Sharper Generalization Bounds for Pairwise Learning

Yunwen Lei[1] **Antoine Ledent**[2*] **Marius Kloft**[2]
[1]School of Computer Science, University of Birmingham, Birmingham B15 2TT, United Kingdom
[2]Department of Computer Science, TU Kaiserslautern, Kaiserslautern 67653, Germany
y.lei@bham.ac.uk ledent@cs.uni-kl.de kloft@cs.uni-kl.de

Abstract

Pairwise learning refers to learning tasks with loss functions depending on a pair of training examples, which includes ranking and metric learning as specific examples. Recently, there has been an increasing amount of attention on the generalization analysis of pairwise learning to understand its practical behavior. However, the existing stability analysis provides suboptimal high-probability generalization bounds. In this paper, we provide a refined stability analysis by developing generalization bounds which can be $\sqrt{n}$-times faster than the existing results, where n is the sample size. This implies excess risk bounds of the order $O(n^{-1/2})$ (up to a logarithmic factor) for both regularized risk minimization and stochastic gradient descent. We also introduce a new on-average stability measure to develop optimistic bounds in a low noise setting. We apply our results to ranking and metric learning, and clearly show the advantage of our generalization bounds over the existing analysis.

1 Introduction

In modern machine learning, we frequently encounter problems where the performance of a model depends on *pairs* of training instances. As examples consider the following. In ranking problems, our aim is to learn a function that can predict the ordering of examples [13, 44]. In metric learning, which plays a key role in clustering problems [9, 28], we wish to learn an adequate distance metric between instances. In AUC maximization, which is deployed to class-imbalanced learning problems, we aim to find a classifier that maximizes the probability of scoring a positive example higher than a negative one [14]. Further examples include learning with minimum error entropy loss functions [27], multiple kernel learning [31], preference learning [22], and gradient learning [41]. All these so-called *pairwise learning* problems involve a loss function based on pairs of training examples. This is in a sharp contrast to classification and regression, where the loss function depends only on a single instance. Those problems are referred to as *pointwise learning* problems.

In machine learning, we frequently build predictive models by optimizing their empirical behavior on training instances, that is, to achieve a small training error. However, a small training error does not imply that the learnt models will generalize well to test examples. Generalization analysis—which is a central topic in statistical learning theory (SLT) [40]—studies the generalization gap between the training and testing errors. There is a large amount of work on the generalization analysis of learning algorithms, largely based on either algorithmic stability [7, 17], complexity analysis of models [3, 52], PAC-Bayesian analysis [38], or integral operators [49, 53]. Most of this work focuses on pointwise learning, while pairwise learning is far less studied. A difficulty occurring in the generalization analysis of pairwise learning is that the objective function is not a sum of identically and independently distributed (i.i.d.) random variables [1, 9, 13, 30]—a fundamental assumption in SLT.

*The first two authors contributed equally

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

In this paper, we employ the methodology of algorithmic stability for generalization analysis of pairwise learning. Appealingly, algorithmic stability considers just the one prediction function output by the learner [7], while methods based on uniform convergence, such as the Rademacher complexity [3], bound the difference of training and testing errors for *all* prediction functions. The latter approach generally involves a square-root dependency on the input dimension [2, 18, 54]. For comparison, algorithmic stability enables dimension-independent generalization bounds [20].

While there is preliminary work on the algorithmic stability of metric learning and ranking, the resulting generalization bounds are not satisfactory. The best existing bounds decay at the order of $O(\gamma\sqrt{n})$ [1, 28, 55], where γ is the uniform-stability constant of the learning algorithm. In regularized risk minimization (RRM), this results in an excess risk bound of order $O(n^{-\frac{1}{4}})$ at best, where n is the number of training examples.

As a main contribution of this paper, we show an improved bound for this setting of order $O(\gamma \log n)$, which translates into $O\big(\log(n)/\sqrt{n}\big)$ for excess risks of RRM. Remarkably, although the bound improves the previously best known rate achieved through stability analysis by a factor of $\sqrt{n}/\log(n)$, it applies more generally: we remove the standard assumption of a bounded loss function used in the prevalent stability analyses [1, 8, 19, 20, 55]. The loss of some of the most commonly used pairwise learning methods—including rankSVM [29] and MPRank [15]—is unbounded, for which we show, for the first time, a stability analysis. Based on our connection between generalization and stability, we also derive, to the best of our knowledge, the first probabilistic generalization bound for stochastic gradient descent (SGD) in pairwise learning. Our result quantifies how to trade-off optimization and generalization to achieve an almost optimal excess risk bound in this setting.

The above bound holds generally for any confidence level, which is informative to understand the variability of the algorithm and is necessary if the algorithm is used many times [20]. Furthermore, we show a sharper bound, but which holds in expectation and in a realizable case (where zero training error is achievable). Such bounds are called *optimistic* bounds in the literature [50]. For this setting, we show an excess risk bound of order $O(n^{-1})$. For the proof, we introduce a new on-average stability measure for pairwise learning and quantify its implication to generalization.

Finally, we consider applications of our general theory to ranking and metric learning, where we obtain generalization bounds with significantly improved dependence on n as compared to the existing stability analysis. Furthermore, our stability analysis also removes the dependency on the complexity of the hypothesis space and the input dimension in the uniform convergence analysis.

Structure. We review related work in Section 2, and give background information in Section 3. We list main results in Section 4 and give applications in Section 5. We conclude the paper in Section 6.

2 Related Work

In this section, we summarize the related work on the generalization analysis of pairwise learning, which we categorize according to the employed proof techniques.

In complexity (uniform convergence) analysis, we view generalization gaps between training and testing errors as U-statistics of order two. We can then bound the supremum of U-statistics over the hypothesis space—the U-process [9, 13, 34, 44, 54, 58, 61]. To this end, decoupling techniques have been introduced to represent the objective function as a summation of i.i.d. random variables plus a degenerate U-statistic [13]. This approach can yield meaningful generalization bounds of the order $O(1/\sqrt{n})$ for several pairwise learning problems, including ranking [13, 44] and metric learning [9, 54]. The authors of [13, 44] show fast-rate generalization bounds under stronger capacity assumptions on the hypothesis space and Bernstein-type of assumptions on the relationship between variances and expectations. Fast generalization bounds were established for metric learning [58], which, however requires a boundedness assumption on the output model, a bounded gradient assumption and the learning models to be linear. The complexity approach ignores the interaction between the learning algorithm and the training dataset in the search of the output model. It therefore implies generalization bounds depending on the complexity of the hypothesis space [3] and the input dimension. As indicated in [2, 13, 54], a square-root dependency on the dimension is generally inevitable for the uniform convergence in metric learning and ranking. This means that such bounds can quickly become uninformative in high dimensions [1]. For hypothesis spaces with an unbounded complexity, uniform convergence bounds cannot be applied at all [1]. The stability analysis is preferable in both cases.

An advantage of uniform convergence approach is that it is able to imply meaningful generalization bounds in a non-convex learning setting [16, 21, 39]. As a comparison, stability analysis requires very small step sizes to enjoy good stability for non-convex problems [25], which inevitably leads to very slow convergence rates of optimization errors.

The second popular approach studies pairwise learning using algorithmic stability, which is a fundamental concept in SLT dating back to 1970s [46]. The modern framework of stability analysis was established in the seminal paper [7], where an important concept called the *uniform stability* was introduced. This stability measure was then extended to study randomized algorithms [17], to investigate the concentration of output models [35], and to exploit the summation structure of the empirical risk [47]. These stability measures have found applications in privacy learning [4], stochastic optimization [5, 10, 25, 32, 33], and structured prediction [36]. The fundamental role of algorithmic stability in SLT was illustrated by establishing its close connection to learnability [42, 47]. Very recently, elegant high-probability bounds were established for uniformly stable algorithms [8, 19, 20, 37]. The above mentioned stability analysis was conducted in the setting of pointwise learning. There is also some interesting work on the stability analysis of pairwise learning. For example, the connection between generalization and algorithmic stability was established for ranking [1, 23]. Furthermore, it was shown there that kernel-based ranking algorithms in a regularization setting enjoy uniform stability. Algorithmic stability was further used to yield dimension-independent bounds for regularized metric learning [28, 55]. The stability and its trade-off with optimization errors were studied for a variant of SGD in pairwise learning [48], inspired by the recent work in the pointwise learning setting [11, 25].

We now briefly mention related work on the generalization analysis of pairwise learning using other proof techniques than complexity analysis or algorithmic stability. Algorithmic robustness was estimated for pairwise learning [12], which in turn implies generalization bounds [6]. Convex analysis was applied to study the regret bounds and generalization bounds of online pairwise learning [30, 56]. The tool of integral operators was used to exploit the structure of the specific least squares loss functions, where the learnt model can be written in a closed-form [59].

3 Background

3.1 Pairwise learning

Assume we are given a training dataset $S = \{z_1, \ldots, z_n\}$ drawn independently from a probability measure ρ defined over a sample space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \subset \mathbb{R}^d$ is an input space of dimension d and $\mathcal{Y} \subset \mathbb{R}$ is an output space. Based on S, we wish to build a model $h : \mathcal{X} \mapsto \mathcal{Y}$ or $h : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ that can be used to do prediction when we are given some testing examples. We consider a parametric model where the model $h_{\mathbf{w}}$ can be parameterized by an index $\mathbf{w} \in \mathcal{W}$, where $\mathcal{W} \subseteq \mathbb{R}^{d'}$ is a parameter space of dimension d' (d' is not necessarily equal to d, and can be infinite). Unlike pointwise learning, a distinctive property of pairwise learning is that the performance of a model should be measured over pairs of training examples. That is, the behavior of $h_{\mathbf{w}}$ over $z, \tilde{z} \in \mathcal{Z}$ is measured by $\ell(\mathbf{w}; z, \tilde{z})$, where $\ell : \mathcal{W} \times \mathcal{Z} \times \mathcal{Z} \mapsto \mathbb{R}_+$ is a loss function. Then, the empirical behavior of $h_{\mathbf{w}}$ can be quantified by the empirical risk $R_S(\mathbf{w})$ defined by

$$R_S(\mathbf{w}) = \frac{1}{n(n-1)} \sum_{i,j \in [n]: i \neq j} \ell(\mathbf{w}; z_i, z_j), \tag{3.1}$$

where we use the notation $[n] = \{1, \ldots, n\}$. We train a predictive model by applying an algorithm A to S, for which some popular choices include empirical risk minimization, regularized/structural risk minimization, (stochastic) gradient descent, etc. An algorithm A can be understood as a mapping from $\mathcal{Z}^n$ to $\mathcal{W}$, with $A(S)$ being the output of A when applied to S. Typically, the output model $A(S)$ would enjoy a small empirical risk since we are often fitting training examples. However, this does not necessarily mean that it also enjoys a small population risk $R(\mathbf{w}) = \mathbb{E}_{z,\tilde{z}}\big[\ell(\mathbf{w}; z, \tilde{z})\big]$, which quantifies the prediction behavior of $\mathbf{w}$ over testing examples. The generalization gap of a model $\mathbf{w}$ is defined as the difference between the population risk and empirical risk, i.e., $R(\mathbf{w}) - R_S(\mathbf{w})$.

We are particularly interested in RRM, where a regularizer $r : \mathcal{W} \mapsto \mathbb{R}_+$ is added into the data-fitting term R_S to increase the regularity of an algorithm. The resulting algorithm then outputs the model by

$$\mathbf{w}_S = \arg\min_{\mathbf{w}\in\mathcal{W}} \Big[F_S(\mathbf{w}) := \frac{1}{n(n-1)} \sum_{i,j\in[n]:i\neq j} f(\mathbf{w}; z_i, z_j)\Big], \tag{3.2}$$

where $f : \mathcal{W} \times \mathcal{Z} \times \mathcal{Z} \mapsto \mathbb{R}_+$ is defined as $f(\mathbf{w}; z, \tilde{z}) = \ell(\mathbf{w}; z, \tilde{z}) + r(\mathbf{w})$. Although the above objective function involves $O(n^2)$ terms in the summand, one can use SGD to achieve sample-size independent convergence rates [45]. Let $\mathbf{w}_1 \in \mathcal{W}$. At the t-th iteration, SGD first randomly selects (i_t, j_t) from the uniform distribution over the set $\{(i, j) : i, j \in [n], i \neq j\}$, and updates the model by

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t\big(\ell'(\mathbf{w}_t; z_{i_t}, z_{j_t}) + r'(\mathbf{w}_t)\big), \tag{3.3}$$

where $\{\eta_t\}_t$ is a step size sequence, and $\ell'(\mathbf{w}_t; z_{i_t}, z_{j_t})$ denotes a subgradient of $\ell(\cdot; z_{i_t}, z_{j_t})$ at $\mathbf{w}_t$.

3.2 Algorithmic stability

Algorithmic stability plays an important role in studying the behavior of a learning algorithm. Intuitively, we say an algorithm $A : \mathcal{Z}^n \mapsto \mathcal{W}$ is stable if the output model $A(S)$ is insensitive to perturbations of S. There are various notions of stability, including uniform stability, hypothesis stability, error stability and on-average stability [7, 17, 47]. A particularly interesting stability measure is uniform stability, which was introduced in [7] and extended in [1] to pairwise learning.

Definition 1 (Uniform Stability)**.** We say a deterministic algorithm $A : \mathcal{Z}^n \mapsto \mathcal{W}$ is γ-uniformly stable if for any training datasets $S, S' \in \mathcal{Z}^n$ that differ by at most a single example we have

$$\sup_{z,\tilde{z}\in\mathcal{Z}} \big|\ell(A(S); z, \tilde{z}) - \ell(A(S'); z, \tilde{z})\big| \leq \gamma.$$

We will use the above notion of uniform stability to develop high-probability generalization bounds. To construct optimistic bounds, we introduce a novel on-average stability for pairwise learning, which is motivated by the recent work on on-average stability for pointwise learning [24, 32, 33, 47]. The difference is that we consider perturbations of a training dataset by two examples.

Definition 2 (On-average stability)**.** Let $S = \{z_1, \ldots, z_n\}, S' = \{z'_1, \ldots, z'_n\}$ be independently drawn from ρ. For any $i < j$, we denote

$$S_{i,j} = \big\{z_1, \ldots, z_{i-1}, z'_i, z_{i+1}, \ldots, z_{j-1}, z'_j, z_{j+1}, \ldots, z_n\big\}. \tag{3.4}$$

We say a deterministic algorithm A is γ-on-average stable if

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

It is clear that on-average stability is weaker than uniform-stability since it involves the expectation over training examples and the average of indices. As a comparison, uniform stability involves a supremum over both training examples and testing examples $z, \tilde{z}$.

4 Main Results

In this section, we present our main results on generalization bounds based on stability. We always let $\|\cdot\|$ be a norm induced by an inner product $\langle\cdot,\cdot\rangle$ in a Hilbert space, i.e., $\|\cdot\|^2 = \langle\cdot,\cdot\rangle$. Then its dual norm is itself. We say a function $g : \mathcal{W} \mapsto \mathbb{R}$ is σ-strongly convex w.r.t. a norm $\|\cdot\|$ if

$$g(\mathbf{w}) - \big(g(\mathbf{w}') + \langle\mathbf{w} - \mathbf{w}', g'(\mathbf{w}')\rangle\big) \geq \sigma\|\mathbf{w} - \mathbf{w}'\|^2/2, \quad \forall \mathbf{w}, \mathbf{w}' \in \mathcal{W}.$$

4.1 Generalization by algorithmic stability

Our first result (Theorem 1) to be proved in Appendix A is a high-probability generalization bound for uniformly stable algorithms in pairwise learning, motivated by the recent analysis in pointwise learning [8, 19, 20, 37]. One of the key tools we use in the analysis is a concentration inequality from [8], which considers a summation of n functions of n independent random variables. However,

this concentration inequality does not fit the structure of pairwise learning. A difficulty is the inter-dependency among the $n(n-1)$ terms in the objective function. Our novelty to tackle this difficulty is to introduce a new decomposition to exploit the structure of the U-statistic in the pairwise objective function (3.1). Below, e denotes the base of the natural logarithm. For any $\alpha \geq 0$, $\lceil \alpha \rceil$ denotes the least integer no smaller than α. For any random variable Z, we denote by $\mathbb{E}_Z[\cdot]$ the conditional expectation with respect to (w.r.t.) Z.

Theorem 1. *Let $A : \mathcal{Z}^n \mapsto \mathcal{W}$ be γ-uniformly stable and $M > 0$. Suppose $\big|\mathbb{E}_S[\ell(A(S); z, \tilde{z})]\big| \leq M$ for all $z, \tilde{z} \in \mathcal{Z}$. Then for all $\delta \in (0, 1/e)$ the following inequality holds with probability $1 - \delta$*

$$|R_S(A(S)) - R(A(S))| \leq 4\gamma + e\Big(12\sqrt{2}M(n-1)^{-\frac{1}{2}}\sqrt{\log(e/\delta)} + 48\sqrt{6}\gamma\lceil\log_2(n-1)\rceil\log(e/\delta)\Big). \tag{4.1}$$

Remark 1. We now compare Theorem 1 with the existing stability analysis. Roughly speaking, Theorem 1 shows that the generalization gap for γ-uniformly stable algorithms decays as $O(\gamma \log n + n^{-\frac{1}{2}})$ with high probability (we ignore $\log(1/\delta)$ for brevity). Under the same conditions, it was shown for pairwise learning that [1, 15, 28, 55]

$$|R_S(A(S)) - R(A(S))| = O\big(\sqrt{n}\gamma + n^{-\frac{1}{2}}\big). \tag{4.2}$$

It is clear that our result significantly improves (4.2) by replacing their dominant term $\sqrt{n}\gamma$ with $\gamma \log n$. Specifically, if $\gamma = O(n^{-\alpha})$ with $\alpha \in (\frac{1}{2}, 1]$ (actually $\gamma = O(1/(n\sigma))$ if F_S is σ-strongly convex [7]), then (4.1) becomes $|R_S(A(S)) - R(A(S))| = O(n^{-\frac{1}{2}})$, while (4.2) becomes $|R_S(A(S)) - R(A(S))| = O(n^{\frac{1}{2}-\alpha})$. The existing complexity analysis for pointwise learning suggests $\sigma = O(n^{-\frac{1}{2}})$ to get an optimal bound [51, eq (14)]. In this case, $\gamma = O(n^{-\frac{1}{2}})$ and our stability analysis implies the nice bound $O(n^{-\frac{1}{2}}\log n)$, while (4.2) implies the vacuous bound $O(1)$.

4.2 Generalization bounds for regularized risk minimization

We now apply Theorem 1 to establish generalization bounds for pairwise learning with strongly convex objective functions. A preliminary step in the application of Theorem 1 is to control $\mathbb{E}_S[\ell(A(S); z, \tilde{z})]$. To this aim, we establish the following lemma to be proved in Appendix B. Let

$$\mathbf{w}^* = \arg\min_{\mathbf{w}\in\mathcal{W}} \big[R(\mathbf{w}) + r(\mathbf{w})\big], \quad \mathbf{w}_R^* = \arg\min_{\mathbf{w}\in\mathcal{W}} R(\mathbf{w}).$$

Lemma 2. *Suppose F_S is σ-strongly convex w.r.t. a norm $\|\cdot\|$. Define the algorithm A as $A(S) = \arg\min_{\mathbf{w}\in\mathcal{W}} F_S(\mathbf{w})$. If A is γ-uniformly stable, then $\mathbb{E}[\|A(S) - \mathbf{w}^*\|^2] \leq 8\gamma/\sigma$.*

In the existing analysis, one often uses the σ-strong convexity of F_S to show $\|A(S)\| = O(1/\sqrt{\sigma})$ [9], which implies a suboptimal bound since the convexity parameter σ is often very small in practice, i.e., $\sigma = O(n^{-\alpha})$ for $\alpha \in (0, 1)$ (σ is roughly the regularization parameter which should decay in this way [19, 51]). As a comparison, Lemma 2 implies that $\mathbb{E}_S[\|A(S) - \mathbf{w}^*\|] = O(\sqrt{\gamma/\sigma})$, which is significantly smaller than $O(1/\sqrt{\sigma})$ since the uniform stability parameter is often very small [7].

We need the following assumption to derive Theorem 3, whose proof is given in Appendix B.

Assumption 1. Let $b, \sigma_0 > 0$. We assume $0 \leq \ell(0; z, \tilde{z}) \leq b$ for all $z, \tilde{z} \in \mathcal{Z}$. We also assume $\mathrm{Var}[\ell(\mathbf{w}^*; Z, \tilde{Z})] < \sigma_0^2$, where $\mathrm{Var}[\ell(\mathbf{w}^*; Z, \tilde{Z})]$ denotes the variance of $\ell(\mathbf{w}^*; Z, \tilde{Z})$.

We use the notation $B \asymp \widetilde{B}$ if there exist constants $c_1, c_2 > 0$ such that $c_1\widetilde{B} < B \leq c_2\widetilde{B}$.

Theorem 3. *Let Assumption 1 hold and $L \in \mathbb{R}_+$. Define A as $A(S) = \arg\min_{\mathbf{w}\in\mathcal{W}} F_S(\mathbf{w})$. Suppose F_S is σ-strongly convex w.r.t. $\|\cdot\|$ for all S. Assume*

$$\big|\ell(\mathbf{w}; z, \tilde{z}) - \ell(\mathbf{w}'; z, \tilde{z})\big| \leq L\|\mathbf{w} - \mathbf{w}'\|, \quad \forall z, \tilde{z} \in \mathcal{Z}, \mathbf{w}, \mathbf{w}' \in \mathcal{W}. \tag{4.3}$$

Then for $\delta \in (0, 1/e)$, with probability $1 - \delta$ the generalization gap $R_S(A(S)) - R(A(S))$ *satisfies*

$$|R_S(A(S)) - R(A(S))| = O\Big((n\sigma)^{-1}\log n\log(1/\delta) + \sqrt{n^{-1}\log(1/\delta)}\Big). \tag{4.4}$$

Furthermore, if $r(\mathbf{w}) = O(\sigma\|\mathbf{w}\|^2)$, $\sigma \asymp n^{-1/2}$, $\sup_{z,z'} \ell(\mathbf{w}_R^*; z, z') = O(\sqrt{n})$ and Assumption 1 holds with $\mathbf{w}^*$ replaced by $\mathbf{w}_R^*$, then with probability at least $1 - \delta$ we have the following bound on excess risk $R(A(S)) - R(\mathbf{w}_R^*)$

$$R(A(S)) - R(\mathbf{w}_R^*) = O(n^{-\frac{1}{2}}\log n\log(1/\delta)). \tag{4.5}$$

Remark 2. We present some comparisons with the existing work. Under similar assumptions and additional boundedness assumptions, the existing stability analysis implies the generalization bound $|R_S(A(S)) - R(A(S))| = O(\sigma^{-1}n^{-\frac{1}{2}})$ for pairwise learning with σ-strongly convex objective functions [1, 15, 28, 55], which can be $\sqrt{n}$-times slower than the bound (4.4). To see this, assume $\sigma \asymp n^{-\alpha}$ with $\alpha \in [0, \frac{1}{2}]$. If $\alpha \in [0, \frac{1}{2})$, then (4.4) implies the bound $O(n^{-\frac{1}{2}})$, while the bounds in [1, 28, 55] become $|R_S(A(S)) - R(A(S))| = O(n^{\alpha-\frac{1}{2}})$. For the special case $\alpha = 1/2$ suggested in the existing analysis of pointwise learning [51], Eq. (4.4) implies the bound $O(n^{-\frac{1}{2}}\log n)$, while the existing bound becomes $O(1)$ [1, 28, 55]. As we will clarify in Remark B.1, the existing stability analysis yields at best the excess risk bound $R(A(S)) - R(\mathbf{w}^*) = O(n^{-\frac{1}{4}})$ no matter how σ changes. As a comparison, our stability analysis yields the bound $R(A(S)) - R(\mathbf{w}^*) = O(n^{-\frac{1}{2}}\log n)$.

Remark 3 (Boundedness assumption)**.** To get the bound $O(n^{-\frac{1}{2}}\log n)$, the existing stability analysis requires a boundedness assumption on loss functions as $0 \leq \ell(A(S); z, \tilde{z}) \leq B$ for a constant $B > 0$ and all $S \in \mathcal{Z}^n, z, z' \in \mathcal{Z}$ (B is treated as a constant absorbed in a big O notation) [1, 8, 19, 20, 55] or a boundness assumption of $\mathcal{W}$ [28]. However, one can only show $\|A(S)\| = O(1/\sqrt{\sigma})$ [1] and therefore the constant B needs to grow as $O(1/\sqrt{\sigma})$ for popular loss functions (e.g., hinge loss and logistic loss), from which the stability analysis in [8] can only imply suboptimal bounds $O((n\sigma)^{-1/2})$ even in the case of *pointwise learning* if one does not impose a boundedness assumption (note σ is often very small). We develop the generalization bound $O(n^{-\frac{1}{2}}\log n)$ by relaxing the boundedness assumption to a variance assumption on $\ell(\mathbf{w}^*; Z, \tilde{Z})$. Note that the expectation of $\ell(\mathbf{w}^*; Z, \tilde{Z})$ is $R(\mathbf{w}^*)$, which is small according to the definition of $\mathbf{w}^*$. Therefore, it is reasonable to assume that the variance of $\ell(\mathbf{w}^*; Z, \tilde{Z})$ is bounded. To achieve this relaxation, we use a novel application of Theorem 1 to $\tilde{\ell}(\mathbf{w}; z, \tilde{z}) = \ell(\mathbf{w}; z, \tilde{z}) - \ell(\mathbf{w}^*; z, \tilde{z})$ instead of $\ell(\mathbf{w}; z, \tilde{z})$ (cf. Line 107 in the Appendix). Moreover, we introduce a novel lemma (Lemma 2) to show $\big|\mathbb{E}_S\big[\tilde{\ell}(A(S); z, \tilde{z})\big]\big| = O(1/(\sqrt{n}\sigma))$ (cf. Line 105 in the Appendix).

4.3 Generalization bounds for stochastic gradient descent

As a further application of Theorem 1, we establish generalization bounds for SGD (3.3) in pairwise learning, which can be considered as a deterministic algorithm if we fix $\{(i_t, j_t)_t\}$ in (3.3). SGD is a highly popular algorithm with wide applications in the big-data era. Note we do not require a strong convexity. We say $g : \mathcal{W} \mapsto \mathbb{R}$ is α-smooth w.r.t. a norm $\|\cdot\|$ if g is differentiable and

$$\|g'(\mathbf{w}) - g'(\mathbf{w}')\| \leq \alpha\|\mathbf{w} - \mathbf{w}'\|, \quad \forall \mathbf{w}, \mathbf{w}' \in \mathcal{W}.$$

Popular smooth loss functions include the logistic loss, the Huber loss, the squared hinge loss and the least squares loss [52]. Note that the logistic loss and the Huber loss is also Lipschitz continuous. The proof is given in Appendix C.

Theorem 4. *Let* (4.3) *hold. Assume for all z, z', $\mathbf{w} \mapsto \ell(\mathbf{w}; z, z')$ is convex and α-smooth w.r.t. the Euclidean norm, and for any $\{(i_t, j_t)_t\}$ we have $\big|\mathbb{E}_S[\ell(\mathbf{w}_T; z, z')]\big| \leq M$, where $\mathbf{w}_T$ is produced by SGD with $\eta_t = c/\sqrt{T}, c \leq 2/\alpha$ and $r(\mathbf{w}) = 0$. For any $\delta \in (0, 1)$ with probability $1 - \delta$ we have*

$$\big|R_S(\mathbf{w}_T) - R(\mathbf{w}_T)\big| = O\Big(\log n \log(1/\delta)\sqrt{T}/n + n^{-\frac{1}{2}}\log n \log^{\frac{3}{2}}(1/\delta)\Big). \tag{4.6}$$

Remark 4. We now show the implication of Theorem 4 on understanding the generalization behavior and implicit regularization of SGD. We can show (the details are given in Remark C.1)

$$R(\mathbf{w}_T) - R(\mathbf{w}_R^*) = \big(R(\mathbf{w}_T) - R_S(\mathbf{w}_T)\big) + \big(R_S(\mathbf{w}_T) - R_S(\mathbf{w}_R^*)\big) + O(n^{-\frac{1}{2}}). \tag{4.7}$$

The first term is the estimation error and the second term is the optimization error. Therefore, Theorem 4 actually gives an estimation error bound. If we further assume $\|\mathbf{w}_t\| \leq B$ for some $B > 0$ and all t, the optimization error was shown to satisfy[2] $R_S(\mathbf{w}_T) - R_S(\mathbf{w}_R^*) = O(T^{-\frac{1}{2}}\log T)$ [26] with high probability. Plugging these estimation and optimization error bounds back into (4.7), we derive with high probability $R(\mathbf{w}_T) - R(\mathbf{w}_R^*) = O\big(\log n\sqrt{T}/n + n^{-\frac{1}{2}}\log n\big) + O(T^{-\frac{1}{2}}\log T)$. It is clear that estimation errors increase as we run more iterations, while optimization errors decrease. One can take an optimal $T \asymp n$ to trade-off the optimization and estimation errors, and get

[2]Although they considered step size $\eta_t = c/\sqrt{t}$ [26], their result also holds for $\eta_t = c/\sqrt{T}$.

$R(\mathbf{w}_T) - R(\mathbf{w}_R^*) = O(n^{-\frac{1}{2}} \log n)$. To the best of our knowledge, this gives the first high-probability generalization bound for SGD in pairwise learning. Although we do not use an explicit regularizer in Theorem 4, our analysis shows that an implicit regularization can be achieved by tuning the number of iterations [25, 57]. We can compare our results with those based on the existing connection (4.2) between generalization and stability. Indeed, if we combine the best known optimization error bounds [26], the uniform stability of SGD established in Lemma C.3 and (4.2) together, we can only derive vacuous excess risk bound $R(\mathbf{w}_T) - R(\mathbf{w}_R^*) = O(1)$ (details are given in Remark C.2), which are significantly improved to $O(n^{-\frac{1}{2}} \log n)$ based on Theorem 1. In Appendix F we show $O(n^{-\frac{1}{2}})$ is minimax optimal for pairwise learning in a general convex setting.

The authors of [48] studied a variant of SGD where the models are updated by

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \frac{\eta_t}{t-1} \sum_{k=1}^{t-1} \ell'(\mathbf{w}_t; z_{i_t}, z_{i_k}), \quad \forall t > 1.$$

Their stability bounds were stated in expectation [48] while we give high-probability analysis.

It should be mentioned that our generalization analysis can be applied to other iterative algorithms for pairwise learning, including gradient descent, Nesterov's accelerated gradient descent, the heavy ball method and stochastic gradient Langevin dynamics [11]. To this aim, it suffices to estimate the uniform stability of these algorithms in pairwise learning.

4.4 Optimistic generalization bounds

Our key idea to derive optimistic bounds is to use the on-average stability in Definition 2, whose connection to generalization is established in the following theorem to be proved in Section D.

Theorem 5. *If A is γ-on-average stable, then $\mathbb{E}\big[R(A(S)) - R_S(A(S))\big] \leq \gamma$.*

We now present an optimistic generalization bound for pairwise learning by exploiting the smoothness of loss functions. By optimistic we mean that the decay rate of generalization bounds depends on the behavior of the best model. That is, we can get faster bounds if $R(\mathbf{w}_R^*) = o(1)$. Optimistic bounds were studied for pointwise learning in the literature [43, 50, 60], which are becoming interesting in the big-data era where models are often powerful enough to achieve a very small training error. For any $\mathbf{w} \in \mathcal{W}$, let $F(\mathbf{w}) = R(\mathbf{w}) + r(\mathbf{w})$. Theorem 6 is proved in Appendix D.

Theorem 6. *Assume for all z, z', the map $\mathbf{w} \mapsto \ell(\mathbf{w}; z, z')$ is α-smooth w.r.t. $\|\cdot\|$. Let $A(S) = \arg\min_{\mathbf{w}\in\mathcal{W}} F_S(\mathbf{w})$ and $\sigma n \geq 8\alpha$. If for all $S \in \mathcal{Z}^n$, F_S is σ-strongly convex w.r.t. $\|\cdot\|$, then*

$$\mathbb{E}\big[F(A(S))\big] - F(\mathbf{w}^*) \leq \mathbb{E}\big[R(A(S)) - R_S(A(S))\big] \leq \Big(\frac{1024\alpha^2}{n^2\sigma^2} + \frac{64\alpha}{n\sigma}\Big)\mathbb{E}\big[R_S(A(S))\big]. \quad (4.8)$$

Furthermore, if $r(\mathbf{w}) = O(\sigma\|\mathbf{w}\|^2)$ we can take some appropriate σ to get

$$\mathbb{E}[R(A(S))] - R(\mathbf{w}_R^*) = O\big(\sqrt{R(\mathbf{w}_R^*)}\|\mathbf{w}_R^*\|n^{-\frac{1}{2}} + \|\mathbf{w}_R^*\|^2 n^{-1}\big). \quad (4.9)$$

Note if $R(\mathbf{w}_R^*) = O(\|\mathbf{w}_R^*\|^2/n)$, the above excess risk bound becomes $\mathbb{E}[R(A(S))] - R(\mathbf{w}_R^*) = O(\|\mathbf{w}_R^*\|^2/n)$. That is, we get a fast excess risk bound if there exists a model with a small population risk.

5 Applications

5.1 Ranking

For ranking we assume real-valued labels indicating a ranking preference on instances, i.e., $y_i < y_j$ means x_i has a lower rank than x_j. We aim to build a function $h_{\mathbf{w}} : \mathcal{X} \mapsto \mathbb{R}$ that ranks instances with larger labels higher than those with smaller labels [1, 13, 44]. The performance of a model $h_{\mathbf{w}}$ at a pair z, z' can be measured by the 0-1 loss $\ell_{0\text{-}1}(\mathbf{w}; z, z') = \mathbb{I}[\text{sgn}(y - y')(h_{\mathbf{w}}(x) - h_{\mathbf{w}}(x')) < 0]$, where $\mathbb{I}[\cdot]$ is the indicator function taking 1 if the argument holds and 0 otherwise, and $\text{sgn}(a)$ denotes the sign of the number a. Since the 0-1 loss leads to an NP-hard problem, we consider loss functions of the form $\ell^\psi(\mathbf{w}; z, z') = \psi(\text{sgn}(y - y')(h_{\mathbf{w}}(x) - h_{\mathbf{w}}(x')))$. Here $\psi : \mathbb{R} \mapsto \mathbb{R}_+$ is convex and

decreasing, for which popular choices include the hinge loss $\psi(t) = \max\{1-t, 0\}$ and the logistic loss $\psi(t) = \log(1+\exp(-t))$. Below we provide bounds for RRM, SGD and optimistic bounds.

Regularized risk minimization. The following proposition follows directly from Theorem 3 by noticing that $r(\mathbf{w}) = \lambda\|\cdot\|^2$ is 2λ-strongly convex w.r.t. $\|\cdot\|$. We omit the proof for simplicity.

Proposition 7. *Let Assumption 1 hold for both* $\mathbf{w}^*$ *and* $\mathbf{w}^*$ *replaced by* $\mathbf{w}_R^*$. *Consider ranking problems with* $f(\mathbf{w}; z, z') = \ell^\psi(\mathbf{w}; z, z') + \lambda\|\mathbf{w}\|^2$. *Assume* ℓ^ψ *is convex w.r.t.* $\mathbf{w}$ *and satisfy* (4.3). *Then for* $A(S) = \arg\min_{\mathbf{w}\in\mathcal{W}} F_S(\mathbf{w})$, $\lambda \asymp n^{-\frac{1}{2}}$ *and any* $\delta \in (0, 1/e)$, *with probability at least* $1-\delta$ *there holds* $R(A(S)) - R(\mathbf{w}_R^*) = O\big(n^{-\frac{1}{2}}\log n \log(1/\delta)\big)$.

Remark 5. High-probability bounds for ranking were developed in the literature under some capacity assumptions on the hypothesis space $\{h_{\mathbf{w}} : \mathbf{w} \in \mathcal{W}\}$ measured by either covering numbers [44, 58] or VC dimension [13]. The arguments there are based on the uniform convergence of empirical risks to population risk and ignore the specific property of the learning algorithm, which inevitably depends on the complexity of the hypothesis space. Furthermore, a dependency on the dimension is necessary if no structural assumptions are imposed [18]. For example, generalization bounds $O(\sqrt{d/n})$ were derived for bipartite ranking (AUC maximization) via the uniform convergence approach [2]. As a comparison, we derive dimension-independent bounds of order $O(n^{-\frac{1}{2}}\log n)$. Furthermore, the existing stability analysis implies $|R_S(A(S)) - R(A(S))| = O\big(\lambda^{-1}\sqrt{1/n}\big)$ [1, 15] and the excess risk bounds $R(A(S)) - R(\mathbf{w}_R^*) = O(n^{-\frac{1}{4}})$, which are worse than the results in Proposition 7.

Stochastic gradient descent. The following proposition is a direct application of Theorem 4.

Proposition 8. *Consider ranking problems with* $f(\mathbf{w}; z, z') = \ell^\psi(\mathbf{w}; z, z')$, *i.e.* $r(\mathbf{w}) = 0$. *Let* (4.3) *hold and assume for all* z, z', *the map* $\mathbf{w} \mapsto \ell^\psi(\mathbf{w}; z, z')$ *is convex and* α*-smooth. Let* $\mathbf{w}_T$ *be produced by SGD with* $\eta_t = c/\sqrt{T}, c \le 2/\alpha$ *and assume* $\big|\mathbb{E}_S[\ell(\mathbf{w}_T; z, z')]\big| \le M$. *Then for any* $\delta \in (0,1)$ *and* $T \asymp n$, *with probability* $1-\delta$ *we have* $\big|R_S(\mathbf{w}_T) - R(\mathbf{w}_T)\big| = O\big(n^{-\frac{1}{2}}\log n \log^{\frac{3}{2}}(1/\delta)\big)$.

Optimistic bounds. Proposition 9 on optimistic bounds is a direct application of Theorem 6.

Proposition 9. *Consider ranking problems with* $f(\mathbf{w}; z, z') = \ell^\psi(\mathbf{w}; z, z') + \lambda\|\mathbf{w}\|^2$. *If* ℓ^ψ *is convex,* α*-smooth and* $A(S) = \arg\min_{\mathbf{w}\in\mathcal{W}} F_S(\mathbf{w})$, *we can choose some* $\lambda \ge 8\alpha/n$ *such that* (4.9) *holds.*

Our results directly apply to bipartite ranking (AUC maximization) [14] with $\mathcal{Y} = \{+1, -1\}$. To see this, bipartite ranking is a specific instance of (3.1) with loss functions of the form $\ell^\psi(\mathbf{w}; z, z') = \psi((h_{\mathbf{w}}(x) - h_{\mathbf{w}}(x')))\mathbb{I}[y = 1, y' = -1]$. We omit this discussion for brevity.

5.2 Metric learning

We consider metric learning for learning a metric to measure the distance between instance pairs. We consider supervised metric learning with $\mathcal{Y} = \{-1, +1\}$, where we want an instance pair to be similar if they have the same class label, and apart from each other if they have different class labels [9, 28, 58]. We consider the Mahalanobis metric $h_{\mathbf{w}}(x, x') = \langle \mathbf{w}, (x-x')(x-x')^\top\rangle$, where $x^\top$ denotes the transpose of x and $\mathbf{w} \in \mathbb{R}^{d\times d}$. The performance of $h_{\mathbf{w}}$ on z, z' can be measured by the 0-1 loss $\ell_{0\text{-}1}(\mathbf{w}; z, z') = \mathbb{I}[\tau(y, y')(1 - h_{\mathbf{w}}(x, x')) \le 0]$, where $\tau(y, y') = 1$ if $y = y'$ and -1 otherwise. We often use a convex surrogate $\psi : \mathbb{R} \mapsto \mathbb{R}_+$, which leads to $\ell^\psi(\mathbf{w}; z, z') = \psi(\tau(y, y')(1 - h_{\mathbf{w}}(x, x')))$. We assume $\sup_{x\in\mathcal{X}} \|x\|_2 \le B$ for some $B > 0$, where $\|\cdot\|_2$ is the Euclidean norm.

Regularized risk minimization. Corollary 10 on RRM is proved in Appendix E.

Corollary 10. *Let Assumption 1 hold for both* $\mathbf{w}^*$ *and* $\mathbf{w}^*$ *replaced by* $\mathbf{w}_R^*$. *Consider metric learning with* $\mathcal{Y} = \{-1, +1\}$. *Consider* $f(\mathbf{w}; z, z') = \psi(\tau(y, y')(1 - h_{\mathbf{w}}(x, x'))) + \lambda\|\mathbf{w}\|^2$, *where* $\|\cdot\|$ *is the Frobenius norm and* $\psi(t) = \max\{0, 1-t\}$. *Then for* $A(S) = \arg\min_{\mathbf{w}\in\mathcal{W}} F_S(\mathbf{w})$, $\lambda \asymp n^{-\frac{1}{2}}$ *and any* $\delta \in (0, 1/e)$, *with probability* $1-\delta$ *it holds* $R(A(S)) - R(\mathbf{w}_R^*) = O\big(n^{-\frac{1}{2}}\log n \log(1/\delta)\big)$.

Remark 6. We make some comparisons. It was shown $R(A(S)) - R_S(A(S)) = O\big(n^{-\frac{1}{2}}\lambda^{-1}\big)$ [9, 28, 55], which leads to the excess risk bound $O(n^{-\frac{1}{4}})$. This is significantly improved to $O(n^{-\frac{1}{2}}\log n)$ in Corollary 10. A uniform convergence rate $O(\sqrt{dn^{-1}})$ was shown for metric learning [54], which is not appealing for high-dimensional problems. It was further indicated that a strong dependence on d is generally necessary for the uniform convergence if one does not impose a structural assumption [54]. As a comparison, our bound in Corollary 10 is dimension-independent.

Stochastic gradient descent. Corollary 11 on bounds for SGD is proved in Appendix E.

Corollary 11. *Consider metric learning with* $\mathcal{Y} = \{-1, +1\}$ *and* $f(\mathbf{w}; z, z') = \psi(\tau(y, y')(1 - h_{\mathbf{w}}(x, x')))$*, where* $\psi(t) = \log(1 + \exp(-t))$*. Let* $\mathbf{w}_T$ *be produced by SGD with* $\eta_t = c/\sqrt{T}, c \leq 1/8B^4$ *and assume* $\left|\mathbb{E}_S[\psi(\tau(y, y')(1 - h_{\mathbf{w}_T}(x, x')))]\right| \leq M$*. Then for any* $\delta \in (0, 1)$ *and* $T \asymp n$*, with probability* $1 - \delta$ *we have* $\left|R_S(\mathbf{w}_T) - R(\mathbf{w}_T)\right| = O\big(n^{-\frac{1}{2}} \log n \log^{\frac{3}{2}}(1/\delta)\big)$.

Optimistic bounds. We also get optimistic bounds for metric learning. We omit the proof for brevity.

Corollary 12. *Let Assumptions of Corollary 10 hold except that we consider the logistic loss* $\psi(t) = \log(1 + \exp(-t))$*. If* $A(S) = \arg\min_{\mathbf{w} \in \mathcal{W}} F_S(\mathbf{w})$ *and* $\|\mathbf{w}_R^*\| = O(1)$*, then we can choose some appropriate* λ *such that Eq.* (4.9) *holds.*

6 Conclusion

We analyze the generalization ability of pairwise learning using the methodology of algorithmic stability. We significantly improve the existing high-probability bounds $O(\sqrt{n}\gamma)$ to $O(\gamma \log n)$ for γ-uniformly stable algorithms. This allows us to improve the previously best excess risk bounds $O(n^{-1/4})$ for RRM and $O(1)$ for SGD to $O(n^{-1/2} \log n)$, which is optimal apart from the log factor. As compared to the uniform convergence analysis, our stability analysis implies the first high-probability risk bound for SGD in pairwise learning, and yields bounds independent of the complexity of models and the input dimension. Furthermore, we introduce an on-average stability to develop optimistic bounds as fast as $O(1/n)$ for learning in a low noise setting. Specific applications are further given to show the advantage of our generalization bounds over the existing analysis.

Below we mention some interesting directions for future research. First, it would be interesting to extend the analysis here to other learning settings, such as distributed learning and online learning for pairwise learning. Second, one could tackle the challenging problem of stability and generalization bounds for non-convex pairwise learning problems, which are popular in modern machine learning.

7 Broader Impact

This work does not present any foreseeable societal consequence.

Acknowledgments

YL acknowledges support by the National Natural Science Foundation of China (Grant Nos. 61806091, 11771012), and by the Alexander von Humboldt Foundation for a Humboldt Research Fellowship. MK acknowledges support by the German Research Foundation (DFG) award KL 2698/2-1 and by the Federal Ministry of Science and Education (BMBF) awards 01IS18051A and 031B0770E.

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] S. Agarwal, T. Graepel, R. Herbrich, S. Har-Peled, and D. Roth. Generalization bounds for the area under the roc curve. *Journal of Machine Learning Research*, 6(Apr):393–425, 2005.

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

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

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

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

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

[8] O. Bousquet, Y. Klochkov, and N. Zhivotovskiy. Sharper bounds for uniformly stable algorithms. *arXiv preprint arXiv:1910.07833*, 2019.

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

[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] C. Cortes, M. Mohri, and A. Rastogi. Magnitude-preserving ranking algorithms. In *Proceedings of the 24th international conference on Machine learning*, pages 169–176, 2007.

[16] D. Davis and D. Drusvyatskiy. Uniform graphical convergence of subgradients in nonconvex optimization and learning. *arXiv preprint arXiv:1810.07590*, 2018.

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

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

[19] V. Feldman and J. Vondrak. Generalization bounds for uniformly stable algorithms. In *Advances in Neural Information Processing Systems*, pages 9747–9757, 2018.

[20] 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.

[21] 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.

[22] J. Fürnkranz and E. Hüllermeier. Preference learning and ranking by pairwise comparison. In *Preference learning*, pages 65–82. Springer, 2010.

[23] W. Gao and Z.-H. Zhou. Uniform convergence, stability and learnability for ranking problems. In *International Joint Conference on Artificial Intelligence*, pages 1337–1343. AAAI Press, 2013.

[24] A. Gonen and S. Shalev-Shwartz. Average stability is invariant to data preconditioning: Implications to exp-concave empirical risk minimization. *The Journal of Machine Learning Research*, 18(1):8245–8257, 2017.

[25] 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.

[26] N. J. Harvey, C. Liaw, Y. Plan, and S. Randhawa. Tight analyses for non-smooth stochastic gradient descent. In *Conference on Learning Theory*, pages 1579–1613, 2019.

[27] 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.

[28] 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.

[29] T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142, 2002.

[30] 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.

[31] A. Kumar, A. Niculescu-Mizil, K. Kavukcoglu, and H. Daumé. A binary classification framework for two-stage multiple kernel learning. In *International Conference on Machine Learning*, pages 1331–1338, 2012.

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

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

[34] Y. Lei, S.-B. Lin, and K. Tang. Generalization bounds for regularized pairwise learning. In *International Joint Conference on Artificial Intelligence*, pages 2376–2382, 2018.

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

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

[37] A. Maurer. A second-order look at stability and generalization. In *Conference on Learning Theory*, pages 1461–1475, 2017.

[38] D. A. McAllester. Some pac-bayesian theorems. *Machine Learning*, 37(3):355–363, 1999.

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

[40] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning*. MIT press, 2012.

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

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

[43] H. W. Reeve and A. Kaban. Optimistic bounds for multi-output prediction. *arXiv preprint arXiv:2002.09769*, 2020.

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

[45] H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, pages 400–407, 1951.

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

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

[48] 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.

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

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

[51] K. Sridharan, S. Shalev-Shwartz, and N. Srebro. Fast rates for regularized objectives. In *Advances in Neural Information Processing Systems*, pages 1545–1552, 2009.

[52] I. Steinwart and A. Christmann. *Support Vector Machines*. Springer Science & Business Media, 2008.

[53] Z. Szabó, B. K. Sriperumbudur, B. Póczos, and A. Gretton. Learning theory for distribution regression. *The Journal of Machine Learning Research*, 17(1):5272–5311, 2016.

[54] N. Verma and K. Branson. Sample complexity of learning mahalanobis distance metrics. In *Advances in Neural Information Processing Systems*, pages 2584–2592, 2015.

[55] 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.

[56] 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.

[57] Y. Yao, L. Rosasco, and A. Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

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

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

[60] 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.

[61] Y. Zhou, H. Chen, R. Lan, and Z. Pan. Generalization performance of regularized ranking with multiscale kernels. *IEEE Transactions on Neural Networks and Learning Systems*, 27(5):993–1002, 2015.

Sharper Generalization Bounds for Pairwise Learning: Supplementary Material

A Proof of Theorem 1

To prove Theorem 1, we need to introduce some lemmas. The following lemma is attributed to [7], which provides far-reaching moment bounds for a summation of weakly dependent and mean-zero random functions with bounded increments under a change of any single coordinate. We denote $S\backslash\{z_i\}$ the set $\{z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n\}$. The L_p-norm of a random variable Z is denoted by $\|Z\|_p := \big(\mathbb{E}[|Z|^p]\big)^{\frac{1}{p}}, p \geq 1$.

Lemma A.1 ([4]). *Let $S = \{z_1, \ldots, z_n\}$ be a set of independent random variables each taking values in $\mathcal{Z}$ and $M > 0$. Let $g_1, \ldots, g_n$ be some functions $g_i : \mathcal{Z}^n \mapsto \mathbb{R}$ such that the following holds for any $i \in [n]$*

- $\big|\mathbb{E}_{S\backslash\{z_i\}}[g_i(S)]\big| \leq M$ *almost surely (a.s.),*
- $\mathbb{E}_{z_i}\big[g_i(S)\big] = 0$ *a.s.,*
- *for any $j \in [n]$ with $j \neq i$, and $z_j'' \in \mathcal{Z}$*

$$\big|g_i(S) - g_i(z_1, \ldots, z_{j-1}, z_j'', z_{j+1}, \ldots, z_n)\big| \leq \beta. \tag{A.1}$$

Then, for any $p \geq 2$

$$\Big\| \sum_{i=1}^{n} g_i(S) \Big\|_p \leq 12\sqrt{6}pn\beta\lceil \log_2 n \rceil + 3\sqrt{2}M\sqrt{pn}.$$

The bounds on moments of random variables can be used to establish concentration inequalities, as shown in the following lemma [4, 16].

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

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

The following lemma controls the change on the output of stable algorithms if we perturb a training dataset by two examples.

Lemma A.3. *Let $A : \mathcal{Z}^n \mapsto \mathcal{W}$ be γ-uniformly stable. Then for any $S' = \{z_1', \ldots, z_n'\}$ and $i \neq j$, we have*

$$\sup_{z,\tilde{z}\in\mathcal{Z}} \big|\ell(A(S); z, \tilde{z}) - \ell(A(S_{i,j}); z, \tilde{z})\big| \leq 2\gamma,$$

where $S_{i,j}$ is defined in (3.4).

Proof. For any $i \in [n]$, introduce

$$S_i = \{z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_n\}. \tag{A.2}$$

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

Note that S, S_i differ only by a single example, and $S_i, S_{i,j}$ differ only by a single example. It then follows from the definition of uniform stability that

$$\begin{aligned}
&\sup_{z,\tilde{z}\in\mathcal{Z}} \big|\ell(A(S);z,\tilde{z}) - \ell(A(S_{i,j});z,\tilde{z})\big| \\
&\leq \sup_{z,\tilde{z}\in\mathcal{Z}} \big|\ell(A(S);z,\tilde{z}) - \ell(A(S_i);z,\tilde{z})\big| + \sup_{z,\tilde{z}\in\mathcal{Z}} \big|\ell(A(S_i);z,\tilde{z}) - \ell(A(S_{i,j});z,\tilde{z})\big| \\
&\leq 2\gamma.
\end{aligned}$$

The proof is complete. □

With these lemmas, we can give the proof of Theorem 1 on high-probability bounds of the generalization gap. The concentration inequality established in Lemma A.1 applies to a summation of n random functions involving n independent random variables, which does not apply to the objective function in pairwise learning since it is a U-statistic. We introduce a novel decomposition to exploit the structure of pairwise learning problems. We abbreviate $\sum_{i,j\in[n]:i\neq j}$ as $\sum_{i\neq j}$.

Proof of Theorem 1. Let $p \geq 2$ be any number. We can decompose the generalization gap associated to $A(S)$ as follows

$$\begin{aligned}
&n(n-1)\mathbb{E}_{Z,\tilde{Z}}\big[\ell(A(S);Z,\tilde{Z})\big] - \sum_{i\neq j}\ell(A(S);z_i,z_j) = \sum_{i\neq j}\mathbb{E}_{Z,\tilde{Z}}\Big[\ell(A(S);Z,\tilde{Z}) - \mathbb{E}_{z_i',z_j'}[\ell(A(S_{i,j});Z,\tilde{Z})]\Big] \\
&+ \sum_{i\neq j}\mathbb{E}_{z_i',z_j'}\Big[\mathbb{E}_{Z,\tilde{Z}}\big[\ell(A(S_{i,j});Z,\tilde{Z})\big] - \ell(A(S_{i,j});z_i,z_j)\Big] + \sum_{i\neq j}\mathbb{E}_{z_i',z_j'}\Big[\ell(A(S_{i,j});z_i,z_j) - \ell(A(S);z_i,z_j)\Big],
\end{aligned}$$

where $S_{i,j}$ is defined in (3.4). According to Lemma A.3, we know

$$\Big|\ell(A(S);Z,\tilde{Z}) - \mathbb{E}_{z_i',z_j'}[\ell(A(S_{i,j});Z,\tilde{Z})]\Big| \leq 2\gamma$$

and

$$\big|\ell(A(S_{i,j});z_i,z_j) - \ell(A(S);z_i,z_j)\big| \leq 2\gamma.$$

Therefore,

$$\Big|n(n-1)\mathbb{E}_{Z,\tilde{Z}}\big[\ell(A(S);Z,\tilde{Z})\big] - \sum_{i\neq j}\ell(A(S);z_i,z_j)\Big| \leq 4n(n-1)\gamma + \big|\sum_{i\neq j} g_{i,j}(S)\big|, \quad \text{(A.3)}$$

where we introduce

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

For any $i,j\in[n]$, we can further decompose $g_{i,j}$ as $g_{i,j} = g_j^{(i)} + \tilde{g}_i^{(j)}$, where (we omit the argument S for brevity)

$$\begin{aligned}
g_j^{(i)} &= \mathbb{E}_{z_i',z_j'}\Big[\mathbb{E}_{Z,\tilde{Z}}\big[\ell(A(S_{i,j});Z,\tilde{Z})\big] - \mathbb{E}_Z[\ell(A(S_{i,j});Z,z_j)]\Big] \\
\tilde{g}_i^{(j)} &= \mathbb{E}_{z_i',z_j'}\Big[\mathbb{E}_Z\big[\ell(A(S_{i,j});Z,z_j\big] - \ell(A(S_{i,j});z_i,z_j)\Big].
\end{aligned}$$

Let us temporarily fix i, and consider $n-1$ random functions $g_1^{(i)},\ldots,g_{i-1}^{(i)},g_{i+1}^{(i)},\ldots,g_n^{(i)}$. According to the assumption $\big|\mathbb{E}_S[\ell(A(S);z,\tilde{z})]\big| \leq M$ for all $z,\tilde{z}$, we know

$$\big|\mathbb{E}_{S\setminus\{z_j\}}[g_j^{(i)}(S)]\big| \leq 2M, \quad \forall j\in[n].$$

For any $j\neq i$, since z_j is independent of $S_{i,j}$ we know

$$\mathbb{E}_{z_j}\Big[\mathbb{E}_{Z,\tilde{Z}}\big[\ell(A(S_{i,j});Z,\tilde{Z})\big] - \mathbb{E}_Z[\ell(A(S_{i,j});Z,z_j)]\Big] = 0.$$

Therefore, $\mathbb{E}_{z_j}[g_j^{(i)}] = 0$. For any $k\neq j$ and any $z_k''\in\mathcal{Z}$, it is clear from the uniform stability of A that

$$\Big|\mathbb{E}_{z_i',z_j'}\mathbb{E}_{Z,\tilde{Z}}\big[\ell(A(S_{i,j});Z,\tilde{Z})\big] - \mathbb{E}_{z_i',z_j'}\mathbb{E}_{Z,\tilde{Z}}\big[\ell(A(S_{i,j}^{(k)});Z,\tilde{Z})\big]\Big| \leq \gamma,$$

where $S_{i,j}^{(k)}$ is the set derived by replacing the k-th element of $S_{i,j}$ with z_k''. Similarly, one have

$$\Big|\mathbb{E}_{z_i',z_j'}\mathbb{E}_Z[\ell(A(S_{i,j});Z,z_j)] - \mathbb{E}_{z_i',z_j'}\mathbb{E}_Z[\ell(A(S_{i,j}^{(k)});Z,z_j)]\Big| \leq \gamma.$$

It then follows from the above two inequalities that $g_j^{(i)}$ satisfies the bounded increment condition (A.1) with $\beta = 2\gamma$ for all $k \neq j$, i.e.,

$$\Big|\mathbb{E}_{z_i',z_j'}\Big[\mathbb{E}_{Z,\tilde{Z}}\big[\ell(A(S_{i,j});Z,\tilde{Z})\big] - \mathbb{E}_Z[\ell(A(S_{i,j});Z,z_j)]\Big] \\ - \mathbb{E}_{z_i',z_j'}\Big[\mathbb{E}_{Z,\tilde{Z}}\big[\ell(A(S_{i,j}^{(k)});Z,\tilde{Z})\big] - \mathbb{E}_Z[\ell(A(S_{i,j}^{(k)});Z,z_j)]\Big]\Big| \leq 2\gamma.$$

Therefore, all the assumptions of Lemma A.1 hold for the random functions $g_1^{(i)},\ldots,g_{i-1}^{(i)},g_{i+1}^{(i)},\ldots,g_n^{(i)}$ with n there replaced by $n-1$ and $\beta = 2\gamma$. We can apply Lemma A.1 to derive

$$\Big\|\sum_{j\in[n],j\neq i} g_j^{(i)}\Big\|_p \leq 24\sqrt{6}p(n-1)\gamma\lceil\log_2(n-1)\rceil + 6\sqrt{2}M\sqrt{p(n-1)}, \quad \forall i\in[n].$$

Similarly, we can also show that

$$\Big\|\sum_{i\in[n],i\neq j} \tilde{g}_i^{(j)}\Big\|_p \leq 24\sqrt{6}p(n-1)\gamma\lceil\log_2(n-1)\rceil + 6\sqrt{2}M\sqrt{p(n-1)}, \quad \forall j\in[n].$$

It then follows from the subadditivity of $\|\cdot\|_p$ and the above two inequalities that

$$\begin{aligned}\|\sum_{i\neq j} g_{i,j}\|_p &\leq \Big\|\sum_{i\neq j} g_j^{(i)}\Big\|_p + \Big\|\sum_{i\neq j}\tilde{g}_i^{(j)}\Big\|_p \\ &\leq \sum_{i\in[n]}\Big\|\sum_{j\in[n],j\neq i} g_j^{(i)}\Big\|_p + \sum_{j\in[n]}\Big\|\sum_{i\in[n],i\neq j}\tilde{g}_i^{(j)}\Big\|_p \\ &\leq 48\sqrt{6}p(n-1)n\gamma\lceil\log_2(n-1)\rceil + 12\sqrt{2}M\sqrt{p(n-1)}n.\end{aligned}$$

We can combine the above p-norm and Lemma A.2 to derive the following inequality with probability at least $1-\delta$

$$\Big|\sum_{i\neq j} g_{i,j}\Big| \leq e\Big(12\sqrt{2}M\sqrt{(n-1)}n\sqrt{\log(e/\delta)} + 48\sqrt{6}(n-1)n\gamma\lceil\log_2(n-1)\rceil\log(e/\delta)\Big).$$

Plugging the above inequality back into (A.3) and using the definition of R_S, R, we derive the following inequality with probability at least $1-\delta$

$$\begin{aligned}|R_S(A(S)) - R(A(S))| &\leq 4\gamma + \frac{1}{n(n-1)}\Big|\sum_{i\neq j} g_{i,j}\Big| \\ &\leq 4\gamma + e\Big(12\sqrt{2}M(n-1)^{-\frac{1}{2}}\sqrt{\log(e/\delta)} + 48\sqrt{6}\gamma\lceil\log_2(n-1)\rceil\log(e/\delta)\Big).\end{aligned}$$

The proof is complete. □

B Proof of Theorem 3

In this section, we prove Theorem 3 on high-probability bounds for learning with strongly convex objective functions. We first prove Lemma 2 on the norm of output model.

Proof of Lemma 2. Since $A(S)$ is the minimizer of F_S, we know there is a $F_S'(A(S)) = 0$ (F_S' is a subgradient of F_S at $A(S)$). This together with the definition of strong convexity implies

$$R_S(\mathbf{w}^*) + r(\mathbf{w}^*) - R_S(A(S)) - r(A(S)) \geq \frac{\sigma}{2}\|A(S) - \mathbf{w}^*\|^2. \tag{B.1}$$

Analogous to (A.3), we know

$$n(n-1)\Big(R(A(S)) - R_S(A(S))\Big) \le 4n(n-1)\gamma + \sum_{i,j\in[n]:i\neq j} g_{i,j},$$

where $g_{i,j}$ is defined in the proof of Theorem 1. In the proof of Theorem 1, we have shown $\mathbb{E}[g_{i,j}] = 0$. It then follows that

$$\mathbb{E}\big[R(A(S)) - R_S(A(S))\big] \le 4\gamma.$$

We can plug the above inequality back into (B.1) to derive

$$\begin{aligned}\frac{\sigma}{2}\mathbb{E}\big[\|A(S)-\mathbf{w}^*\|^2\big] &\le \mathbb{E}\big[R_S(\mathbf{w}^*) + r(\mathbf{w}^*) - R_S(A(S)) - r(A(S))\big] \\ &\le \mathbb{E}\big[R_S(\mathbf{w}^*) + r(\mathbf{w}^*) - R(A(S)) - r(A(S))\big] + 4\gamma \\ &= \mathbb{E}\big[R(\mathbf{w}^*) + r(\mathbf{w}^*) - R(A(S)) - r(A(S))\big] + 4\gamma \le 4\gamma,\end{aligned}$$

where the last inequality holds since $\mathbf{w}^*$ minimizes $F = R + r$. The stated inequality then follows and finishes the proof. □

To prove Theorem 3, we introduce some lemmas.

Lemma B.1. *For any $S \in \mathcal{Z}^n$, define A as $A(S) = \arg\min_{\mathbf{w}\in\mathcal{W}} F_S(\mathbf{w})$. For any $k \in [n]$, let S_k be defined by* (A.2)*. Then*

$$\begin{aligned}&F_S(A(S_k)) - F_S(A(S)) \le \\ &\frac{1}{n(n-1)}\sum_{i\in[n]:i\neq k}\Big(\big(\ell(A(S_k);z_i,z_k) - \ell(A(S);z_i,z_k)\big) + \big(\ell(A(S_k);z_k,z_i) - \ell(A(S);z_k,z_i)\big) \\ &\qquad + \big(\ell(A(S);z_i,z_k') - \ell(A(S_k);z_i,z_k')\big) + \big(\ell(A(S);z_k',z_i) - \ell(A(S_k);z_k',z_i)\big)\Big).\end{aligned}$$

Proof. Without loss of generality, we can assume $k = n$. Since $A(S_n)$ is a minimizer of F_{S_n}, we know

$$\begin{aligned}&F_S(A(S_n)) - F_S(A(S)) \\ &= F_S(A(S_n)) - F_{S_n}(A(S_n)) + F_{S_n}(A(S_n)) - F_{S_n}(A(S)) + F_{S_n}(A(S)) - F_S(A(S)) \\ &\le F_S(A(S_n)) - F_{S_n}(A(S_n)) + F_{S_n}(A(S)) - F_S(A(S)).\end{aligned} \tag{B.2}$$

By the definition of F_S and F_{S_n}, we know

$$\begin{aligned}&n(n-1)\big(F_S(A(S_n)) - F_{S_n}(A(S_n))\big) = \sum_{i,j\in[n]:i\neq j} f(A(S_n);z_i,z_j) \\ &- \Big(\sum_{i,j\in[n-1]:i\neq j} f(A(S_n);z_i,z_j) + \sum_{i\in[n-1]} f(A(S_n);z_i,z_n') + \sum_{i\in[n-1]} f(A(S_n);z_n',z_i)\Big) \\ &= \sum_{i\in[n-1]}\Big(f(A(S_n);z_i,z_n) + f(A(S_n);z_n,z_i) - f(A(S_n);z_i,z_n') - f(A(S_n);z_n',z_i)\Big).\end{aligned}$$

Similarly, we know

$$\begin{aligned}&n(n-1)\big(F_{S_n}(A(S)) - F_S(A(S))\big) = \\ &\qquad\sum_{i\in[n-1]}\Big(f(A(S);z_i,z_n') + f(A(S);z_n',z_i) - f(A(S);z_i,z_n) - f(A(S);z_n,z_i)\Big).\end{aligned}$$

Therefore, we can combine the above two identities to derive

$$\begin{aligned}&n(n-1)\big(F_S(A(S_n)) - F_{S_n}(A(S_n)) + F_{S_n}(A(S)) - F_S(A(S))\big) = \\ &\sum_{i\in[n-1]}\Big(\big(f(A(S_n);z_i,z_n) - f(A(S);z_i,z_n)\big) + \big(f(A(S_n);z_n,z_i) - f(A(S);z_n,z_i)\big) + \\ &\qquad\big(f(A(S);z_i,z_n') - f(A(S_n);z_i,z_n')\big) + \big(f(A(S);z_n',z_i) - f(A(S_n);z_n',z_i)\big)\Big).\end{aligned}$$

This together with the structure of f ($f = \ell + r$ with r depending only on $\mathbf{w}$) implies

$$
\begin{aligned}
&n(n-1)\big(F_S(A(S_n)) - F_{S_n}(A(S_n)) + F_{S_n}(A(S)) - F_S(A(S))\big) = \\
&\quad \sum_{i\in[n-1]} \Big(\big(\ell(A(S_n); z_i, z_n) - \ell(A(S); z_i, z_n)\big) + \big(\ell(A(S_n); z_n, z_i) - \ell(A(S); z_n, z_i)\big) + \\
&\qquad\qquad \big(\ell(A(S); z_i, z_n') - \ell(A(S_n); z_i, z_n')\big) + \big(\ell(A(S); z_n', z_i) - \ell(A(S_n); z_n', z_i)\big)\Big).
\end{aligned}
$$

Plugging the above identity back into (B.2), we derive

$$
\begin{aligned}
&F_S(A(S_n)) - F_S(A(S)) \\
&\leq \frac{1}{n(n-1)} \sum_{i\in[n-1]} \Big(\big(\ell(A(S_n); z_i, z_n) - \ell(A(S); z_i, z_n)\big) + \big(\ell(A(S_n); z_n, z_i) - \ell(A(S); z_n, z_i)\big) \\
&\qquad + \big(\ell(A(S); z_i, z_n') - \ell(A(S_n); z_i, z_n')\big) + \big(\ell(A(S); z_n', z_i) - \ell(A(S_n); z_n', z_i)\big)\Big).
\end{aligned}
$$

The proof is complete. □

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

Lemma B.2. *Define A as $A(S) = \arg\min_{\mathbf{w}\in\mathcal{W}} F_S(\mathbf{w})$. Suppose F_S is σ-strongly convex w.r.t. $\|\cdot\|$. Assume for all $z, \tilde{z}$ we have* (4.3)*. Then A is $\frac{8L^2}{n\sigma}$-uniformly stable.*

Proof. Let S, S' be two sets that differ by a single example and let $\mathbf{w}_S = A(S)$ and $\mathbf{w}_{S'} = A(S')$. Without loss of generality, we can assume $S' = \{z_1, \ldots, z_{n-1}, z_n'\}$, i.e., S and S' differ by the last example.

Since $\mathbf{w}_S$ is a minimizer of F_S we know there is a subgradient $F_S'(\mathbf{w}_S) = 0$, which together with the σ-strong convexity of F_S, implies

$$
F_S(\mathbf{w}_{S'}) - F_S(\mathbf{w}_S) \geq \frac{\sigma}{2}\|\mathbf{w}_{S'} - \mathbf{w}_S\|^2. \tag{B.3}
$$

According to (4.3) and Lemma B.1, we know

$$
F_S(\mathbf{w}_{S'}) - F_S(\mathbf{w}_S) \leq \frac{4(n-1)L\|\mathbf{w}_S - \mathbf{w}_{S'}\|}{n(n-1)}.
$$

which, together with (B.3), implies

$$
\|\mathbf{w}_S - \mathbf{w}_{S'}\| \leq \frac{8L}{n\sigma}.
$$

This further together with (4.3) implies the $\frac{8L^2}{n\sigma}$-uniform stability of A. The proof is complete. □

To obtain tight control on the term $R(\mathbf{w}^*) - R_S(\mathbf{w}^*)$, we will need a version of Bernstein's inequality for U-statistics. The following theorem is attributed to [10], and can be found in [5] (inequality A.1 on page 868), and in [12] (Theorem 2). A complete proof is provided in [13] (page 4).

Lemma B.3 (Bernstein's inequality for U-Statistic [10, 13])**.** *Let $Z_1, \ldots, \ldots, Z_n$ 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 $1 - \delta$*

$$
\Big|\frac{1}{n(n-1)} \sum_{i,j\in[n]:i\neq j} q(Z_i, Z_j) - \mathbb{E}_{Z,\tilde{Z}}[q(Z, \tilde{Z})]\Big| \leq \frac{2b\log(1/\delta)}{3\lfloor n/2\rfloor} + \sqrt{\frac{2\sigma_0^2\log(1/\delta)}{\lfloor n/2\rfloor}}. \tag{B.4}
$$

We now give the proof of Theorem 3.

Proof of Theorem 3. According to Lemma B.2, we know that A is $\frac{8L^2}{n\sigma}$-uniformly stable. Using this together with Lemma 2 we derive $\mathbb{E}_S[\|\mathbf{w}^* - A(S)\|^2] \leq \frac{64L^2}{n\sigma^2}$ and therefore

$$\mathbb{E}_S[\|\mathbf{w}^* - A(S)\|] \leq \Big(\mathbb{E}_S[\|\mathbf{w}^* - A(S)\|^2]\Big)^{\frac{1}{2}} \leq \frac{8L}{\sqrt{n}\sigma}. \tag{B.5}$$

For any $\mathbf{w} \in \mathcal{W}$ and $z, \tilde{z}$, define

$$\tilde{\ell}(\mathbf{w}; z, \tilde{z}) = \ell(\mathbf{w}; z, \tilde{z}) - \ell(\mathbf{w}^*; z, \tilde{z}).$$

Then it is clear from Lemma B.2 that A is also $\frac{8L^2}{n\sigma}$-uniformly stable when measured by the "loss" $\tilde{\ell}$, i.e., for any S, S' differing by one example

$$\begin{aligned}
&\sup_{z,\tilde{z}} \big|\tilde{\ell}(A(S); z, \tilde{z}) - \tilde{\ell}(A(S'); z, \tilde{z})\big| \\
&= \sup_{z,\tilde{z}} \big|\ell(A(S); z, \tilde{z}) - \ell(A(S'); z, \tilde{z}) - \ell(\mathbf{w}^*; z, \tilde{z}) + \ell(\mathbf{w}^*; z, \tilde{z})\big| \\
&= \sup_{z,\tilde{z}} \big|\ell(A(S); z, \tilde{z}) - \ell(A(S'); z, \tilde{z})\big| \leq 8L^2/(n\sigma).
\end{aligned}$$

Furthermore, by the Lipschitz continuity (4.3) and (B.5), we know the following inequality for all $z, \tilde{z} \in \mathcal{Z}$

$$\begin{aligned}
\Big|\mathbb{E}_S\big[\tilde{\ell}(A(S); z, \tilde{z})\big]\Big| &= \Big|\mathbb{E}_S\big[\ell(A(S); z, \tilde{z}) - \ell(\mathbf{w}^*; z, \tilde{z})\big]\Big| \\
&\leq L\mathbb{E}_S\big[\|\mathbf{w}^* - A(S)\|\big] \leq \frac{8L^2}{\sqrt{n}\sigma}.
\end{aligned}$$

We can now apply Theorem 1, with $\gamma = \frac{8L^2}{n\sigma}$, $M = 8L^2/(\sqrt{n}\sigma)$ and ℓ replaced by $\tilde{\ell}$, and show the following inequality with probability $1 - \delta/2$

$$\begin{aligned}
\Big|\frac{1}{n(n-1)}\sum_{i\neq j}\tilde{\ell}(A(S); z_i, z_j) - \mathbb{E}_{z,\tilde{z}}\big[\tilde{\ell}(A(S); z, \tilde{z})\big]\Big| &\leq \frac{32L^2}{n\sigma} \\
&+ e\Big(\frac{96\sqrt{2}L^2\sqrt{\log(2e/\delta)}}{\sqrt{n(n-1)}\sigma}\Big) + \frac{384\sqrt{6}L^2\lceil\log_2 n\rceil\log(2e/\delta)}{n\sigma},
\end{aligned}$$

from which we derive the following inequality with probability $1 - \delta/2$

$$\begin{aligned}
\big|R_S(A(S)) - R(A(S))\big| &\leq \Big|\frac{1}{n(n-1)}\sum_{i\neq j}\ell(\mathbf{w}^*; z_i, z_j) - \mathbb{E}_{z,\tilde{z}}\big[\ell(\mathbf{w}^*; z, \tilde{z})\big]\Big| \\
&+ \frac{32L^2}{n\sigma}\Big(1 + 3\sqrt{\frac{2n\log(2e/\delta)}{n-1}} + 12\sqrt{6}\lceil\log_2 n\rceil\log(2e/\delta)\Big).
\end{aligned} \tag{B.6}$$

By to the definition of $\mathbf{w}^*$ ($R'(\mathbf{w}^*) + r'(\mathbf{w}^*) = 0$), the σ-strong convexity and Assumption 1 ($0 \leq \ell(0; z, \tilde{z})$), we know

$$\frac{\sigma\|\mathbf{w}^*\|^2}{2} \leq R(0) + r(0) - R(\mathbf{w}^*) - r(\mathbf{w}^*) \Longrightarrow \|\mathbf{w}^*\| \leq \sqrt{\frac{2(R(0) + r(0))}{\sigma}}.$$

It then follows from the Lipschitz continuity (4.3) that

$$\begin{aligned}
|\ell(\mathbf{w}^*; z, \tilde{z})| &= \big|\ell(\mathbf{w}^*; z, \tilde{z}) - \ell(0; z, \tilde{z}) + \ell(0; z, \tilde{z})\big| \leq L\|\mathbf{w}^*\| + \sup_{z,\tilde{z}}\ell(0; z, \tilde{z}) \\
&\leq L\sqrt{\frac{2(R(0) + r(0))}{\sigma}} + \sup_{z,\tilde{z}}\ell(0; z, \tilde{z}).
\end{aligned}$$

According to Bernstein's inequality (B.4), we derive the following inequality with probability $1 - \delta/2$ that

$$\begin{aligned}
\Big|\frac{1}{n(n-1)}\sum_{i\neq j}\ell(\mathbf{w}^*; z_i, z_j) - \mathbb{E}_{z,\tilde{z}}[\ell(\mathbf{w}^*; z, \tilde{z})]\Big| &\leq \\
\frac{2\big(L\sqrt{2(R(0) + r(0))/\sigma} + b\big)\log(2/\delta)}{3\lfloor n/2\rfloor} &+ \sqrt{\frac{2\sigma_0^2\log(2/\delta)}{\lfloor n/2\rfloor}}.
\end{aligned} \tag{B.7}$$

Plugging the above inequality back into (B.6), we derive the following inequality with probability at least $1-\delta$

$$\big|R_S(A(S)) - R(A(S))\big| \leq \frac{2\big(L\sqrt{2(R(0)+r(0))/\sigma}+b\big)\log(2/\delta)}{3\lfloor n/2\rfloor} + \sqrt{\frac{2\sigma_0^2\log(2/\delta)}{\lfloor n/2\rfloor}} + \frac{32L^2}{n\sigma}\Big(1+3\sqrt{\frac{2n\log(2e/\delta)}{n-1}}+12\sqrt{6}\lceil\log_2 n\rceil\log(2e/\delta)\Big).$$

The above inequality can be written as the stated bound (4.4).

We now turn to (4.5). According to the definition of R and F, we can decompose the excess risk $R(A(S)) - R(\mathbf{w}_R^*)$ as follows

$$\begin{aligned}
&R(A(S)) - R(\mathbf{w}_R^*)\\
&= R(A(S)) - R_S(A(S)) + R_S(\mathbf{w}_R^*) - R(\mathbf{w}_R^*) + R_S(A(S)) - R_S(\mathbf{w}_R^*)\\
&= R(A(S)) - R_S(A(S)) + R_S(\mathbf{w}_R^*) - R(\mathbf{w}_R^*) + F_S(A(S)) - F_S(\mathbf{w}_R^*) + r(\mathbf{w}_R^*) - r(A(S))\\
&\leq R(A(S)) - R_S(A(S)) + R_S(\mathbf{w}_R^*) - R(\mathbf{w}_R^*) + O(\sigma\|\mathbf{w}_R^*\|^2) - r(A(S)), \qquad \text{(B.8)}
\end{aligned}$$

where we have used the inequality $F_S(A(S)) \leq F_S(\mathbf{w}_R^*)$ due to the definition of $A(S)$ and the assumption $r(\mathbf{w}) = O(\sigma\|\mathbf{w}\|^2)$ in the last step. Analogous to (B.7), one can use Bernstein's inequality (Lemma B.3) to show with probability at least $1-\delta/2$ that (under a very mild assumption $\sup_{z,z'} \ell(\mathbf{w}_R^*; z, z') = O(\sqrt{n})$)

$$R_S(\mathbf{w}_R^*) - R(\mathbf{w}_R^*) = O\Big(\frac{\log(1/\delta)}{\sqrt{n}} + \sqrt{\frac{\sigma_0^2\log(1/\delta)}{n}}\Big). \qquad \text{(B.9)}$$

Plugging the above inequality and (4.4) back into (B.8) shows the following inequality with probability at least $1-\delta$

$$R(A(S)) - R(\mathbf{w}_R^*) = O\Big((n\sigma)^{-1}\log n\log(1/\delta) + n^{-\frac{1}{2}}\log(1/\delta)\Big) + O(\sigma\|\mathbf{w}_R^*\|^2).$$

The stated bound (4.5) follows with $\sigma \asymp n^{-1/2}$. The proof is complete. □

Remark B.1. We show here that the existing stability bound (eq. (4.2) with $\gamma = O(1/(n\sigma))$) [1, 6, 11, 17]

$$|R_S(A(S)) - R(A(S))| = O(\sigma^{-1}n^{-\frac{1}{2}}) \qquad \text{(B.10)}$$

yields at best the excess risk bound $R(A(S)) - R(\mathbf{w}_R^*) = O(n^{-\frac{1}{4}})$. Indeed, plugging (B.10) and (B.9) back into (B.8), we derive the following inequality with high probability

$$R(A(S)) - R(\mathbf{w}_R^*) = O(\sigma^{-1}n^{-\frac{1}{2}}) + O(\sigma).$$

We can balance the above two terms by taking $\sigma \asymp n^{-\frac{1}{4}}$ and get

$$R(A(S)) - R(\mathbf{w}_R^*) = O(n^{-\frac{1}{4}}).$$

C Proof of Theorem 4

To prove Theorem 4, we first introduce some lemmas. Lemma C.1 shows the non-expansiveness of the gradient-update operator, which plays a key role in establishing the stability of SGD. Lemma C.2 is a Chernoff's bound for a summation of independent Bernoulli random variables [2]. In this section, we let $\|\cdot\|_2$ be the Euclidean norm.

Lemma C.1 ([8]). *Assume for all $z \in \mathcal{Z}$, the function $\mathbf{w} \mapsto \ell(\mathbf{w}; z, z')$ is convex and α-smooth. Then for all $\eta \leq 2/\alpha$ and $z, z' \in \mathcal{Z}$ there holds*

$$\|\mathbf{w} - \eta\ell'(\mathbf{w}; z, z') - \mathbf{w}' + \eta\ell'(\mathbf{w}'; z, z')\|_2 \leq \|\mathbf{w} - \mathbf{w}'\|_2.$$

Lemma C.2 (Chernoff's Bound). *Let $X_1, \ldots, X_T$ be independent random variables taking values in $\{0, 1\}$. Let $X = \sum_{t=1}^T X_t$ and $\mu = \mathbb{E}[X]$. Then for any $\epsilon \in (0, 1)$ with probability at least $1 - \exp\big(-\mu\epsilon^2/3\big)$ we have $X \leq (1+\epsilon)\mu$.*

We now establish the uniform stability of SGD. The randomness of SGD can be characterized by $\{\{(i_t, j_t)_t\} : i_t, j_t \in [n], i_t \neq j_t\}$. Therefore, SGD can be considered as a deterministic algorithm if $\{\{(i_t, j_t)_t\} : i_t, j_t \in [n], i_t \neq j_t\}$ is fixed. For simplicity, we consider two datasets that differ by the last example. However, our discussion directly extends to the general case where two datasets differ by a single example. Notice that the Lipschitz continuity (4.3) implies $\|\ell'(\mathbf{w}; z, z')\|_2 \leq L$.

Lemma C.3. *Consider fixed $\{\{(i_t, j_t)_t\} : i_t, j_t \in [n], i_t \neq j_t\}$. Let $S = \{z_1, \ldots, z_n\}$ and $S' = \{z'_1, \ldots, z'_n\}$ be two datasets that differ only by the last example, i.e., $z_i = z'_i$ if $i \in [n-1]$. Suppose for all $z, z' \in \mathcal{Z}$ the function $\mathbf{w} \mapsto \ell(\mathbf{w}; z, z')$ is convex, α-smooth and L-Lipschitz w.r.t. $\|\cdot\|_2$. Let $\{\mathbf{w}_t\}, \{\mathbf{w}'_t\}$ be produced by SGD on S and S' respectively with $\eta_t \leq 2/\alpha$, i.e.,* (3.3) *with $r(\mathbf{w}) = 0$. Then SGD with t iterations is γ-uniformly stable with*

$$\gamma \leq 2L^2 \sum_{k=1}^{t} \eta_k \mathbb{I}[i_k = n \text{ or } j_k = n].$$

Proof. Let us consider two cases. We first consider the case $i_t \in [n-1]$ and $j_t \in [n-1]$. In this case, according to the SGD update (3.3) with $r(\mathbf{w}) = 0$ we know

$$\begin{aligned}\mathbf{w}_{t+1} - \mathbf{w}'_{t+1} &= \mathbf{w}_t - \eta_t \ell'(\mathbf{w}_t; z_{i_t}, z_{j_t}) - \mathbf{w}'_t + \eta_t \ell'(\mathbf{w}'_t; z'_{i_t}, z'_{j_t}) \\ &= \mathbf{w}_t - \eta_t \ell'(\mathbf{w}_t; z_{i_t}, z_{j_t}) - \mathbf{w}'_t + \eta_t \ell'(\mathbf{w}'_t; z_{i_t}, z_{j_t}).\end{aligned}$$

It then follows from Lemma C.1 that

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

We now consider the case that either $i_t = n$ or $j_t = n$. In this case, we know

$$\begin{aligned}\|\mathbf{w}_{t+1} - \mathbf{w}'_{t+1}\|_2 &= \big\|\mathbf{w}_t - \eta_t \ell'(\mathbf{w}_t; z_{i_t}; z_{j_t}) - \mathbf{w}'_t + \eta_t \ell'(\mathbf{w}'_t; z'_{i_t}, z'_{j_t})\big\|_2 \\ &\leq \|\mathbf{w}_t - \mathbf{w}'_t\|_2 + \big\|\eta_t \ell'(\mathbf{w}'_t; z'_{i_t}, z'_{j_t}) - \eta_t \ell'(\mathbf{w}_t; z_{i_t}, z_{j_t})\big\|_2 \\ &\leq \|\mathbf{w}_t - \mathbf{w}'_t\|_2 + 2\eta_t L,\end{aligned}$$

where we have used $\|\ell'(\mathbf{w}; z, z')\|_2 \leq L$ due to the L-Lipschitzness. As a combination of the above two cases, we derive

$$\|\mathbf{w}_{t+1} - \mathbf{w}'_{t+1}\|_2 \leq \|\mathbf{w}_t - \mathbf{w}'_t\|_2 + 2\eta_t L \mathbb{I}[i_t = n \text{ or } j_t = n],$$

where $\mathbb{I}[\cdot]$ is the indicator function taking 1 if the argument holds and 0 otherwise. Taking a summation of the above inequality gives ($\mathbf{w}_1 = \mathbf{w}'_1$)

$$\|\mathbf{w}_{t+1} - \mathbf{w}'_{t+1}\|_2 \leq 2L \sum_{k=1}^{t} \eta_k \mathbb{I}[i_k = n \text{ or } j_k = n].$$

This together with the Lipschitz continuity of ℓ implies the following inequality for all $z, z' \in \mathcal{Z}$

$$\begin{aligned}\big|\ell(\mathbf{w}_{t+1}; z, z') - \ell(\mathbf{w}'_{t+1}; z, z')\big| &\leq L\|\mathbf{w}_{t+1} - \mathbf{w}'_{t+1}\|_2 \\ &\leq 2L^2 \sum_{k=1}^{t} \eta_k \mathbb{I}[i_k = n \text{ or } j_k = n].\end{aligned}$$

The proof is complete. □

We now apply the above uniform stability bounds and Theorem 1 to prove Theorem 4.

Proof of Theorem 4. We can apply Theorem 1 with $A(S) = \mathbf{w}_T$ and the uniform stability bounds in Lemma C.3 to show with probability at least $1 - \delta/2$ that

$$\big|R_S(\mathbf{w}_T) - R(\mathbf{w}_T)\big| = O\Big((\log n \log(1/\delta)) \sum_{t=1}^{T} \eta \mathbb{I}[i_t = n \text{ or } j_t = n]\Big) + O(n^{-\frac{1}{2}}\sqrt{\log(1/\delta)}), \tag{C.1}$$

where $\eta = c/\sqrt{T}$. Let $X_t = \mathbb{I}[i_t = n \text{ or } j_t = n]$. It is clear that

$$\mathbb{E}[X_t] = \Pr\{i_t = n \text{ or } j_t = n\} \leq \Pr\{i_t = n\} + \Pr\{j_t = n\} = 2/n.$$

Applying Lemma C.2 with $X_t = \mathbb{I}[i_t = n \text{ or } j_t = n]$ then gives with probability $1 - \delta/2$ that

$$\sum_{t=1}^{T} X_t \leq \Big(1 + \sqrt{3\mu^{-1}\log(1/\delta)}\Big)\mu,$$

where $\mu = \sum_{t=1}^{T} \mathbb{E}[X_t] \leq 2T/n$. It then follows with probability $1 - \delta/2$ that

$$\sum_{t=1}^{T} X_t \leq \frac{2T}{n}\Big(1 + \sqrt{2nT^{-1}\log(1/\delta)}\Big). \tag{C.2}$$

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

$$\begin{aligned}\big|R_S(\mathbf{w}_T) - R(\mathbf{w}_T)\big| = O(n^{-\frac{1}{2}}\sqrt{\log(1/\delta)} + T\eta(\log n \log(1/\delta))n^{-1} \\ + \eta \log n \log(1/\delta)\sqrt{n^{-1}T\log(1/\delta)}).\end{aligned}$$

The proof is complete with $\eta = c/\sqrt{T}$. □

Remark C.1. We now give details on deriving excess risk bounds based on the estimation error bounds in Theorem 4. We can decompose the excess risk into optimization errors and estimation errors as follows (we omit $\log(1/\delta)$) [3]

$$\begin{aligned}R(\mathbf{w}_T) - R(\mathbf{w}_R^*) &= R(\mathbf{w}_T) - R_S(\mathbf{w}_T) + R_S(\mathbf{w}_T) - R_S(\mathbf{w}_R^*) + R_S(\mathbf{w}_R^*) - R(\mathbf{w}_R^*) \\ &= \big(R(\mathbf{w}_T) - R_S(\mathbf{w}_T)\big) + \big(R_S(\mathbf{w}_T) - R_S(\mathbf{w}_R^*)\big) + O(n^{-\frac{1}{2}}),\end{aligned} \tag{C.3}$$

where we have used (B.9). The first term is the estimation error and comes from the approximation of testing errors by training errors. The second term is the optimization error which comes since the optimization algorithm may not output the exact minimizer. Then Theorem 4 actually presents estimation error bounds. If we further assume $\|\mathbf{w}_t\| \leq B$ for some $B > 0$ and all t, then it was shown with high probability that [9]

$$R_S(\mathbf{w}_T) - R_S(\mathbf{w}_R^*) = O(T^{-\frac{1}{2}}\log T). \tag{C.4}$$

We can plug the above optimization error bounds and the estimation error bounds in Theorem 4 into (C.3), and get with high probability

$$R(\mathbf{w}_T) - R(\mathbf{w}_R^*) = O\Big(\log n\sqrt{T}/n + n^{-\frac{1}{2}}\log n\Big) + O(T^{-\frac{1}{2}}\log T).$$

One can take an optimal $T \asymp n$ to trade-off the optimization and estimation errors, and get

$$R(\mathbf{w}_T) - R(\mathbf{w}_R^*) = O(n^{-\frac{1}{2}}\log n).$$

Remark C.2. If we plug the uniform stability bounds in Lemma C.3 into the existing connection between stability and generalization established in (4.2), we get with high probability that

$$\big|R_S(\mathbf{w}_T) - R(\mathbf{w}_T)\big| = O\Big(\sqrt{n}\sum_{t=1}^{T}\eta_t\mathbb{I}[i_t = n \text{ or } j_t = n] + n^{-\frac{1}{2}}\Big).$$

This together with (C.2) shows the following inequality with high probability ($\eta_t = \eta = O(1/\sqrt{T})$)

$$\big|R_S(\mathbf{w}_T) - R(\mathbf{w}_T)\big| = O\Big(\frac{T\eta}{\sqrt{n}}\big(1 + \sqrt{n/T}\big) + n^{-\frac{1}{2}}\Big) = O\Big(\frac{\sqrt{T}}{\sqrt{n}}\big(1 + \sqrt{n/T}\big) + n^{-\frac{1}{2}}\Big).$$

We can plug the above estimation error bound, the optimization error bound (C.4) back into (C.3), and derive the following excess risk bound with high probability

$$R(\mathbf{w}_T) - R(\mathbf{w}_R^*) = O\Big(\frac{\log T}{\sqrt{T}} + \frac{\sqrt{T}}{\sqrt{n}}\big(1 + \sqrt{n/T}\big) + n^{-\frac{1}{2}}\Big) = O(1).$$

D Proofs on Optimistic Bounds

In this section, we prove optimistic bounds in Theorem 6 by using the smoothness of loss functions. We first prove Theorem 5 on the connection between generalization and on-average stability.

Proof of Theorem 5. For all $i, j \in [n]$, let $S_{i,j}$ be defined by (3.4). Due to the symmetry, we know $\mathbb{E}[R(A(S))] = \mathbb{E}[R(A(S_{i,j}))]$ for all $i, j \in [n]$ with $i \neq j$ and therefore

$$\begin{aligned}\mathbb{E}\big[R(A(S)) - R_S(A(S))\big] &= \frac{1}{n(n-1)} \sum_{i,j\in[n]:i\neq j} \mathbb{E}\big[R(A(S_{i,j})) - R_S(A(S))\big] \\ &= \frac{1}{n(n-1)} \sum_{i,j\in[n]:i\neq j} \mathbb{E}\Big[\ell\big(A(S_{i,j}); z_i, z_j\big) - \ell\big(A(S); z_i, z_j\big)\Big] \leq \gamma,\end{aligned}$$

where the second identity holds since $A(S_{i,j})$ is independent of z_i and z_j. The proof is complete. □

We then introduce some basic properties of smooth functions. For a α-smooth and non-negative function g, we have the following self-bounding property [14]

$$\|g'(\mathbf{w})\|^2 \leq 2\alpha g(\mathbf{w}), \quad \forall \mathbf{w} \in \mathcal{W} \tag{D.1}$$

and the following elementary inequality

$$g(\mathbf{w}) \leq g(\mathbf{w}') + \langle g'(\mathbf{w}'), \mathbf{w} - \mathbf{w}'\rangle + \frac{\alpha\|\mathbf{w} - \mathbf{w}'\|^2}{2}, \quad \forall \mathbf{w}, \mathbf{w}' \in \mathcal{W}. \tag{D.2}$$

We then present a useful lemma.

Lemma D.1. *Let S, S' be defined in Definition 2. Assume for all z, z', $\ell(\cdot, z, z')$ is α-smooth w.r.t. a norm. For all $i \in [n]$, let S_i be defined as* (A.2) *and $\epsilon > 0$. Then*

$$\mathbb{E}\big[R(A(S)) - R_S(A(S))\big] \leq \frac{\alpha\mathbb{E}[R_S(A(S))]}{\epsilon} + \frac{2(\epsilon+\alpha)}{n} \sum_{i\in[n]} \mathbb{E}\big[\|A(S_i) - A(S)\|^2\big].$$

Proof. For all $i, j \in [n]$, let $S_{i,j}$ be defined by (3.4). According to (D.2), the Cauchy-Schwartz inequality and (D.1), for all $i, j \in [n]$ we know

$$\begin{aligned}\ell(A(S_{i,j}); z_i, z_j) - \ell(A(S); z_i; z_j) &\leq \langle \ell'(A(S); z_i, z_j), A(S_{i,j}) - A(S)\rangle + \frac{\alpha}{2}\|A(S_{i,j}) - A(S)\|^2 \\ &\leq \big\|\ell'(A(S); z_i, z_j)\big\|\big\|A(S_{i,j}) - A(S)\big\| + \frac{\alpha}{2}\|A(S_{i,j}) - A(S)\|^2 \\ &\leq \frac{\|\ell'(A(S); z_i, z_j)\|^2}{2\epsilon} + \frac{\epsilon+\alpha}{2}\|A(S_{i,j}) - A(S)\|^2 \\ &\leq \frac{\alpha\ell(A(S); z_i, z_j)}{\epsilon} + \frac{\epsilon+\alpha}{2}\|A(S_{i,j}) - A(S)\|^2.\end{aligned}$$

We can plug the above inequality into Theorem 5 to derive

$$\begin{aligned}&\mathbb{E}\big[R(A(S)) - R_S(A(S))\big] \\ &\leq \frac{\alpha}{\epsilon n(n-1)} \sum_{i\neq j} \mathbb{E}\big[\ell(A(S); z_i, z_j)\big] + \frac{\epsilon+\alpha}{2n(n-1)} \sum_{i\neq j} \mathbb{E}\big[\|A(S_{i,j}) - A(S)\|^2\big] \\ &= \frac{\alpha\mathbb{E}[R_S(A(S))]}{\epsilon} + \frac{\epsilon+\alpha}{2n(n-1)} \sum_{i\neq j} \mathbb{E}\big[\|A(S_{i,j}) - A(S)\|^2\big].\end{aligned} \tag{D.3}$$

By the elementary inequality $(a+b)^2 \leq 2a^2 + 2b^2$, we get the following inequality for all $i \neq j$

$$\begin{aligned}\mathbb{E}\big[\|A(S_{i,j}) - A(S)\|^2\big] &\leq 2\mathbb{E}\big[\|A(S_{i,j}) - A(S_i)\|^2\big] + 2\mathbb{E}\big[\|A(S_i) - A(S)\|^2\big] \\ &= 2\mathbb{E}\big[\|A(S_i) - A(S)\|^2\big] + 2\mathbb{E}\big[\|A(S_j) - A(S)\|^2\big],\end{aligned}$$

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

$$\mathbb{E}\big[\|A(S_{i,j}) - A(S_i)\|^2\big] = \mathbb{E}\big[\|A(S_j) - A(S)\|^2\big].$$

Plugging the above inequality back into (D.3), we know

$$\begin{aligned}\mathbb{E}\big[R(A(S)) - R_S(A(S))\big] \leq \frac{\alpha\mathbb{E}[R_S(A(S))]}{\epsilon} \\ + \frac{\epsilon+\alpha}{n(n-1)}\sum_{i\neq j}\Big(\mathbb{E}\big[\|A(S_i) - A(S)\|^2\big] + \mathbb{E}\big[\|A(S_j) - A(S)\|^2\big]\Big).\end{aligned}$$

This yields the stated inequality and finishes the proof. □

Proof of Theorem 6. According to Lemma B.1 and the α-smoothness of ℓ, we know the following inequality for any k

$$\begin{aligned}n(n-1)\big(F_S(A(S_k)) - F_S(A(S))\big) \leq \sum_{i\in[n]:i\neq k}\bigg(\Big\langle \ell'(A(S);z_i,z_k) + \ell'(A(S);z_k,z_i) \\ - \ell'(A(S_k);z_i,z_k') - \ell'(A(S_k);z_k',z_i); A(S_k) - A(S)\Big\rangle + \frac{4\alpha\|A(S_k) - A(S)\|^2}{2}\bigg).\end{aligned}$$

It then follows from the Cauchy-Schwartz inequality that

$$\begin{aligned}n(n-1)\big(F_S(A(S_k)) - F_S(A(S))\big) \leq \sum_{i\in[n]:i\neq k}\Big(\big\|\ell'(A(S);z_i,z_k)\big\| + \big\|\ell'(A(S);z_k,z_i)\big\| \\ + \big\|\ell'(A(S_k);z_i,z_k')\big\| + \big\|\ell'(A(S_k);z_k',z_i)\big\|\Big)\|A(S_k) - A(S)\| + 2\alpha(n-1)\|A(S_k) - A(S)\|^2.\end{aligned}$$

This, together with (D.1) and (B.3), implies

$$\begin{aligned}\frac{\sigma n(n-1)\|A(S_k) - A(S)\|^2}{2} \leq \sqrt{2\alpha}\sum_{i\in[n]:i\neq k}\Big(\sqrt{\ell(A(S);z_i,z_k)} + \sqrt{\ell(A(S);z_k,z_i)} \\ + \sqrt{\ell(A(S_k);z_i,z_k')} + \sqrt{\ell(A(S_k);z_k',z_i)}\Big)\|A(S_k) - A(S)\| + 2\alpha(n-1)\|A(S_k) - A(S)\|^2\end{aligned}$$

and further

$$\begin{aligned}\frac{\sigma n(n-1)\|A(S_k) - A(S)\|}{2} \leq \sqrt{2\alpha}\sum_{i\in[n]:i\neq k}\Big(\sqrt{\ell(A(S);z_i,z_k)} + \sqrt{\ell(A(S);z_k,z_i)} \\ + \sqrt{\ell(A(S_k);z_i,z_k')} + \sqrt{\ell(A(S_k);z_k',z_i)}\Big) + 2\alpha(n-1)\|A(S_k) - A(S)\|.\end{aligned}$$

Since $2\alpha \leq \sigma n/4$, we further get

$$\begin{aligned}\frac{\sigma n(n-1)\|A(S_k) - A(S)\|}{4} \leq \sqrt{2\alpha}\sum_{i\in[n]:i\neq k}\Big(\sqrt{\ell(A(S);z_i,z_k)} + \sqrt{\ell(A(S);z_k,z_i)} \\ + \sqrt{\ell(A(S_k);z_i,z_k')} + \sqrt{\ell(A(S_k);z_k',z_i)}\Big).\end{aligned}$$

Taking a square over both sides and using the standard inequality $\big(\sum_{i=1}^{n-1} a_i\big)^2 \leq (n-1)\sum_{i=1}^{n-1} a_i^2$, we derive

$$\begin{aligned}\frac{\sigma^2 n^2(n-1)^2\|A(S_k) - A(S)\|^2}{16} \leq 2\alpha(n-1)\sum_{i\in[n]:i\neq k}\Big(\sqrt{\ell(A(S);z_i,z_k)} + \sqrt{\ell(A(S);z_k,z_i)} \\ + \sqrt{\ell(A(S_k);z_i,z_k')} + \sqrt{\ell(A(S_k);z_k',z_i)}\Big)^2.\end{aligned}$$

This, further together with the inequality $(a+b+c+d)^2 \leq 4(a^2+b^2+c^2+d^2)$, implies

$$\sigma^2 n^2(n-1)\|A(S_k)-A(S)\|^2 \leq 128\alpha \sum_{i\in[n]:i\neq k} \Big(\ell(A(S);z_i,z_k)+\ell(A(S);z_k,z_i) + \ell(A(S_k);z_i,z_k')+\ell(A(S_k);z_k',z_i)\Big).$$

Taking a summation of the above inequality from $k=1$ to n, we get

$$\sigma^2 n^2(n-1)\sum_{k=1}^n \|A(S_k)-A(S)\|^2 \leq 128\alpha \sum_{i,k\in[n]:i\neq k} \Big(\ell(A(S);z_i,z_k)+\ell(A(S);z_k,z_i) + \ell(A(S_k);z_i,z_k')+\ell(A(S_k);z_k',z_i)\Big). \quad \text{(D.4)}$$

Due to the symmetry, we know

$$\mathbb{E}\big[\ell(A(S_k);z_i,z_k')\big] = \mathbb{E}\big[\ell(A(S);z_i,z_k)\big], \quad \forall i\neq k.$$

It then follows that

$$\begin{aligned}
&\sum_{i,k\in[n]:i\neq k} \mathbb{E}\Big[\ell(A(S);z_i,z_k)+\ell(A(S);z_k,z_i)+\ell(A(S_k);z_i,z_k')+\ell(A(S_k);z_k',z_i)\Big] \\
&= \sum_{i,k\in[n]:i\neq k} \mathbb{E}\Big[\ell(A(S);z_i,z_k)+\ell(A(S);z_k,z_i)+\ell(A(S);z_i,z_k)+\ell(A(S);z_k,z_i)\Big] \\
&= 4n(n-1)\mathbb{E}\big[R_S(A(S))\big].
\end{aligned}$$

We can plug the above inequality back into (D.4) and derive that

$$\sigma^2 n\sum_{k=1}^n \mathbb{E}\big[\|A(S_k)-A(S)\|^2\big] \leq 512\alpha\mathbb{E}\big[R_S(A(S))\big].$$

We now plug the above inequality back into Lemma D.1 and derive that the following inequality for all $\epsilon>0$

$$\mathbb{E}\big[R(A(S))-R_S(A(S))\big] \leq \frac{\alpha\mathbb{E}[R_S(A(S))]}{\epsilon} + \frac{1024(\epsilon+\alpha)\alpha}{n^2\sigma^2}\mathbb{E}\big[R_S(A(S))\big].$$

We can take $\epsilon = \frac{n\sigma}{32}$ to derive

$$\mathbb{E}\big[R(A(S))-R_S(A(S))\big] \leq \Big(\frac{1024\alpha^2}{n^2\sigma^2}+\frac{64\alpha}{n\sigma}\Big)\mathbb{E}\big[R_S(A(S))\big].$$

Furthermore, according to the definition of $A(S)$ we know ($\mathbf{w}^*$ is independent of S)

$$\begin{aligned}
\mathbb{E}[F(A(S))]-F(\mathbf{w}^*) &= \mathbb{E}\big[F(A(S))-F_S(A(S))\big]+\mathbb{E}\big[F_S(A(S))-F_S(\mathbf{w}^*)\big] \\
&\leq \mathbb{E}\big[F(A(S))-F_S(A(S))\big] = \mathbb{E}\big[R(A(S))-R_S(A(S))\big].
\end{aligned}$$

This finishes the proof of (4.8).

We now turn to the bound of $\mathbb{E}[R(A(S))]-R(\mathbf{w}_R^*)$. Analogous to (B.8), we know

$$\begin{aligned}
\mathbb{E}\big[R(A(S))-R(\mathbf{w}_R^*)\big] &\leq \mathbb{E}\big[R(A(S))-R_S(A(S))\big]+O(\sigma\|\mathbf{w}_R^*\|^2) \\
&= O\Big(\frac{1}{n\sigma}\Big)\mathbb{E}\big[R_S(A(S))\big]+O(\sigma\|\mathbf{w}_R^*\|^2), \quad \text{(D.5)}
\end{aligned}$$

where we have used (4.8) in the last step. According to the definition of $A(S)$, we further know

$$R_S(A(S))+r(A(S)) \leq R_S(\mathbf{w}_R^*)+r(\mathbf{w}_R^*) = R_S(\mathbf{w}_R^*)+O(\sigma\|\mathbf{w}_R^*\|^2).$$

Since $\mathbf{w}_R^*$ is independent of S, we can take expectation to derive

$$\mathbb{E}[R_S(A(S))] = R(\mathbf{w}_R^*)+O(\sigma\|\mathbf{w}_R^*\|^2).$$

We can plug the above inequality back into (D.5), and derive

$$\mathbb{E}\big[R(A(S)) - R(\mathbf{w}_R^*)\big] = O\Big(\frac{R(\mathbf{w}_R^*)}{n\sigma} + O\big(n^{-1} + \sigma\big)\|\mathbf{w}_R^*\|^2\Big).$$

We can take

$$\sigma = \max\Big\{\frac{8\alpha}{n}, \sqrt{\frac{R(\mathbf{w}_R^*)}{n\|\mathbf{w}_R^*\|^2}}\Big\}$$

and derive

$$\mathbb{E}\big[R(A(S)) - R(\mathbf{w}_R^*)\big] = O\Big(\frac{\sqrt{R(\mathbf{w}_R^*)}\|\mathbf{w}_R^*\|}{\sqrt{n}} + \frac{\|\mathbf{w}_R^*\|^2}{n}\Big).$$

This establishes (4.9) and finishes the proof.

□

E Proofs on Applications

In this section, we present proofs for applications of our general results to metric learning.

Proof of Corollary 10. It is well known that F_S is 2λ-strongly convex w.r.t. $\|\cdot\|$. To apply Theorem 3, we require to check (4.3). For all $\mathbf{w}, \mathbf{w}', z, z'$, we know

$$\begin{aligned}
&\big|\ell^\psi(\mathbf{w}; z, z') - \ell^\psi(\mathbf{w}'; z, z')\big| \\
&= \Big|\max\big\{0, 1 - \tau(y, y')(1 - h_{\mathbf{w}}(x, x'))\big\} - \max\big\{0, 1 - \tau(y, y')(1 - h_{\mathbf{w}'}(x, x'))\big\}\Big| \\
&\le |\tau(y, y')|\big|h_{\mathbf{w}}(x, x') - h_{\mathbf{w}'}(x, x')\big| \le \big|\langle \mathbf{w} - \mathbf{w}', (x - x')(x - x')\rangle\big| \\
&\le 4B^2\|\mathbf{w} - \mathbf{w}'\|.
\end{aligned}$$

Therefore, (4.3) holds with $L = 4B^2$. The proof then completes by applying Theorem 3. □

Proof of Corollary 11. To apply Theorem 4, it suffices to show the smoothness of the loss function. The gradient of ℓ^ψ w.r.t. $\mathbf{w}$ can be calculated by

$$\nabla\ell^\psi(\mathbf{w}; z, z') = -\psi'\big(\tau(y, y')(1 - h_{\mathbf{w}}(x, x'))\big)\tau(y, y')(x - x')(x - x')^\top.$$

Then, for any $\mathbf{w}$ and $\mathbf{w}' \in \mathcal{W}$ we have

$$\begin{aligned}
&\big\|\nabla\ell^\psi(\mathbf{w}; z, z') - \nabla\ell^\psi(\mathbf{w}'; z, z')\big\|_K \\
&\le \big\|\tau(y, y')(x - x')(x - x')^\top\big\|\big|\psi'\big(\tau(y, y')(1 - h_{\mathbf{w}}(x, x'))\big) - \psi'\big(\tau(y, y')(1 - h_{\mathbf{w}'}(x, x'))\big)\big| \\
&\le 4B^2\big|\psi'\big(\tau(y, y')(1 - h_{\mathbf{w}}(x, x'))\big) - \psi'\big(\tau(y, y')(1 - h_{\mathbf{w}'}(x, x'))\big)\big| \\
&\le 4B^2|\tau(y, y')|\big|(1 - h_{\mathbf{w}}(x, x')) - (1 - h_{\mathbf{w}'}(x, x'))\big| \\
&= 4B^2\big|\langle \mathbf{w} - \mathbf{w}', (x - x')(x - x')^\top\rangle\big| \\
&\le 16B^4\|\mathbf{w} - \mathbf{w}'\|,
\end{aligned}$$

where we have used the 1-smoothness of the logistic loss in the third step. That is, ℓ^ψ is $(16B^4)$-smooth w.r.t. the Frobenius norm. The stated bound then follows from Theorem 4. □

F Minimax Optimal Excess Risk Bounds for Pairwise Learning

Here we explain that the bound $O(n^{-\frac{1}{2}})$ is minimax optimal for the excess risks in pairwise learning. To see this, we consider pairwise loss functions which do not depend on the second example, i.e., $\ell(\mathbf{w}; z, z') = \ell(\mathbf{w}; z, \tilde{z}')$ for all $z', \tilde{z}' \in \mathcal{Z}$. Then it is clear that R_S defined in (3.1) becomes

$$R_S(\mathbf{w}) = \frac{1}{n}\sum_{i\in[n]}\frac{1}{n-1}\sum_{j\in[n]:j\neq i}\ell(\mathbf{w}; z_i, z_j) = \frac{1}{n}\sum_{i\in[n]}\ell(\mathbf{w}; z_i, z_0) := \widetilde{R}_S(\mathbf{w}),$$

where z_0 is any fixed point in $\mathcal{Z}$. This is actually an objective function for pointwise learning. We know that for any estimator we can find a pointwise learning problem such that this estimator has the excess risk bound $O(n^{-\frac{1}{2}})$ [15]. Then, for any estimator we can build a pairwise learning problem such that this estimator has at best the excess risk bound $O(n^{-\frac{1}{2}})$. Furthermore, we can construct such a pairwise learning problem with the loss function independent of the second example.

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] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford university press, 2013.

[3] O. Bousquet and L. Bottou. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems*, pages 161–168, 2008.

[4] O. Bousquet, Y. Klochkov, and N. Zhivotovskiy. Sharper bounds for uniformly stable algorithms. *arXiv preprint arXiv:1910.07833*, 2019.

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

[6] C. Cortes, M. Mohri, and A. Rastogi. Magnitude-preserving ranking algorithms. In *Proceedings of the 24th international conference on Machine learning*, pages 169–176, 2007.

[7] 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.

[8] 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.

[9] N. J. Harvey, C. Liaw, Y. Plan, and S. Randhawa. Tight analyses for non-smooth stochastic gradient descent. In *Conference on Learning Theory*, pages 1579–1613, 2019.

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

[11] 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.

[12] T. Peel, S. Anthoine, and L. Ralaivola. Empirical bernstein inequalities for u-statistics. In *Advances in Neural Information Processing Systems*, pages 1903–1911. 2010.

[13] Y. Pitcan. A note on concentration inequalities for u-statistics, 2017.

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

[15] A. B. Tsybakov. *Introduction to nonparametric estimation*. Springer Science & Business Media, 2008.

[16] R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[17] 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.