Trajectory-Dependent Generalization Bounds for Pairwise Learning with φ-Mixing Samples

Liyuan Liu[1], **Hong Chen**[1,2,*], **Weifu Li**[1,2], **Tieliang Gong**[3], **Hao Deng**[1] and **Yulong Wang**[1]
[1]College of Informatics, Huazhong Agricultural University, Wuhan 430070, China
[2]Engineering Research Center of Intelligent Technology for Agriculture, Ministry of Education, Wuhan 430070, China
[3]School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China
liulymt@foxmail.com, chenh@mail.hzau.edu.cn

Abstract

Recently, the mathematical tool from fractal geometry (i.e., fractal dimension) has been employed to investigate optimization trajectory-dependent generalization ability for some pointwise learning models with independent and identically distributed (i.i.d.) observations. This paper goes beyond the limitations of pointwise learning and i.i.d. samples, and establishes generalization bounds for pairwise learning with uniformly strong mixing samples. The derived theoretical results fill the gap of trajectory-dependent generalization analysis for pairwise learning, and can be applied to wide learning paradigms, e.g., metric learning, ranking and gradient learning. Technically, our framework brings concentration estimation with Rademacher complexity and trajectory-dependent fractal dimension together in a coherent way for felicitous learning theory analysis. In addition, the efficient computation of fractal dimension can be guaranteed for random algorithms (e.g., stochastic gradient descent algorithm for deep neural networks) by bridging topological data analysis tools and the trajectory-dependent fractal dimension.

1 Introduction

As a popular learning paradigm, pairwise learning has attracted much attention in the machine learning community, where the loss function is associated with two samples simultaneously [Lei *et al.*, 2018; Huang *et al.*, 2023]. Typical frameworks of pairwise learning include metric learning [Cao *et al.*, 2016], gradient learning [Mukherjee and Zhou, 2006; Feng *et al.*, 2015], AUC maximization [Cortes and Mohri, 2003; Gao and Zhou, 2015], and ranking [Rejchel, 2012; Huang *et al.*, 2023]. In theory, the generalization performance of pairwise learning has been well investigated from various analysis routines, such as Rademacher complexity [Agarwal and Niyogi, 2009; Lei *et al.*, 2021], algorithmic stability [Jin *et al.*, 2009; Lei *et al.*, 2021], integral operators [Mukherjee and Zhou, 2006; Ying and Zhou, 2016; Zhao *et al.*, 2017], etc. However, all the existing results focus on the uniform convergence analysis independent of the real optimization algorithm, which often ignores the crucial impact of computing trajectory on the capacity of hypothesis space [Şimşekli *et al.*, 2021].

Recently, the computing trajectory of learning model has been addressed increasingly in learning theory studies due to its close relationship with generalization performance. For deep neural networks (DNNs), the existing generalization analysis often focuses on the final trained network [Neyshabur *et al.*, 2017], where the computing trajectory is ignored usually. In view of this situation, Şimşekli *et al.* [2021] have demonstrated that algorithmic trajectories can exhibit a fractal structure when a stochastic optimization algorithm is employed for implementing deep learning models. Moreover, fractal dimensions w.r.t optimization trajectories have been integrated into the generalization error analysis, where the generalization bounds rely on the intrinsic computing-dependent dimension rather than the hypothesis space dimension [Camuto *et al.*, 2021; Birdal *et al.*, 2021; Hodgkinson *et al.*, 2022]. To illustrate the recent progress clearly, we summarize their main contributions in Table 1.

Indeed, these trajectory-dependent generalization guarantees provide some explanation for the phenomena: why over-parameterized networks may not exhibit overfitting in practical scenarios [Dupuis *et al.*, 2023]. When the loss function ℓ is L-Lipschitz and uniformly bounded by a constant B, the previous results in Şimşekli *et al.* [2021], Camuto *et al.* [2021], Birdal *et al.* [2021] and Sachs *et al.* [2023] stated that the generalization error of learning models can be bounded by $LB\sqrt{\frac{M^d(\mathcal{W}_{\mathbf{Z},U})\log(n)+\log(1/\zeta)}{n}}$ with probability $1-\zeta$, where n is the sample size and $M^d(\mathcal{W}_{\mathbf{Z},U})$ is the fractal dimension with trajectory-dependent hypothesis space $\mathcal{W}_{\mathbf{Z},U}$ (see Section 3 for detail definitions). Moreover, [Dupuis *et al.*, 2023] got the refined generalization bounds based on data-depend fractal dimensions without Lipschitz assumption for loss function. Although these algorithm-dependent bounds have brought a new perspective on understanding generalization, they are limited to the pointwise learning with i.i.d. samples which often violated in real-world applications [Kohler and Krzyzak, 2020; Kurisu *et al.*, 2022].

This paper focuses on the scenario where the observations are drawn from a stationary φ-mixing (or uniformly strong mixing) process, which is a commonly used assumption in learning theory [Yu, 1994; Meir, 2000; He *et al.*, 2016]. The-

*Corresponding author.

Ref.	Main contribution
Şimşekli *et al.* [2021]	Introduce the fractal tools to generalization analysis
Camuto *et al.* [2021]	Study the generalization bound under stationary distribution
Birdal *et al.* [2021]	Use the persistent homology to calculate fractal dimension
Sachs *et al.* [2023]	Propose an algorithm-dependent Rademacher complexity
Dupuis *et al.* [2023]	Propose a pseudo-metric to remove the Lipschitz condition
Ours	Study the pairwise learning with φ-mixing samples

Table 1: Generalization analysis with Fractal dimension.

oretically, we establish generalization bounds for pairwise learning with φ-mixing samples with the help of trajectory-based fractal dimensions, which assures the generalization error of our estimator can achieve the polynomial decay rate $\mathcal{O}\big(\Psi\sqrt{\log(n)/n}\big)$. Here n is the sample size and Ψ is the sum of the φ-mixing coefficients (see Definition 1 in Section 3.1). To the best of our knowledge, there is no theoretical characterization of the generalization guarantees of pairwise learning with φ-mixing samples before. The main contributions of this paper are twofold:

- *Optimization-dependent generalization bounds.* We employ the fractal tool to bound the trajectory-dependent generalization error for pairwise learning, where the tight estimations can be derived benefit from much smaller hypothesis space. This change has enabled us to transform our focus from concerning the size of covering numbers (independently of optimization processes, e.g., Rejchel [2012] and Huang *et al.* [2023]) to fractal dimensions by topological data analysis tools Şimşekli *et al.* [2021] and Dupuis *et al.* [2023]. Our results reduce the gap between the theoretical properties and algorithmic implementation of pairwise learning models, which improve the matching degree of generalization bounds to real applications.
- *Improved analysis techniques.* Without the requirement of independent observations, the previous analysis techniques in Lei *et al.* [2018] and Huang *et al.* [2023] can not be applied to the pairwise learning with φ-mixing samples directly. This obstacle is conquered through McDiarmid's inequality for φ-mixing processes [Mohri and Rostamizadeh, 2010] and extending Lemma A.1 in Clémençon *et al.* [2008] to stationary processes. In particular, our analysis framework removes the Lipschitz condition of pairwise loss [Cao *et al.*, 2016; Lei *et al.*, 2018] by incorporating a data-dependent pseudo-metric for capacity estimation.

2 Related Work

Now we review the related work about generalization bounds for pairwise learning and learning theory with φ-mixing samples. To better evaluate our work, we compare it with the related results in Table 2 from the aspects of analysis tool, learning models, and learning rate.

2.1 Generalization Bounds for Pairwise Learning

Stability Analysis. Due to the intrinsic relationship between generalization and stability, there are some generalization guarantees of pairwise learning in terms of algorithmic stability tools, see e.g., Jin *et al.* [2009], Agarwal and Niyogi [2009] and Lei *et al.* [2021]. The derived results demonstrate that, under the strong convexity of the objective function, the related pairwise learning models enjoy the dimensional-independent generalization bounds with the decay rate at $\mathcal{O}(n^{-\frac{1}{2}})$. All of the existing studies require that the learned pairwise model would change slightly if a single training example is replaced by another one, and include ranking algorithm [Agarwal and Niyogi, 2009], the iterative localized algorithm for pairwise learning [Lei *et al.*, 2021] as special examples.

Uniform Convergence Analysis. The concentration estimation techniques in U-statistic were developed to tackle error analysis of regularized metric learning [Cao *et al.*, 2016] and ranking [Rejchel, 2012; Huang *et al.*, 2023]. The basic idea is to control the generalization bound via a Rademacher complexity [Lei *et al.*, 2018] which uses the symmetry of the U-statistic [Clémençon *et al.*, 2008]. Moreover, there is also a classic method using Hoeffding decomposition [Hoeffding, 1992] to decompose the pairwise learning problem into i.i.d. terms and degenerate U-statistic terms [Huang *et al.*, 2023]. Under proper capacity assumptions of hypothesis function space and independence of samples, convergence rate with $\mathcal{O}(n^{-\frac{1}{2}})$ is obtained for the above pairwise learning models. However, it is often difficult to verify the independence assumptions directly in real applications, which limits the adaptivity of learning theory analysis. For online pairwise learning, [Wang *et al.*, 2012] and [Kar *et al.*, 2013] established regret bounds by developing analysis techniques associated with covering numbers and Rademacher complexity, respectively. Furthermore, the convergence rates for the final iteration of online pairwise learning algorithms have been examined through the lens of convex analysis [Ying and Zhou, 2016; Guo *et al.*, 2016; Lin *et al.*, 2017]. Despite rapid progress, all the uniform convergence analysis doesn't address the optimization trajectories and just states the computation-independent error bounds.

Operator Approximation Analysis. The operator approximation technique, from functional analysis, is introduced to characterize the generalization ability of least square regularized ranking [Chen, 2012] and gradient learning [Mukherjee and Zhou, 2006]. Compared with the uniform convergence depending heavily on the capacity assumption of hypothesis space, the learning theory analysis framework of operator approximation just requires that the minimizer of optimization objectives can be expressed by the empirical version of integral operator [Mukherjee and Zhou, 2006; Kriukova *et al.*, 2016]. Explicitly, for the least square ranking, [Chen, 2012] gave the first operator approximation analysis on its convergence rate and Chen *et al.* [2013] stated the decay rate $\mathcal{O}(n^{-\frac{1}{4}})$ for the excess risk of stochastic gradient descent ranking under mild parameter conditions. Moreover,

Ref.	Technique	Model	Samples	Convergence rate
Mukherjee and Zhou [2006]	Integral operator	Gradient learning	i.i.d.	$\mathcal{O}\big(n^{-\tau/2(n+2+3\tau)}\big)$
Jin *et al.* [2009]	Algorithmic stability	Metric learning	i.i.d.	$\mathcal{O}\Big(1/\sqrt{n}\Big)$
Agarwal and Niyogi [2009]	Algorithmic stability	Ranking	i.i.d.	$\mathcal{O}\Big(1/\sqrt{n}\Big)$
Rejchel [2012]	Rademacher complexity	Ranking	i.i.d.	$\mathcal{O}\Big(1/\sqrt{n}\Big)$
Wang *et al.* [2012]	Covering number	Online learning	i.i.d.	$\mathcal{O}\Big(1/\sqrt{n}\Big)$
Ying and Zhou [2016]	Integral operator	Online learning	i.i.d.	$\mathcal{O}\Big(1/n\Big)$
Guo *et al.* [2016]	Rademacher complexity	Online learning	i.i.d.	$\mathcal{O}\Big(\log n/n\Big)$
Cao *et al.* [2016]	Rademacher complexity	Metric learning	i.i.d.	$\mathcal{O}\Big(1/\sqrt{n}\Big)$
Zhao *et al.* [2017]	Integral operator	Ranking	i.i.d.	$\mathcal{O}\Big(1/\sqrt{n}\Big)$
Lei *et al.* [2018]	Rademacher complexity	Pairwise learning	i.i.d.	$\mathcal{O}\Big(l/\sqrt{n\beta}\Big)$
Lei *et al.* [2021]	Algorithmic stability	Pairwise learning	i.i.d.	$\mathcal{O}\Big(1/\sqrt{n}\Big)$
Huang *et al.* [2023]	Hoeffding decomposition	Ranking	i.i.d.	$\mathcal{O}\Big(\big(\log^2 n/n\big)^{\frac{r(\theta+1)}{2d+r(\theta+2)}}\Big)$
Ours	Rademacher complexity	Pairwise learning	φ-mixing	$\mathcal{O}\big(\Psi\sqrt{\log(n)/n}\big)$

Table 2: Convergence analysis of pairwise learning $\big(n$: sample size, β: strong convex coefficients, $\Psi = 1 + \sum_{i=1}^{n} \varphi(k)$, $\varphi(k)$: the φ-mixing coefficients, r, d, θ: the parameter of hypothesis space, τ: the parameter of data distribution $\big)$

[Kriukova *et al.*, 2016] obtained the improved estimations $\mathcal{O}(n^{-\frac{1}{2}})$ under general source conditions, and [Zhao *et al.*, 2017] got the same convergence rate by integrating Hoeffding's decomposition and regression estimation techniques.

2.2 Learning Theory With φ-Mixing Samples

Stationary Mixing Processes. The pioneering work of Yu *et al.* [1994] led to VC-dimension bounds under the assumption of stationary φ-mixing observations. Later, [Meir, 2000] derived the upper bounds of generalization error in terms of covering numbers and reached the convergence rate with $\mathcal{O}((\log n/n)^{\frac{\gamma_0}{2(1+\gamma_0)}})$, where γ_0 is the parameter in φ-mixing coefficient. Moreover, [Mohri and Rostamizadeh, 2010] established the data-dependent learning bounds in terms of algorithmic stability, where the blocks of points for Hoeffding's decomposition are not necessarily of equal size and the generalized McDiarmid's inequality in Leo and Kavita [2008] plays a crucial role. [Ralaivola *et al.*, 2010] and [Alquier and Wintenberger, 2012] provided PAC-Bayesian learning bounds under the φ-mixing assumptions. Additionally, Ralaivola *et al.* [2010] employed fractional covers to address the dependencies within the given set of random variables, and proposed a strategy to construct subsets of independent random variables, enabling the application of the standard i.i.d. PAC-Bayes bound. For the pairwise ranking, [He *et al.*, 2016] provided the generalization bounds with φ-mixing samples by combining the independent block technique and the algorithmic stability analysis together. However, the information of computing trajectory isn't addressed sufficiently in the previous analysis for dependent observations.

Non-stationary Mixing Processes. For the the asymptotically stationary process, [Agarwal and Duchi, 2012] presented generalization bounds of stable online learning algorithms. Moreover, [Kuznetsov and Mohri, 2017] provided the first generalization bounds for time series prediction with a non-stationary φ-mixing stochastic process. For the Lasso estimator with φ-mixing samples, [Peng *et al.*, 2023] employed the non-asymptotic concentration inequalities to the error estimations. As far as known, there is no generalization guarantee specifically addressing pairwise learning with φ-mixing samples based on algorithmic trajectory.

3 Preliminaries

This section introduces the framework of pairwise learning in φ-mixing process and recalls the definition of Minkowski dimension.

3.1 Pairwise Learning in φ-Mixing Process

Let the input space $\mathcal{X} \subset \mathbb{R}^p$ be a compact space and the output set $\mathcal{Y}$ be a subset of $\mathbb{R}$. The sample set

$$\mathbf{Z} = \{\mathbf{z}_i = (\mathbf{x}_i, y_i) \in \mathcal{Z}\}_{i=1}^{n}$$

is drawn from an unknown distribution. Usually, the objective of a learning algorithm is to learn a prediction function $f_{\mathbf{w}}$ based on $\mathbf{Z}$. We consider parametric models where $f_{\mathbf{w}}$ can be indexed by a vector element $\mathbf{w} \in \mathcal{W} \subset \mathbb{R}^d$. Recognizing the limitations of the independent assumption in practical scenarios [Mohri and Rostamizadeh, 2010], we opt to discard it and instead investigate the set $\mathbf{Z}$ satisfying the exponentially φ-mixing condition.

Definition 1. *[Yu, 1994] The process $\{\mathbf{z}_i\}_{i=1}^{\infty}$ is said to be exponentially φ-mixing or uniformly strong mixing if there*

exist some constants $b_0 > 0, c_0 \geq 0, \gamma_0 > 0$ *such that* φ*-mixing coefficient* $\varphi(k)$ *satisfying*

$$\varphi(k) = \sup_{A\in\sigma_{m+k}^{\infty}, B\in\sigma_1^m} |\mathbb{P}(A \mid B) - \mathbb{P}(A)| \leq b_0 \exp(-c_0 k^{\gamma_0})$$

where σ_i^j *is the* σ*-algebra generated by* $\mathbf{z}_i, \cdots, \mathbf{z}_j$.

A classical φ-mixing example can be seen in *Technical appendix A.1*. In this paper, we focus on the pairwise learning problem on an example pair $(\mathbf{z}_i, \mathbf{z}_j)$, in which the effectiveness of the prediction function $f_{\mathbf{w}}$ relies on a non-negative loss function $\ell(\mathbf{w}, \mathbf{z}_i, \mathbf{z}_j)$. To evaluate the ability of $\mathbf{w}$ learned from set $\mathbf{Z}$, we define the empirical risk

$$R_n(\mathbf{w}, \mathbf{Z}) = \frac{1}{n(n-1)} \sum_{i\neq j} \ell(\mathbf{w}, \mathbf{z}_i, \mathbf{z}_j)$$

and its population version

$$R(\mathbf{w}) = \int_{\mathcal{Z}} \int_{\mathcal{Z}} \ell(\mathbf{w}, \mathbf{z}, \hat{\mathbf{z}}) \mathrm{d}\mathbf{z}\mathrm{d}\hat{\mathbf{z}}.$$

Typically, the optimal parameter $\mathbf{w}^\star$ can be found using a learning algorithm $\mathcal{A}$ to minimize the empirical risk $R_n(\mathbf{w}, \mathbf{Z})$. As mentioned in Molchanov [2005], the algorithm $\mathcal{A}$ can be thought as a measurable map that generates the *trajectories* $\mathcal{W}_{\mathbf{Z},U}$ (hypothesis space) from set $\mathbf{Z}$ and an external random variable $U \in \mathcal{U}$. In general, U is assumed to be independent of $\mathbf{Z}$ and accounting for the randomness of the learning algorithm (such as the batch indices in training).

Definition 2. *The trajectories of the algorithm* $\mathcal{A}$ *after* T *iterations under the sample set* $\mathbf{Z}$ *can be defined as*

$$\mathcal{W}_{\mathbf{Z},U} = \{\mathbf{w}_t\}_{t=0}^{T},$$

where the parameter $\mathbf{w}_t$ *is returned by* $\mathcal{A}$ *at time* t.

A common example of $\mathcal{A}$ is the stochastic gradient descent (SGD) algorithm, which can be viewed as a discretization of a stochastic differential equation [Mandt *et al.*, 2016; Jastrzebski *et al.*, 2017; Chaudhari and Soatto, 2018]:

$$\mathrm{d}\mathbf{w}_t = -\nabla_{\mathbf{w}} R_n(\mathbf{w}_t, \mathbf{Z})\mathrm{d}t + \Sigma(\mathbf{w}_t)\, \mathrm{d}B_t$$

where the second term accounts for randomness U coming from Brownian motion B_t and diffusion coefficient $\Sigma : \mathbb{R}^d \to \mathbb{R}^{d\times d}$. As considered in various studies [Mandt *et al.*, 2016; Şimşekli *et al.*, 2021; Chaudhari and Soatto, 2018], the *trajectories* $\mathcal{W}_{\mathbf{Z},U}$ of SGD algorithm is the set $\{\mathbf{w}_t\}_{t\geq 0}$ with the parameter $\mathbf{w}_t$ returned by $\mathcal{A}$ at time t.

On this basis, we define the worst-case generalization error which reflects the generalization ability

$$\mathcal{G}(\mathbf{w}, \mathbf{Z}) = \sup_{\mathbf{w}\in\mathcal{W}_{\mathbf{Z},U}} \big(R(\mathbf{w}) - R_n(\mathbf{w}, \mathbf{Z})\big), \tag{1}$$

which is a trajectory-dependent definition of generalization and widely used in the literature [Bousquet and Elisseeff, 2002; Dupuis *et al.*, 2023].

3.2 Minkowski Dimension

In this paper, we employ the Minkowski dimension to measure the capacity of the hypothesis space $\mathcal{W}_{\mathbf{Z},U}$ and bound the worst-case generalization error (1). Before proceeding, we first propose a data-dependent pseudo-metric $d_{\mathbf{Z}}(\cdot,\cdot)$ on space $\mathcal{W}_{\mathbf{Z},U}$ as follows:

$$d_{\mathbf{Z}}(\mathbf{w}, \hat{\mathbf{w}}) = \max_{\pi\in\Pi} \left\lceil \frac{2}{n} \right\rceil \sum_{i=1}^{\lfloor n/2 \rfloor} \big|\ell\big(\mathbf{w}, \mathbf{z}_{\pi(i)}, \mathbf{z}_{\pi(\lfloor n/2\rfloor + i)}\big) - \ell\big(\hat{\mathbf{w}}, \mathbf{z}_{\pi(i)}, \mathbf{z}_{\pi(\lfloor n/2\rfloor + i)}\big)\big|, \tag{2}$$

where $\lfloor n/2 \rfloor$ rounds down $n/2$, $\lceil n/2 \rceil$ rounds up $n/2$ and Π is the set of all permutations of $\{1, \cdots, n\}$. The fundamental requirements of pseudo-metric (2) which include the triangle inequality, symmetry, and non-negativity, have been verified and demonstrated in *Technical appendix A.1*. Based on this pseudo-metric, we present the definition of the covering number.

Definition 3. *[Şimşekli* et al.*, 2021] A set* $\{\mathbf{w}_l\}_{l=1}^N$ *is the* δ*-cover of pseudo-metric space* $(\mathcal{W}_{\mathbf{Z},U}, d_{\mathbf{Z}})$*, if* $\forall \mathbf{w} \in \mathcal{W}_{\mathbf{Z},U}$*, there exists* $\mathbf{w}_k \in \{\mathbf{w}_l\}_{l=1}^N$ *such that*

$$d_{\mathbf{Z}}(\mathbf{w}_k, \mathbf{w}) \leq \delta.$$

The covering number $\mathcal{N}(\mathcal{W}_{\mathbf{Z},U}, d_{\mathbf{Z}}, \delta)$ *is the minimum cardinal of all* δ*-covers.*

As demonstrated in Dupuis *et al.* [2023], the δ-cover induced by pseudo-metric $d_{\mathbf{Z}}(\cdot,\cdot)$ is measurable which can induce the correct Minkowski dimension.

Definition 4. *[Falconer, 2004] For a bounded pseudo-metric space* $(\mathcal{W}_{\mathbf{Z},U}, d_{\mathbf{Z}})$*, the Minkowski dimension is defined as:*

$$M^{d_{\mathbf{Z}}}(\mathcal{W}_{\mathbf{Z},U}) = \lim_{\delta\to 0} \frac{\log(\mathcal{N}(\mathcal{W}_{\mathbf{Z},U}, d_{\mathbf{Z}}, \delta))}{\log(1/\delta)}. \tag{3}$$

Besides, the upper Minkowski dimension and the lower Minkowski dimension are respectively defined as :

$$\overline{M}^{d_{\mathbf{Z}}}(\mathcal{W}_{\mathbf{Z},U}) = \limsup_{\delta\to 0} \frac{\log(\mathcal{N}(\mathcal{W}_{\mathbf{Z},U}, d_{\mathbf{Z}}, \delta))}{\log(1/\delta)},$$

and

$$\underline{M}^{d_{\mathbf{Z}}}(\mathcal{W}_{\mathbf{Z},U}) = \liminf_{\delta\to 0} \frac{\log(\mathcal{N}(\mathcal{W}_{\mathbf{Z},U}, d_{\mathbf{Z}}, \delta))}{\log(1/\delta)}$$

which will be used in Section 4.2.

4 Main Results

This section presents our main theoretical results. The following assumptions are involved in our error analysis.

Assumption 1. *The loss function* $\ell : \mathcal{W}_{\mathbf{Z},U} \times \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ *is continuous and bounded by* B.

Assumption 1 requires the continuity of loss function, which ensures that the quantities like (1) are well-defined random variables [Molchanov, 2005]. The bounded condition is standard in many generalization error estimations, see e.g., Dupuis *et al.* [1999] and Schmidt-Hieber [2017]. Different from Şimşekli *et al.* [2021] and Camuto *et al.* [2021], we eliminate the necessity of the Lipschitz condition with respect to $\mathbf{w}$ for loss function ℓ due to the utilization of pseudo-metric (2).

Assumption 2. *The strictly stationary process $\{\mathbf{z}_i\}_{i\geq 1}$ satisfies the exponentially φ-mixing condition.*

Assumption 2 imposes the dependence constraint among input variables. To handle this type of data, the generalized McDiarmid's inequality for φ-mixing observations is introduced in Leo and Kavita [2008] and Mohri and Rostamizadeh [2010].

Assumption 3. *[Dupuis* et al.*, 2023] Let $C \subset \mathbb{R}^d$ be any closed set, $\delta > 0, \mathbf{Z} \in \mathcal{Z}^n$ and $\mathbf{Z}' \in \mathcal{Z}^m$. We can construct minimal δ-coverings $\mathcal{N}(C \cap \mathcal{W}_{\mathbf{Z},U}, d_{\mathbf{Z}}, \delta)$, which are random finite sets with respect to the product σ-algebra $\mathcal{F}_{\mathbf{Z}} \otimes \mathcal{F}_{\mathbf{Z}'} \otimes \mathcal{F}_U$ (measurability with respect to $\mathbf{Z}, \mathbf{Z}', U$).*

Assumption 3 can be cast as a selection property. Indeed, there may be a wide range of possible minimal coverings for each realization of $(\mathbf{Z}, \mathbf{Z}', U)$. This paper assumes that we can select one of them satisfying that the obtained random set is measurable. In addition, since the Minkowski dimension (3) can be expressed as a countable limit, Assumption 3 implies that $M^{d_{\mathbf{Z}}}(\mathcal{W}_{\mathbf{Z},U})$ is a random variable [Dupuis *et al.*, 2023]. Please see Dupuis *et al.* [2023] for more detailed discussions.

In our analysis, the capacity of searching space is measured by the Rademacher complexity and Gaussian complexity, which are defined as follows.

Definition 5. *[Dupuis* et al.*, 2023] Given the sample set $\mathbf{Z} = \{\mathbf{z}_i\}_{i=1}^n$, the parameter space $\mathcal{W}$, and the function $l_{\mathbf{w}}$ indexed by $\mathbf{w} \in \mathcal{W}$, the empirical Rademacher complexity over $\mathcal{W}$ on $\mathbf{Z}$ can be defined as*

$$\mathfrak{R}(\mathbf{Z}, \mathcal{W}) = \frac{1}{n} E_{\boldsymbol{\varepsilon}} \left[\sup_{\mathbf{w}\in\mathcal{W}} \sum_{i=1}^{n} \varepsilon_i l_{\mathbf{w}}(\mathbf{z}_i) \right] \tag{4}$$

where the set $\boldsymbol{\varepsilon} = \{\varepsilon_i\}_{i=1}^n$ is composed of Rademacher random variables satisfying $\mathbb{P}(\varepsilon_i = 1) = \mathbb{P}(\varepsilon_i = -1) = 0.5$. Similarly, the empirical Gaussian complexity over $\mathcal{W}$ on $\mathbf{Z}$ is defined as

$$\Gamma(\mathbf{Z}, \mathcal{W}) = \frac{1}{n} E_{\mathbf{g}} \left[\sup_{\mathbf{w}\in\mathcal{W}} \sum_{i=1}^{n} g_i l_{\mathbf{w}}(\mathbf{z}_i) \right] \tag{5}$$

where $g_1, \cdots, g_n$ are independent Gaussian random variables.

4.1 Upper Bound of Generalization Error

Now we introduce the following stepping-stone lemma, which is an extension of Lemma A.1. in Clémençon *et al.* [2008] and plays a crucial role in our error estimations.

Lemma 1. *Let $q_\tau : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ be real-valued functions indexed by τ belonging to a set T. If $\{\mathbf{z}_i\}_{i=1}^n$ is a stationary process, then*

$$\begin{aligned} &E \sup_{\tau\in T} \frac{1}{n(n-1)} \sum_{i\neq j} q_\tau(\mathbf{z}_i, \mathbf{z}_j) \\ \leq &\max_{\pi\in\Pi} E \sup_{\tau\in T} \left\lfloor \frac{2}{n} \right\rfloor \sum_{i=1}^{\lfloor n/2 \rfloor} q_\tau\left(\mathbf{z}_{\pi(i)}, \mathbf{z}_{\pi(\lfloor n/2 \rfloor + i)}\right), \end{aligned}$$

where Π is the set of all permutations of $\{1, \cdots, n\}$.

The proof of Lemma 1 is shown in *Technical appendix A.2*. Then we establish the upper bound of generalization error for general pairwise learning with φ-mixing samples.

Theorem 1. *Under Assumptions 1-3, for any $\epsilon, \gamma, \zeta > 0$ and $n \in \mathbb{N}_+$ there exists $\delta_{n,\gamma,\epsilon} > 0$ such that with probability at least $1 - 2\zeta - \gamma$, we have:*

$$\begin{aligned} \mathcal{G}(\mathbf{w}, \mathbf{Z}) \leq &9B\Psi \sqrt{\frac{\log(2/\zeta) + (M^{d_{\mathbf{Z}}}(\mathcal{W}_{\mathbf{Z},U}) + \epsilon)\log(1/\delta)}{n}} \\ &+ 2\delta, \forall \delta < \delta_{n,\gamma,\epsilon} \end{aligned}$$

where $\Psi = 1 + 2\sum_{i=1}^{n} \varphi(i)$.

The proof of Theorem 1 is given in *Technical appendix A.2*, where essential improvement of analysis techniques [Dudley, 2014] are required due to the non-i.i.d. nature of the sequences.

Remark 1. *We employ Rademacher complexity and McDiarmid's inequality to deal with the difficulties brought by U-statistic [Rejchel, 2012; Cao* et al.*, 2016; Lei* et al.*, 2018]. Generally, traditional methods require Lemma A.1 in Clémençon et al. [2008] to transform the pairwise learning to the single index problem. However, it is not working for φ-mixing samples. To overcome this difficulty, we introduce Lemma 1 for rearranging the training samples and the generalized McDiarmid's inequality proved in Mohri and Rostamizadeh [2010] for handling the φ-mixing samples.*

For learning models with complex hypothesis space (e.g., deep neural networks), it often leads to trivial bound when directly applying the uniform convergence analysis framework independently of computing algorithms [Zhang et al.*, 2021]. To address this issue, we integrate the fractal tools into error analysis of pairwise learning, where the refined bound benefits from trajectory-dependent capacity estimation of hypothesis space.*

Remark 2. *Our capacity estimation is similar to the classical bounds for pointwise learning based on topological tools [Camuto* et al.*, 2021; Birdal* et al.*, 2021; Hodgkinson* et al.*, 2022; Dupuis* et al.*, 2023]. To our knowledge, our results are the first endeavor on applying topological tools to pairwise learning. In particular, the Lipschitz condition for loss function is abrogated using a data-dependent pseudo-metric. The derived result in Theorem 1 illustrates the impact of mixing samples (via the coefficient Ψ) on the convergence rate. As the dependency among samples increases, the convergence rate will slow down.*

Remark 3. *It should be pointed out that $M^{d_{\mathbf{Z}}}(\mathcal{W}_{\mathbf{Z},U})$ can be calculated by topological data analysis tools [Pérez* et al.*, 2021]. And it always can be bounded by a positive constant in SGD algorithm (if $\{\mathbf{w}_t\}_{t\geq 0}$ is a family of Feller processes) [Şimşekli* et al.*, 2021].*

It is well known that i.i.d. process can be viewed as a special case of φ-mixing process where the φ-mixing coefficients $\varphi(k) = 0, k = 1, 2, \cdots$. As a byproduct, we also state the corresponding result of Theorem 1 for i.i.d. samples.

Corollary 1. *Under Assumptions 1 and 3, for i.i.d. samples and any $\epsilon, \gamma, \zeta > 0$, there exists $\delta_{n,\gamma,\epsilon} > 0$ such that with*

probability at least $1-2\zeta-\gamma$*, we have*

$$\mathcal{G}(\mathbf{w},\mathbf{Z}) \leq 9B\sqrt{\frac{\log(2/\zeta)+(M^{d_{\mathbf{Z}}}(\mathcal{W}_{\mathbf{Z},U})+\epsilon)\log(1/\delta)}{n}} + 2\delta, \forall \delta < \delta_{n,\gamma,\epsilon}.$$

Theorem 1 shows that the generalization error depends on the Minkowski dimension of algorithmic trajectories, which extends the previous results of pointwise learning with i.i.d samples [Şimşekli *et al.*, 2021; Camuto *et al.*, 2021; Birdal *et al.*, 2021; Sachs *et al.*, 2023] to the pairwise setting with mixing observations. From Lemma A.15, we know that $M^{d_{\mathbf{Z}}}(\mathcal{W}_{\mathbf{Z},U})$ can be computed efficiently by tools of topological data analysis.

Remark 4. *Typically, we can set* $\delta = 1/\sqrt{n}$*. From Egoroff's Theorem in* Technical appendix A.1*, there exists a* $\gamma \geq 0$ *such that with probability at* $1-\gamma$*,* $\mathcal{N}(\mathcal{W}_{\mathbf{Z},U}, d_{\mathbf{Z}}, 1/\sqrt{n})$ *uniformly converges to* $M^{d_{\mathbf{Z}}}(\mathcal{W}_{\mathbf{Z},U})$ *with respect to* $1/\sqrt{n} \to 0$*. Thus, we have*

$$\log \mathcal{N}(\mathcal{W}_{\mathbf{Z},U}, d_{\mathbf{Z}}, 1/\sqrt{n}) \leq (M^{d_{\mathbf{Z}}}(\mathcal{W}_{\mathbf{Z},U})+1)$$

for a sufficiently large n*. Then under Assumptions 1-3, for any* $\zeta > 0$ *and a sufficiently large* n*, there exists a* $\gamma \geq 0$ *such that with probability at least* $1-2\zeta-\gamma$*, we have*

$$\mathcal{G}(\mathbf{w},\mathbf{Z}) \leq 9B\Psi\sqrt{\frac{\log(2/\zeta)+\frac{1}{2}(M^{d_{\mathbf{Z}}}(\mathcal{W}_{\mathbf{Z},U})+1)\log(n)}{n}} + \frac{2}{\sqrt{n}}.$$

Here, we recover a convergence rate $\sqrt{\log(n)/n}$ with i.i.d. setting ($\Psi = 1$) which is analogous to results relied on fractal tools in pointwise learning [Şimşekli *et al.*, 2021; Camuto *et al.*, 2021; Dupuis *et al.*, 2023]. To ensure a fair comparison with the works of Şimşekli *et al.* [2021] and Camuto *et al.* [2021], we present the generalization bound under the assumption that the loss function satisfies the Lipschitz condition.

Corollary 2. *Under Assumptions 1-3 and the L-Lipschitz assumption for the loss function* ℓ*, for any* $\gamma, \zeta > 0$ *and a sufficiently large* n *such that with probability at least* $1-2\zeta-\gamma$*, we have:*

$$\mathcal{G}(\mathbf{w},\mathbf{Z}) \leq 9B\Psi\sqrt{\frac{\log(2/\zeta)+\frac{1}{2}(M^{d_{\mathbf{w}}}(\mathcal{W}_{\mathbf{Z},U}))\log(L^2 n)}{n}} + \frac{2}{\sqrt{n}}$$

where $d_{\mathbf{w}}$ *is the Euclidean metric.*

The proof of Corollary 2 is shown in Technical appendix A.4. This result is similar to Şimşekli *et al.* [2021]. Moreover, it can be observed that in the absence of the Lipschitz condition, we have shifted the analytical challenge to $M^{d_{\mathbf{Z}}}(\mathcal{W}_{\mathbf{Z},U})$.

4.2 Lower Bound of Generalization Error

As an additional theoretical outcome, we endeavor to establish a lower bound by incorporating the introduced data-dependent fractal dimension. The next theorem is based on the classical arguments involving Sudakov's theorem [Vershynin, 2018] and Gaussian complexity. Here, we extend the technique for analyzing lower bound from the pointwise learning [Dupuis *et al.*, 2023] to the pairwise learning with φ-mixing samples. The lower bound considered here requires a slightly different definition of the worst-case generalization error:

$$\hat{\mathcal{G}}(\mathbf{w},\mathbf{Z}) = \sup_{\mathbf{w}\in\mathcal{W}_{\mathbf{Z},U}} \big|R_n(\mathbf{w},\mathbf{Z}) - R(\mathbf{w})\big|. \tag{6}$$

Furthermore, we define a new pseudo-metric

$$d_{g,\mathbf{Z}}(\mathbf{w},\hat{\mathbf{w}}) = \sqrt{E\left[(G_{\mathbf{w},\mathbf{Z}} - G_{\hat{\mathbf{w}},\mathbf{Z}})^2\right]},\ \mathbf{w},\hat{\mathbf{w}} \in \mathcal{W}_{\mathbf{Z},U},$$

where $\{G_{\mathbf{w},\mathbf{Z}}\}_{\mathbf{w}\in\mathcal{W}_{\mathbf{Z},U}}$ is a Gaussian process related to $\mathbf{Z}$ (see *Technical appendix A.1* for more details). In this Section, we replace $d_{\mathbf{Z}}(\cdot,\cdot)$ in Assumption 3 with $d_{g,\mathbf{Z}}(\cdot,\cdot)$. Now we can provide the lower bound of generalization error for pairwise learning with φ-mixing samples.

Theorem 2. *Under Assumptions 1-3, for any* $\epsilon, \zeta > 0$*, there exist a constant* $c > 0$ *and* $\delta_{n,\epsilon,\zeta} > 0$ *such that*

$$\hat{\mathcal{G}}(\mathbf{w},\mathbf{Z}) \geq \frac{c\delta}{4}\sqrt{\frac{2\log(1/\delta)\underline{M}^{d_{g,\mathbf{Z}}}(\mathcal{W}_{\mathbf{Z},U})}{n\big(\log(n)-\log(2)\big)}} - 9B\Psi\sqrt{\frac{2\log(2)+2\log(2/\zeta)}{\lfloor n/2\rfloor}}, \forall \delta \leq \delta_{n,\epsilon,\zeta}$$

holds true with probability at least $1-\zeta-\epsilon$*.*

Remark 5. *Theorem 2 states a novel lower bound for pairwise learning through constructing a pseudo-metric* $d_{g,\mathbf{Z}}(\cdot,\cdot)$*, whose effectiveness is guaranteed by Assumption 3. The main differences between our proposed lower bound and the existing result [Dupuis* et al.*, 2023] are the additional* φ*-mixing coefficient and the different fractal dimensions, which are induced by the mixing samples and the* U*-statistic associated with pairwise loss.*

5 Experiment

In this section, we conduct several numerical experiments to investigate the relationship between the Minkowski dimension and generalization bound. Subsequently, we sought to validate our theoretical findings regarding the convergence properties through these experiments. More experiment details can be found in *Technical appendix A.6*.

5.1 Experimental Setup

Our experimental design is similar to Birdal *et al.* [2021] and Dupuis *et al.* [2023]. Given a fixed positive integer m_0, we generate the random series $\{\mathbf{e}_i\}_{i\geq 1}$, which are i.i.d. drawn from Gaussian distribution $\mathcal{N}(\mathbf{0}_p, \mathbf{I}_{p\times p})$. For any $i \geq 1$, let

$$\mathbf{x}_i = (\mathbf{x}_i^1, \cdots, \mathbf{x}_i^p)^\top = \sum_{j=0}^{m_0} \mathbf{e}_{i+j} \in \mathbb{R}^p,$$

then the sequence $\{\mathbf{x}_i\}_{i\geq 1}$ is an m_0-dependent process and hence φ-mixing [Peng *et al.*, 2023]. Consider the intrinsic

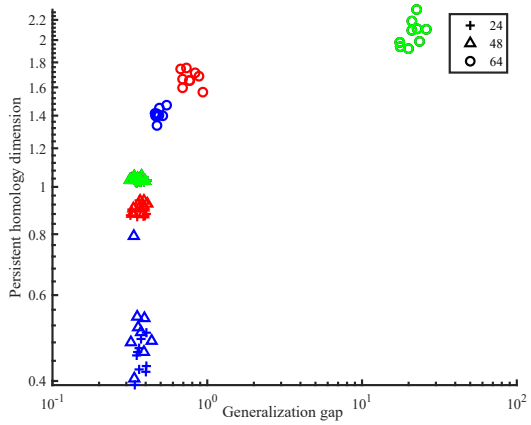

Figure 1: Generalization gap versus PH dimension for a seven-layer FCN with 2756 φ-mixing samples. Different colors indicate different learning rates (green: 0.1; red: 0.01; blue: 0.001) and different markers indicate different batch sizes.

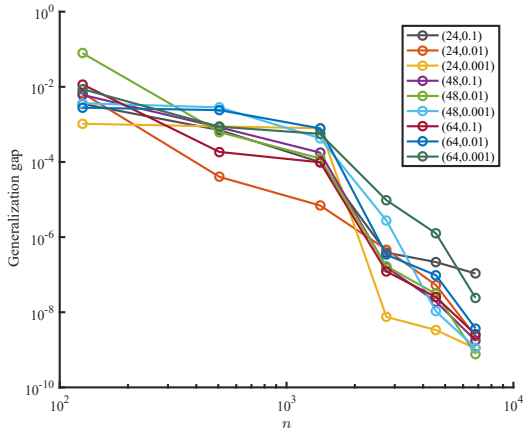

Figure 2: Generalization gap versus the sample size n across diverse scenarios for a seven-layer FCN.

relationship between the input and its response defined as Mukherjee and Zhou [2006],

$$y_i = \sum_{q=1}^{p} \big(\cos(\mathbf{x}_i^q) + \sin(\mathbf{x}_i^q) \big), \ i = 1, \cdots, n.$$

During the experiment, we focus on the case of gradient learning model (more details for gradient learning model can be found in *Technical appendix A.6*). It is easy to verify that pairwise loss $\ell(\mathbf{w}; \mathbf{z}_i, \mathbf{z}_j)$ satisfies the Assumption 1 for the aforementioned data generating process. Now we describe the implementation details of the empirical evaluation. Given the dataset $\mathbf{Z} = \{\mathbf{z}_1, \cdots, \mathbf{z}_n\}$, we employ the SGD algorithm to train three fully connected networks (FCN) with different layers under different parameter settings (batch size: 24, 48, 64, learning rate: 0.1, 0.01, 0.001, sample size n: 225, 900, 1406, 2025, 2756, 4556, 5625, 6806). During the training, the softmax activation function is employed. Subsequently, we consider the optimization trajectory near the local minimum discovered by SGD as the hypothesis set $\mathcal{W}_{\mathbf{Z},U}$. To be specific, we assume the SGD algorithm reaches a local minimum after $k^\star$ iterations. Then, we continue running this algorithm for 1000 iterations and set $\mathcal{W}_{\mathbf{Z},U}$ to be $\{\mathbf{w}_{k^\star+1}, \cdots, \mathbf{w}_{k^\star+1000}\}$.

Indeed, the Minkowski dimension often can be calculated by the persistent homology (PH) dimension $\dim_{PH_0}^{d_\mathbf{Z}} \mathcal{W}_{\mathbf{Z},U}$, which can be further approximated numerically in terms of the PH software provided by Pérez *et al.* [2021]. For more detailed information on PH, please refer to *Technical appendix A.5*.

5.2 Experimental Analysis

To ensure the fairness and impartiality, each experiment is repeated nine times for different scenarios. Figure 1 presents the PH dimension $\dim_{PH_0}^{d_\mathbf{Z}} \mathcal{W}_{\mathbf{Z},U}$ and the generalization gap of a seven-layer FCN (64, 32, 16, 8, 8, 8, 4) under various settings. Among them, different colors indicate different learning rates (green: 0.1; red: 0.01; blue: 0.001) and different markers indicate different batch sizes (plus: 24, triangle: 48, circle: 64). We observe a strong correlation between the PH dimension and the generalization gap in the case of a batch size 64. Besides, we find that the smaller batch size seems to show less correlation, which is consistent with the previous observation in Birdal *et al.* [2021] and Dupuis *et al.* [2023]. Intuitively, this phenomena maybe caused by the increased noise (such as the randomness associated with selecting k samples as the batch) in $\mathcal{W}_{\mathbf{Z},U}$ which may lead to more complex fractal structures. As a result, more points are required for precise computation of the PH dimension in such cases.

Figure 2 shows the average generalization gaps of nine trials with different sample sizes under nine various scenarios (batch size, learning rate). As the sample size n increases, we observe a gradual decrease in the average generalization gaps. This observation aligns with our theoretical results and provides empirical evidence to support our findings.

The simulation experiments are also conducted on various settings using a five-layers FCN $(64, 32, 16, 8, 4)$ and a nine-layers FCN $(64, 32, 16, 8, 8, 8, 8, 8, 4)$. Please refer to *Technical appendix A.6* for the experimental results.

6 Conclusions

This paper established the trajectory-dependent generalization analysis for pairwise learning with φ-mixing observations. For pairwise learning models, the derived theoretical results alleviate their sampling assumption on training data (i.e., i.i.d. observations) and match their real implementing algorithms tightly by employing the fractal theory to measure the computing trajectories in the hypothesis space. As far as we know, the current analysis is the first touch of optimization trajectory-dependent generalization characterization for general pairwise learning. In the future, it would be interesting to extend our analysis to much general dependent process, e.g., the τ-mixing process [Dedecker and Prieur, 2005; Liu *et al.*, 2025].

Acknowledgments

This work is supported by the National Natural Science Foundation of China (NSFC) (Nos. 62376104, 12426512, 12301651) and the Open Research Fund of Engineering Research Center of Intelligent Technology for Agriculture, Ministry of Education (No. ERCITA-KF002).

References

[Agarwal and Duchi, 2012] Alekh Agarwal and John C Duchi. The generalization ability of online algorithms for dependent data. *IEEE Transactions on Information Theory*, 59(1):573–587, 2012.

[Agarwal and Niyogi, 2009] Shivani Agarwal and Partha Niyogi. Generalization bounds for ranking algorithms via algorithmic stability. *Journal of Machine Learning Research*, 10(2), 2009.

[Alquier and Wintenberger, 2012] Pierre Alquier and Olivier Wintenberger. Model selection for weakly dependent time series forecasting. *Bernoulli*, 18(3):883–913, 2012.

[Birdal *et al.*, 2021] Tolga Birdal, Aaron Lou, Leonidas J Guibas, and Umut Şimşekli. Intrinsic dimension, persistent homology and generalization in neural networks. *Advances in Neural Information Processing Systems*, 34:6776–6789, 2021.

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

[Camuto *et al.*, 2021] Alexander Camuto, George Deligiannidis, Murat A Erdogdu, Mert Gurbuzbalaban, Umut Şimşekli, and Lingjiong Zhu. Fractal structure and generalization properties of stochastic optimization algorithms. *Advances in neural information processing systems*, 34:18774–18788, 2021.

[Cao *et al.*, 2016] Qiong Cao, Zheng-Chu Guo, and Yiming Ying. Generalization bounds for metric and similarity learning. *Machine Learning*, 102(1):115–132, 2016.

[Chaudhari and Soatto, 2018] Pratik Chaudhari and Stefano Soatto. Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. In *Information Theory and Applications Workshop*, pages 1–10. IEEE, 2018.

[Chen *et al.*, 2013] Hong Chen, Yi Tang, Luoqing Li, Yuan Yuan, Xuelong Li, and Yuanyan Tang. Error analysis of stochastic gradient descent ranking. *IEEE transactions on cybernetics*, 43(3):898–909, 2013.

[Chen, 2012] Hong Chen. The convergence rate of a regularized ranking algorithm. *Journal of Approximation Theory*, 164(12):1513–1519, 2012.

[Clémençon *et al.*, 2008] Stéphan Clémençon, Gábor Lugosi, and Nicolas Vayatis. Ranking and empirical minimization of u-statistics. *The Annals of Statistics*, pages 844–874, 2008.

[Cortes and Mohri, 2003] Corinna Cortes and Mehryar Mohri. Auc optimization vs. error rate minimization. *Advances in neural information processing systems*, 16, 2003.

[Dedecker and Prieur, 2005] Jérôme Dedecker and Clémentine Prieur. New dependence coefficients. examples and applications to statistics. *Probability Theory and Related Fields*, 132:203–236, 2005.

[Dudley, 2014] Richard M Dudley. *Uniform central limit theorems*, volume 142. Cambridge university press, 2014.

[Dupuis *et al.*, 2023] Benjamin Dupuis, George Deligiannidis, and Umut Şimşekli. Generalization bounds using data-dependent fractal dimensions. In *International Conference on Machine Learning*, page 8922–8968, 2023.

[Falconer, 2004] Kenneth Falconer. *Fractal geometry: mathematical foundations and applications*. John Wiley & Sons, 2004.

[Feng *et al.*, 2015] Yunlong Feng, Yuning Yang, and Johan AK Suykens. Robust gradient learning with applications. *IEEE transactions on neural networks and learning systems*, 27(4):822–835, 2015.

[Gao and Zhou, 2015] Wei Gao and Zhi-Hua Zhou. On the consistency of auc pairwise optimization. In *International Conference on Artificial Intelligence*, pages 939–945, 2015.

[Guo *et al.*, 2016] Zheng-Chu Guo, Yiming Ying, and Ding-Xuan Zhou. Online regularized learning with pairwise loss functions. *Advances in Computational Mathematics*, 28(4):1–24, 2016.

[He *et al.*, 2016] Fangchao He, Ling Zuo, and Hong Chen. Stability analysis for ranking with stationary φ-mixing samples. *Neurocomputing*, 171:1556–1562, 2016.

[Hodgkinson *et al.*, 2022] Liam Hodgkinson, Umut Simsekli, Rajiv Khanna, and Michael Mahoney. Generalization bounds using lower tail exponents in stochastic optimizers. In *International Conference on Machine Learning*, pages 8774–8795. PMLR, 2022.

[Hoeffding, 1992] Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *Breakthroughs in Statistics: Foundations and Basic Theory*, pages 308–334, 1992.

[Huang *et al.*, 2023] Shuo Huang, Junyu Zhou, Han Feng, and Ding-Xuan Zhou. Generalization analysis of pairwise learning for ranking with deep neural networks. *Neural Computation*, pages 1–24, 2023.

[Jastrzebski *et al.*, 2017] Stanislaw Jastrzebski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Three factors influencing minima in sgd. *arXiv preprint arXiv:1711.04623*, 2017.

[Jin *et al.*, 2009] Rong Jin, Shijun Wang, and Yang Zhou. Regularized distance metric learning: Theory and algorithm. *Advances in neural information processing systems*, 22, 2009.

[Kar *et al.*, 2013] Purushottam Kar, Bharath Sriperumbudur, Prateek Jain, and Harish Karnick. On the generalization ability of online learning algorithms for pairwise loss functions. In *International Conference on Machine Learning*, pages 441–449, 2013.

[Kohler and Krzyzak, 2020] Michael Kohler and Adam Krzyzak. On the rate of convergence of a deep recurrent neural network estimate in a regression problem with dependent data. *arXiv preprint arXiv:2011.00328*, 2020.

[Kriukova *et al.*, 2016] Galyna Kriukova, Sergei V Pereverzyev, and Pavlo Tkachenko. On the convergence rate and some applications of regularized ranking algorithms. *Journal of Complexity*, 33:14–29, 2016.

[Kurisu *et al.*, 2022] Daisuke Kurisu, Riku Fukami, and Yuta Koike. Adaptive deep learning for nonparametric time series regression. *arXiv preprint arXiv:2207.02546*, 2022.

[Kuznetsov and Mohri, 2017] Vitaly Kuznetsov and Mehryar Mohri. Generalization bounds for non-stationary mixing processes. *Machine Learning*, 106(1):93–117, 2017.

[Lei *et al.*, 2018] Yunwen Lei, Shao-Bo Lin, and Ke Tang. Generalization bounds for regularized pairwise learning. In *International Joint Conference on Artificial Intelligence*, pages 2376–2382, 2018.

[Lei *et al.*, 2021] Yunwen Lei, Mingrui Liu, and Yiming Ying. Generalization guarantee of sgd for pairwise learning. *Advances in Neural Information Processing Systems*, 34:21216–21228, 2021.

[Leo and Kavita, 2008] Kontorovich Leo and Ramanan Kavita. Concentration inequalities for dependent random variables via the martingale method. *Annals of Probability*, 36(6):2126–2158, 2008.

[Lin *et al.*, 2017] Junhong Lin, Yunwen Lei, Bo Zhang, and Ding-Xuan Zhou. Online pairwise learning algorithms with convex loss functions. *Information Sciences*, 406:57–70, 2017.

[Liu *et al.*, 2025] Liyuan Liu, Yaohui Chen, Weifu Li, Yingjie Wang, Bin Gu, Feng Zheng, and Hong Chen. Generalization bounds of deep neural networks with τ -mixing samples. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–15, 2025.

[Mandt *et al.*, 2016] Stephan Mandt, Matthew Hoffman, and David Blei. A variational analysis of stochastic gradient algorithms. In *International conference on machine learning*, pages 354–363. PMLR, 2016.

[McAllester, 1999] David A McAllester. Pac-bayesian model averaging. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, page 164–170, 1999.

[Meir, 2000] Ron Meir. Nonparametric time series prediction through adaptive model selection. *Machine learning*, 39:5–34, 2000.

[Mohri and Rostamizadeh, 2010] Mehryar Mohri and Afshin Rostamizadeh. Stability bounds for stationary φ-mixing and β-mixing processes. *Journal of Machine Learning Research*, 11(2), 2010.

[Molchanov, 2005] Ilya S Molchanov. *Theory of random sets*, volume 19. Springer, 2005.

[Mukherjee and Zhou, 2006] Sayan Mukherjee and Ding-Xuan Zhou. Learning coordinate covariances via gradients. *Journal of Machine Learning Research*, 7(3), 2006.

[Neyshabur *et al.*, 2017] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. *Advances in neural information processing systems*, 30, 2017.

[Peng *et al.*, 2023] Ling Peng, Yan Zhu, and Wenxuan Zhong. Lasso regression in sparse linear model with φ-mixing errors. *Metrika*, 86(1):1–26, 2023.

[Pérez *et al.*, 2021] Julián Burella Pérez, Sydney Hauke, Umberto Lupo, Matteo Caorsi, and Alberto Dassatti. giotto-ph: A python library for high-performance computation of persistent homology of vietoris-rips filtrations. *arXiv preprint arXiv:2107.05412*, 2021.

[Ralaivola *et al.*, 2010] Liva Ralaivola, Marie Szafranski, and Guillaume Stempfel. Chromatic pac-bayes bounds for non-iid data: Applications to ranking and stationary β-mixing processes. *The Journal of Machine Learning Research*, 11:1927–1956, 2010.

[Rejchel, 2012] Wojciech Rejchel. On ranking and generalization bounds. *Journal of Machine Learning Research*, 13(5), 2012.

[Sachs *et al.*, 2023] Sarah Sachs, Tim van Erven, Liam Hodgkinson, Rajiv Khanna, and Umut Şimşekli. Generalization guarantees via algorithm-dependent rademacher complexity. In *Annual Conference on Learning Theory*, pages 4863–4880. PMLR, 2023.

[Schmidt-Hieber, 2017] Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. *Annals of Statistics*, 48, 2017.

[Şimşekli *et al.*, 2021] Umut Şimşekli, Ozan Sener, George Deligiannidis, and Murat A Erdogdu. Hausdorff dimension, heavy tails, and generalization in neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124014, 2021.

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

[Wang *et al.*, 2012] Yuyang Wang, Roni Khardon, Dmitry Pechyony, and Rosie Jones. Generalization bounds for online learning algorithms with pairwise loss functions. In *Conference on Learning Theory*, pages 13–1. JMLR Workshop and Conference Proceedings, 2012.

[Ying and Zhou, 2016] Yiming Ying and Ding-Xuan Zhou. Online pairwise learning algorithms. *Neural Computation*, 28(4):743–777, 2016.

[Yu, 1994] Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, pages 94–116, 1994.

[Zhang *et al.*, 2021] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

[Zhao *et al.*, 2017] Yulong Zhao, Jun Fan, and Lei Shi. Learning rates for regularized least squares ranking algorithm. *Analysis and Applications*, 15(06):815–836, 2017.