Learning Rates for Nonconvex Pairwise Learning

Shaojie Li and Yong Liu

***Abstract*—Pairwise learning is receiving increasing attention since it covers many important machine learning tasks, e.g., metric learning, AUC maximization, and ranking. Investigating the generalization behavior of pairwise learning is thus of great significance. However, existing generalization analysis mainly focuses on the convex objective functions, leaving the nonconvex pairwise learning far less explored. Moreover, the current learning rates of pairwise learning are mostly of slower order. Motivated by these problems, we study the generalization performance of nonconvex pairwise learning and provide improved learning rates. Specifically, we develop different uniform convergence of gradients for pairwise learning under different assumptions, based on which we characterize empirical risk minimizer, gradient descent, and stochastic gradient descent. We first establish learning rates for these algorithms in a general nonconvex setting, where the analysis sheds insights on the trade-off between optimization and generalization and the role of early-stopping. We then derive faster learning rates of order $\mathcal{O}(1/n)$ for nonconvex pairwise learning with a gradient dominance curvature condition, where n is the sample size. Provided that the optimal population risk is small, we further improve the learning rates to $\mathcal{O}(1/n^2)$, which, to the best of our knowledge, are the first $\mathcal{O}(1/n^2)$ rates for pairwise learning.**

***Index Terms*—Generalization performance, learning rates, nonconvex optimization, pairwise learning.**

I. Introduction

PAIRWISE learning focuses on learning tasks with loss functions depending on a pair of training examples, and thus has a great advantage in modeling relative relationships between paired samples. As an important field of modern machine learning, pairwise learning instantiates many well-known learning tasks, for instance, similarity and metric learning [10], [30], [45], [55], AUC maximization [15], [16], [21], [42], [52], [77], [83], [86], [91], bipartite ranking [1], [12], [13], [57], gradient learning [60], [61], [85], minimum error entropy principle [23], [28], multiple kernel learning [35], and preference learning [20], etc.

Since its significance, there has been an increasing interest in the generalization performance analysis of pairwise learning to understand why it performs well in practice. Generalization analysis investigates how the predictive models learned from training samples behave on the testing samples, which is one of the primary interests in the machine learning community [6], [34], [43], [54], [80]. In contrast to the classical pointwise learning problems where the loss function involves single instances, pairwise learning loss contains pairs of training samples. This coupled construction leads to the fact that the empirical risk of pairwise learning has $\mathcal{O}(n^2)$ dependent terms if there are n training samples [38]. The fundamental assumption of independent and identical distributed (i.i.d.) random variables for sample is thus violated for the empirical risk of pairwise learning, which, unfortunately, renders the standard generalization analysis in the i.i.d. case not applicable in this context.

There are many existing studies on the generalization performance of pairwise learning, but most of them have the following limitations. First, they mostly study specific instantiations, for instance, metric learning, bipartite ranking or AUC maximization [37]. On the contrary, there is far less work studying the general framework of pairwise learning [36], [38]. Second, they typically require convexity conditions [38]. In the related work of studying the general pairwise framework, [31], [49], [78] investigate online pairwise learning, which is different from the offline setting of this paper. And [64], [71] study the variants of stochastic gradient descent (SGD). The most related works to this paper are [36], [37], [38]. In [37], the authors study the generalization performance of regularized empirical risk minimizer (RRM) via a peeling technology in uniform convergence. In [36], the authors establish the relationship between the generalization measure and algorithmic stability, and then use this connection to study the generalization performance of RRM and SGD. While in [38], the authors conduct a systematic generalization analysis of SGD under milder assumptions via algorithmic stability and uniform convergence of gradients. However, the above works [31], [36], [37], [38], [49], [64], [71] are almost limited to convex learning, and even often require the restrictive strong convexity condition. An exception is [38], where nonconvex learning is involved. Third, in [38], the authors only investigate the SGD, where there are two learning rates derived for nonconvex pairwise learning. One is of order $\mathcal{O}(\sqrt{d/n})$, provided with high probability under general nonconvex assumptions, while another is of order $\mathcal{O}(n^{-\frac{2}{3}})$, provided in expectation under an extra gradient dominated assumption [38], where n is the sample size and d is the dimension of parameter space. However, one can see that these rates are of slower order.

Motivated by these limitations, we provide a systematic and improved generalization analysis for nonconvex pairwise learning. Our contributions are summarized as follows.

Manuscript received 9 November 2021; revised 24 June 2022; accepted 13 February 2023. This work was supported in part by the National Natural Science Foundation of China under Grants 62076234, 61703396, and 62106257, in part by the Beijing Outstanding Young Scientist Program under Grant BJJWZYJH012019100020098. Recommended for acceptance by K.M. Lee (EIC). *(Corresponding author: Yong Liu.)*

The authors are with the Gaoling School of Artificial Intelligence, Renmin University of China, Beijing 100872, China (e-mail: 2020000277@ruc.edu.cn; liuyonggsai@ruc.edu.cn).

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TPAMI.2023.3259324.

Digital Object Identifier 10.1109/TPAMI.2023.3259324

0162-8828 © 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

IEEE Proof

- We study the generalization performance of the rarely explored nonconvex pairwise learning problems. Our analysis is performed on the general pairwise learning framework and spans empirical risk minimizer (ERM), gradient descent (GD), and stochastic gradient descent (SGD).
- We first consider the general nonconvex learning and obtain learning rates for these algorithms. Our analysis reveals that the optimization and generalization should be balanced to achieve good learning rates, which sheds insights on the role of early-stopping. The derived learning rates are based on our developed uniform convergences of gradients for pairwise learning, which may be of independent interest.
- We then study the nonconvex learning with a commonly used curvature condition, i.e., the gradient dominance assumption. We establish faster learning rates of order $\mathcal{O}(1/n)$. If the optimal population risk is small, we further improve this learning rate to $\mathcal{O}(1/n^2)$. To our best knowledge, the $\mathcal{O}(1/n)$ rate is the first for nonconvex pairwise learning, and the $\mathcal{O}(1/n^2)$ rate is the first for pairwise learning, whether in convex learning or nonconvex learning. In summary, this work provides a comprehensive and systematical analysis on the generalization properties of nonconvex pairwise learning.

This paper is organized as follows. The related work is reviewed in Section II. In Section III, we introduce the notations and present our main results. We provide the proofs in Section IV. Section V concludes this paper. Some discussions and proofs are deferred to the Appendix, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2023.3259324, including a systematic comparison with the related work.

II. Related Work

This section introduces the related work on generalization performance analysis of pairwise learning based on different approaches.

Algorithmic stability is a popular approach to study the generalization performance of pairwise learning. It is also a fundamental concept in statistical learning theory [8], [9], [33], which has a deep connection with learnability [65], [68], [70]. A training algorithm is stable if small changes in the training set result in small differences in the output predictions of the trained model [8]. [1], [22] establish the relationship between generalization and stability for ranking. [30], [76] study the regularized metric learning based on stability. [29], [81] consider differential privacy problems in pairwise setting. [71] uses stability to study the trade-off between the generalization error and optimization error for a variant of pairwise SGD. [36] starts the studying of pairwise learning framework via algorithmic stability. They provide an improved stability analysis based on [9], and further use it to establish learning rates for RRM and SGD. [38] further provides generalization guarantees for pairwise SGD under milder assumptions. Although algorithmic stability has been widely employed in pairwise learning, it generally requires convexity assumptions [38], which means that the above studies are mostly limited to convex learning. Moreover, the strong convexity condition is often required when establishing faster learning rates. However, it is known that the strong convexity condition is too restrictive [32].

Another popular approach employed for pairwise learning is uniform convergence [4], [5], [46], [56]. An advantage of uniform convergence is that it can imply meaningful learning rates for nonconvex learning [17], [19], [36], [38], [58]. In the related work of uniform convergence, [10], [12], [13], [42], [45], [52], [57], [67], [74], [83], [84], [86], [92] focus on the specific instantiations of pairwise learning, i.e., metric learning, ranking or AUC maximization. They often bound the generalization gap by its supremum over the whole (or a subset) of the hypothesis space. Then, some space complexity measures, including VC dimension, covering number, and Rademacher complexity, can be adapted to prove the learning rates. Although some work above doesn't require the convexity condition, they don't study the pairwise learning framework. [37] studies the pairwise learning framework via the uniform convergence technique. But they require a strong convexity assumption. In a very recent work, [38] develops uniform convergence of gradients for pairwise learning based on [39], and further uses it to investigate the learning rates of SGD in nonconvex pairwise learning. The uniform convergence of gradients has recently drawn increasing attention in nonconvex learning [17], [19], [39], [58], [79] and stochastic optimization [53], [88], [89], which is a gap between the gradients of the population risk and the gradients of the empirical risk. However, these works are limited to the pointwise learning setting. In this paper, we study the more complex pairwise learning and provide improved uniform convergence of gradients than [38], based on which we investigate the learning rates for generalization performance of nonconvex pairwise learning. As discussed before, the dependency in the empirical risk hinders the standard i.i.d technique. To overcome this difficulty, we need to decouple this dependency so that the standard generalization analysis established for independent data can be applied to this context. Furthermore, we develop different uniform convergence of gradients under different assumptions. For the demand of the proof, we also create two more general forms of the Bernstein inequality of pairwise learning, which may be of independent interest and benefit the Bernstein inequality's broader applicability (please refer to Appendix B, available in the online supplemental material, for details).

Except for the algorithmic stability and uniform convergence, convex analysis is employed in online pairwise learning [31], [78]. The tool of integral operator is also used to study the generalization of pairwise learning, but is often limited to the specific least square loss functions [23], [87].

III. Main Results

A. Preliminaries

Let P be a probability measure defined over a sample space $\mathcal{Z}$ and P_n be the corresponding empirical probability measure. Let $f(\cdot, z, z') : \mathcal{W} \mapsto R$ be a random objective function depending on random variables $z, z' \in \mathcal{Z}$, where $\mathcal{W}$ is a parameter space of dimension d. In pairwise learning, we aim to minimize the

IEEE Proof

IEEE Proof

following expected risk

$$F(\mathbf{w}) = \mathbb{E}_{z,z'}\left[f(\mathbf{w}; z, z')\right], \tag{1}$$

where $\mathbb{E}_{z,z'}$ denotes the expectation with respect to (w.r.t.) $z, z' \sim P$. In (1), $F(\mathbf{w})$ is also referred to as population risk. z and z' can be considered as samples, $\mathbf{w}$ can be interpreted as a model or hypothesis, and $f(\cdot,\cdot,\cdot)$ can be viewed as a loss function.

A well-known example of (1) is the pairwise supervised learning. Specifically, in the supervised learning, $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ with $\mathcal{X} \subset \mathbb{R}^{d'}$ being the input space and $\mathcal{Y} \subset \mathbb{R}$ being the output space (d' may not equal to d). Let $S = \{z_1, \ldots, z_n\}$ be a training dataset drawn independently according to P, based on which we wish to build a prediction function $h : \mathcal{X} \mapsto \mathbb{R}$ or $h : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$. Considering the parametric models, in which the predictor $h_{\mathbf{w}}$ can be indexed by a parameter $\mathbf{w} \in \mathcal{W}$, and defining $\ell(\mathbf{w}; z, z')$ as the loss that measures the quality of $h_{\mathbf{w}}$ over $z, z' \in \mathcal{Z}$, where $\ell : \mathcal{W} \times \mathcal{Z} \times \mathcal{Z} \mapsto \mathbb{R}$, the corresponding expected risk of supervised learning can be written as

$$F(\mathbf{w}) = \mathbb{E}_{z,z'}\left[\ell(\mathbf{w}; z, z')\right]. \tag{2}$$

In contrast to the traditional pointwise learning problems where the quality of a model parameter $\mathbf{w}$ is measured over an individual point, a distinctive property of (2) is that the performance of $h_{\mathbf{w}}$ should be quantified on pairs of data samples. Note that the minimization of (1) is more general than supervised learning in (2) and could be more challenging to handle [68], [70].

From (1), we know that the population risk $F(\mathbf{w})$ measures the prediction performance of $\mathbf{w}$ over the underlying distribution. However, P is typically not available and what we get is only a set of i.i.d. training samples S. In practice, we minimize the following empirical risk as an approximation [75]

$$F_S(\mathbf{w}) = \frac{1}{n(n-1)} \sum_{i,j\in[n], i\neq j} f(\mathbf{w}; z_i, z_j), \tag{3}$$

where $[n] = \{1, \ldots, n\}$. In optimizing (3), some popular algorithms are proposed including empirical risk minimizer (ERM), gradient descent (GD), and stochastic gradient descent (SGD). For this reason, we will provide generalization analysis for these algorithms. We now introduce some notations used in this paper. Denote $\|\cdot\|$ to be the L_2 norm in $\mathbb{R}^d$, i.e., $\|\mathbf{w}\| = \left(\sum_{i=1}^d |w_i|^2\right)^{\frac{1}{2}}$. Let $\mathbf{w}^*$ be the best parameter within $\mathcal{W}$, satisfying $\mathbf{w}^* \in \arg\min_{\mathcal{W}} F(\mathbf{w})$. Let $B(\mathbf{w}_0, R) := \{\mathbf{w} \in \mathbb{R}^d : \|\mathbf{w} - \mathbf{w}_0\| \leq R\}$ denote a ball with center $\mathbf{w}_0 \in \mathbb{R}^d$ and radius R. We assume that there is a radius R_1 such that $\mathcal{W} \subseteq B(\mathbf{w}^*, R_1)$. Let e be the base of the natural logarithm.

For a better understanding of the pairwise learning framework (1)–(3), we provide two examples to explain it.

- *Bipartite ranking.* In ranking problems, we aim to learn a good estimator $h_{\mathbf{w}} : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ which can correctly predict the ordering of pairs of binary labeled samples, i.e., predicting $y > y'$ if $h_{\mathbf{w}}(x, x') > 0$. The performance of $h_{\mathbf{w}}$ at examples (z, z') can be measured by choosing the $0-1$ loss. However, the $0-1$ loss is hard to be optimized in practice, one often employs surrogate losses [14]. By considering the convex surrogate losses $\ell : \mathbb{R} \mapsto \mathbb{R}_+$, the loss function of ranking is of the form $f(\mathbf{w}; z, z') = \ell(sign(y - y')h_{\mathbf{w}}(x, x'))$, where $sign(x)$ is the sign of x. Common choices of the surrogate loss ℓ include the hinge loss and the logistic loss [59].
- *Metric learning.* Let's consider the supervised metric learning with the label space $\mathcal{Y} = \{-1, +1\}$. Under this setting, we want to learn a distance metric function $h_{\mathbf{w}}(x, x') = \langle \mathbf{w}, (x - x')(x - x')^T \rangle$ such that a pair (x, x') of inputs from the same class $(y = y')$ are close to each other while a pair from different classes $(y \neq y')$ have a large distance $h_{\mathbf{w}}(x, x')$ [38], where x^T denotes the transpose of $x \in \mathbb{R}^d$ and $\mathbf{w} \in \mathbb{R}^{d\times d}$. Similarly, considering the convex surrogate loss function ℓ, a common choice of the loss function in supervised metric learning is of the form $f(\mathbf{w}; z, z') = \ell(yy'(1 - h_{\mathbf{w}}(x, x')))$ [30], [38]. Moreover, one can refer to [45] for examples of unsupervised metric learning, where the authors study the similarity-based clustering learning under the framework of pairwise learning.

B. Uniform Convergence of Gradients

Uniform convergence of gradients measures the deviation between the population gradients ∇F and the empirical gradients ∇F_S, where ∇ denotes the gradient operator. In this subsection, we aim to provide improved uniform convergence of gradients than the associated one in [38]. Before providing the main theorems, we first introduce a crucial assumption.

Assumption 1. For all $\mathbf{w}_1, \mathbf{w}_2 \in \mathcal{W}$, we assume that $\frac{\nabla f(\mathbf{w}_1; z, z') - \nabla f(\mathbf{w}_2; z, z')}{\|\mathbf{w}_1 - \mathbf{w}_2\|}$ is a γ-sub-exponential random vector, i.e., for any unit vector $\mathbf{u} \in B(0, 1)$ and $\mathbf{w}_1, \mathbf{w}_2 \in \mathcal{W}$,

$$\mathbb{E}\left\{\exp\left(\frac{|\mathbf{u}^T(\nabla f(\mathbf{w}_1; z, z') - \nabla f(\mathbf{w}_2; z, z'))|}{\gamma\|\mathbf{w}_1 - \mathbf{w}_2\|}\right)\right\} \leq 2,$$

where $\gamma > 0$.

Remark 1. This assumption is stronger than the smoothness of the population risk, but much milder than the uniform smoothness condition (Assumption 4). Please refer to Section IV-A for the proof.

Based on Assumption 1, we have the first theorem on uniform convergence of gradients.

Theorem 1. Suppose Assumption 1 holds. Then for any $\delta \in (0, 1)$, with probability $1 - \delta$, for all $\mathbf{w} \in \mathcal{W}$, we have

$$\begin{aligned} &\|(\nabla F(\mathbf{w}) - \nabla F_S(\mathbf{w})) - (\nabla F(\mathbf{w}^*) - \nabla F_S(\mathbf{w}^*))\| \\ &\leq c\gamma \max\left\{\|\mathbf{w} - \mathbf{w}^*\|, \frac{1}{n}\right\}\left(\sqrt{\frac{d + \log\frac{4\log_2(\sqrt{2}R_1 n + 1)}{\delta}}{n}}\right. \\ &\qquad \left. + \frac{d + \log\frac{4\log_2(\sqrt{2}R_1 n + 1)}{\delta}}{n}\right), \end{aligned}$$

where c is an absolute constant.

Remark 2. Uniform convergence of gradients is first studied in convex learning [88], [89]. Recently, uniform convergence of gradients of nonconvex learning is also proposed based on different techniques. Specifically, [58] is based on covering numbers, [19] is based on a chain rule for vector-valued Rademacher complexity, [39] is based on Rademacher chaos complexity, [17] is based on the gradient of the Moreau envelops, and [79] is based on a novel uniform localized convergence technique. However, the above-mentioned works are limited to the pointwise learning case. In Theorem 1, we present the uniform convergence of gradients for the more complex pairwise learning. As discussed in Section II, a key difference between pointwise learning and pairwise learning is that the gradient of the empirical risk in pairwise learning (see (3)) involves $\mathcal{O}(n^2)$ dependent terms, which makes the proof of Theorem 1 more challenging.

We now introduce a Bernstein condition at the optimal point, based on which we will show Theorem 2.

Assumption 2. The gradient at $\mathbf{w}^*$ satisfies the Bernstein condition, i.e., there exists $D_* > 0$ such that for all $2 \le k \le n$,

$$\mathbb{E}\left[\|\nabla f(\mathbf{w}^*; z, z')\|^k\right] \le \frac{k!}{2}\mathbb{E}\left[\|\nabla f(\mathbf{w}^*; z, z')\|^2\right] D_*^{k-2}.$$

Remark 3. Assumption 2 is pretty mild since $D_* > 0$ only depends on gradients at $\mathbf{w}^*$. Moreover, the Bernstein condition is milder than the bounded assumption of random variables and is also satisfied by various unbounded variables [75]. Please refer to [75] for more discussions on this assumption.

Theorem 2. Suppose Assumptions 1 and 2 hold. For any $\delta > 0$, with probability at least $1 - \delta$, for all $\mathbf{w} \in \mathcal{W}$, we have

$$\|\nabla F(\mathbf{w}) - \nabla F_S(\mathbf{w})\| \le c\gamma \max\left\{\|\mathbf{w} - \mathbf{w}^*\|, \frac{1}{n}\right\}$$
$$\times \left(\sqrt{\frac{d + \log\frac{8\log_2(\sqrt{2}R_1 n+1)}{\delta}}{n}} + \frac{d + \log\frac{8\log_2(\sqrt{2}R_1 n+1)}{\delta}}{n}\right)$$
$$+ \frac{4D_*\log\frac{4}{\delta}}{n} + \sqrt{\frac{8\mathbb{E}\left[\|\nabla f(\mathbf{w}^*; z, z')\|^2\right]\log\frac{4}{\delta}}{n}},$$

where c is an absolute constant.

Remark 4. There is only one existing result guaranteeing uniform convergence of gradients for pairwise learning, developed in [38]. We now compare our uniform convergence of gradients with [38]. Under uniformly smooth assumption (Assumption 4), [38] shows that with probability at least $1 - \delta$

$$\sup_{\mathbf{w}\in B(0,R)} \|\nabla F(\mathbf{w}) - \nabla F_S(\mathbf{w})\|$$
$$\le \frac{c(\beta R + b)}{\sqrt{n}}\left(2 + \sqrt{96e(\log 2 + d\log(3e))} + \sqrt{\log(1/\delta)}\right), \tag{4}$$

where $b = \sup_{z,z'\in\mathcal{Z}} \|\nabla f(0; z, z')\|$. Compared with (4), we successfully relax the uniform smoothness assumption to a milder Assumptions 1. Moreover, the factor in (4) is $c(\beta R + b)$, while in Theorem 2 is $c\gamma\max\{\|\mathbf{w} - \mathbf{w}^*\|, \frac{1}{n}\}$, not involving a term $\sup_{z,z'\in\mathcal{Z}} \|\nabla f(0; z, z')\|$ that may be very large. And we emphasize that it is the construction of the factor that allows us to derive improved learning rates when considering Assumption 3. The proof techniques of bounding the term $\sup_{\mathbf{w}\in B(0,R)} \|\nabla F(\mathbf{w}) - \nabla F_S(\mathbf{w})\|$ in [38] rely on the McDiarmid's inequality and the global Rademacher complexity. Different from the technique in [38], we use the uniform localized convergence (localized complexity technique) proposed in [79], i.e., Lemma 1 in the Appendix, available in the online supplemental material. However, [79] studies the pointwise setting. We study the uniform convergence of gradients for the more complex pairwise learning. The influence is that, for instance, in the proof of Theorem 1, after obtaining the sub-exponential random variable of (12) by following the proof of [79], we need Bernstein inequalities of pairwise learning for the unbounded random variable, which is different from the commonly used Bernstein inequalities for the bounded random variable. As discussed in Section II, the loss structure of pairwise learning hinders the standard i.i.d technique. To proceed, we need to decouple the dependency that emerged in pairwise learning. Please see Lemmas 6 and 8 in the appendix, available in the online supplemental material, for details. Then, using the generic chaining technique and Lemma 1 in the Appendix, available in the online supplemental material, we finish the proof.

In the following, we further provide an improved uniform convergence of gradients when the PL curvature condition (gradient dominance condition) is satisfied.

Assumption 3. Fix a set $\mathcal{W}$. For any function $f : \mathcal{W} \mapsto \mathbb{R}$, let $f^* = \min_{\mathbf{w}\in\mathcal{W}} f(\mathbf{w})$. f satisfies the Polyak-Łojasiewicz (PL) condition with parameter $\mu > 0$ on $\mathcal{W}$ if

$$f(\mathbf{w}) - f^* \le \frac{1}{2\mu}\|\nabla f(\mathbf{w})\|^2, \quad \forall \mathbf{w} \in \mathcal{W}.$$

Remark 5. PL condition is also referred to as "gradient dominance condition" [19]. This condition means that the suboptimality of function values can be bounded by the squared magnitude of gradients, which can be used to bound how far away the nearest minimizer is in terms of the optimality gap. It is one of the weakest curvature conditions and is widely employed in nonconvex learning [11], [32], [38], [39], [41], [66], [79], [93], to mention but a few. Under suitable assumptions on the input, many popular nonconvex objective functions satisfy PL condition, including neural networks with one hidden layer [48], ResNets with linear activations [24], robust regression [50], linear dynamical systems [25], matrix factorization [50], phase retrieval [73], blind deconvolution [47], mixture of two Gaussians [3], etc. Furthermore, the PL condition is assumed on the parameter $\mathbf{w}$, not the sample. Thus, the PL condition of pointwise learning can be easily extended to pairwise learning. We now take AUC maximization as an example to illustrate this point. Specifically, AUC maximization aims to rank positive instances above negative ones which involves a loss $f(\mathbf{w}; (x, y), (x,' y')) = (1 - \mathbf{w}^T(x - x'))_+\mathbb{I}_{[y=1 \wedge y'=-1]}$ with $x, x' \in \mathcal{X} \subseteq \mathbb{R}^d$ and $y, y' \in \mathcal{Y} = \{\pm 1\}$. Consider the problem of learning a generalized linear model with the square loss, the loss of pointwise learning is $f(\mathbf{w}; x, y) = (y - logit(\mathbf{w}^T x))^2$, where $logit(t) = (1 + \exp(-t))^{-1}$. In Section III of [19], it was shown that this loss satisfies the PL condition. In this case, the loss function for the problem of AUC maximization becomes $f(\mathbf{w}; (x, y), (x,' y')) = (1 - logit(\mathbf{w}^T(x - x')))^2\mathbb{I}_{[y=1 \wedge y'=-1]}$. Since the PL condition focuses on the parameter $\mathbf{w}$, this loss of AUC maximization also satisfies the PL condition, as shown in [82]. Moreover, AUC maximization problem with the classifier given by a one hidden layer network

IEEE Proof

satisfies the PL condition as shown in Theorem 4 in [51], corresponding to the pointwise learning in [48]. Additionally, under technical restrictions, such as the smoothness of Assumption 4, many other well-known conditions including strong convexity, one-point convexity, star convexity and τ-star convexity imply the PL condition [32].

Theorem 3. Assume Assumptions 1 and 2 hold. Suppose the population risk F satisfies Assumption 3 with parameter μ. Then for any $\delta > 0$, when $n \geq \frac{c\gamma^2\left(d+\log\frac{8\log_2(\sqrt{2}R_1 n+1)}{\delta}\right)}{\mu^2}$, with probability at least $1-\delta$, for all $\mathbf{w} \in \mathcal{W}$, we have

$$\|\nabla F(\mathbf{w}) - \nabla F_S(\mathbf{w})\| \leq \|\nabla F_S(\mathbf{w})\| + \frac{\mu}{n} + \frac{8D_* \log(4/\delta)}{n} + 4\sqrt{\frac{2\mathbb{E}\left[\|\nabla f(\mathbf{w}^*; z, z')\|^2\right]\log(4/\delta)}{n}}, \quad (5)$$

where c is an absolute constant.

Remark 6. Note that $\mathbf{w}^*$ cannot be any minimizer of F. $\mathbf{w}^*$ should be the projection of $\mathbf{w}$ onto the minimizer of F. It depends on $\mathbf{w}$. For Theorem 3, it is clear that (5) implies

$$\|\nabla F(\mathbf{w})\| \leq 2\|\nabla F_S(\mathbf{w})\| + \frac{\mu}{n} + \frac{8D_* \log(4/\delta)}{n} + 4\sqrt{\frac{2\mathbb{E}[\|\nabla f(\mathbf{w}^*; z, z')\|^2]\log(4/\delta)}{n}}. \quad (6)$$

Typically, we call $\|\nabla F_S(\mathbf{w})\|^2$ the optimization error and $\|\nabla F_S(\mathbf{w}) - \nabla F(\mathbf{w})\|^2$ the statistical error (or generalization error) [39], since the former is related to the optimization algorithm to optimize F_S, and the latter is related to approximating the true gradient with its empirical form. In Theorem 3, $\|\nabla F_S(\mathbf{w})\|$ can be tiny since the optimization algorithms, such as GD and SGD, can optimize it to be small enough. $\mathbb{E}\left[\|\nabla f(\mathbf{w}^*; z, z')\|^2\right]$ may be also small since it depends on the gradient on the optima $\mathbf{w}^*$ and involves an expectation operator. First, the bound in (4) scales with $\sup_{z,z'\in\mathcal{Z}} \|\nabla f(0; z, z')\|$, which depends on the worst case of the sample space $\sup_{z,z'\in\mathcal{Z}}$ and may be very large, while $\mathbb{E}\left[\|\nabla f(\mathbf{w}^*; z, z')\|^2\right]$ involves an expectation operator. Second, from (35), one can see that if f is nonnegative and β-smooth, we have $\mathbb{E}[\|\nabla f(\mathbf{w}^*; z, z')\|^2] \leq 4\beta F(\mathbf{w}^*)$. For the overparametrized models, such as the deep learning models, the population risk at the optima $\mathbf{w}^*$, i.e., the optimal population risk $F(\mathbf{w}^*)$, is generally very small. In the latter application in Sections III-C, III-D, and III-E, we assume $\mathbb{E}\left[\|\nabla f(\mathbf{w}^*; z, z')\|^2\right] = \mathcal{O}\left(\frac{1}{n}\right)$ or $F(\mathbf{w}^*) = \mathcal{O}\left(\frac{1}{n}\right)$ just to show that we can get improved bounds under the low noise condition. The two terms should be independent of n. It is notable that the assumption $F(\mathbf{w}^*) = \mathcal{O}\left(\frac{1}{n}\right)$, even $F(\mathbf{w}^*) = \mathbf{0}$, is common and can be found in [36], [38], [40], [53], [72], [88], [89], which is natural since $F(\mathbf{w}^*)$ is the minimal population risk. Moreover, even without the low noise condition, the bounds with a fast rate established in this paper are still sharper than the results in the related work. Therefore, compared with Theorems 1 and 2, and (4), this uniform convergence of gradients is clearly tighter. Moreover, the fact that our established convergence of gradients scales tightly with the optimal parameter, i.e., the gradient norms at the optima $\mathbf{w}^*$, largely contributes to derive faster $\mathcal{O}(1/n^2)$ rates of this paper, which is a remarkable advance compared to (4). The appearance of $\mathbb{E}\left[\|\nabla f(\mathbf{w}^*; z, z')\|^2\right]$ requires technical analysis. Additionally, an obvious shortcoming of uniform convergence is that it often implies learning rates with a square-root dependency on the dimension d when considering general problems [18], as shown in (4), and Theorems 1 and 2. Another distinctive improvement of Theorem 3 is that we successfully remove the dimension d when the population risk F satisfies the PL condition and the sample size n is large enough. Based on Theorem 3, we will provide dimension-independent learning rates for ERM, GD, and SGD. In addition to these algorithms, the uniform convergence of gradients in this paper can be employed to study other optimization algorithms, such as variance reduction variants and momentum-based optimization algorithms [62], which would also be very interesting.

C. Empirical Risk Minimizer

Generalization performance means the generalization behavior of the trained model on testing examples. Let $\mathbf{w}(S)$ be the learned model produced by some algorithms on the training set S. In Sections III-C, III-D, and III-E, we first consider the general nonconvex learning problems and present the learning rate for the gradient norm of the population risk, i.e., $\|\nabla F(\mathbf{w}(S))\|$. After that, we study the noconvex learning with the PL condition and provide learning rates for the generalization performance gap $F(\mathbf{w}(S)) - F(\mathbf{w}^*)$, where $\mathbf{w}^* \in \arg\min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w})$. In this section, we consider the ERM problem. In ERM, we focus on the optima $\hat{\mathbf{w}}^*$ of the empirical risk F_S, i.e., $\hat{\mathbf{w}}^* \in \arg\min_{\mathbf{w}\in\mathcal{W}} F_S(\mathbf{w})$.

Theorem 4. Suppose the empirical risk minimizers $\hat{\mathbf{w}}^*$ exists. Assume Assumptions 1 and 2 hold. For any $\delta \in (0,1)$, with probability at least $1-\delta$, we have

$$\|\nabla F(\hat{\mathbf{w}}^*)\| = \mathcal{O}\left(\sqrt{\frac{d + \log\frac{\log n}{\delta}}{n}}\right).$$

Remark 7. When Assumptions 1 and 2 hold, Theorem 4 shows that the learning rate of $\|\nabla F(\hat{\mathbf{w}}^*)\|$ is of order $\mathcal{O}\left(\sqrt{\frac{d+\log\frac{1}{\delta}}{n}}\right)$ ($\log n$ is small and can be ignored typically). Note that this bound does not require the uniform smoothness condition (Assumption 4). Although it is hard to find $\hat{\mathbf{w}}^*$ in nonconvex learning, this learning rate is meaningful by assuming the ERM has been found. Moreover, this learning rate may be comparable to the classical one $\mathcal{O}\left(\sqrt{\frac{d\log n \log(d/\delta)}{n}}\right)$ in the stochastic convex optimization [69], without requiring the convexity condition.

Theorem 5. Suppose Assumptions 1 and 2 hold, and the population risk $F(\mathbf{w})$ statisfies Assumption 3 with parameter μ. For any $\delta \in (0,1)$, with probability at least $1-\delta$, when $n \geq \frac{c\gamma^2\left(d+\log\left(\frac{8\log(\sqrt{2}nR_1+1)}{\delta}\right)\right)}{\mu^2}$, we have

$$F(\hat{\mathbf{w}}^*) - F(\mathbf{w}^*) = \mathcal{O}\left(\frac{\log^2\frac{1}{\delta}}{n^2} + \frac{\mathbb{E}\left[\|\nabla f(\mathbf{w}^*; z, z')\|^2\right]\log\frac{1}{\delta}}{n}\right).$$

IEEE Proof

Algorithm 1: GD for Pairwise Learning.

Input: initial point $\mathbf{w}_1 = 0$, step sizes $\{\eta_t\}_t$, and dataset $S = \{z_1, \ldots, z_n\}$

for $t = 1, \ldots, T$ **do**

update $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla F_S(\mathbf{w}_t)$

end for

If further assume $\mathbb{E}\left[\|\nabla f(\mathbf{w}^*; z, z')\|^2\right] = \mathcal{O}\left(\frac{1}{n}\right)$, we have

$$F(\hat{\mathbf{w}}^*) - F(\mathbf{w}^*) = \mathcal{O}\left(\frac{\log^2(1/\delta)}{n^2}\right).$$

Remark 8. Theorem 5 shows that when population risk $F(\mathbf{w})$ satisfies the PL condition, we can provide much faster learning rate than Theorem 4. The learning rate can even up to $\mathcal{O}\left(\frac{1}{n^2}\right)$. We now compare our result with the most related work [37], [41]. [37] studies the learning rate of generalization performance gap of regularized empirical risk minimizers (RRM) via uniform convergence technique. Under the Lipschitz continuity condition and the strong convexity condition, Theorems 1 and 2 in [37] provide $\mathcal{O}\left(\frac{\log(1/\delta)}{n}\right)$ order rates. [41] studies the generalization performance gap of RRM via algorithmic stability. Under the Lipschitz continuity and strong convexity conditions, Theorem 3 in [41] provides $\mathcal{O}\left(\frac{\log n \log(1/\delta)}{\sqrt{n}}\right)$ order rates. By the comparison, we have established much faster learning rates, significantly, under a nonconvex learning setting.

D. Gradient Descent

We now analyze the generalization performance of gradient descent of pairwise learning, where the algorithm is shown in Algorithm 1. Denote $A \asymp B$ if there exists universal constants $C_1, C_2 > 0$ such that $C_1 A \leq B \leq C_2 A$. Similarly, we first introduce a necessary assumption.

Assumption 4 (Smoothness). Let $\beta > 0$. For any sample $z, z' \in \mathcal{Z}$ and $\mathbf{w}_1, \mathbf{w}_2 \in \mathcal{W}$, there holds that

$$\|\nabla f(\mathbf{w}_1; z, z') - \nabla f(\mathbf{w}_2; z, z')\| \leq \beta \|\mathbf{w}_1 - \mathbf{w}_2\|.$$

Remark 9. The uniform smoothness condition is commonly used in nonconvex learning [17], [19], [26], [38], [39], [58]. As discussed in Section IV-A, Assumption 4 implies Assumption 1. Thus, the established uniform convergences of gradients is also correct under Assumption 4. In the following, we require this assumption to derive the optimization error bound, i.e., $\|\nabla F_S(\mathbf{w}(S))\|$.

Theorem 6. Suppose Assumptions 2 and 4 hold and the objective function f is nonnegative. Let $\{\mathbf{w}_t\}_t$ be the sequence produced by Algorithm 1 with $\eta_t = \eta_1 t^{-\theta}, \theta \in (0,1)$ and $\eta_1 \leq 1/\beta$. For any $\delta \in (0,1)$, with probability at least $1-\delta$, when $T \asymp (nd^{-1})^{\frac{1}{2(1-\theta)}}$, we have

$$\frac{1}{\sum_{t=1}^T \eta_t} \sum_{t=1}^T \eta_t \|\nabla F(\mathbf{w}_t)\|^2 \leq \mathcal{O}\left(\frac{d + \log \frac{\log n}{\delta}}{\sqrt{nd}}\right).$$

Remark 10. To our best knowledge, this is the first work that investigates the learning rates of GD for nonconvex pairwise learning. Theorem 6 shows that for pairwise GD, one should select an appropriate iterative number for early-stopping to achieve a good learning rate. In the proof, (28) reveals that we should balance the optimization error (optimization) and the statistical error (generalization), which demonstrates the reason for early-stopping. According to Theorem 6, the optimal iterative number should be chosen as $T \asymp (nd^{-1})^{\frac{1}{2(1-\theta)}}$ for polynomially decaying step sizes.

Algorithm 2: SGD for Pairwise Learning.

Input: initial point $\mathbf{w}_1 = 0$, step sizes $\{\eta_t\}_t$, and dataset $S = \{z_1, \ldots, z_n\}$

for $t = 1, \ldots, T$ **do**

draw (i_t, j_t) from the uniform distribution over the set $\{(i,j) : i, j \in [n], i \neq j\}$

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

end for

Theorem 7. Suppose Assumptions 2 and 4 hold and the objective function f is nonnegative. Assume the empirical risk F_S and the population risk F satisfy Assumption 3 with parameter μ. Let $\{\mathbf{w}_t\}_t$ be the sequence produced by Algorithm 1 with $\eta_t = 1/\beta$. For any $\delta \in (0,1)$, with probability at least $1-\delta$, when $n \geq \frac{c\beta^2\left(d + \log\left(\frac{16\log(\sqrt{2}nR_1+1)}{\delta}\right)\right)}{\mu^2}$, we have

$$F(\mathbf{w}_{T+1}) - F(\mathbf{w}^*) \leq \mathcal{O}\left((1-\frac{\mu}{\beta})^T\right) + \mathcal{O}\left(\frac{\log^2(1/\delta)}{n^2} + \frac{F(\mathbf{w}^*)\log(1/\delta)}{n}\right).$$

If further assume $F(\mathbf{w}^*) = \mathcal{O}\left(\frac{1}{n}\right)$ and choose $T \asymp \log n$, we have

$$F(\mathbf{w}_{T+1}) - F(\mathbf{w}^*) = \mathcal{O}\left(\frac{\log^2(1/\delta)}{n^2}\right).$$

Remark 11. For brevity, we show Theorem 7 with a step size $\eta_t = 1/\beta$. Indeed, Theorem 7 is correct for any $0 < \eta_t \leq 1/\beta$. Theorem 7 reveals that when the PL condition is satisfied, the generalization performance gap of GD is of the order $\mathcal{O}\left(\frac{F(\mathbf{w}^*)\log(1/\delta)}{n}\right)$, faster than the result of Theorem 6. If we suppose the optimal population risk is small as assumed in [36], [38], [40], [53], [72], [88], [89], we further obtain faster learning rate of order $\mathcal{O}(\frac{\log^2(1/\delta)}{n^2})$.

E. Stochastic Gradient Descent

Stochastic gradient descent optimization algorithm has found wide application in machine learning due to its simplicity in implementation, low memory requirement and low computational complexity per iteration, as well as good practical behavior [2], [7], [27], [90]. The description of SGD of pairwise learning is shown in Algorithm 2. We also first introduce a necessary assumption.

Assumption 5. Assume the existence of $G > 0$ and $\sigma > 0$ satisfying

$$\sqrt{\eta_t}\|\nabla f(\mathbf{w}_t; z, z')\| \leq G, \forall t \in \mathbb{N}, z, z' \in \mathcal{Z}, \quad (7)$$

IEEE Proof

$$\mathbb{E}_{i_t,j_t}\left[\|\nabla f(\mathbf{w}_t; z_{i_t}, z_{j_t}) - \nabla F_S(\mathbf{w}_t)\|^2\right] \le \sigma^2, \forall t \in \mathbb{N}, \quad (8)$$

where $\mathbb{E}_{i_t,j_t}$ denotes the expectation w.r.t. i_t and j_t.

Remark 12. In Assumption 5, (7) is much milder than the bounded gradient assumption (see Appendix A, available in the online supplemental material) since η_t is typically small [38], such as the setting of this paper. (8) is a common assumption in the generalization performance analysis of SGD [38], [44], [93].

Theorem 8. Suppose Assumptions 2, 4 and 5 hold and the objective function f is nonnegative. Let $\{\mathbf{w}_t\}_t$ be the sequence produced by Algorithm 2 with $\eta_t = \eta_1 t^{-\theta}, \theta \in (0, 1)$ and $\eta_1 \le \frac{1}{2\beta}$. Then, for any $\delta > 0$, with probability $1 - \delta$, when $T \asymp (nd^{-1})^{\frac{1}{2-2\theta}}$, we have

$$\begin{aligned}&\left(\sum_{t=1}^{T}\eta_t\right)^{-1}\sum_{t=1}^{T}\eta_t\|\nabla F(\mathbf{w}_t)\|^2\\ &= \begin{cases}\mathcal{O}\left(\left(\sqrt{\frac{d}{n}}\right)^{\frac{\theta}{1-\theta}}\log^3(1/\delta)\right), & \text{if } \theta < 1/2,\\ \mathcal{O}\left(\sqrt{\frac{d}{n}}\log(T/\delta)\log^3(1/\delta)\right), & \text{if } \theta = 1/2,\\ \mathcal{O}\left(\sqrt{\frac{d}{n}}\log^3(1/\delta)\right), & \text{if } \theta > 1/2.\end{cases}\end{aligned}$$

Remark 13. Similar to Theorem 6, Theorem 8 also implies a trade-off between the optimization error (optimization) and the statistical error (generalization) for SGD, as revealed in (36)–(38). Theorem 8 suggests that we achieve similar fast learning rates for polynomially decaying step size with $\theta \in [1/2, 1)$. While for the varying $T \asymp (nd^{-1})^{\frac{1}{2-2\theta}}$, the optimal iterative number should be chosen with $\theta = 1/2$ or closing to $1/2$. We compare Theorem 8 with the most related work [38]. [38] also studies SGD of nonconvex pairwise learning, and provide $\mathcal{O}\left(n^{-\frac{1}{2}\log^2(1/\delta)}(d + \log(1/\delta))^{\frac{1}{2}}\right)$ order rates, which has the same order $\mathcal{O}(\sqrt{\frac{d}{n}})$ as ours. However, the proof technique between Theorem 8 and [38] is different. Another difference is that [38] studies the case $\eta_t = \eta/\sqrt{T}$ with $\eta \le \sqrt{T}/(2\beta)$, while Theorem 8 studies with different step sizes. Theorem 8 is thus served as an important complementary result for nonconvex pairwise learning.

Theorem 9. Suppose Assumptions 2, 4 and 5 hold, and the objective function f is nonnegative. Suppose the empirical risk F_S and the population risk F satisfy Assumption 3 with parameter 2μ. Let $\{\mathbf{w}_t\}_t$ be the sequence produced by Algorithm 2 with $\eta_t = \frac{2}{\mu(t+t_0)}$ such that $t_0 \ge \max\{\frac{4\beta}{\mu}, 1\}$ for all $t \in \mathbb{N}$. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over the sample S, when $n \ge \frac{c\beta^2\left(d+\log\left(\frac{16\log(\sqrt{2}nR_1+1)}{\delta}\right)\right)}{\mu^2}$ and $T \asymp n^2$, we have

$$F(\mathbf{w}_{T+1}) - F(\mathbf{w}^*) = \mathcal{O}\left(\frac{\log n \log^3(\frac{1}{\delta})}{n^2} + \frac{F(\mathbf{w}^*)\log\frac{1}{\delta}}{n}\right).$$

If further assume $F(\mathbf{w}^*) = \mathcal{O}(\frac{1}{n})$, we have

$$F(\mathbf{w}_{T+1}) - F(\mathbf{w}^*) = \mathcal{O}\left(\frac{\log n \log^3(1/\delta)}{n^2}\right).$$

Remark 14. Theorem 9 reveals that under the PL condition, the learning rate of SGD can be significantly improved compared to Theorem 8. In the related work, if f is nonnegative, Lipschitz continuous and smooth, F_S satisfies the PL condition, and Assumption 5 hold, the learning rate derived for $\mathbb{E}[F(\mathbf{w}_{T+1}) - F(\mathbf{w}^*)]$ in [38] is at most of order $\mathcal{O}\left(n^{-\frac{2}{3}}\right)$. By a comparison, one can see that our learning rates are derived with high probability and are significantly faster than the results in [38]. The generalization performance gap is also studied for pairwise SGD in [36] via algorithmic stability. However, their learning rate is limited to convex learning. Specifically, if f is convex and smooth, $F(\mathbf{w}_{T+1}) - F(\mathbf{w}^*)$ is of order $\mathcal{O}\left(\log n\sqrt{T}/n + n^{-\frac{1}{2}}\right) + \mathcal{O}\left(T^{-\frac{1}{2}}\log T\right)$. By taking the optimal $T \asymp n$, the learning rate becomes $\mathcal{O}\left(n^{-\frac{1}{2}}\log n\right)$, which is much slower than results of Theorem 9. To our best knowledge, the $\mathcal{O}\left(\frac{1}{n}\right)$ rate is the first for SGD in nonconvex pairwise learning, and the $\mathcal{O}\left(\frac{1}{n^2}\right)$ rate is also the first whether in convex or nonconvex pairwise learning. Additionally, when we take $T \asymp n$, the learning rate of the generalization performance gap of Theorem 9 is of order $\frac{\log n \log^3(\frac{1}{\delta})}{n}$, which is still faster than the existing rates in the related work. Furthermore, please refer to Table I in Appendix A, available in the online supplemental material, for a systematic comparison with the related work.

Remark 15. In conclusion, this paper studies two cases: the general nonconvex learning and then the PL condition. The results of the general nonconvex learning are general enough to be extended to other nonconvex settings. When deriving the fast rate, we need the PL condition. The fast rate cannot be achieved for free. PL condition is a simple condition that is sufficient to show a global linear convergence rate for gradient descent. Moreover, in terms of showing a global linear convergence rate to the optimal solution, the PL condition is weaker than most existing conditions [32]. How to relax the PL condition so that the results can be extended to more nonconvex settings is an interesting problem and worth further study.

IV. Proofs

In this section, we provide proofs of theorems in Section III.

A. Proof of Remark 1

Proof. According to the uniform smoothness condition, for any sample $z, z' \in \mathcal{Z}$ and $\mathbf{w}_1, \mathbf{w}_2 \in \mathcal{W}$, there holds

$$\|\nabla f(\mathbf{w}_1; z, z') - \nabla f(\mathbf{w}_2; z, z')\| \le \beta\|\mathbf{w}_1 - \mathbf{w}_2\|.$$

Then, for any unit vector $\mathbf{u} \in B(0, 1)$, we have

$$\begin{aligned}&|\mathbf{u}^T(\nabla f(\mathbf{w}_1; z, z') - \nabla f(\mathbf{w}_2; z, z'))|\\ &\le \|\mathbf{u}\|\|\nabla f(\mathbf{w}_1; z, z') - \nabla f(\mathbf{w}_2; z, z')\| \le \beta\|\mathbf{w}_1 - \mathbf{w}_2\|,\end{aligned}$$

IEEE Proof

which implies

$$\frac{|\mathbf{u}^T(\nabla f(\mathbf{w}_1;z,z') - \nabla f(\mathbf{w}_2;z,z'))|}{\beta\|\mathbf{w}_1 - \mathbf{w}_2\|} \leq 1.$$

Then we get

$$\mathbb{E}\left\{\exp\left(\frac{\ln 2|\mathbf{u}^T(\nabla f(\mathbf{w}_1;z,z') - \nabla f(\mathbf{w}_2;z,z'))|}{\beta\|\mathbf{w}_1 - \mathbf{w}_2\|}\right)\right\} \leq 2.$$

So we obtain that $\frac{\nabla f(\mathbf{w}_1;z,z') - \nabla f(\mathbf{w}_2;z,z')}{\|\mathbf{w}_1 - \mathbf{w}_2\|}$ is a $\frac{\beta}{\ln 2}$-sub-exponential random vector, for all $\mathbf{w}_1, \mathbf{w}_2 \in \mathcal{W}$.

Furthermore, when Assumption 1 holds, according to Jensen's inequality, we can derive that

$$\exp\left\{\mathbb{E}\left(\frac{|\mathbf{u}^T(\nabla f(\mathbf{w}_1;z,z') - \nabla f(\mathbf{w}_2;z,z'))|}{\beta\|\mathbf{w}_1 - \mathbf{w}_2\|}\right)\right\} \leq 2,$$

which means

$$\begin{aligned}\mathbb{E}\|\nabla f(\mathbf{w}_1;z,z') - \nabla f(\mathbf{w}_2;z,z')\| &\leq (\ln 2)\beta\|\mathbf{w}_1 - \mathbf{w}_2\| \\ &\leq \beta\|\mathbf{w}_1 - \mathbf{w}_2\|.\end{aligned}$$

Further by Jensen's inequality, we obtain

$$\|\nabla F(\mathbf{w}_1) - \nabla F(\mathbf{w}_2)\| \leq \beta\|\mathbf{w}_1 - \mathbf{w}_2\|.$$

The proof is complete. □

B. *Proof of Theorem 1*

The proof is inspired by the recent breakthrough work [79]. To prove Theorem 1, we need many preliminaries on generic chaining and two more general forms of the Bernstein inequality of pairwise learning. Considering the length limit, we leave the introduction of this part to Appendix B, available in the online supplemental material.

Proof. We define $\mathcal{V} = \{\mathbf{v} \in \mathbb{R}^d : \|\mathbf{v}\| \leq \max\{R_1, \frac{1}{n}\}\}$. For all $(\mathbf{w}, \mathbf{v}) \in \mathcal{W} \times \mathcal{V}$, let $g_{(\mathbf{w},\mathbf{v})} = (\nabla f(\mathbf{w};z,z') - \nabla f(\mathbf{w}^*;z,z'))^T\mathbf{v}$. Also, for any $(\mathbf{w}_1, \mathbf{v}_1)$ and $(\mathbf{w}_2, \mathbf{v}_2) \in \mathcal{W} \times \mathcal{V}$, we define the following norm on the product space $\mathcal{W} \times \mathcal{V}$,

$$\|(\mathbf{w}_1, \mathbf{v}_1) - (\mathbf{w}_2, \mathbf{v}_2)\|_{\mathcal{W}\times\mathcal{V}} = (\|\mathbf{w}_1 - \mathbf{w}_2\|^2 + \|\mathbf{v}_1 - \mathbf{v}_2\|^2)^{\frac{1}{2}}.$$

Define a ball $B(\sqrt{r}) = \{(\mathbf{w}, \mathbf{v}) \in \mathcal{W} \times \mathcal{V} : \|\mathbf{w} - \mathbf{w}^*\|^2 + \|\mathbf{v}\|^2 \leq r\}$. Given any $(\mathbf{w}_1, \mathbf{v}_1)$ and $(\mathbf{w}_2, \mathbf{v}_2) \in B(\sqrt{r})$, we make the following decomposition

$$\begin{aligned}&g_{(\mathbf{w}_1,\mathbf{v}_1)}(z,z') - g_{(\mathbf{w}_2,\mathbf{v}_2)}(z,z') \\ &= (\nabla f(\mathbf{w}_1;z,z') - \nabla f(\mathbf{w}^*;z,z'))^T\mathbf{v}_1 \\ &\quad - (\nabla f(\mathbf{w}_2;z,z') - \nabla f(\mathbf{w}^*;z,z'))^T\mathbf{v}_2 \\ &= (\nabla f(\mathbf{w}_1;z,z') - \nabla f(\mathbf{w}^*;z,z'))^T(\mathbf{v}_1 - \mathbf{v}_2) \\ &\quad + (\nabla f(\mathbf{w}_1;z,z') - \nabla f(\mathbf{w}_2;z,z'))^T\mathbf{v}_2.\end{aligned}$$

Since $(\mathbf{w}_1, \mathbf{v}_1)$ and $(\mathbf{w}_2, \mathbf{v}_2) \in B(\sqrt{r})$, there holds that

$$\begin{aligned}\|\mathbf{w}_1 - \mathbf{w}^*\|\|\mathbf{v}_1 - \mathbf{v}_2\| &\leq \sqrt{r}\|\mathbf{v}_1 - \mathbf{v}_2\| \\ &\leq \sqrt{r}\|(\mathbf{w}_1, \mathbf{v}_1) - (\mathbf{w}_2, \mathbf{v}_2)\|_{\mathcal{W}\times\mathcal{V}}.\end{aligned} \tag{9}$$

And, according to Assumption 1, we know that $\frac{\nabla f(\mathbf{w}_1,z,z') - \nabla f(\mathbf{w}_2,z,z')}{\|\mathbf{w}_1 - \mathbf{w}_2\|}$ is a γ-sub-exponential random vector for all $\mathbf{w}_1, \mathbf{w}_2 \in \mathcal{W}$, which means that

$$\mathbb{E}\left\{\exp\left(\frac{(\nabla f(\mathbf{w}_1;z,z') - \nabla f(\mathbf{w}^*;z,z'))^T(\mathbf{v}_1 - \mathbf{v}_2)}{\gamma\|\mathbf{w}_1 - \mathbf{w}^*\|\|\mathbf{v}_1 - \mathbf{v}_2\|}\right)\right\} \leq 2. \tag{10}$$

Now, combined with (10) and (9), and according to Definition 1 of Appendix B, available in the online supplemental material, we know $(\nabla f(\mathbf{w}_1;z,z') - \nabla f(\mathbf{w}^*;z,z'))^T(\mathbf{v}_1 - \mathbf{v}_2)$ is $\gamma\sqrt{r}\|(\mathbf{w}_1, \mathbf{v}_1) - (\mathbf{w}_2, \mathbf{v}_2)\|_{\mathcal{W}\times\mathcal{V}}$-sub-exponential. Similarly, we can derive that

$$\begin{aligned}\|\mathbf{w}_1 - \mathbf{w}_2\|\|\mathbf{v}_2\| &\leq \sqrt{r}\|\mathbf{w}_1 - \mathbf{w}_2\| \\ &\leq \sqrt{r}\|(\mathbf{w}_1, \mathbf{v}_1) - (\mathbf{w}_2, \mathbf{v}_2)\|_{\mathcal{W}\times\mathcal{V}}.\end{aligned}$$

Also, there holds that

$$\mathbb{E}\left\{\exp\left(\frac{(\nabla f(\mathbf{w}_1;z,z') - \nabla f(\mathbf{w}_2;z,z'))^T(\mathbf{v}_2)}{\gamma\|\mathbf{w}_1 - \mathbf{w}_2\|\|\mathbf{v}_2\|}\right)\right\} \leq 2.$$

Thus, we know $(\nabla f(\mathbf{w}_1;z,z') - \nabla f(\mathbf{w}_2;z,z'))^T\mathbf{v}_2$ is also $\gamma\sqrt{r}\|(\mathbf{w}_1, \mathbf{v}_1) - (\mathbf{w}_2, \mathbf{v}_2)\|_{\mathcal{W}\times\mathcal{V}}$-sub-exponential.

Till here, for any $(\mathbf{w}_1, \mathbf{v}_1)$ and $(\mathbf{w}_2, \mathbf{v}_2) \in B(\sqrt{r})$, we obtain

$$\begin{aligned}&\mathbb{E}\left\{\exp\left(\frac{g_{(\mathbf{w}_1,\mathbf{v}_1)}(z,z') - g_{(\mathbf{w}_2,\mathbf{v}_2)}(z,z')}{2\gamma\sqrt{r}\|(\mathbf{w}_1, \mathbf{v}_1) - (\mathbf{w}_2, \mathbf{v}_2)\|_{\mathcal{W}\times\mathcal{V}}}\right)\right\} \\ &\leq \mathbb{E}\left\{\frac{1}{2}\exp\left(\frac{(\nabla f(\mathbf{w}_1;z,z') - \nabla f(\mathbf{w}^*;z,z'))^T(\mathbf{v}_1 - \mathbf{v}_2)}{\gamma\sqrt{r}\|(\mathbf{w}_1, \mathbf{v}_1) - (\mathbf{w}_2, \mathbf{v}_2)\|_{\mathcal{W}\times\mathcal{V}}}\right)\right\} \\ &+ \mathbb{E}\left\{\frac{1}{2}\exp\left(\frac{(\nabla f(\mathbf{w}_1;z,z') - \nabla f(\mathbf{w}_2;z,z'))^T(\mathbf{v}_2)}{\gamma\sqrt{r}\|(\mathbf{w}_1, \mathbf{v}_1) - (\mathbf{w}_2, \mathbf{v}_2)\|_{\mathcal{W}\times\mathcal{V}}}\right)\right\} \leq 2,\end{aligned} \tag{11}$$

where the first inequality follows from Jensen's inequality. And (11) means that $g_{(\mathbf{w}_1,\mathbf{v}_1)}(z,z') - g_{(\mathbf{w}_2,\mathbf{v}_2)}(z,z')$ is a $2\gamma\sqrt{r}\|(\mathbf{w}_1, \mathbf{v}_1) - (\mathbf{w}_2, \mathbf{v}_2)\|_{\mathcal{W}\times\mathcal{V}}$-sub-exponential random variable, that is

$$\begin{aligned}&\|g_{(\mathbf{w}_1,\mathbf{v}_1)}(z,z') - g_{(\mathbf{w}_2,\mathbf{v}_2)}(z,z')\|_{Orlicz-1} \\ &\leq 2\gamma\sqrt{r}\|(\mathbf{w}_1, \mathbf{v}_1) - (\mathbf{w}_2, \mathbf{v}_2)\|_{\mathcal{W}\times\mathcal{V}}.\end{aligned} \tag{12}$$

Then, the next step is to apply the Bernstein inequality of pairwise learning (Lemma 10 of Appendix B, available in the online supplemental material) to $g_{(\mathbf{w}_1,\mathbf{v}_1)}(z,z') - g_{(\mathbf{w}_2,\mathbf{v}_2)}(z,z')$. From (12), we know that the Bernstein parameters of sub-exponential $g_{(\mathbf{w}_1,\mathbf{v}_1)}(z,z') - g_{(\mathbf{w}_2,\mathbf{v}_2)}(z,z')$ are $2\gamma\sqrt{r}\|(\mathbf{w}_1, \mathbf{v}_1) - (\mathbf{w}_2, \mathbf{v}_2)\|_{\mathcal{W}\times\mathcal{V}}$ (see Lemma 13 of Appendix B, available in the online supplemental material). Now, we can derive that

$$\begin{aligned}&Pr\Bigg(\left|(P - P_n)[g_{(\mathbf{w}_1,\mathbf{v}_1)}(z,z') - g_{(\mathbf{w}_2,\mathbf{v}_2)}(z,z')]\right| \\ &\geq 2\gamma\sqrt{r}\|(\mathbf{w}_1, \mathbf{v}_1) - (\mathbf{w}_2, \mathbf{v}_2)\|_{\mathcal{W}\times\mathcal{V}}\sqrt{\frac{2u}{\lfloor\frac{n}{2}\rfloor}} \\ &\quad + \frac{2\gamma\sqrt{r}\|(\mathbf{w}_1, \mathbf{v}_1) - (\mathbf{w}_2, \mathbf{v}_2)\|_{\mathcal{W}\times\mathcal{V}}}{\lfloor\frac{n}{2}\rfloor}u\Bigg) \leq 2e^{-u},\end{aligned} \tag{13}$$

IEEE Proof

where $\lfloor \frac{n}{2} \rfloor$ is the largest integer no greater than $\frac{n}{2}$ and "Pr" means probability. According to Definition 3 of Appendix B, available in the online supplemental material, (13) implies that the process $(P-P_n)[g_{(\mathbf{w},\mathbf{v})}(z,z')]$ has a mixed sub-Gaussian-sub-exponential increments w.r.t. the metric pair $\left(\frac{2\gamma\sqrt{r}\|\cdot\|_{\mathcal{W}\times\mathcal{V}}}{\lfloor \frac{n}{2} \rfloor}, 2\gamma\|\cdot\|_{\mathcal{W}\times\mathcal{V}}\sqrt{\frac{2r}{\lfloor \frac{n}{2} \rfloor}}\right)$. Hence, from the generic chaining for a process with mixed tail increments in Lemma 7 of Appendix B, available in the online supplemental material, for all $\delta \in (0,1)$, with probability at least $1-\delta$, we have

$$\begin{aligned}
&\sup_{\|\mathbf{w}-\mathbf{w}^*\|^2+\|\mathbf{v}\|^2 \le r} |(P-P_n)[g_{(\mathbf{w},\mathbf{v})}(z,z')]| \\
&\le C\left(\gamma_2\left(B(\sqrt{r}), 2\gamma\|\cdot\|_{\mathcal{W}\times\mathcal{V}}\sqrt{\frac{2r}{\lfloor \frac{n}{2} \rfloor}}\right)\right. \\
&\left. +\gamma_1\left(B(\sqrt{r}), \frac{2\gamma\sqrt{r}\|\cdot\|_{\mathcal{W}\times\mathcal{V}}}{\lfloor \frac{n}{2} \rfloor}\right) + \gamma r\frac{\log\frac{1}{\delta}}{\lfloor \frac{n}{2} \rfloor} + \gamma r\sqrt{\frac{\log\frac{1}{\delta}}{\lfloor \frac{n}{2} \rfloor}}\right).
\end{aligned}$$

From Lemma 6 of Appendix B, available in the online supplemental material, the γ_1 functional and the γ_2 functional can be bounded by the Dudley's integral, which implies that there exists an absolute constant C such that for any $\delta \in (0,1)$, with probability at least $1-\delta$

$$\begin{aligned}
&\sup_{\|\mathbf{w}-\mathbf{w}^*\|^2+\|\mathbf{v}\|^2 \le r} |(P-P_n)[g_{(\mathbf{w},\mathbf{v})}(z,z')]| \\
&\le C\gamma r\left(\sqrt{\frac{d+\log\frac{1}{\delta}}{\lfloor \frac{n}{2} \rfloor}} + \frac{d+\log\frac{1}{\delta}}{\lfloor \frac{n}{2} \rfloor}\right),
\end{aligned} \tag{14}$$

where the inequality follows from (B.3) of [79]. Till here, the next step is to apply Lemma 5 of Appendix B, available in the online supplemental material, to (14).

We set $T(f) = \|\mathbf{w}-\mathbf{w}^*\|^2 + \|\mathbf{v}\|^2$, $\psi(r;\delta) = C\gamma r\left(\sqrt{\frac{d+\log\frac{1}{\delta}}{\lfloor \frac{n}{2} \rfloor}} + \frac{d+\log\frac{1}{\delta}}{\lfloor \frac{n}{2} \rfloor}\right)$. Since $\|\mathbf{w}-\mathbf{w}^*\|^2 + \|\mathbf{v}\|^2 \le R_1^2 + R_1^2 + \frac{1}{n^2}$, we set $R = 2R_1^2 + \frac{1}{n^2}$. And let $r_0 = \frac{2}{n^2}$. Applying Lemma 5, we obtain that for any $\delta \in (0,1)$, with probability at least $1-\delta$, for all $\mathbf{w} \in \mathcal{W}$ and $\mathbf{v} \in \mathcal{V}$,

$$\begin{aligned}
&(P-P_n)[g_{(\mathbf{w},\mathbf{v})}(z,z')] \\
&= (P-P_n)\left[(\nabla f(\mathbf{w};z,z') - \nabla f(\mathbf{w}^*;z,z'))^T\mathbf{v}\right] \\
&\le \psi\left(\max\left\{\|\mathbf{w}-\mathbf{w}^*\|^2 + \|\mathbf{v}\|^2, \frac{2}{n^2}\right\}; \frac{\delta}{2\log_2(Rn^2)}\right) \\
&= C\gamma\max\left\{\|\mathbf{w}-\mathbf{w}^*\|^2 + \|\mathbf{v}\|^2, \frac{2}{n^2}\right\} \\
&\quad \times \left(\sqrt{\frac{d+\log\frac{2\log_2(Rn^2)}{\delta}}{\lfloor \frac{n}{2} \rfloor}} + \frac{d+\log\frac{2\log_2(Rn^2)}{\delta}}{\lfloor \frac{n}{2} \rfloor}\right).
\end{aligned} \tag{15}$$

Now, we choose $\mathbf{v}$ as $\max\left\{\|\mathbf{w}-\mathbf{w}^*\|, \frac{1}{n}\right\} \frac{(P-P_n)(\nabla f(\mathbf{w};z,z') - \nabla f(\mathbf{w}^*;z,z'))}{\|(P-P_n)(\nabla f(\mathbf{w};z,z') - \nabla f(\mathbf{w}^*;z,z'))\|}$. It is clear that $\|\mathbf{v}\| = \max\{\|\mathbf{w}-\mathbf{w}^*\|, \frac{1}{n}\} \le \max\{R_1, \frac{1}{n}\}$, which belongs to the space $\mathcal{V}$. Plugging this $\mathbf{v}$ into (15), we obtain that for any $\delta \in (0,1)$, with probability at least $1-\delta$, for all $\mathbf{w} \in \mathcal{W}$,

$$\begin{aligned}
&\|(P-P_n)(\nabla f(\mathbf{w};z,z') - \nabla f(\mathbf{w}^*;z,z'))\| \\
&\le C\gamma\max\left\{\|\mathbf{w}-\mathbf{w}^*\|, \frac{1}{n}\right\} \\
&\quad \times \left(\sqrt{\frac{d+\log\frac{2\log_2(Rn^2)}{\delta}}{\lfloor \frac{n}{2} \rfloor}} + \frac{d+\log\frac{2\log_2(Rn^2)}{\delta}}{\lfloor \frac{n}{2} \rfloor}\right) \\
&\le C\gamma\max\left\{\|\mathbf{w}-\mathbf{w}^*\|, \frac{1}{n}\right\} \\
&\quad \times \left(\sqrt{\frac{d+\log\frac{2\log_2(Rn^2)}{\delta}}{n}} + \frac{d+\log\frac{2\log_2(Rn^2)}{\delta}}{n}\right).
\end{aligned} \tag{16}$$

Since $R = 2R_1^2 + \frac{1}{n^2}$, (16) thus implies that

$$\begin{aligned}
&\|(P-P_n)(\nabla f(\mathbf{w};z,z') - \nabla f(\mathbf{w}^*;z,z'))\| \\
&\le C\gamma\max\left\{\|\mathbf{w}-\mathbf{w}^*\|, \frac{1}{n}\right\} \\
&\quad \times \left(\sqrt{\frac{d+\log\frac{4\log_2(\sqrt{2}R_1n+1)}{\delta}}{n}} + \frac{d+\log\frac{4\log_2(\sqrt{2}R_1n+1)}{\delta}}{n}\right).
\end{aligned}$$

The proof is complete. $\square$

C. Proof of Theorem 2

Proof. From Theorem 1, we have

$$\begin{aligned}
&\|\nabla F(\mathbf{w}) - \nabla F_S(\mathbf{w})\| \\
&\le \|\nabla F(\mathbf{w}^*) - \nabla F_S(\mathbf{w}^*)\| + C\gamma\max\left\{\|\mathbf{w}-\mathbf{w}^*\|, \frac{1}{n}\right\} \\
&\quad \times \left(\sqrt{\frac{d+\log\frac{4\log_2(\sqrt{2}R_1n+1)}{\delta}}{n}} + \frac{d+\log\frac{4\log_2(\sqrt{2}R_1n+1)}{\delta}}{n}\right),
\end{aligned} \tag{17}$$

where the inequality follows from that $\|\nabla F(\mathbf{w}) - \nabla F_S(\mathbf{w})\| - \|\nabla F(\mathbf{w}^*) - \nabla F_S(\mathbf{w}^*)\| \le \|(\nabla F(\mathbf{w}) - \nabla F_S(\mathbf{w})) - (\nabla F(\mathbf{w}^*) - \nabla F_S(\mathbf{w}^*))\|$. Denote $\xi_{n,R_1,d,\delta} = \sqrt{\frac{d+\log\frac{4\log_2(\sqrt{2}R_1n+1)}{\delta}}{n}} + \frac{d+\log\frac{4\log_2(\sqrt{2}R_1n+1)}{\delta}}{n}$. We are now to prove the bound of $\|\nabla F(\mathbf{w}^*) - \nabla F_S(\mathbf{w}^*)\|$.

It is clear that $\nabla F(\mathbf{w}^*) = 0$. From Lemma 12 of Appendix B, available in the online supplemental material, and Assumption 2, we have the following inequality for any $\delta > 0$, with probability at least $1-\delta$

$$\begin{aligned}
&\|\nabla F(\mathbf{w}^*) - \nabla F_S(\mathbf{w}^*)\| \\
&\le \sqrt{\frac{2\mathbb{E}[\|\nabla f(\mathbf{w}^*;z,z')\|^2]\log\frac{2}{\delta}}{\lfloor \frac{n}{2} \rfloor}} + \frac{D_*\log\frac{2}{\delta}}{\lfloor \frac{n}{2} \rfloor}.
\end{aligned} \tag{18}$$

IEEE Proof

Plugging (18) into (17), we obtain that for any $\delta > 0$, with probability at least $1-\delta$

$$\begin{aligned}\|\nabla F(\mathbf{w}) - \nabla F_S(\mathbf{w})\| &\leq C\gamma \max\left\{\|\mathbf{w}-\mathbf{w}^*\|, \frac{1}{n}\right\}\xi_{n,R_1,d,\frac{\delta}{2}} \\ &+ \sqrt{\frac{2\mathbb{E}[\|\nabla f(\mathbf{w}^*;z,z')\|^2]\log\frac{4}{\delta}}{\lfloor\frac{n}{2}\rfloor}} + \frac{D_*\log\frac{4}{\delta}}{\lfloor\frac{n}{2}\rfloor} \\ &\leq \sqrt{\frac{8\mathbb{E}[\|\nabla f(\mathbf{w}^*;z,z')\|^2]\log\frac{4}{\delta}}{n}} + \frac{4D_*\log\frac{4}{\delta}}{n} \\ &+ C\gamma\max\left\{\|\mathbf{w}-\mathbf{w}^*\|, \frac{1}{n}\right\}\xi_{n,R_1,d,\frac{\delta}{2}}.\end{aligned}$$

The proof is complete. □

D. *Proof of Theorem 3*

Proof. Denote $\xi_{n,R_1,d,\delta} = \sqrt{\frac{d+\log\frac{8\log_2(\sqrt{2}R_1n+1)}{\delta}}{n}} + \frac{d+\log\frac{8\log_2(\sqrt{2}R_1n+1)}{\delta}}{n}$. According to Theorem 2, for any $\delta\in(0,1)$, with probability at least $1-\delta$, we have the following inequality

$$\begin{aligned}\|\nabla F(\mathbf{w}) - \nabla F_S(\mathbf{w})\| &\leq \sqrt{\frac{8\mathbb{E}[\|\nabla f(\mathbf{w}^*;z,z')\|^2]\log\frac{4}{\delta}}{n}} \\ &+ \frac{4D_*\log\frac{4}{\delta}}{n} + C\gamma\max\left\{\|\mathbf{w}-\mathbf{w}^*\|, \frac{1}{n}\right\}\xi_{n,R_1,d,\delta}. \quad (19)\end{aligned}$$

This implies that

$$\begin{aligned}\|\nabla F(\mathbf{w})\| - \|\nabla F_S(\mathbf{w})\| &\leq C\gamma\max\left\{\|\mathbf{w}-\mathbf{w}^*\|, \frac{1}{n}\right\}\xi_{n,R_1,d,\delta} \\ &+ \frac{4D_*\log\frac{4}{\delta}}{n} + \sqrt{\frac{8\mathbb{E}[\|\nabla f(\mathbf{w}^*;z,z')\|^2]\log\frac{4}{\delta}}{n}}.\end{aligned}$$

According to Remark 1, Assumption 1 implies the population risk $F(\mathbf{w})$ is γ-smooth. Moreover, when $F(\mathbf{w})$ is smooth and satisfies the PL condition, there holds the following error bound property (refer to Theorem 2 in [32])

$$\|\nabla F(\mathbf{w})\| \geq \mu\|\mathbf{w}-\mathbf{w}^*\|.$$

Thus, we have

$$\begin{aligned}\mu\|\mathbf{w}-\mathbf{w}^*\| &\leq \|\nabla F(\mathbf{w})\| \leq \|\nabla F_S(\mathbf{w})\| \\ &+ \sqrt{\frac{8\mathbb{E}[\|\nabla f(\mathbf{w}^*;z,z')\|^2]\log\frac{4}{\delta}}{n}} + \frac{4D_*\log\frac{4}{\delta}}{n} \\ &+ C\gamma\max\left\{\|\mathbf{w}-\mathbf{w}^*\|, \frac{1}{n}\right\}\xi_{n,R_1,d,\delta}. \quad (20)\end{aligned}$$

And according to [63], there holds the following property for γ-smooth functions f:

$$\frac{1}{2\gamma}\|\nabla f(\mathbf{w})\|^2 \leq f(\mathbf{w}) - \inf_{\mathbf{w}\in\mathcal{W}} f(\mathbf{w}). \quad (21)$$

Thus we have

$$\frac{1}{2\gamma}\|\nabla F(\mathbf{w})\|^2 \leq F(\mathbf{w}) - F(\mathbf{w}^*) \leq \frac{\|\nabla F(\mathbf{w})\|^2}{2\mu}, \quad (22)$$

which means that $\frac{\mu}{\gamma}\leq 1$. Let $c=\max\{4C^2,1\}$. When

$$n \geq \frac{c\gamma^2\left(d+\log\frac{8\log_2(\sqrt{2}R_1n+1)}{\delta}\right)}{\mu^2},$$

we have $C\gamma\xi_{n,R_1,d,\delta}\leq\frac{\mu}{2}$, followed from the fact that $\frac{\mu}{\gamma}\leq 1$.

Plugging $C\gamma\xi_{n,R_1,d,\delta}\leq\frac{\mu}{2}$ into (20), we can derive that

$$\begin{aligned}\|\mathbf{w}-\mathbf{w}^*\| &\leq \frac{2}{\mu}\Bigg(\|\nabla F_S(\mathbf{w})\| + \frac{4D_*\log(4/\delta)}{n} \\ &+ \sqrt{\frac{8\mathbb{E}[\|\nabla f(\mathbf{w}^*;z,z')\|^2]\log(4/\delta)}{n}} + \frac{\mu}{2n}\Bigg). \quad (23)\end{aligned}$$

Then, substituting (23) into (19), we derive that for all $\mathbf{w}\in\mathcal{W}$, when $n\geq\frac{c\gamma^2\left(d+\log\frac{8\log_2(\sqrt{2}R_1n+1)}{\delta}\right)}{\mu^2}$, with probability at least $1-\delta$

$$\begin{aligned}\|\nabla F(\mathbf{w}) - \nabla F_S(\mathbf{w})\| &\leq \|\nabla F_S(\mathbf{w})\| \\ &+ \frac{\mu}{n} + 2\frac{4D_*\log(4/\delta)}{n} + 2\sqrt{\frac{8\mathbb{E}[\|\nabla f(\mathbf{w}^*;z,z')\|^2]\log(4/\delta)}{n}}.\end{aligned}$$

The proof is complete. □

E. *Proof of Theorem 4*

Proof. Plugging $\hat{\mathbf{w}}^*$ into Theorem 2, we have

$$\begin{aligned}&\|\nabla F(\hat{\mathbf{w}}^*)\| - \|\nabla F_S(\hat{\mathbf{w}}^*)\| \\ &\leq \sqrt{\frac{8\mathbb{E}[\|\nabla f(\mathbf{w}^*;z,z')\|^2]\log\frac{4}{\delta}}{n}} + \frac{4D_*\log\frac{4}{\delta}}{n} \\ &+ C\gamma\max\left\{\|\hat{\mathbf{w}}^*-\mathbf{w}^*\|, \frac{1}{n}\right\} \\ &\times\left(\sqrt{\frac{d+\log\frac{8\log_2(\sqrt{2}R_1n+1)}{\delta}}{n}} + \frac{d+\log\frac{8\log_2(\sqrt{2}R_1n+1)}{\delta}}{n}\right).\end{aligned}$$

Since $\hat{\mathbf{w}}^*$ is the ERM of F_S, there holds that $\nabla F_S(\hat{\mathbf{w}}^*)=0$. Thus, we can derive that

$$\begin{aligned}\|\nabla F(\hat{\mathbf{w}}^*)\| &\leq \sqrt{\frac{8\mathbb{E}[\|\nabla f(\mathbf{w}^*;z,z')\|^2]\log\frac{4}{\delta}}{n}} \\ &+ \frac{4D_*\log\frac{4}{\delta}}{n} + C\gamma\left(R_1+\frac{1}{n}\right) \\ &\times\left(\sqrt{\frac{d+\log\frac{8\log_2(\sqrt{2}R_1n+1)}{\delta}}{n}} + \frac{d+\log\frac{8\log_2(\sqrt{2}R_1n+1)}{\delta}}{n}\right).\end{aligned}$$

The proof is complete. □

IEEE Proof

F. Proof of Theorem 5

Proof. Since $F(\mathbf{w})$ satisfies the PL condition with parameter μ, we have

$$F(\mathbf{w}) - F(\mathbf{w}^*) \leq \frac{\|\nabla F(\mathbf{w})\|^2}{2\mu}, \quad \forall \mathbf{w} \in \mathcal{W}.$$

Therefore, to bound the excess risk $F(\hat{\mathbf{w}}^*) - F(\mathbf{w}^*)$, we need to bound the term $\|\nabla F(\hat{\mathbf{w}}^*)\|^2$. Plugging $\hat{\mathbf{w}}^*$ into Theorem 3 and (6), for any $\delta > 0$, when $n \geq \frac{c\gamma^2\left(d+\log\frac{8\log_2(\sqrt{2}R_1 n+1)}{\delta}\right)}{\mu^2}$, with probability at least $1-\delta$,

$$\|\nabla F(\hat{\mathbf{w}}^*)\| \leq 2\|\nabla F_S(\hat{\mathbf{w}}^*)\| + \frac{\mu}{n} + \frac{8D_* \log(4/\delta)}{n} + 4\sqrt{\frac{2\mathbb{E}[\|\nabla f(\mathbf{w}^*; z, z')\|^2]\log(4/\delta)}{n}},$$

Since $\nabla F_S(\hat{\mathbf{w}}^*) = 0$, we have $\|\nabla F_S(\hat{\mathbf{w}}^*)\| = 0$. We can derive that

$$F(\hat{\mathbf{w}}^*) - F(\mathbf{w}^*) \leq \frac{12D_*^2\log^2(4/\delta)}{\mu n^2} + \frac{6\mathbb{E}[\|\nabla f(\mathbf{w}^*; z, z')\|^2\log(4/\delta)}{\mu n} + \frac{2\mu}{n^2}.$$

The proof is complete. □

G. Proof of Theorem 6

Proof. According to Assumption 4 and $\eta_t \leq 1/\beta$, we can derive that

$$\begin{aligned} &F_S(\mathbf{w}_{t+1}) - F_S(\mathbf{w}_t) \\ &\leq \langle \mathbf{w}_{t+1} - \mathbf{w}_t, \nabla F_S(\mathbf{w}_t)\rangle + \frac{\beta}{2}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2 \\ &= -\eta_t\|\nabla F_S(\mathbf{w}_t)\|^2 + \frac{\beta}{2}\eta_t^2\|\nabla F_S(\mathbf{w}_t)\|^2 \\ &= \left(\frac{\beta}{2}\eta_t^2 - \eta_t\right)\|\nabla F_S(\mathbf{w}_t)\|^2 \\ &\leq -\frac{1}{2}\eta_t\|\nabla F_S(\mathbf{w}_t)\|^2, \end{aligned} \tag{24}$$

which implies that

$$\eta_t\|\nabla F_S(\mathbf{w}_t)\|^2 \leq -2(F_S(\mathbf{w}_{t+1}) - F_S(\mathbf{w}_t)).$$

Take a summation from $t = 1$ to T, we have

$$\sum_{t=1}^T \eta_t\|\nabla F_S(\mathbf{w}_t)\|^2 \leq 2(F_S(\mathbf{w}_1) - F_S(\mathbf{w}_{T+1})). \tag{25}$$

Furthermore, we derive that

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

which implies that with probability at least $1-\delta$

$$\begin{aligned} &\frac{1}{\sum_{t=1}^T \eta_t}\sum_{t=1}^T \eta_t\|\nabla F(\mathbf{w}_t)\|^2 \\ &\leq 2\max_{t=1,\ldots,T}\|\nabla F(\mathbf{w}_t) - \nabla F_S(\mathbf{w}_t)\|^2 + \left(\sum_{t=1}^T \eta_t\right)^{-1}\mathcal{O}(1) \\ &\leq \left(\sum_{t=1}^T \eta_t\right)^{-1}\mathcal{O}(1) + 2\max_{t=1,\ldots,T}\Bigg[C\beta\max\left\{\|\mathbf{w}_t - \mathbf{w}^*\|, \frac{1}{n}\right\} \\ &\times\left(\sqrt{\frac{d+\log\frac{4\log_2(\sqrt{2}R_1 n+1)}{\delta}}{n}} + \frac{d+\log\frac{4\log_2(\sqrt{2}R_1 n+1)}{\delta}}{n}\right) \\ &+ \frac{4D_*\log\frac{4}{\delta}}{n} + \sqrt{\frac{8\mathbb{E}\left[\|\nabla f(\mathbf{w}^*; z, z')\|^2\right]\log\frac{4}{\delta}}{n}}\Bigg]^2, \end{aligned} \tag{26}$$

where $\mathcal{O}(1)$ in the first inequality is due to (25) and the nonnegative property of f, and where the second inequality holds since Theorem 2 and that Assumption 4 implies Assumption 1 (see Remark 1).

We now to prove the bound of $\|\mathbf{w}_t - \mathbf{w}^*\|$. Since $\mathbf{w}_1 = 0$ and $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t\nabla F_S(\mathbf{w}_t)$, we have $\mathbf{w}_{t+1} = \sum_{k=1}^t -\eta_k\nabla F_S(\mathbf{w}_k)$. And according to Schwarz's inequality, we have

$$\begin{aligned} &\left\|\sum_{k=1}^t \eta_k\nabla F_S(\mathbf{w}_k)\right\|^2 \leq \left(\sum_{k=1}^t \eta_k\|\nabla F_S(\mathbf{w}_k)\|\right)^2 \\ &\leq \left(\sum_{k=1}^t \eta_k\right)\left(\sum_{k=1}^t \eta_k\|\nabla F_S(\mathbf{w}_k)\|^2\right) \leq \left(\sum_{k=1}^t \eta_k\right)\mathcal{O}(1). \end{aligned}$$

Then we have

$$\begin{aligned} &\|\mathbf{w}_{t+1} - \mathbf{w}^*\| \leq \|\mathbf{w}_{t+1}\| + \|\mathbf{w}^*\| \\ &= \left\|\sum_{k=1}^t \eta_k\nabla F_S(\mathbf{w}_k)\right\| + \|\mathbf{w}^*\| = \mathcal{O}\left(\left(\sum_{k=1}^t \eta_k\right)^{\frac{1}{2}}\right). \end{aligned}$$

If $\theta \in (0,1)$, then $\sum_{k=1}^t k^{-\theta} \leq t^{1-\theta}/(1-\theta)$. Thus, we have the following result uniformly for all $t = 1, \ldots, T$

$$\|\mathbf{w}_{t+1} - \mathbf{w}^*\| = \mathcal{O}\left(T^{\frac{1-\theta}{2}}\right) \quad \text{if } \theta \in (0,1). \tag{27}$$

Therefore, plugging (27) into (26), we get that with probability at least $1-\delta$

$$\begin{aligned} &\frac{1}{\sum_{t=1}^T \eta_t}\sum_{t=1}^T \eta_t\|\nabla F(\mathbf{w}_t)\|^2 \leq \left(\sum_{t=1}^T \eta_t\right)^{-1}\mathcal{O}(1) \\ &+ \mathcal{O}\Bigg(\frac{d+\log\frac{4\log_2(\sqrt{2}R_1 n+1)}{\delta}}{n}T^{1-\theta} \\ &+ \frac{\log^2\frac{4}{\delta}}{n^2} + \frac{\mathbb{E}[\|\nabla f(\mathbf{w}^*; z, z')\|^2]\log\frac{4}{\delta}}{n}\Bigg) \end{aligned}$$

$$\leq \mathcal{O}\bigg(\frac{1}{T^{1-\theta}}\bigg) + \mathcal{O}\bigg(\frac{d + \log\frac{\log n}{\delta}}{n}T^{1-\theta} + \frac{\log^2\frac{4}{\delta}}{n^2} + \frac{\mathbb{E}[\|\nabla f(\mathbf{w}^*; z, z')\|^2]\log\frac{4}{\delta}}{n}\bigg). \tag{28}$$

If we choose $T \asymp (nd^{-1})^{\frac{1}{2(1-\theta)}}$, then we derive that

$$\frac{1}{\sum_{t=1}^T \eta_t}\sum_{t=1}^T \eta_t\|\nabla F(\mathbf{w}_t)\|^2 \leq \mathcal{O}\left(\frac{d^{\frac{1}{2}} + d^{-\frac{1}{2}}\log\frac{\log n}{\delta}}{n^{\frac{1}{2}}} + \frac{\log^2\frac{4}{\delta}}{n^2} + \frac{\mathbb{E}[\|\nabla f(\mathbf{w}^*; z, z')\|^2]\log\frac{4}{\delta}}{n}\right)$$
$$\leq \mathcal{O}\bigg(\frac{d^{\frac{1}{2}} + d^{-\frac{1}{2}}\log\frac{\log n}{\delta}}{n^{\frac{1}{2}}}\bigg),$$

where the second inequality holds because $\frac{d^{\frac{1}{2}} + d^{-\frac{1}{2}}\log\frac{\log n}{\delta}}{n^{\frac{1}{2}}}$ is the dominant term. The proof is complete. □

H. *Proof of Theorem 7*

Proof. By (24) and the PL condition of F_S, we can prove that

$$F_S(\mathbf{w}_{t+1}) - F_S(\mathbf{w}_t) \leq -\frac{1}{2}\eta_t\|\nabla F_S(\mathbf{w}_t)\|^2$$
$$\leq -\mu\eta_t(F_S(\mathbf{w}_t) - F_S(\hat{\mathbf{w}}^*)),$$

which implies that

$$F_S(\mathbf{w}_{t+1}) - F_S(\hat{\mathbf{w}}^*) \leq (1 - \mu\eta_t)(F_S(\mathbf{w}_t) - F_S(\hat{\mathbf{w}}^*)).$$

If $\eta_t \leq \frac{1}{\beta}$, then $0 < 1 - \mu\eta_t < 1$ since $\frac{\mu}{\beta} \leq 1$ according to (22). Taking over T iterations, we get

$$F_S(\mathbf{w}_{T+1}) - F_S(\hat{\mathbf{w}}^*) \leq (1 - \mu\eta_t)^T(F_S(\mathbf{w}_1) - F_S(\hat{\mathbf{w}}^*)). \tag{29}$$

If $\eta_t = 1/\beta$, combined with (29), the smoothness of F_S (see (21)), and the nonnegative property of f, it can be derived that

$$\|\nabla F_S(\mathbf{w}_{T+1})\|^2 = \mathcal{O}\left((1 - \frac{\mu}{\beta})^T\right). \tag{30}$$

Furthermore, since F satisfies the PL assumption with parameter μ, we have

$$F(\mathbf{w}_{T+1}) - F(\mathbf{w}^*) \leq \frac{\|\nabla F(\mathbf{w}_{T+1})\|^2}{2\mu}, \quad \forall \mathbf{w} \in \mathcal{W}. \tag{31}$$

So to bound $F(\mathbf{w}_{T+1}) - F(\mathbf{w}^*)$, we need to bound the term $\|\nabla F(\mathbf{w}_{T+1})\|^2$. And there holds

$$\|\nabla F(\mathbf{w}_{T+1})\|^2$$
$$\leq 2\,\|\nabla F(\mathbf{w}_{T+1}) - \nabla F_S(\mathbf{w}_{T+1})\|^2 + 2\|\nabla F_S(\mathbf{w}_{T+1})\|^2. \tag{32}$$

For the first term $\|\nabla F(\mathbf{w}_{T+1}) - \nabla F_S(\mathbf{w}_{T+1})\|^2$, from Theorem 3, for all $\mathbf{w} \in \mathcal{W}$, when $n \geq \frac{c\beta^2\left(d + \log\frac{8\log_2(\sqrt{2}R_1 n+1)}{\delta}\right)}{\mu^2}$, with probability at least $1 - \delta$, there holds

$$\|\nabla F(\mathbf{w}_{T+1}) - \nabla F_S(\mathbf{w}_{T+1})\| \leq \|\nabla F_S(\mathbf{w}_{T+1})\| + \frac{\mu}{n} + \frac{8D_*\log(4/\delta)}{n} + 4\sqrt{\frac{2\mathbb{E}[\|\nabla f(\mathbf{w}^*; z, z')\|^2]\log(4/\delta)}{n}}. \tag{33}$$

Therefore, plugging (33), (30) and (32) into (31), we derive with probability at least $1 - \delta$

$$F(\mathbf{w}_{T+1}) - F(\mathbf{w}^*) \leq \mathcal{O}\left((1 - \frac{\mu}{\beta})^T\right) + \mathcal{O}\left(\frac{\log^2(1/\delta)}{n^2} + \frac{\mathbb{E}[\|\nabla f(\mathbf{w}^*; z, z')\|^2]\log(1/\delta)}{n}\right). \tag{34}$$

When f is nonnegative and β-smooth, from Lemma 4.1 of [72], we have

$$\|\nabla f(\mathbf{w}^*; z, z')\|^2 \leq 4\beta f(\mathbf{w}^*; z, z'),$$

thus we have

$$\mathbb{E}[\|\nabla f(\mathbf{w}^*; z, z')\|^2] \leq 4\beta\mathbb{E}f(\mathbf{w}^*; z, z') = 4\beta F(\mathbf{w}^*). \tag{35}$$

By (35), (34) implies

$$F(\mathbf{w}_{T+1}) - F(\mathbf{w}^*) \leq \mathcal{O}\left((1 - \frac{\mu}{\beta})^T\right) + \mathcal{O}\left(\frac{\log^2(1/\delta)}{n^2} + \frac{F(\mathbf{w}^*)\log(1/\delta)}{n}\right).$$

The proof is complete. □

I. *Proof of Theorem 8*

We first introduce some necessary lemmas on the empirical risk. Note that the proof of the following lemmas of SGD (Algorithm 2) for pairwise learning is the same as that for pointwise learning.

Lemma 1. [44] Let $\{\mathbf{w}_t\}_t$ be the sequence produced by Algorithm 2 with $\eta_t \leq \frac{1}{2\beta}$ for all $t \in \mathbb{N}$. Suppose Assumptions 4 and 5 hold. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$, there holds that

$$\sum_{k=1}^t \eta_k\|\nabla F_S(\mathbf{w}_k)\|^2 = \mathcal{O}\left(\log\frac{1}{\delta} + \sum_{k=1}^t \eta_k^2\right).$$

Lemma 2. [44] Let $\{\mathbf{w}_t\}_t$ be the sequence produced by Algorithm 2 with $\eta_t \leq \frac{1}{2\beta}$ for all $t \in \mathbb{N}$. Suppose Assumptions 4 and 5 hold. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$, there holds uniformly for all $t = 1, .., T$

$$\|\mathbf{w}_{t+1} - \mathbf{w}^*\|$$
$$= \mathcal{O}\left(\left(\sum_{k=1}^T \eta_k^2\right)^{1/2} + 1\right)\left(\left(\sum_{k=1}^t \eta_k\right)^{1/2} + 1\right)\log\left(\frac{1}{\delta}\right).$$

Lemma 3. [44] Let $\{\mathbf{w}_t\}_t$ be the sequence produced by Algorithm 2 with $\eta_t = \frac{2}{\mu(t+t_0)}$ such that $t_0 \geq \max\{\frac{4\beta}{\mu}, 1\}$ for all $t \in \mathbb{N}$. Suppose Assumptions 4 and 5 hold, and suppose F_S satisfies Assumption 3 with parameter 2μ. Then, for any $\delta > 0$, with probability at least $1 - \delta$, there holds that

$$F_S(\mathbf{w}_{T+1}) - F_S(\hat{\mathbf{w}}^*) = \mathcal{O}\left(\frac{\log(T)\log^3(1/\delta)}{T}\right).$$

IEEE Proof

Lemma 4. [39] Let e be the base of the natural logarithm. There holds the following elementary inequalities.

a) If $\theta \in (0,1)$, then $\sum_{k=1}^{t} k^{-\theta} \leq t^{1-\theta}/(1-\theta)$;

b) If $\theta = 1$, then $\sum_{k=1}^{t} k^{-\theta} \leq \log(et)$;

c) If $\theta > 1$, then $\sum_{k=1}^{t} k^{-\theta} \leq \frac{\theta}{\theta-1}$.

Now, we begin to prove Theorem 8.

Proof. Similar to the proof of Theorem 6. First, we have

$$
\begin{aligned}
&\sum_{t=1}^{T} \eta_t \|\nabla F(\mathbf{w}_t)\|^2 \\
&\leq 2\sum_{t=1}^{T} \eta_t \|\nabla F(\mathbf{w}_t) - \nabla F_S(\mathbf{w}_t)\|^2 + 2\sum_{t=1}^{T} \eta_t \|\nabla F_S(\mathbf{w}_t)\|^2 \\
&\leq 2\sum_{t=1}^{T} \eta_t \max_{t=1,\ldots,T} \|\nabla F(\mathbf{w}_t) - \nabla F_S(\mathbf{w}_t)\|^2 \\
&\quad + \mathcal{O}\left(\sum_{t=1}^{T} \eta_t^2 + \log\left(\frac{1}{\delta}\right)\right)
\end{aligned}
$$

with probability at least $1-\delta/3$, which also implies that with probability at least $1-2\delta/3$,

$$
\begin{aligned}
&\left(\sum_{t=1}^{T} \eta_t\right)^{-1} \sum_{t=1}^{T} \eta_t \|\nabla F(\mathbf{w}_t)\|^2 \\
&\leq 2 \max_{t=1,\ldots,T} \|\nabla F(\mathbf{w}_t) - \nabla F_S(\mathbf{w}_t)\|^2 \\
&\quad + \left(\sum_{t=1}^{T} \eta_t\right)^{-1} \mathcal{O}\left(\sum_{t=1}^{T} \eta_t^2 + \log\left(\frac{1}{\delta}\right)\right) \\
&\leq \left(\sum_{t=1}^{T} \eta_t\right)^{-1} \mathcal{O}\left(\sum_{t=1}^{T} \eta_t^2 + \log\left(\frac{1}{\delta}\right)\right) \\
&\quad + 2\max_{t=1,\ldots,T} \left[C\beta \max\left\{\|\mathbf{w}_t - \mathbf{w}^*\|, \frac{1}{n}\right\} \right. \\
&\quad \times \left(\sqrt{\frac{d + \log\frac{12\log_2(\sqrt{2}R_1 n+1)}{\delta}}{n}} \right. \\
&\quad \left.\left. + \frac{d + \log\frac{12\log_2(\sqrt{2}R_1 n+1)}{\delta}}{n}\right)\right]^2. \qquad (36)
\end{aligned}
$$

According to Lemma 2 and Lemma 4, with probability $1-\delta/3$, we have the following inequality uniformly for all $t = 1,..,T$

$$
\|\mathbf{w}_{t+1} - \mathbf{w}^*\| = \begin{cases} \mathcal{O}(\log(1/\delta))T^{\frac{2-3\theta}{2}}, & \text{if } \theta < 1/2 \\ \mathcal{O}(\log(1/\delta))T^{\frac{1}{4}}\log^{1/2} T, & \text{if } \theta = 1/2 \\ \mathcal{O}(\log(1/\delta))T^{\frac{1-\theta}{2}}, & \text{if } \theta > 1/2. \end{cases} \qquad (37)
$$

Moreover, according to Lemma 4, we have

$$
\begin{aligned}
&\left(\sum_{t=1}^{T} \eta_t\right)^{-1} \mathcal{O}\left(\sum_{t=1}^{T} \eta_t^2 + \log\left(\frac{1}{\delta}\right)\right) \\
&= \begin{cases} \mathcal{O}(\log(1/\delta)T^{-\theta}), & \text{if } \theta < 1/2 \\ \mathcal{O}(\log(T/\delta)T^{-\frac{1}{2}}), & \text{if } \theta = 1/2 \\ \mathcal{O}(\log(1/\delta)T^{\theta-1}), & \text{if } \theta > 1/2. \end{cases} \qquad (38)
\end{aligned}
$$

Denote $\xi_{n,d,\delta} = \frac{d + \log\frac{\log n}{\delta}}{n}\log^2(1/\delta)$. Plugging (37) and (38) into (36), we finally get that with probability $1-\delta$

$$
\begin{aligned}
&\left(\sum_{t=1}^{T} \eta_t\right)^{-1} \sum_{t=1}^{T} \eta_t \|\nabla F(\mathbf{w}_t)\|^2 \\
&= \begin{cases} \mathcal{O}(\xi_{n,d,\delta})T^{2-3\theta} + \mathcal{O}(\log(1/\delta)T^{-\theta}), & \text{if } \theta < 1/2 \\ \mathcal{O}(\xi_{n,d,\delta})T^{\frac{1}{2}}\log T + \mathcal{O}(\log(T/\delta)T^{-\frac{1}{2}}), & \text{if } \theta = 1/2 \\ \mathcal{O}(\xi_{n,d,\delta})T^{1-\theta} + \mathcal{O}(\log(1/\delta)T^{\theta-1}), & \text{if } \theta > 1/2, \end{cases}
\end{aligned}
$$

If $\theta < 1/2$, we choose $T \asymp (nd^{-1})^{\frac{1}{2(1-\theta)}}$. If $\theta = 1/2$, we set $T \asymp nd^{-1}$. While if $\theta > 1/2$, we set $T \asymp (nd^{-1})^{\frac{1}{2(1-\theta)}}$. Then we can prove the learning rates of Theorem 8. The proof is complete. □

J. Proof of Theorem 9

Proof. Since F satisfies the PL assumption with parameter 2μ, we have

$$
F(\mathbf{w}) - F(\mathbf{w}^*) \leq \frac{\|\nabla F(\mathbf{w})\|^2}{4\mu}, \quad \forall \mathbf{w} \in \mathcal{W}. \qquad (39)
$$

So to bound $F(\mathbf{w}_{T+1}) - F(\mathbf{w}^*)$, we need to bound the term $\|\nabla F(\mathbf{w}_{T+1})\|^2$. And there holds that

$$
\begin{aligned}
\|\nabla F(\mathbf{w}_{T+1})\|^2 &\leq 2\|\nabla F(\mathbf{w}_{T+1}) - \nabla F_S(\mathbf{w}_{T+1})\|^2 \\
&\quad + 2\|\nabla F_S(\mathbf{w}_{T+1})\|^2. \qquad (40)
\end{aligned}
$$

From Theorem 3, if Assumptions 2 and 4 hold and F satisfies Assumption 3, for all $\mathbf{w} \in \mathcal{W}$ and any $\delta > 0$, with probability at least $1-\delta/2$, when $n \geq \frac{c\beta^2\left(d + \log\frac{16\log_2(\sqrt{2}R_1 n+1)}{\delta}\right)}{\mu^2}$, there holds

$$
\begin{aligned}
&\|\nabla F(\mathbf{w}_{T+1}) - \nabla F_S(\mathbf{w}_{T+1})\| \leq \|\nabla F_S(\mathbf{w}_{T+1})\| + \frac{2\mu}{n} \\
&+ \frac{8D_*\log(8/\delta)}{n} + 4\sqrt{\frac{8\beta F(\mathbf{w}^*)\log(8/\delta)}{n}}, \qquad (41)
\end{aligned}
$$

where $F(\mathbf{w}^*)$ follows from (35). For the second term $\|\nabla F_S(\mathbf{w}_{T+1})\|^2$, according to the smoothness property of F_S (see (21)) and Lemma 3, it can be derived that with probability at least $1-\delta/2$

$$
\|\nabla F_S(\mathbf{w}_{T+1})\|^2 = \mathcal{O}\left(\frac{\log(T)\log^3(1/\delta)}{T}\right). \qquad (42)
$$

Plugging (42) into (41), we can derive that

$$
\begin{aligned}
&\|\nabla F(\mathbf{w}_{T+1}) - \nabla F_S(\mathbf{w}_{T+1})\|^2 \\
&= \mathcal{O}\left(\frac{\log T\log^3(1/\delta)}{T}\right) + \mathcal{O}\left(\frac{\log^2(1/\delta)}{n^2} + \frac{F(\mathbf{w}^*)\log(1/\delta)}{n}\right). \qquad (43)
\end{aligned}
$$

Therefore, substituting (43) and (42) into (40), we derive that

$$
\|\nabla F(\mathbf{w}_{T+1})\|^2
$$

IEEE Proof

$$= \mathcal{O}\left(\frac{\log T \log^3(1/\delta)}{T}\right) + \mathcal{O}\left(\frac{\log^2(1/\delta)}{n^2} + \frac{F(\mathbf{w}^*)\log(1/\delta)}{n}\right). \tag{44}$$

Further substituting (44) into (39) and choosing $T \asymp n^2$, we finally obtain with probability at least $1-\delta$

$$F(\mathbf{w}_{T+1}) - F(\mathbf{w}^*) = \mathcal{O}\left(\frac{\log n \log^3\left(\frac{1}{\delta}\right)}{n^2} + \frac{F(\mathbf{w}^*)\log\left(\frac{1}{\delta}\right)}{n}\right).$$

The proof is complete. □

V. Conclusion

We studied the generalization performance of nonconvex pairwise learning given that it was rarely studied. We established several uniform convergences of gradients, based on which we provided a series of learning rates for ERM, GD, and SGD. We first investigated the general nonconvex setting and then the nonconvex learning with a gradient dominance curvature condition. Former demonstrated how the optimal iterative numbers should be selected to balance the generalization and optimization, shed insights on the role of early-stopping, and the latter highlight the established learning rates which are significantly faster than the state-of-the-art, even up to $\mathcal{O}(1/n^2)$. Overall, we provide a relatively systematic study of nonconvex pairwise learning.

Acknowledgments

We sincerely appreciate the associate editor and the anonymous reviewers for their invaluable and constructive comments.

References

[1] S. Agarwal and P. Niyogi, "Generalization bounds for ranking algorithms via algorithmic stability," *J. Mach. Learn. Res.*, vol. 10, no. 16, pp. 441–474, 2009.
[2] F. Bach and E. Moulines, "Non-strongly-convex smooth stochastic approximation with convergence rate $\mathcal{O}(1/n)$," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2013, pp. 773–781.
[3] S. Balakrishnan, M. J. Wainwright, and B. Yu, "Statistical guarantees for the em algorithm: From population to sample-based analysis," *Ann. Statist.*, vol. 45, no. 1, pp. 77–120, 2017.
[4] P. L. Bartlett, O. Bousquet, and S. Mendelson, "Local rademacher complexities," *Ann. Statist.*, vol. 33, no. 4, pp. 1497–1537, 2005.
[5] P. L. Bartlett and S. Mendelson, "Rademacher and Gaussian complexities: Risk bounds and structural results," *J. Mach. Learn. Res.*, vol. 3, no. Nov., pp. 463–482, 2002.
[6] W. Bian and D. Tao, "Asymptotic generalization bound of Fisher's linear discriminant analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 12, pp. 2325–2337, Dec. 2014.
[7] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *SIAM Rev.*, vol. 60, no. 2, pp. 223–311, 2018.
[8] O. Bousquet and A. Elisseeff, "Stability and generalization," *J. Mach. Learn. Res.*, vol. 2, no. 3, pp. 499–526, 2002.
[9] O. Bousquet, Y. Klochkov, and N. Zhivotovskiy, "Sharper bounds for uniformly stable algorithms," in *Proc. Conf. Learn. Theory*, 2020, pp. 610–626.
[10] Q. Cao, Z.-C. Guo, and Y. Ying, "Generalization bounds for metric and similarity learning," *Mach. Learn.*, vol. 102, no. 1, pp. 115–132, 2016.
[11] Z. B. Charles and D. S. Papailiopoulos, "Stability and generalization of learning algorithms that converge to global optima," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 744–753.
[12] S. Clémençon, G. Lugosi, and N. Vayatis, "Ranking and scoring using empirical risk minimization," in *Proc. Conf. Learn. Theory*, 2005, pp. 1–15.
[13] S. Clémençon, G. Lugosi, and N. Vayatis, "Ranking and empirical minimization of U-statistics," *Ann. Statist.*, vol. 36, no. 2, pp. 844–874, 2008.
[14] C. Cortes, V. Kuznetsov, M. Mohri, and S. Yang, "Structured prediction theory based on factor graph complexity," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 2514–2522.
[15] C. Cortes and M. Mohri, "AUC optimization vs. error rate minimization," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2003, pp. 313–320.
[16] Z. Dang, X. Li, B. Gu, C. Deng, and H. Huang, "Large-scale nonlinear AUC maximization via triply stochastic gradients," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1385–1398, Mar. 2022.
[17] D. Davis and D. Drusvyatskiy, "Graphical convergence of subgradients in nonconvex optimization and learning," *Math. Operations Res.*, vol. 47, pp. 209–231, 2022.
[18] V. Feldman, "Generalization of ERM in stochastic convex optimization: The dimension strikes back," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 3576–3584.
[19] D. J. Foster, A. Sekhari, and K. Sridharan, "Uniform convergence of gradients for non-convex learning and optimization," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 8745–8756.
[20] J. Fürnkranz and E. Hüllermeier, "Preference learning and ranking by pairwise comparison," in *Preference Learning*. Berlin, Germany: Springer, 2010, pp. 65–82.
[21] W. Gao, R. Jin, S. Zhu, and Z.-H. Zhou, "One-pass AUC optimization," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 906–914.
[22] W. Gao and Z.-H. Zhou, "Uniform convergence, stability and learnability for ranking problems," in *Proc. Int. Joint Conf. Artif. Intell.*, 2013, pp. 1337–1343.
[23] X. Guo, T. Hu, and Q. Wu, "Distributed minimum error entropy algorithms," *J. Mach. Learn. Res.*, vol. 21, no. 126, pp. 1–31, 2020.
[24] M. Hardt and T. Ma, "Identity matters in deep learning," in *Proc. Int. Conf. Learn. Representations*, 2016. Q2
[25] M. Hardt, T. Ma, and B. Recht, "Gradient descent learns linear dynamical systems," *J. Mach. Learn. Res.*, vol. 19, no. 29, pp. 1–44, 2018.
[26] M. Hardt, B. Recht, and Y. Singer, "Train faster, generalize better: Stability of stochastic gradient descent," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1225–1234.
[27] N. J. A. Harvey, C. Liaw, Y. Plan, and S. Randhawa, "Tight analyses for non-smooth stochastic gradient descent," in *Proc. Conf. Learn. Theory*, 2019, pp. 1579–1613.
[28] T. Hu, J. Fan, Q. Wu, and D.-X. Zhou, "Learning theory approach to minimum error entropy criterion," *J. Mach. Learn. Res.*, vol. 14, no. 1, pp. 377–397, 2013.
[29] M. Huai, D. Wang, C. Miao, J. Xu, and A. Zhang, "Pairwise learning with differential privacy guarantees," in *Proc. Nat. Conf. Artif. Intell.*, 2020, pp. 694–701.
[30] R. Jin, S. Wang, and Y. Zhou, "Regularized distance metric learning:theory and algorithm," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2009, pp. 862–870.
[31] P. Kar, B. Sriperumbudur, P. Jain, and H. Karnick, "On the generalization ability of online learning algorithms for pairwise loss functions," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 441–449.
[32] H. Karimi, J. Nutini, and M. Schmidt, "Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition," in *Proc. Eur. Conf. Mach. Learn. Knowl. Discov. Databases*, 2016, pp. 795–811.
[33] Y. Klochkov and N. Zhivotovskiy, "Stability and deviation optimal risk bounds with convergence rate $\mathcal{O}(1/n)$," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 5065–5076.
[34] B. Krishnapuram, L. Carin, M. Figueiredo, and A. Hartemink, "Sparse multinomial logistic regression: Fast algorithms and generalization bounds," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 6, pp. 957–968, Jun. 2005.
[35] A. Kumar, A. Niculescu-mizil, K. Kavukcuoglu, and H. Daume, "A binary classification framework for two-stage multiple kernel learning," in *Proc. Int. Conf. Mach. Learn.*, 2012, pp. 1331–1338.
[36] Y. Lei, A. Ledent, and M. Kloft, "Sharper generalization bounds for pairwise learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 21236–21246.
[37] Y. Lei, S.-B. Lin, and K. Tang, "Generalization bounds for regularized pairwise learning," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 2376–2382.

IEEE Proof

[38] Y. Lei, M. Liu, and Y. Ying, "Generalization guarantee of SGD for pairwise learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 21216–21228.

[39] Y. Lei and K. Tang, "Learning rates for stochastic gradient descent with nonconvex objectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 12, pp. 4505–4511, Dec. 2021.

[40] Y. Lei and Y. Ying, "Fine-grained analysis of stability and generalization for stochastic gradient descent," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 5809–5819.

[41] Y. Lei and Y. Ying, "Sharper generalization bounds for learning with gradient-dominated objective functions," in *Proc. Int. Conf. Learn. Representations*, 2021.

[42] Y. Lei and Y. Ying, "Stochastic proximal AUC maximization," *J. Mach. Learn. Res.*, vol. 22, no. 61, pp. 1–45, 2021.

[43] S. Li, K. Jia, Y. Wen, T. Liu, and D. Tao, "Orthogonal deep neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 4, pp. 1352–1368, Apr. 2021.

[44] S. Li and Y. Liu, "Improved learning rates for stochastic optimization: Two theoretical viewpoints," 2021, *arXiv:2107.08686*.

[45] S. Li and Y. Liu, "Sharper generalization bounds for clustering," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 6392–6402.

[46] S. Li and Y. Liu, "Towards sharper generalization bounds for structured prediction," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 26844–26857.

[47] X. Li, S. Ling, T. Strohmer, and K. Wei, "Rapid, robust, and reliable blind deconvolution via nonconvex optimization," *Appl. Comput. Harmon. Anal.*, vol. 47, no. 3, pp. 893–934, 2019.

[48] Y. Li and Y. Yuan, "Convergence analysis of two-layer neural networks with ReLU activation," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 597–607.

[49] J. Lin, Y. Lei, B. Zhang, and D.-X. Zhou, "Online pairwise learning algorithms with convex loss functions," *Inf. Sci.*, vol. 406, pp. 57–70, 2017.

[50] H. Liu, W. Wu, and A. M.-C. So, "Quadratic optimization with orthogonality constraints: Explicit Lojasiewicz exponent and linear convergence of line-search methods," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1158–1167.

[51] M. Liu, Z. Yuan, Y. Ying, and T. Yang, "Stochastic AUC maximization with deep neural networks," in *Proc. Int. Conf. Learn. Representations*, 2020.

[52] M. Liu, X. Zhang, Z. Chen, X. Wang, and T. Yang, "Fast stochastic AUC maximization with $\mathcal{O}(1/n)$-convergence rate," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 3189–3197.

[53] M. Liu, X. Zhang, L. Zhang, R. Jin, and T. Yang, "Fast rates of ERM and stochastic approximation: Adaptive to error bound conditions," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 4678–4689.

[54] T. Liu, D. Tao, M. Song, and S. J. Maybank, "Algorithm-dependent generalization bounds for multi-task learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 2, pp. 227–241, Feb. 2017.

[55] Y. Liu, "Refined learning bounds for kernel and approximate k-means," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 6142–6154.

[56] Y. Liu, S. Liao, S. Jiang, L. Ding, H. Lin, and W. Wang, "Fast cross-validation for kernel-based algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 5, pp. 1083–1096, May 2020.

[57] Y. Liu, S. Liao, H. Lin, Y. Yue, and W. Wang, "Generalization analysis for ranking using integral operator," in *Proc. Nat. Conf. Artif. Intell.*, 2017, pp. 2273–2279.

[58] S. Mei, Y. Bai, and A. Montanari, "The landscape of empirical risk for nonconvex losses," *Ann. Statist.*, vol. 46, pp. 2747–2774, 2018.

[59] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*. Cambridge, MA, USA: MIT Press, 2012.

[60] S. Mukherjee and Q. Wu, "Estimation of gradients and coordinate covariation in classification," *J. Mach. Learn. Res.*, vol. 7, no. 88, pp. 2481–2514, 2006.

[61] S. Mukherjee and D.-X. Zhou, "Learning coordinate covariances via gradients," *J. Mach. Learn. Res.*, vol. 7, no. 18, pp. 519–549, 2006.

[62] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, "Robust stochastic approximation approach to stochastic programming," *SIAM J. Optim.*, vol. 19, no. 4, pp. 1574–1609, 2008.

[63] I. E. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*. Berlin, Germany: Springer, 2014.

[64] G. Papa, S. Clémençon, and A. Bellet, "SGD algorithms based on incomplete u-statistics: Large-scale minimization of empirical risk," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 1027–1035.

[65] A. Rakhlin, S. Mukherjee, and T. Poggio, "Stability results in learning theory," *Anal. Appl.*, vol. 3, no. 4, pp. 397–417, 2005.

[66] S. J. Reddi, A. Hefny, S. Sra, B. Póczós, and A. Smola, "Stochastic variance reduction for nonconvex optimization," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 314–323.

[67] W. Rejchel, "On ranking and generalization bounds," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 1373–1392, 2012.

[68] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2015.

[69] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan, "Stochastic convex optimization," in *Proc. Conf. Learn. Theory*, 2009.

[70] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan, "Learnability, stability and uniform convergence," *J. Mach. Learn. Res.*, vol. 11, no. 90, pp. 2635–2670, 2010.

[71] W. Shen, Z. Yang, Y. Ying, and X. Yuan, "Stability and optimization error of stochastic gradient descent for pairwise learning," *Anal. Appl.*, vol. 18, no. 5, pp. 887–927, 2020.

[72] N. Srebro, K. Sridharan, and A. Tewari, "Optimistic rates for learning with a smooth loss," 2010, *arXiv:1009.3896*.

[73] J. Sun, Q. Qu, and J. Wright, "A geometric analysis of phase retrieval," *Found. Comput. Math.*, vol. 18, no. 5, pp. 1131–1198, 2018.

[74] N. Verma and K. Branson, "Sample complexity of learning Mahalanobis distance metrics," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 2584–2592.

[75] M. J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge, U.K.: Cambridge Univ. Press, 2019.

[76] B. Wang, H. Zhang, P. Liu, Z. Shen, and J. Pineau, "Multitask metric learning: Theory and algorithm," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2019, pp. 3362–3371.

[77] P. Wang, Z. Yang, Y. Lei, Y. Ying, and H. Zhang, "Differentially private empirical risk minimization for AUC maximization," *Neurocomputing*, vol. 461, pp. 419–437, 2021.

[78] Y. Wang, R. Khardon, D. Pechyony, and R. Jones, "Generalization bounds for online learning algorithms with pairwise loss functions," in *Proc. 25th Annu. Conf. Learn. Theory*, 2012, pp. 13.1–13.22.

[79] Y. Xu and A. Zeevi, "Towards optimal problem dependent generalization error bounds in statistical learning theory," 2020, *arXiv:2011.06186*.

[80] Y. Xu and A. Zeevi, "Upper counterfactual confidence bounds: A new optimism principle for contextual bandits," 2020, *arXiv:2007.07876*.

[81] Z. Yang, Y. Lei, S. Lyu, and Y. Ying, "Stability and differential privacy of stochastic gradient descent for pairwise learning with non-smooth loss," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2021, pp. 2026–2034.

[82] Z. Yang, Y. Lei, P. Wang, T. Yang, and Y. Ying, "Simple stochastic and online gradient descent algorithms for pairwise learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 20160–20171.

[83] Z. Yang, Q. Xu, S. Bao, X. Cao, and Q. Huang, "Learning with multiclass AUC: Theory and algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7747–7763, Nov. 2022.

[84] H.-J. Ye, D.-C. Zhan, and Y. Jiang, "Fast generalization rates for distance metric learning," *Mach. Learn.*, vol. 108, no. 2, pp. 267–295, 2019.

[85] Y. Ying and C. Campbell, "Learning coordinate gradients with multi-task kernels," in *Proc. 21st Annu. Conf. Learn. Theory*, 2008, pp. 217–228.

[86] Y. Ying, L. Wen, and S. Lyu, "Stochastic online AUC maximization," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 451–459.

[87] Y. Ying and D.-X. Zhou, "Online pairwise learning algorithms," *Neural Computation*, vol. 28, no. 4, pp. 743–777, 2016.

[88] L. Zhang, T. Yang, and R. Jin, "Empirical risk minimization for stochastic convex optimization: $\mathcal{O}(1/n)$- and $\mathcal{O}(1/n^2)$-type of risk bounds," in *Proc. Annu. Conf. Learn. Theory*, 2017, pp. 1954–1979.

[89] L. Zhang and Z.-H. Zhou, "Stochastic approximation of smooth and strongly convex functions: Beyond the $\mathcal{O}(1/t)$ convergence rate," in *Proc. Annu. Conf. Learn. Theory*, 2019, pp. 3160–3179.

[90] T. Zhang, "Solving large scale linear prediction problems using stochastic gradient descent algorithms," in *Proc. Int. Conf. Mach. Learn.*, 2004, Art. no. 116.

[91] P. Zhao, R. Jin, T. Yang, and S. C. Hoi, "Online AUC maximization," in *Proc. Int. Conf. Mach. Learn.*, 2011, pp. 233–240.

[92] Y. Zhou, H. Chen, R. Lan, and Z. Pan, "Generalization performance of regularized ranking with multiscale kernels," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 5, pp. 993–1002, May 2016.

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

IEEE Proof

Shaojie Li is currently working toward the PhD degree with the Gaoling School of Artificial Intelligence, Renmin University of China, Beijing. His research interests include statistical learning theory, optimization, and deep learning. He has first-authored several academic papers in top-tier international conferences including ICML/NeurIPS/ICLR/AAAI. He serves as a reviewer for ICML and NeurIPS.

Yong Liu received the PhD degree in computer science from Tianjin University, in 2016. He is currently an associate professor with the Beijing Key Laboratory of Big Data Management and Analysis Methods, Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China. His research interests are mainly about machine learning, with special attention to large-scale machine learning, AutoML, statistical machine learning theory, etc. He has published more than 40 papers on top-tier conferences and journals in artificial intelligence, e.g., *IEEE Transactions on Pattern Analysis and Machine Intelligence*, NeurIPS, ICML, ICLR, IJCAI, AAAI, *IEEE Transactions on Image Processing*, *IEEE Transactions on Neural Networks and Learning Systems*, etc. He received the "Outstanding Scholar of Renmin University of China," the "Youth Innovation Promotion Association" of CAS and the "Excellent Talent Introduction" of Institute of Information Engineering, CAS.

IEEE Proof