}T^{Lc(1-2/n)}\sum_{t=t_0+1}^{T} t^{-Lc(1-2/n)-1} \\ &\leq \frac{4Gc}{n}T^{Lc(1-2/n)}\int_{t_0}^{T} x^{-Lc(1-2/n)-1}dx \leq \frac{1}{Lc(1 - 2/n)}\frac{4Gc}{n}\Big(\frac{T}{t_0}\Big)^{Lc(1-2/n)},\end{aligned}$$

where we have used

$$\eta_t = \frac{2t+1}{2\beta(t+1)^2} \leq c/t, \quad c := 1/\beta.$$

We can combine the above bound and Lemma J.1 together, and get

$$\mathbb{E}\big[|f(\mathbf{w}_T; z, z') - f(\mathbf{w}'_T; z, z')|\big] = O\Big(\frac{t_0}{n} + \frac{G^2}{nL}\Big(\frac{T}{t_0}\Big)^{Lc}\Big).$$

We can choose $t_0 \asymp T^{\frac{Lc}{Lc+1}}$ and get the following stability bounds

$$\mathbb{E}\big[|f(\mathbf{w}_T; z, z') - f(\mathbf{w}'_T; z, z')|\big] = O\Big(\frac{T^{\frac{Lc}{Lc+1}}}{n}\Big).$$

We can plug the above stability bounds into Part (a) of Theorem 1, and get the following generalization bounds

$$\mathbb{E}\big[F(\mathbf{w}_T) - F_S(\mathbf{w}_T)\big] = O\Big(\frac{T^{\frac{L/\beta}{L/\beta+1}}}{n}\Big).$$

Furthermore, according to (D.8) we have the following optimization error bounds

$$\mathbb{E}_A[F_S(\mathbf{w}_T)] - \inf_{\mathbf{w}}[F_S(\mathbf{w})] = O\big(1/(T\beta^2)\big).$$

We can plug the above generalization and optimization error bounds into (3.1), and get (5.7). The proof is complete. □

K Examples of Pairwise Learning

In this section, we give some specific examples of pairwise learning: metric learning, ranking and AUC maximization. We denote $(t)_+ := \max(t, 0)$ and $x^\top$ the transpose of $x \in \mathbb{R}^d$. Let sign(t) denote the sign of $t \in \mathbb{R}$.

Supervised metric learning. In supervised metric learning, we assume $\mathcal{Y} = \{\pm 1\}$ and aim to find a distance metric such that examples in the same class are similar while examples in different classes are apart from each other under this metric. A typical choice is the Mahalanobis metric of the form $h_{\mathbf{w}}(x_i, x_j) = \langle \mathbf{w}, (x_i - x_j)(x_i - x_j)^\top \rangle, \mathbf{w} \in \mathbb{S}^{d\times d}$, where $\mathbb{S}^{d\times d}$ denotes the set of positive semi-definite matrices in $\mathbb{R}^{d\times d}$. A common loss function in metric learning for $\mathbf{w}$ on $z = (x, y), z' = (x', y')$ takes the form [8]

$$f(\mathbf{w}; z, z') = g(yy'(1 - h_{\mathbf{w}}(x, x'))),$$

where $g : \mathbb{R} \to \mathbb{R}_+$ is a convex function for which some typical choices include the hinge loss $g(t) = (1 - t)_+$ and the logistic loss $g(t) = \log(1 + \exp(-t))$.

Ranking. For ranking problems, the output reflects the ordering between instances, i.e., the instance x is considered to be better than x' if $y > y'$. Our task is to predict the ordering between the objects based on observations by constructing ranking rules $h_{\mathbf{w}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and predict $y > y'$ if $h_{\mathbf{w}}(x, x') > 0$ [3]. A common pairwise loss function used in ranking problems takes the form

$$f(\mathbf{w}; z, z') = g(\text{sign}(y - y')h_{\mathbf{w}}(x, x')),$$

where $g : \mathbb{R} \to \mathbb{R}_+$ is a convex function for which some typical choices are the exponential loss $g(t) = \exp(-t)$, the logistic loss $g(t) = \log(1 + \exp(-t))$ and the hinge loss $g(t) = (1 - t)_+$ [3].

AUC maximization. AUC is a widely used metric for measuring the performance of machine learning algorithms in imbalanced classification. If $\mathcal{Y} = \{\pm 1\}$, the AUC score of a model $h_{\mathbf{w}} : \mathcal{X} \mapsto \mathcal{Y}$ measures its probability of giving a larger value to a positive instance than to a negative instance. The problem of AUC maximization can be formulated as a pairwise learning problem with the following loss function [4, 22]

$$f(\mathbf{w}; z, z') = g(\mathbf{w}^\top (x - x'))\mathbb{I}_{[y=1, y'=-1]}, \tag{K.1}$$

where $g : \mathbb{R} \to \mathbb{R}_+$ is a convex function for which some typical choices are the least square loss $g(t) = (1 - t)^2$, the logistic loss $g(t) = \log(1 + \exp(-t))$ and the hinge loss $g(t) = (1 - t)_+$.

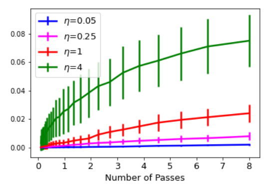
L Experimental Results

In this section, we present some experimental results to support our theory on the stability bounds. We consider AUC maximization with the loss function of the form of (K.1). We consider several datasets available at the LIBSVM site [2], whose information is summarized in Table L.1. We transform datasets with multiple class labels into datasets with binary class labels by grouping the first half of class labels into positive labels, and grouping the remaining class labels into negative labels. We randomly choose 80 percents of each dataset as the training set S, from which we perturb a single example in S to create a neighboring dataset S'. We apply SGD (3.2) with the same parameters to S and S', and get two sequence of iterates $\{\mathbf{w}_t\}$ and $\{\mathbf{w}'_t\}$. We then calculate the Euclidean distance $\triangle_t = \|\mathbf{w}_t - \mathbf{w}'_t\|_2$ at each iteration to verify the stability of SGD. We consider step sizes of the form $\eta_t = \eta/\sqrt{T}$ with $\eta \in \{0.05, 0.25, 1, 4\}$, and report $\triangle_t$ as a function of the number of passes (the iteration number t divided by the sample size n). We repeat the experiments 100 times and report the average as well as the standard deviation. Since we develop stability bounds for both smooth and nonsmooth loss functions, we consider two representative loss functions: the smooth logistic loss (i.e., Eq. (K.1) with $g(t) = \log(1 + \exp(-t))$) and the nonsmooth hinge loss (i.e., Eq. (K.1) with $g(t) = (1 - t)_+$).

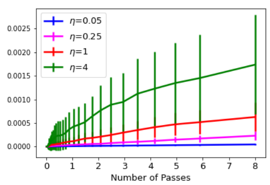
In Figure L.1, we report the Euclidean distance $\triangle_t$ for AUC maximization with the hinge loss and the 4 stepsize sequences, while in Figure L.2, we report $\triangle_t$ for AUC maximization with the logistic loss. It is clear that $\triangle_t$ is an increasing function of both t and η, which is consistent with our stability bounds in Theorem 3 and Theorem 6. It is also clear that the Euclidean distances for the logistic loss are significantly smaller than those for the hinge loss if we consider the same stepsize sequence, which is also consistent with Remark 4 on the comparison of stability bounds for SGD with smooth and nonsmooth problems.

Table L.1: Description of the datasets used in the experiments.

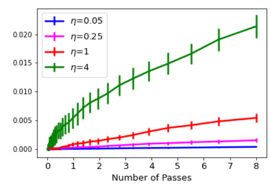
datasets	# inst	# feat	datasets	# inst	# feat	datasets	# inst	# feat	datasets	# inst	# feat
a3a	3185	122	acoustic	78823	50	cifar10	50000	3072	gisette	7000	5000
madelon	2600	500	mnist	60000	780	usps	7291	256	webspam_u	350000	254



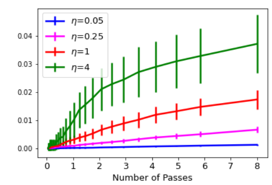
(a) a3a



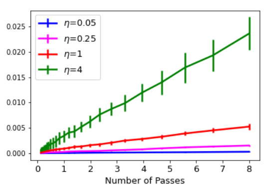
(b) acoustic



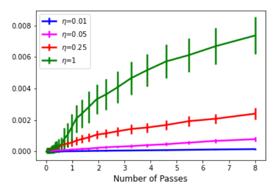
(c) cifar10



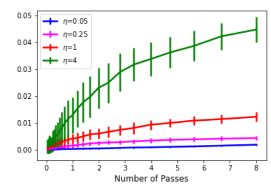
(d) gisette



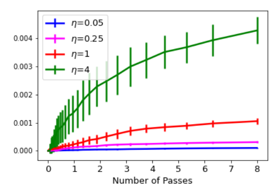
(e) madelon



(f) mnist

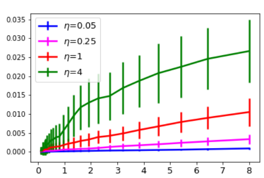


(g) usps

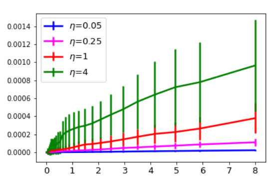


(h) webspam_u

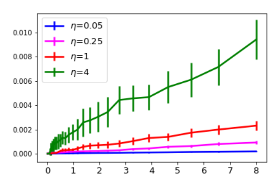
Figure L.1: Euclidean distance $\triangle_t$ as a function of the number of passes for the hinge loss.



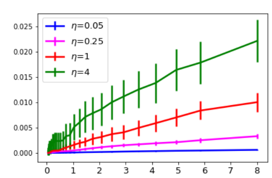
(a) a3a



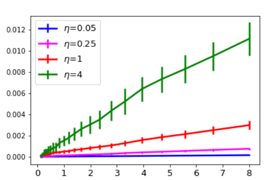
(b) acoustic



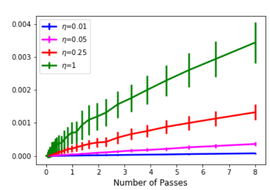
(c) cifar10



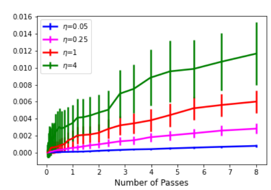
(d) gisette



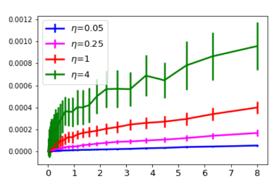
(e) madelon



(f) mnist



(g) usps



(h) webspam_u

Figure L.2: Euclidean distance $\triangle_t$ as a function of the number of passes for the logistic loss.

References

[1] O. Bousquet, Y. Klochkov, and N. Zhivotovskiy. Sharper bounds for uniformly stable algorithms. In *Conference on Learning Theory*, pages 610–626, 2020.

[2] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27, 2011.

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

[4] C. Cortes and M. Mohri. AUC optimization vs. error rate minimization. In *Advances in Neural Information Processing Systems*, pages 313–320, 2004.

[5] V. Feldman, T. Koren, and K. Talwar. Private stochastic convex optimization: optimal rates in linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 439–449, 2020.

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

[7] N. J. Harvey, C. Liaw, and S. Randhawa. Simple and optimal high-probability bounds for strongly-convex stochastic gradient descent. *arXiv preprint arXiv:1909.00843*, 2019.

[8] R. Jin, S. Wang, and Y. Zhou. Regularized distance metric learning: Theory and algorithm. In *Advances in Neural Information Processing Systems*, pages 862–870, 2009.

[9] H. Karimi, J. Nutini, and M. Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *European Conference on Machine Learning*, pages 795–811, 2016.

[10] Y. Klochkov and N. Zhivotovskiy. Stability and deviation optimal risk bounds with convergence rate O(1/n). *arXiv preprint arXiv:2103.12024*, 2021.

[11] J. Kulkarni, Y. T. Lee, and D. Liu. Private non-smooth empirical risk minimization and stochastic convex optimization in subquadratic steps. *arXiv preprint arXiv:2103.15352*, 2021.

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

[13] Y. Lei and L. Ding. Refined rademacher chaos complexity bounds with applications to the multikernel learning problem. *Neural Computation*, 26(4):1–22, 2014.

[14] Y. Lei and K. Tang. Learning rates for stochastic gradient descent with nonconvex objectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. doi: 10.1109/TPAMI.2021.3068154.

[15] Y. Lei and Y. Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *International Conference on Machine Learning*, pages 5809–5819, 2020.

[16] Y. Lei, A. Ledent, and M. Kloft. Sharper generalization bounds for pairwise learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 21236–21246, 2020.

[17] 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.

[18] G. Pisier. *The Volume of Convex Bodies and Banach Space Geometry*, volume 94. Cambridge University Press, 1999.

[19] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[20] N. Srebro, K. Sridharan, and A. Tewari. Smoothness, low noise and fast rates. In *Advances in Neural Information Processing Systems*, pages 2199–2207, 2010.

[21] Y. Ying and C. Campbell. Rademacher chaos complexities for learning the kernel problem. *Neural computation*, 22(11):2858–2886, 2010.

[22] Y. Ying, L. Wen, and S. Lyu. Stochastic online AUC maximization. In *Advances in Neural Information Processing Systems*, pages 451–459, 2016.

Checklist

1. For all authors...

(a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
(b) Did you describe the limitations of your work? [Yes]
(c) Did you discuss any potential negative societal impacts of your work? [N/A]
(d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]