Stability-Based Generalization Analysis for Mixtures of Pointwise and Pairwise Learning

Jiahuan Wang[1], Jun Chen[2], Hong Chen[1, 3, 4,*], Bin Gu[5], Weifu Li[1, 3, 4], Xin Tang[6]
[1]College of Science, Huazhong Agricultural University, Wuhan 430070, China
[2]College of Informatics, Huazhong Agricultural University, Wuhan 430070, China
[3]Engineering Research Center of Intelligent Technology for Agriculture, Ministry of Education, Wuhan 430070, China
[4]Key Laboratory of Smart Farming for Agricultural Animals, Wuhan 430070, China
[5]Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
[6]Ping An Property & Casualty Insurance Company, Shenzhen, China
chenh@mail.hzau.edu.cn

Abstract

Recently, some mixture algorithms of pointwise and pairwise learning (PPL) have been formulated by employing the hybrid error metric of "pointwise loss + pairwise loss" and have shown empirical effectiveness on feature selection, ranking and recommendation tasks. However, to the best of our knowledge, the learning theory foundation of PPL has not been touched in the existing works. In this paper, we try to fill this theoretical gap by investigating the generalization properties of PPL. After extending the definitions of algorithmic stability to the PPL setting, we establish the high-probability generalization bounds for uniformly stable PPL algorithms. Moreover, explicit convergence rates of stochastic gradient descent (SGD) and regularized risk minimization (RRM) for PPL are stated by developing the stability analysis technique of pairwise learning. In addition, the refined generalization bounds of PPL are obtained by replacing uniform stability with on-average stability.

Introduction

There are mainly two paradigms to formulate machine learning systems including pointwise learning and pairwise learning. Usually, the former aims to train models under the error metric associated with single sample, while the latter concerns the relative relationships between objects measured by the loss related to the pair of samples. Besides wide applications, the theoretical foundations of the above paradigms have been well established from the viewpoint of statistical learning theory, e.g., pointwise stability analysis (Bousquet and Elisseeff 2002; London, Huang, and Getoor 2016; Sun, Li, and Wang 2021), pairwise stability analysis (Agarwal and Niyogi 2009; Lei, Ledent, and Kloft 2020; Lei, Liu, and Ying 2021), and uniform convergence analysis (Clémençon, Lugosi, and Vayatis 2008; Rejchel 2012; Cao, Guo, and Ying 2016; Ying, Wen, and Lyu 2016; Ying and Zhou 2016).

It is well known that pointwise (or pairwise) learning enjoys certain advantages and limitations for real-world data analysis. For the same number of samples, pointwise learning has computation feasibility due to its low model complexity, while pairwise learning can mine valuable information in terms of the intrinsic relationship among samples. As illustrated by Wang et al. (2016), the degraded performance may occur for the pointwise learning with the ambiguity of some labels and for the pairwise learning as samples in different categories have similar features. Therefore, it is natural to consider the middle modality of the above paradigms to alleviate their drawbacks. Along this line, some learning algorithms have been proposed under the pointwise and pairwise learning (PPL) framework, where the pointwise loss and the pairwise loss are employed jointly (Liu and Zhang 2015; Wang et al. 2016; Lei et al. 2017; Zhuo et al. 2022; Wang et al. 2022a). In PPL, its pointwise part concerns the fitting ability to empirical observations and its pairwise part addresses the stability or robustness of learning models (Liu and Zhang 2015). While the studies on algorithmic design and applications are increasing, there are far fewer results to investigate the generalization ability of PPL in theory.

As one of the main routines of learning theory analysis, the algorithmic stability tools have advantages in some aspects, such as dimensional independence and adaptivity for broad learning paradigms (Bousquet and Elisseeff 2002; Shalev-Shwartz et al. 2010; Hardt, Recht, and Singer 2016; Feldman and Vondrak 2018, 2019). Specially, the stability and generalization have been well understood recently for stochastic gradient descent (SGD) and regularized risk minimization (RRM) under both the pointwise learning (Hardt, Recht, and Singer 2016; Lei and Ying 2020) and the pairwise learning setting (Lei, Ledent, and Kloft 2020; Lei, Liu, and Ying 2021). Inspired by the recent progress, in this paper, we try to fill this theoretical gap of PPL by establishing its generalization bounds in terms of the algorithmic stability technique. To the best of our knowledge, this is the first theoretical understanding of generalization properties for PPL.

The main work of this paper is two-fold: One is to establish a relationship between uniform stability and estimation error for the mixture setting, which can be considered as natural extension of the related results in (Lei, Ledent, and Kloft 2020; Lei, Liu, and Ying 2021). The other is to characterize the stability-based generalization bounds for some PPL algorithms (i.e. SGD and RRM) under mild conditions.

*Corresponding author.

Objective function	Reference	Optimization	Application	Generalization analysis
Pointwise loss +Pairwise loss +Regularization	Liu and Zhang (2015)	ACG	Feature Selection	No
	Wang et al. (2016)	ADMM	Image Classification	No
	Lei et al. (2017)	SGD	Ranking	No
	Wang et al. (2022a)	SGD	Recommendation	No

Table 1: Summary of pointwise and pairwise learning (ACG: Accelerated Proximal Gradient; ADMM: Alternating Direction Method of Multipliers).

Related Work

To better evaluate our theoretical results, we review the related works on PPL and generalization analysis.

Pointwise and pairwise learning (PPL). In recent years, some algorithms of PPL have been designed for learning tasks such as feature selection, image classification, ranking, and recommendation systems. Liu and Zhang (2015) proposed a pairwise constraint-guided sparse learning method for feature selection, where the pairwise constraint is used for improving robustness. For image classification tasks, Wang et al. (2016) designed a novel joint framework, called pointwise and pairwise image label prediction, to predict both pointwise and pairwise labels and achieved superior performance. For emphasis selection tasks, Huang et al. (2020) employed a pointwise regression loss and a pairwise ranking loss simultaneously to fit models. Recently, Lei et al. (2017) proposed an alternating pointwise-pairwise ranking to improve decision performance. Zhuo et al. (2022) and Wang et al. (2022a) formulated hybrid learning models for the recommendation systems. Although the empirical effectiveness has been validated for the above PPL algorithms, their theoretical foundations (e.g., generalization guarantee) have not been investigated before. To further highlight the gap in generalization analysis, we summarize the basic properties of PPL models in Table 1.

Generalization analysis. From the viewpoint of statistical learning theory, generalization analysis is crucial since it provides the statistical theory support for the empirical performance of trained models. Usually, model training is a process of calculating loss based on data and then seeking an optimal function in the predetermined hypothetical function space through an optimization algorithm. Naturally, the generalization performance of learning systems can be investigated from the perspectives of hypothetical function space (Smale and Zhou 2007; Yin, Kannan, and Bartlett 2019; Lei and Tang 2021; Wang et al. 2020) and data (Bousquet and Elisseeff 2002; Elisseeff et al. 2005; Shalev-Shwartz et al. 2010), respectively. The former is often called uniform convergence analysis and the latter is realized by stability analysis. In essential, the uniform convergence analysis considers the capacity of hypothesis space (e.g., via VC-dimension (Vapnik, Levin, and Le Cun 1994), covering numbers (Zhou 2002; Chen et al. 2017, 2021), Rademacher complexity (Yin, Kannan, and Bartlett 2019)), while the stability analysis concerns the change of model parameters caused by the change of training data (Bousquet and Elisseeff 2002; Lei, Liu, and Ying 2021). Algorithmic stability has shown remarkable effectiveness in deriving dimension-independent generalization bounds for wide learning frameworks. A classic framework for stability analysis is developed by Bousquet and Elisseeff (2002), in which the uniform stability and hypothesis stability are introduced. Subsequently, the uniform stability measure was extended to study stochastic algorithms (Elisseeff et al. 2005; Hardt, Recht, and Singer 2016) and inspired several other stability concepts including uniform argument stability (Liu et al. 2017), locally elastic stability (Deng, He, and Su 2021), on-average loss stability (Lei, Ledent, and Kloft 2020; Lei and Ying 2020; Lei, Liu, and Ying 2021) and on-average argument stability (Shalev-Shwartz et al. 2010; Lei, Liu, and Ying 2021).

From the lens of learning paradigms, generalization guarantees have been established for various pointwise learning algorithms (Bousquet and Elisseeff 2002; London, Huang, and Getoor 2016; Hardt, Recht, and Singer 2016; Lei and Ying 2020; Sun, Li, and Wang 2021; Klochkov and Zhivotovskiy 2021) and pairwise learning models (Agarwal and Niyogi 2009; Lei, Ledent, and Kloft 2020; Lei, Liu, and Ying 2021; Yang et al. 2021). Therefore, it is natural to explore the generalization properties of PPL by the means of the stability analysis technique.

Preliminaries

This section introduces the problem formulation of PPL and the definitions of algorithmic stability.

Pointwise and Pairwise Learning

Consider a training dataset $S := \{z_i\}_{i=1}^n$, where each z_i is independently drawn from a probability measure ρ defined over a sample space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Here, $\mathcal{X} \subset \mathbb{R}^d$ is an input space of dimension d and $\mathcal{Y} \subset \mathbb{R}$ is an output space. Let $\mathcal{W}$ be a given parameter space of learning models. The goal of pointwise learning is to find a parameter $\mathbf{w}$ based model such that the population risk (or expected risk), defined as

$$R^{point}(\mathbf{w}) = \mathbb{E}_z[f(\mathbf{w}; z)],$$

is as small as possible, where $f : \mathcal{W} \times \mathcal{Z} \to [0, \infty)$ is a pointwise loss and $\mathbb{E}_z$ denotes the expectation with respect to $z \sim \rho$. For brevity, we also use $\mathbf{w}$ to denote the parameter $\mathbf{w}$ based model in the sequel.

However, we can't get the minimizer of $R^{point}(\mathbf{w})$ directly since the intrinsic distribution ρ is unknown. As a natural surrogate, for algorithmic design, we often consider the corresponding empirical risk defined as

$$R_S^{point}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} f\left(\mathbf{w}; z_i\right).$$

Unlike the pointwise learning, the pairwise learning model $\mathbf{w}$ is measured by

$$R^{pair}(\mathbf{w}) = \mathbb{E}_{z,\tilde{z}}[g(\mathbf{w}; z, \tilde{z})],$$

where $g : \mathcal{W} \times \mathcal{Z} \times \mathcal{Z} \to [0, \infty)$ is a pairwise loss function and $\mathbb{E}_{z,\tilde{z}}$ denotes the expectation with respect to $z, \tilde{z} \sim \rho$. In pairwise learning models, $R^{pair}(\mathbf{w})$ is approximately characterized by the empirical risk

$$R_S^{pair}(\mathbf{w}) = \frac{1}{n(n-1)} \sum_{i,j\in[n]:i\neq j} g\left(\mathbf{w}; z_i, z_j\right),$$

where $z_i, z_j \sim \rho$ and $[n] := \{1, \ldots, n\}$.

In this paper, we consider a mixture paradigm of pointwise learning and pairwise learning, called pointwise and pairwise learning (PPL). The population risk of $\mathbf{w}$ in PPL is

$$R(\mathbf{w}) = \tau R^{point}(\mathbf{w}) + (1-\tau)R^{pair}(\mathbf{w}),$$

where $\tau \in [0, 1]$ is a tuning parameter. Given training set S, the corresponding empirical version of $R(\mathbf{w})$ is

$$R_S(\mathbf{w}) = \tau R_S^{point}(\mathbf{w}) + (1-\tau)R_S^{pair}(\mathbf{w}). \tag{1}$$

For brevity, $A(S)$ denotes the derived model by applying algorithm A (e.g., SGD and RRM) on S. In the process of training and adjustment of parameters, the output model $A(S)$ can be a small empirical risk since we often can fit training examples perfectly. However, the empirical effectiveness of $A(S)$ can not assure the small population risk. In statistical learning theory, the difference between the population risk and empirical risk

$$R(\mathbf{w}) - R_S(\mathbf{w}) \tag{2}$$

is called the generalization error of learning model $\mathbf{w}$. It is key concern of this paper to bound this gap in theory.

Algorithmic Stability

Algorithmic stability is an important concept in statistical learning, which measures the sensitivity of an algorithm to the perturbation of training sets. This paper focuses on the analysis techniques associated with the algorithmic uniform stability (Bousquet and Elisseeff 2002; Elisseeff et al. 2005; Agarwal and Niyogi 2009; Hardt, Recht, and Singer 2016), on-average loss stability (Lei, Ledent, and Kloft 2020; Lei and Ying 2020; Lei, Liu, and Ying 2021), and on-average argument stability (Shalev-Shwartz et al. 2010; Lei, Liu, and Ying 2021). To match the generalization analysis of PPL algorithms, we firstly extend the definitions of algorithmic stability (e.g., the pointwise uniform stability (Bousquet and Elisseeff 2002) and pairwise uniform stability (Lei, Ledent, and Kloft 2020)) to the PPL setting.

Let $S = \{z_1, \ldots, z_n\}$ and $S' = \{z'_1, \ldots, z'_n\}$ be independently drawn from ρ. For any $i < j, i, j \in [n]$, denote

$$S_i = \{z_1, \ldots, z_{i-1}, z'_i, z_{i+1}, \ldots, z_n\} \tag{3}$$

and

$$S_{i,j} = \left\{z_1, \ldots, z_{i-1}, z'_i, z_{i+1}, \ldots, z_{j-1}, z'_j, z_{j+1}, \ldots, z_n\right\}. \tag{4}$$

Definition 1. *(PPL Uniform Stability). Assume that $f(\cdot; z)$ is a pointwise loss function and $g(\cdot; z, \tilde{z})$ is a pairwise loss function. We say $A : \mathcal{Z}^n \mapsto \mathcal{W}$ is PPL γ-uniformly stable, if for any training datasets $S, S_i \in \mathcal{Z}^n$*

$$\max\{U_{point}, U_{pair}\} \leq \gamma, \forall i \in [n],$$

where $U_{point} = \sup_{z\in\mathcal{Z}} |f(A(S); z) - f(A(S_i); z)|$ and $U_{pair} = \sup_{z,\tilde{z}\in\mathcal{Z}} |g(A(S); z, \tilde{z}) - g(A(S_i); z, \tilde{z})|$.

Remark 1. *Denote*

$$\ell(A(S_i); z, \tilde{z}) := \tau f(A(S_i); z) + (1-\tau)g(A(S_i); z, \tilde{z})$$

for simplicity, and call it as the PPL loss function. Then, we can define the weaker stability measure by replacing the maximum in Definition 1 with

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

Following the mixture stability associated with τ, we can also get the similar generalization results as Theorems 1 and 2 in the next section.

We then introduce the definitions of PPL on-average loss stability and PPL on-average argument stability described as follows.

Definition 2. *(PPL On-average Loss Stability). Let $f(\cdot; z)$ be a pointwise loss function and let $g(\cdot; z, \tilde{z})$ be a pairwise loss function. We say $A : \mathcal{Z}^n \mapsto \mathcal{W}$ is PPL γ-on-average loss stable if, for any training datasets $S, S_i, S_{i,j} \in \mathcal{Z}^n$,*

$$\max\{V_{point}, V_{pair}\} \leq \gamma, \forall i < j \in [n],$$

where

$$V_{point} = \frac{1}{n}\sum_{i\in[n]} \mathbb{E}_{S,S'}\left[f\left(A\left(S_i\right); z_i\right) - f\left(A(S); z_i\right)\right]$$

and

$$V_{pair} = \frac{1}{n(n-1)} \sum_{i,j\in[n]:i\neq j} \mathbb{E}_{S,S'}\big[g\left(A\left(S_{i,j}\right); z_i, z_j\right) - g\left(A(S); z_i, z_j\right)\big]$$

Remark 2. *Definition 2 is built by combining the on-average loss stability for pointwise learning (Shalev-Shwartz et al. 2010; Lei and Ying 2020) with the one for pairwise learning (Lei, Ledent, and Kloft 2020; Lei, Liu, and Ying 2021). The requirement of PPL on-average loss stability is milder than the PPL uniform stability, where the stabilization is not measured by the changes of all samples but the mean of training sets.*

Definition 3. *(PPL On-average Argument Stability). We say $A : \mathcal{Z}^n \mapsto \mathcal{W}$ is PPL ℓ_1 γ-on-average argument stable if, for any training datasets $S, S_i \in \mathcal{Z}^n$,*

$$\max\{\mathbb{E}_{S,S',A}[H_{point}], \mathbb{E}_{S,S',A}[H_{pair}]\} \leq \gamma, \tag{5}$$

where $H_{point} = \frac{1}{n}\sum_{i=1}^{n} \|A(S) - A(S_i)\|_2$ and $H_{pair} = \frac{1}{n(n-1)}\sum_{i,j\in[n]:i\neq j} \|A(S) - A(S_{i,j})\|_2$.
We say A is PPL ℓ_2 γ-on-average argument stable if, for any training datasets $S, S_i \in \mathcal{Z}^n$,

$$\mathbb{E}_{S,S',A}\Big[\frac{1}{n}\sum_{i=1}^{n} \|A(S) - A(S_i)\|_2^2\Big] \leq \gamma^2.$$

Remark 3. *The ℓ_2 on-average argument stability of PPL is similar with that of pointwise learning (Lei and Ying 2020) and pairwise learning (Lei, Liu, and Ying 2021). Differently from Definition 1, which relied on the drift of loss functions, Definition 2, 3 measures the stability in terms of the changes of the model $A(S)$.*

Main Results

In this section, we present our main results on the generalization bounds of PPL algorithms based on uniform stability and on-average stability.

Uniform Stability-Based Generalization

This subsection establishes the relationship between the generalization ability and the uniform stability for PPL. In the sequel, e represents the base of the natural logarithm, $\lceil a \rceil$ means the smallest integer which is no less than a. *Supplementary Material B.1* provides the detailed proof of the following theorem.

Theorem 1. *Let $A : \mathcal{Z}^n \mapsto \mathcal{W}$ be PPL γ-uniformly stable. Assume that*

$$\max_{z,\tilde{z}\in\mathcal{Z}} \{ |\mathbb{E}_S[f(A(S);z)]| \, , |\mathbb{E}_S[g(A(S);z,\tilde{z})]| \} \leq M$$

for some positive constant M. Then, for all $\tau \in [0,1]$ and $\delta \in (0, 1/e)$, we have, with probability $1-\delta$,

$$\begin{aligned} &|R_S(A(S)) - R(A(S))| \\ \leq &(4-2\tau)\gamma + e\Big(4M(4-3\tau)n^{-\frac{1}{2}}\sqrt{\log(e/\delta)} \\ &+ 24\sqrt{2}(2-\tau)\gamma \lceil \log_2(n) \rceil \log(e/\delta)\Big) . \end{aligned}$$

Remark 4. *Theorem 1 is a high-probability generalization bound for uniformly stable PPL algorithms, motivated by the recent analyses in the pointwise learning (Hardt, Recht, and Singer 2016) and the pairwise learning (Lei, Ledent, and Kloft 2020; Lei, Liu, and Ying 2021). Similar to Lei, Ledent, and Kloft (2020), the error estimations of the pointwise part and the pairwise part of PPL are obtained by applying and developing the concentration inequality (Bousquet, Klochkov, and Zhivotovskiy 2020). Ignoring the constants (e.g. M, τ, $\log(e/\delta)$), we can get the convergence order $O(n^{-\frac{1}{2}} + \gamma \log_2 n)$ from Theorem 1. Due the generality and flexibility of PPL induced by $\tau \in [0,1]$, the derived connection, between uniform stability and generalization, contains the previous results for pairwise learning (Lei, Ledent, and Kloft 2020) as special example.*

Remark 5. *To better understand the stability-based generalization bound, we summarize the main results about the relationships between generalization and various definitions of algorithmic stability in Table 2. In this table, for feasibility, we denote the generalization error as*

$$Gen := R(A(S)) - R_S(A(S))$$

and denote the expected generalization error as

$$\mathbb{E}_{Gen} := \mathbb{E}_{S,A}[R(A(S)) - R_S(A(S))].$$

Table 2 demonstrates our characterized relations are comparable with the existing results.

Generalization Bounds of SGD for PPL

As a popular computing strategy, SGD has been employed for PPL as shown in Table 1. The SGD for PPL can be regarded as an elastic net version of pointwise SGD and pairwise SGD, which involves the gradients of the pointwise loss function f and the pairwise loss function g. At the t-th iteration, (i_t, j_t) is taken from the uniform distribution over $[n]$ randomly, which requires $i_t \neq j_t$. The SGD for the PPL model is updated by

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \left(\tau \nabla f\left(\mathbf{w}_t; z_{i_t}\right) + (1-\tau)\nabla g\left(\mathbf{w}_t; z_{i_t}, z_{j_t}\right)\right), \tag{6}$$

where $\{\eta_t\}_t$ is a step size sequence, and $\nabla f\left(\mathbf{w}_t; z_{i_t}\right)$ and $\nabla g\left(\mathbf{w}_t; z_{i_t}, z_{j_t}\right)$ denote the subgradients of $f\left(\cdot; z_{i_t}\right)$ and $g\left(\cdot; z_{i_t}, z_{j_t}\right)$ at $\mathbf{w}_t$, respectively.

To bound the gradient update process of SGD, it is necessary to presume some properties of the loss functions. For brevity, we just recall some conditions for pointwise loss function (Hardt, Recht, and Singer 2016; Lei and Ying 2020) since the definitions of pairwise setting are analogous.

Definition 4. *A loss function $f : \mathcal{W} \times \mathcal{Z} \to [0, \infty)$ is σ-strongly convex if*

$$f(u) \geq f(v) + \langle \nabla f(v), u - v \rangle + \frac{\sigma}{2}\|u-v\|_2^2, \forall u, v \in \mathcal{W}.$$

Specially, f is convex if $\sigma = 0$.

Clearly, a strongly convex loss function must be convex, but the contrary may not be true. It is well known that convexity is crucial for some optimization analyses of learning algorithms (Hardt, Recht, and Singer 2016; Harvey et al. 2019).

Definition 5. *A loss function $f : \mathcal{W} \times \mathcal{Z} \to [0, \infty)$ is L-Lipschitz if*

$$|f(u) - f(v)| \leq L\|u - v\|_2, \forall u, v \in \mathcal{W}.$$

The above inequality is equivalent to the gradient boundedness of f, i.e. $\|\nabla f(x)\|_2 \leq L$. Thus, the L-Lipschitz continuity assures the boundedness of the gradient update.

Definition 6. *A loss function $f : \mathcal{W} \times \mathcal{Z} \to [0, \infty)$ is β-smooth if*

$$\|\nabla f(u) - \nabla f(v)\|_2 \leq \beta\|u - v\|_2, \forall u, v \in \mathcal{W}.$$

Following the steps in (Hardt, Recht, and Singer 2016; Lei, Ledent, and Kloft 2020), we can verify that the gradient update is non-expansive when f is convex and β-smooth.

Now we present the generalization bounds of SGD for PPL. The proof is given in *Supplementary Material B.2*.

Theorem 2. *Suppose for any $z, \tilde{z} \in \mathcal{Z}$, $f(\mathbf{w}; z)$ and $g(\mathbf{w}; z, \tilde{z})$ are convex, β-smooth and L-Lipschitz with respect to $\mathbf{w} \in \mathcal{W}$. Without loss of generality, let S and S' be different only in the last example. If $\eta_t \leq 2/\beta$, then SGD for PPL with t iterations is PPL γ-uniformly stable with*

$$\gamma \leq 2L^2 \sum_{k=1}^{t} \eta_k \mathbb{I}\left[i_k = n\right] + 2L^2(1-\tau) \sum_{k=1}^{t} \eta_k \mathbb{I}\left[j_k = n\right],$$

where $\mathbb{I}[\cdot]$ is the indicator function.

Type	Reference	Algorithmic stability (γ)	Relationship
Pointwise Learning	Bousquet and Elisseeff (2002)	Uniform Stability	$\mathbb{E}_{Gen} \leq \gamma$.
	Bousquet and Elisseeff (2002)	Hypothesis Stability	$Gen = O(\sqrt{\gamma})$.
	Shalev-Shwartz et al. (2010)	On-average Stability	$\mathbb{E}_{Gen} \leq \gamma$.
	Hardt, Recht, and Singer (2016)	Uniform Stability	$\mathbb{E}_{Gen} \leq \gamma$.
	Feldman and Vondrak (2018)	Uniform Stability	$Gen = O(\sqrt{\gamma} + n^{-\frac{1}{2}})$.
	Feldman and Vondrak (2019)	Uniform Stability	$Gen = O\left(\gamma(\log_2 n)^2 + n^{-\frac{1}{2}}\right)$.
	Lei and Ying (2020)	On-average Stability	$\mathbb{E}_{Gen} \leq L\gamma$.
Pairwise Learning	Lei, Ledent, and Kloft (2020)	Uniform Stability	$Gen = 4\gamma + O(n^{-\frac{1}{2}} + \gamma \log_2 n)$.
	Lei, Liu, and Ying (2021)	ℓ_1 On-average Argument Stability	$\mathbb{E}_{Gen} = O(\gamma^2 + \gamma/n)$.
	Lei, Liu, and Ying (2021)	ℓ_2 On-average Argument Stability	$\mathbb{E}_{Gen} = O(\gamma^2 + \gamma/\sqrt{n})$.
PPL	Ours (Theorem 1)	Uniform Stability	$Gen = (4-2\tau)\gamma + O((4-3\tau)n^{-\frac{1}{2}} + (2-\tau)\gamma \log_2 n)$.
	Ours (Theorem 4)	On-average Loss Stability	$\mathbb{E}_{Gen} \leq \gamma$.
	Ours (Corollary 2)	ℓ_1 On-average Argument Stability	$\mathbb{E}_{Gen} \leq L\gamma$.
	Ours (Theorem 5)	ℓ_2 On-average Argument Stability	$\mathbb{E}_{Gen} = O(\frac{1}{\gamma} + (2-\frac{3}{2}\tau)\gamma)$.

Table 2: Summary of stability-based generalization bounds (γ-stability parameter; Gen-generalization error; $\mathbb{E}_{Gen}$-expected generalization error).

Let $\{\mathbf{w}_t\}$, $\{\mathbf{w}'_t\}$ be generated by SGD on S and S' with $\eta_t = \eta$. Then, for all $\delta \in (0, 1/e)$, the following inequality holds with probability $1-\delta$

$$\begin{aligned}&\left\|\mathbf{w}_{t+1} - \mathbf{w}'_{t+1}\right\|_2\\ \leq& 2L\eta(2-\tau)\Big(\frac{t}{n} + \log(1/\delta) + \sqrt{2n^{-1}t\log(1/\delta)}\Big).\end{aligned}$$

Theorem 2 characterizes the impact of the change of training set on the training loss and the model parameter, which extends the previous related results of pointwise (or pairwise) SGD to the general PPL setting.

We now apply Theorem 1 with $A(S) = \mathbf{w}_T$ where T is the index of the last iteration and the uniform stability bounds to derive the following result.

Corollary 1. *Suppose that $f(\mathbf{w}; z)$ and $g(\mathbf{w}; z, \tilde{z})$ are convex, β-smooth and L-Lipschitz with respect to $\mathbf{w}$, and*

$$\max_{z,\tilde{z}\in\mathcal{Z}} \{|\mathbb{E}_S[f(\mathbf{w}_T; z)]|, |\mathbb{E}_S[g(\mathbf{w}_T; z, \tilde{z})]|\} \leq M$$

for some positive constant M, where $\mathbf{w}_T$ is produced by SGD (6) *at T-th iteration with $\eta_t \equiv c/\sqrt{T} \leq 2/\beta$. Then, for any $\delta \in (0, 1/e)$, the following inequality holds with probability $1-\delta$*

$$\begin{aligned}|R_S(\mathbf{w}_T) - R(\mathbf{w}_T)| =& O\Big(\sqrt{\frac{\log(\frac{1}{\delta})}{n}} + \frac{\sqrt{T}}{n}\log_2 n \log(\frac{1}{\delta})\Big)\\ &+O\Big(T^{-\frac{1}{2}}\log_2 n \log^2(\frac{1}{\delta}) + n^{-\frac{1}{2}}\log_2 n \log^{\frac{3}{2}}(\frac{1}{\delta})\Big).\end{aligned}$$

Remark 6. *Let*

$$\mathbf{w}_R^* = \arg\min_{\mathbf{w}\in\mathcal{W}} R(\mathbf{w}). \tag{7}$$

The excess risk of SGD for PPL is defined as

$$\begin{aligned}R(\mathbf{w}_T) - R(\mathbf{w}_R^*) =& [R(\mathbf{w}_T) - R_S(\mathbf{w}_T)] + [R_S(\mathbf{w}_T)\\ &- R_S(\mathbf{w}_R^*)] + [R_S(\mathbf{w}_R^*) - R(\mathbf{w}_R^*)],\end{aligned}$$

where the first term and the second term of right side are called estimation error (or generalization error) and optimization error, respectively. Theorem 3 provides the upper bound of the first term and the results of (Harvey et al. 2019) imply the bound $O(T^{-\frac{1}{2}}\log_2 T)$ for the optimization error. For the third term $R_S(\mathbf{w}_R^*) - R(\mathbf{w}_R^*)$, we can bound it by Bernstein's inequality. When $T = O(n)$, with probability $1-\delta$ we have

$$R(\mathbf{w}_T) - R(\mathbf{w}_R^*) = O(n^{-\frac{1}{2}}\log_2 n),$$

which is comparable with the convergence analysis of SGD for pairwise learning (Lei, Ledent, and Kloft 2020).

Generalization Bounds of RRM for PPL

Let $r : \mathcal{W} \to [0, \infty)$ be a regularization term for achieving sparsity or preventing over-fitting of learning algorithms associated with $R_S(w)$ defined in (1). The RRM for PPL aims to search the mininizer of

$$F_S(\mathbf{w}) := R_S(\mathbf{w}) + r(\mathbf{w}) \tag{8}$$

over $\mathbf{w} \in \mathcal{W}$. Let

$$\mathbf{w}^* = \arg\min_{\mathbf{w}\in\mathcal{W}} [R(\mathbf{w}) + r(\mathbf{w})] \tag{9}$$

and

$$A(S) = \arg\min_{\mathbf{w}\in\mathcal{W}} F_S(\mathbf{w}). \tag{10}$$

We can verify the uniform stability of PPL with a strongly convex loss function, which is proved in *Supplementary Material B.3*.

Lemma 1. *Assume that A is defined by* (10). *Suppose F_S is σ-strongly convex and the pointwise loss function $f(\cdot; z)$ and pairwise loss function $g(\cdot; z, \tilde{z})$ are both L-Lipschitz. Then, A is $\frac{4L^2}{n\sigma}(2-\tau)$-uniformly stable.*

The following lemma shows the distance between the empirical optimal solution (the best algorithm learned in the training set) and the theoretically optimal solution in expectation.

Lemma 2. *Assume that F_S is σ-strongly convex. If the algorithm A defined in* (10) *is PPL γ-uniformly stable, then*

$$\mathbb{E}_S \|A(S) - \mathbf{w}^*\|_2^2 \leq 4\gamma(2-\tau)/\sigma.$$

A mixed version of Bernstein's inequality from (Hoeffding 1963; Pitcan 2017; Lei, Ledent, and Kloft 2020) is also introduced here, which is used in our error analysis.

Lemma 3. *Assume that*

$$\min_{z,\tilde{z}\in\mathcal{Z}} \{f(\mathbf{w}^*; z), g(\mathbf{w}^*; z, \tilde{z})\} \geq 0,$$

$$\max_{z,\tilde{z}\in\mathcal{Z}} \{f(\mathbf{w}^*; z), g(\mathbf{w}^*; z, \tilde{z})\} \leq b$$

for some constants $b, \theta > 0$, and

$$\max\{Var[f(\mathbf{w}^*; Z)], Var[g(\mathbf{w}^*; Z, \tilde{Z})]\} \leq \theta,$$

where $Var(a)$ denotes the variance of a and $\mathbf{w}^*$ is defined by (9). *Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$ we have*

$$\left|R(\mathbf{w}^*) - R_S(\mathbf{w}^*)\right| \leq \frac{2(1-\tau)b\log(1/\delta)}{3\lfloor n/2\rfloor} + \frac{2\tau b\log(1/\delta)}{3\lfloor n\rfloor} + (1-\tau)\sqrt{\frac{2\theta\log(1/\delta)}{\lfloor n/2\rfloor}} + \tau\sqrt{\frac{2\theta\log(1/\delta)}{\lfloor n\rfloor}},$$

where $\lfloor a\rfloor$ is the biggest integer no more than a.

Next, we firstly derive the upper bounds of the pointwise loss function and the pairwise loss function, and then apply Theorem 1 to get the generalization bounds for PPL with strongly convex objective functions.

Theorem 3. *Assume that $F_S(\mathbf{w})$ is σ-strongly convex, $f(\cdot; z)$ and $g(\cdot; z, \tilde{z})$ are both L-Lipschitz. Under the assumptions of Lemma 3, for the RRM algorithm A defined by* (10) *and any $\delta \in (0, 1/e)$, with probability $1-\delta$ we have*

$$\begin{aligned}
&|R_S(A(S)) - R(A(S))| \\
\leq & \frac{2b\log(1/\delta)}{3\lfloor n\rfloor} + \sqrt{\frac{2\theta\log(1/\delta)}{\lfloor n\rfloor}} + \frac{8L^2}{n\sigma}(2-\tau)^2 \\
&+ e\left(\frac{16L^2}{n\sigma}(2-\tau)(4-3\tau)\sqrt{\log(e/\delta)}\right. \\
&\left. + \frac{96\sqrt{2}L^2}{n\sigma}(2-\tau)^2\lceil\log_2(n)\rceil\log(e/\delta)\right).
\end{aligned}$$

Remark 7. *Note that the excess risk*

$$\begin{aligned}
&R(A(S)) - R(\mathbf{w}_R^*) \\
=&[R(A(S)) - R_S(A(S))] + [R_S(A(S) - R_S(\mathbf{w}_R^*)] \\
&+ [R_S(\mathbf{w}_R^*) - R(\mathbf{w}_R^*)] \\
=&[R(A(S)) - R_S(A(S))] + [R_S(\mathbf{w}_R^*) - R(\mathbf{w}_R^*)] \\
&+ [F_S(A(S) - F_S(\mathbf{w}_R^*)] + r(\mathbf{w}_R^*) - r(A(S)) \\
\leq&[R(A(S)) - R_S(A(S))] + [R_S(\mathbf{w}_R^*) - R(\mathbf{w}_R^*)] \\
&+ r(\mathbf{w}_R^*) - r(A(S)),
\end{aligned}$$

where $\mathbf{w}_R^*$ is defined by (7).

Following the similar proof strategy of Theorem 4, we derive

$$R_S(\mathbf{w}_R^*) - R(\mathbf{w}_R^*) = O\left(\frac{\log(1/\delta)}{\sqrt{n}} + \sqrt{\frac{\theta\log(1/\delta)}{n}}\right)$$

with probability $1-\delta$. When $r(\mathbf{w}_R^*) = O(\sigma\|\mathbf{w}_R^*\|_2^2)$, $\sigma = O(n^{-\frac{1}{2}})$, and

$$\max\{\sup_z (f(\mathbf{w}_R^*; z)), \sup_{z,\tilde{z}} (g(\mathbf{w}_R^*; z, \tilde{z}))\} = O(\sqrt{n}),$$

we have

$$R(A(S)) - R(\mathbf{w}_R^*) = O\left(n^{-\frac{1}{2}}\log_2 n \log(1/\delta)\right)$$

with probability $1-\delta$ based on Theorem 3 and the above decomposition of excess risk.

Remark 8. *We now apply Theorem 3 to the pairwise constraint-guided sparse model (Liu and Zhang 2015), which is inspired from the ℓ_1-penalty and $\ell_{2,1}$-penalty used in Lasso (Tibshirani 2011) and its variants (Zou 2006; Yuan and Lin 2006; Simon et al. 2013; Friedman, Hastie, and Tibshirani 2010). The optimization objective of (Liu and Zhang 2015) can be formulated as $\frac{1}{n}\sum_{i\in[n]} f(\mathbf{w}; z_i) + \frac{\lambda_1}{n(n-1)}\sum_{i,j\in[n]:i\neq j} g(\mathbf{w}; z_i, z_j) + \lambda_2\|\mathbf{w}\|_1$, where $f(\mathbf{w}; z_i)$ is the general least square loss and the pairwise part is measured by*

$$\sum_{(x_i,x_j)\in\mathbf{M}} (\mathbf{w}^T x_i - \mathbf{w}^T x_j)^2 - \lambda_3 \sum_{(x_i,x_j)\in\mathbf{C}} (\mathbf{w}^T x_i - \mathbf{w}^T x_j)^2.$$

Here, $\mathbf{M}$ and $\mathbf{C}$ denote the must-link set and the cannot-link set respectively, and $\lambda_1, \lambda_2, \lambda_3$ are tuneable parameters. It is straightforward to verify that the above objective function is strongly-convex and Lipschitz. Therefore, our theoretical analysis provides the generalization bounds of the PPL model (Liu and Zhang 2015).

Optimistic Generalization Bounds

This subsection further investigates the refined generalization bounds with the help of on-average loss stability in Definition 2 and on-average argument stability in Definition 3. The related proofs can be found in *Supplementary Material B.4*.

Theorem 4. *If A is PPL γ-on-average loss stable, then*

$$\mathbb{E}_S[R(A(S)) - R_S(A(S))] \leq \gamma.$$

As illustrated in Table 2, this quantitative relation is consistent with the previous results for pointwise learning (Lei and Ying 2020) and pairwise learning (Lei, Liu, and Ying 2021). A similar result for ℓ_1 on-average argument stability is stated as follows.

Corollary 2. *Assume that A is PPL ℓ_1 γ-on-average argument stable. If the pointwise loss function $f(\mathbf{w};z)$ and the pairwise loss function $g(\mathbf{w};z,\tilde{z})$ are L-Lipschitz with respect to $\mathbf{w}$, then*

$$\mathbb{E}_{S,A}[R(A(S)) - R_S(A(S))] \leq L\gamma.$$

For completeness, we introduce the following result of ℓ_2 on-average argument stability, which is a natural extension of Theorem 2 part (b) (Lei and Ying 2020) and Theorem 1 (Lei, Liu, and Ying 2021), and removes the requirement on the L-Lipschitz condition of loss functions for PPL.

Theorem 5. *(Lei and Ying 2020; Lei, Liu, and Ying 2021) Let A be PPL ℓ_2 γ-on-average argument stable and $\epsilon > 0$. If $f(\mathbf{w};z)$ and $g(\mathbf{w};z,\tilde{z})$ are nonnegative and β-smooth with respect to $\mathbf{w}$, then*

$$\begin{aligned}&\mathbb{E}_{S,A}[R(A(S)) - R_S(A(S))]\\ \leq&\frac{\beta}{\gamma}\left(\mathbb{E}_{S,A}\left[\tau R_S^{point}(A(S)) + (1-\tau)R_S^{pair}(A(S))\right]\right)\\ &+(\beta+\epsilon)\gamma\left(2-\frac{3}{2}\tau\right).\end{aligned}$$

In the expectation viewpoint, the generalization error can be bounded by the empirical risk and the drift of model parameters induced by the changes of training data, which is illustrated in the following lemma.

Lemma 4. *Assume that the pointwise loss function $f(\mathbf{w};z)$ and the pairwise loss function $g(\mathbf{w};z,\tilde{z})$ are β-smooth with respect to $\mathbf{w}$. Let $\epsilon > 0$ and $\tau \in [0,1]$. Then,*

$$\begin{aligned}&\mathbb{E}\left[R(A(S)) - R_S(A(S))\right]\\ \leq&\frac{\beta\tau\mathbb{E}\left[R_S^{point}(A(S))\right]}{\epsilon} + \frac{\beta(1-\tau)\mathbb{E}\left[R_S^{pair}(A(S))\right]}{\epsilon}\\ &+\frac{(\epsilon+\beta)}{n}\left(2-\frac{3}{2}\tau\right)\sum_{i=1}^{n}\mathbb{E}\left[\|A(S_i) - A(S)\|_2^2\right].\end{aligned}$$

Theorem 6. *Assume that the pointwise loss function $f(\mathbf{w};z)$ and the pairwise loss function $g(\mathbf{w};z,\tilde{z})$ are β-smooth with respect to the first argument. Suppose A is defined by (10) and $\mathbf{w}^*$ is defined by (9). If F_S is σ-strongly convex and $\beta \leq \sigma n/4(2-\tau)$, then*

$$\begin{aligned}&\mathbb{E}_S[F(A(S))] - F(\mathbf{w}^*)\\ \leq&\mathbb{E}_S[R(A(S)) - R_S(A(S))]\\ \leq&\frac{\beta\tau\mathbb{E}\left[R_S^{point}(A(S))\right]}{\epsilon} + \frac{\beta(1-\tau)\mathbb{E}\left[R_S^{pair}(A(S))\right]}{\epsilon}\\ &+\frac{384\tau^2(\epsilon+\beta)\beta}{\sigma^2 n^2}\left(2-\frac{3}{2}\tau\right)\mathbb{E}\left[R_S^{point}(A(S))\right]\\ &+\frac{768(1-\tau)^2(\epsilon+\beta)\beta}{\sigma^2 n^2}\left(2-\frac{3}{2}\tau\right)\mathbb{E}\left[R_S^{pair}(A(S))\right].\end{aligned}$$

Remark 9. *If $r(\mathbf{w}) = O\left(\sigma\|\mathbf{w}\|_2^2\right)$, we can get*

$$\begin{aligned}&\mathbb{E}_S[R(A(S))] - R(\mathbf{w}_R^*)\\ =&O\left(\frac{1}{n\sigma}\left(R^{point}(\mathbf{w}_R^*) + R^{pair}(\mathbf{w}_R^*)\right)\right)\\ &+O\left((\frac{1}{n}+\sigma)\|\mathbf{w}_R^*\|_2^2\right),\end{aligned}$$

where $\mathbf{w}_R^*$ is defined by (7). Furthermore, taking

$$\sigma = \max\left\{\frac{12\beta}{n}, \sqrt{\frac{R^{point}(\mathbf{w}_R^*) + R^{pair}(\mathbf{w}_R^*)}{n\|\mathbf{w}_R^*\|_2^2}}\right\},$$

we can conclude that

$$\begin{aligned}&\mathbb{E}_S[R(A(S))] - R(\mathbf{w}_R^*)\\ =&O\left(\frac{\|\mathbf{w}_R^*\|_2}{\sqrt{n}}\sqrt{R^{point}(\mathbf{w}_R^*) + R^{pair}(\mathbf{w}_R^*)} + \frac{\|\mathbf{w}_R^*\|_2^2}{n}\right).\end{aligned}$$

Moreover, when

$$\max\{R^{point}(\mathbf{w}_R^*), R^{pair}(\mathbf{w}_R^*)\} = O\left(\frac{\|\mathbf{w}_R^*\|_2^2}{n}\right),$$

we get the fast convergence rate

$$\mathbb{E}_S[R(A(S))] - R(\mathbf{w}_R^*) = O\left(\frac{\|\mathbf{w}_R^*\|_2^2}{n}\right).$$

The derived rate $O(n^{-1})$ often is considered as tightness enough in statistical learning theory (Shalev-Shwartz et al. 2010; Hardt, Recht, and Singer 2016).

Remark 10. *It should be noticed that the current result is consistent with the pointwise setting (Lei and Ying 2020) as $\tau = 1$, with the pairiwise setting (Lei, Ledent, and Kloft 2020) as $\tau = 0$. Our convergence analysis of PPL setting covers more complicated learning algorithms (e.g., algorithms described in Table 1) due to the flexibility of $\tau \in [0,1]$.*

Conclusion

This paper focuses on establishing the generalization bounds of PPL by means of algorithmic stability analysis. After characterizing the quantitative relationship between generalization error and algorithmic stability, we establish the upper bounds of excess risk of SGD and RRM for PPL. Our stability-based analysis fills the gap of statistical learning theory in part for the related PPL algorithms. In the future, it is interesting to further investigate the stability-based generalization of SGD for PPL under non-i.i.d sampling, e.g., Markov chain sampling (Sun, Sun, and Yin 2018; Wang et al. 2022b).

Acknowledgments

This work was supported in part by National Natural Science Foundation of China under Grant No. 12071166 and by the Fundamental Research Funds for the Central Universities of China under Grant 2662020LXQD002. We are grateful to the anonymous AAAI reviewers for their constructive comments.

References

Agarwal, S.; and Niyogi, P. 2009. Generalization bounds for ranking algorithms via algorithmic stability. *Journal of Machine Learning Research*, 10: 441–474.

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

Bousquet, O.; Klochkov, Y.; and Zhivotovskiy, N. 2020. Sharper bounds for uniformly stable algorithms. In *Conference on Learning Theory (COLT)*, 610–626.

Cao, Q.; Guo, Z.-C.; and Ying, Y. 2016. Generalization bounds for metric and similarity learning. *Machine Learning*, 102: 115–132.

Chen, H.; Wang, X.; Deng, C.; and Huang, H. 2017. Group sparse additive machine. In *Advances in Neural Information Processing Systems (NeurIPS)*, 198–208.

Chen, H.; Wang, Y.; Zheng, F.; Deng, C.; and Huang, H. 2021. Sparse modal additive model. *IEEE Transactions on Neural Networks and Learning Systems*, 32(6): 2373–2387.

Clémençon, S.; Lugosi, G.; and Vayatis, N. 2008. Ranking and empirical minimization of U-statistics. *The Annals of Statistics*, 36: 844–874.

Deng, Z.; He, H.; and Su, W. 2021. Toward better generalization bounds with locally elastic stability. In *International Conference on Machine Learning (ICML)*, 2590–2600.

Elisseeff, A.; Evgeniou, T.; Pontil, M.; and Kaelbing, L. P. 2005. Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6: 55–79.

Feldman, V.; and Vondrak, J. 2018. Generalization bounds for uniformly stable algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 9770–9780.

Feldman, V.; and Vondrak, J. 2019. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Conference on Learning Theory (COLT)*, 1270–1279.

Friedman, J. H.; Hastie, T.; and Tibshirani, R. 2010. A note on the group lasso and a sparse group lasso. arXiv:1001.0736.

Hardt, M.; Recht, B.; and Singer, Y. 2016. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning (ICML)*, 1225–1234.

Harvey, N. J.; Liaw, C.; Plan, Y.; and Randhawa, S. 2019. Tight analyses for non-smooth stochastic gradient descent. In *Conference on Learning Theory (COLT)*, 1579–1613.

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

Huang, Z.; Feng, S.; Su, W.; Chen, X.; Wang, S.; Liu, J.; Ouyang, X.; and Sun, Y. 2020. ERNIE at SemEval-2020 Task 10: Learning word emphasis selection by pre-trained language model. In *International Committee on Computational Linguistics (ICCL)*, 1456–1461.

Klochkov, Y.; and Zhivotovskiy, N. 2021. Stability and deviation optimal risk bounds with convergence rate $O(1/n)$. In *Advances in Neural Information Processing Systems (NeurIPS)*, 5065–5076.

Lei, Y.; Ledent, A.; and Kloft, M. 2020. Sharper generalization bounds for pairwise learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 21236–21246.

Lei, Y.; Li, W.; Lu, Z.; and Zhao, M. 2017. Alternating pointwise-pairwise learning for personalized item ranking. In *Conference on Information and Knowledge Management*, 2155–2158.

Lei, Y.; Liu, M.; and Ying, Y. 2021. Generalization guarantee of SGD for pairwise learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 21216–21228.

Lei, Y.; and Tang, K. 2021. Learning rates for stochastic gradient descent with nonconvex objectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43: 4505–4511.

Lei, Y.; and Ying, Y. 2020. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *International Conference on Machine Learning (ICML)*, 5809–5819.

Liu, M.; and Zhang, D. 2015. Pairwise constraint-guided sparse learning for feature selection. *IEEE Transactions on Cybernetics*, 46: 298–310.

Liu, T.; Lugosi, G.; Neu, G.; and Tao, D. 2017. Algorithmic stability and hypothesis complexity. In *International Conference on Machine Learning (ICML)*, 2159–2167.

London, B.; Huang, B.; and Getoor, L. 2016. Stability and generalization in structured prediction. *Journal of Machine Learning Research*, 17: 7808–7859.

Pitcan, Y. 2017. A note on concentration inequalities for U-statistics. arXiv:1712.06160.

Rejchel, W. 2012. On ranking and generalization bounds. *Journal of Machine Learning Research*, 13: 1373–1392.

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

Simon, N.; Friedman, J.; Hastie, T.; and Tibshirani, R. 2013. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22: 231–245.

Smale, S.; and Zhou, D.-X. 2007. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26: 153–172.

Sun, T.; Li, D.; and Wang, B. 2021. Stability and generalization of decentralized stochastic gradient descent. In *AAAI Conference on Artificial Intelligence*, 9756–9764.

Sun, T.; Sun, Y.; and Yin, W. 2018. On Markov chain gradient descent. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Tibshirani, R. 2011. Regression shrinkage and selection via the lasso: A retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73: 273–282.

Vapnik, V.; Levin, E.; and Le Cun, Y. 1994. Measuring the VC-dimension of a learning machine. *Neural Computation*, 6: 851–876.

Wang, M.; Guo, Y.; Zhao, Z.; Hu, G.; Shen, Y.; Gong, M.; and Torr, P. 2022a. MP2: A momentum contrast approach for recommendation with pointwise and pairwise learning. arXiv:2204.08326.

Wang, P.; Lei, Y.; Ying, Y.; and Zhou, D.-X. 2022b. Stability and generalization for Markov chain stochastic gradient methods. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Wang, Y.; Chen, H.; Zheng, F.; Xu, C.; Gong, T.; and Chen, Y. 2020. Multi-task additive models for robust estimation and automatic structure discovery. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Wang, Y.; Wang, S.; Tang, J.; Liu, H.; and Li, B. 2016. PPP: Joint pointwise and pairwise image label prediction. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 6005–6013.

Yang, Z.; Lei, Y.; Wang, P.; Yang, T.; and Ying, Y. 2021. Simple Stochastic and Online Gradient Descent Algorithms for Pairwise Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 20160–20171.

Yin, D.; Kannan, R.; and Bartlett, P. 2019. Rademacher complexity for adversarially robust generalization. In *International Conference on Machine Learning (ICML)*, 7085–7094.

Ying, Y.; Wen, L.; and Lyu, S. 2016. Stochastic online AUC maximization. In *Advances in Neural Information Processing Systems (NIPS)*.

Ying, Y.; and Zhou, D.-X. 2016. Online pairwise learning algorithms. *Neural Computation*, 28: 743–777.

Yuan, M.; and Lin, Y. 2006. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68: 49–67.

Zhou, D.-X. 2002. The covering number in learning theory. *Journal of Complexity*, 18: 739–767.

Zhuo, J.; Zhu, Q.; Yue, Y.; and Zhao, Y. 2022. Learning explicit user interest boundary for recommendation. In *ACM Web Conference*, 193–202.

Zou, H. 2006. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101: 1418–1429.