Error Analysis Affected by Heavy-Tailed Gradients for Non-Convex Pairwise Stochastic Gradient Descent

Jun Chen[1], **Hong Chen**[1, 2, 3*], **Bin Gu**[4], **Guodong Liu**[5], **Yingjie Wang**[6], **Weifu Li**[1, 2, 3*]
[1]College of Informatics, Huazhong Agricultural University, Wuhan, China
[2]Engineering Research Center of Intelligent Technology for Agriculture, Ministry of Education, Wuhan, China
[3]Key Laboratory of Smart Farming for Agricultural Animals, Wuhan, China
[4]School of Artificial Intelligence, Jilin University, Jilin, China
[5]University of Pittsburgh
[6]College of Control Science and Engineering, China University of Petroleum (East China), Qingdao, China
chenh@mail.hzau.edu.cn; liweifu@mail.hzau.edu.cn

Abstract

In recent years, there have been a growing number of works studying the generalization properties of stochastic gradient descent (SGD) from the perspective of algorithmic stability. However, few of them devote to simultaneously studying the generalization and optimization for the non-convex setting, especially pairwise SGD with heavy-tailed gradient noise. This paper considers the impact of the heavy-tailed gradient noise obeying sub-Weibull distribution on the stability-based learning guarantees for non-convex pairwise SGD by investigating its generalization and optimization jointly. Specifically, based on two novel pairwise uniform model stability tools, we firstly bound the generalization error of pairwise SGD in the general non-convex setting after bridging the quantitative relationships between stability and generalization error. Then, we further consider the practical heavy-tailed sub-Weibull gradient noise condition to establish a refined generalization bound without the bounded gradient condition. Finally, sharper error bounds for generalization and optimization are built by introducing the gradient dominance condition. Comparing these results reveals that sub-Weibull gradient noise brings some positive dependencies on the heavy-tailed strength for generalization and optimization. Furthermore, we extend our analysis to the corresponding pairwise minibatch SGD and derive the first stability-based near-optimal generalization and optimization bounds which are consistent with many empirical observations.

Introduction

Pairwise learning has attracted much attention in machine learning literature, where its prediction performance is measured by pairwise loss function. Typical paradigms of pairwise learning include metric learning (Xing et al. 2002; Jin, Wang, and Zhou 2009), ranking (Clémençon, Lugosi, and Vayatis 2008; Agarwal and Niyogi 2009), AUC maximization (Cortes and Mohri 2003; Gao et al. 2013), gradient learning (Mukherjee and Zhou 2006), and learning under the minimum error entropy criterion (Príncipe 2010). Despite enjoying the benefits of particular contrastive motivations, pairwise learning often suffers from a heavy computational burden as its optimization objective involves $\mathcal{O}\left(n^2\right)$ terms for the problems with n training samples.

It is well known that stochastic gradient descent (SGD) is ubiquitous and popular for deploying learning systems due to its low time complexity (Lei, Liu, and Ying 2021; Wang et al. 2022; Tang et al. 2024) and high adaptability to big data (Bottou and Bousquet 2007; Lei and Ying 2020). As a natural extension of SGD, minibatch SGD iteratively updates the model parameter based on several selected samples rather than a single sample, which can further reduce the variance and accelerate algorithmic convergence (Cotter et al. 2011; Dekel et al. 2012). Therefore, it is natural to employ SGD and minibatch SGD to formulate the procedure of pairwise learning. Along with the wide applications of SGD and minibatch SGD in pairwise learning, there are some theoretical progresses focusing on their generalization guarantees recently (Lei, Liu, and Ying 2021; Lei, Ledent, and Kloft 2020; Shen et al. 2019; Yang et al. 2021). However, most of the existing theoretical results are limited to convex losses, which can not cover typical non-convex pairwise learning algorithms, e.g., neural networks-based pairwise learning (Köppel et al. 2019; Li et al. 2022), let alone the ones with heavy-tailed gradient noise.

Moreover, most of the previous theoretical works on pairwise SGD and its variants require the bounded variance condition (Zhou, Liang, and Zhang 2022) and the sub-Gaussian tail assumption limiting the tail performance of the gradient noise (Simsekli et al. 2019; Simsekli, Sagun, and Gürbüzbalaban 2019). However, these assumptions may be too idealistic in practice. Indeed, SGD may only involve the bounded p-th moment for some $p \in (1, 2]$ rather than the bounded variance (Cutkosky and Mehta 2021), and it often shows the heavier-tailed gradient noise in many learning problems (Madden, Dall'Anese, and Becker 2020; Lei and Tang 2021; Li and Liu 2022). For example, Simsekli, Sagun, and Gürbüzbalaban (2019) statistically characterized the gradient noise of SGD and stated that the gradient noise expresses a heavy-tailed behavior under an isotropic model (also see Zhou et al. (2020); Zhang et al. (2020b)). The heavy-tailed gradient noise may degrade the generalization performance of SGD methods significantly (Nguyen et al. 2019; Hodgkinson and Mahoney 2021). However, Raj et al. (2023a,b) found that heavy tails of gradient noise can help

*Corresponding authors.

with generalization in pointwise SGD. Therefore, it is crucial to investigate the theoretical guarantees of non-convex pairwise SGD with heavy tails. As far as we know, this issue has not been touched for pairwise SGD.

Bottou and Bousquet (2007) demonstrated that the model performance depends on the joint influence of generalization error and optimization error. The generalization error is used to evaluate the performance of a trained model to some unseen inputs (Vapnik 1998) and the optimization error concerns the gap between the actual empirical risk and the optimal empirical risk (Li and Liu 2022). Hence, it is necessary to investigate both generalization error and optimization error for a better understanding of the learning guarantees of SGD. Following this line, some error analysis of SGD can be found in Lei, Ledent, and Kloft (2020); Lei, Hu, and Tang (2021). Compared with uniform convergence analysis for SGD methods (Lei, Liu, and Ying 2021; Lei and Tang 2021; Li and Liu 2022), algorithmic stability analysis often enjoys promising properties on adaptability (Agarwal and Niyogi 2009; Hardt, Recht, and Singer 2016) and flexibility (Lei, Liu, and Ying 2021; Lei and Ying 2020). In particular, the stability-based theory analysis is suitable for wide application scenarios and independent of the capacity of hypothesis function space (Zhou, Liang, and Zhang 2022; Hardt, Recht, and Singer 2016; Bousquet and Elisseeff 2002; Shalev-Shwartz et al. 2010).

At present, stability and generalization have been characterized for the non-convex pointwise SGD (Zhou, Liang, and Zhang 2022; Hardt, Recht, and Singer 2016). This paper develops the previous analysis techniques (Lei, Liu, and Ying 2021; Lei and Ying 2020; Madden, Dall'Anese, and Becker 2020; Li and Liu 2022) to explore the impacts of sub-Weibull gradient noise for the non-convex pairwise cases with the consideration of generalization and optimization errors simultaneously. The main contributions of this paper are summarized as follows:

- *Generalization bounds of non-convex pairwise SGD with sub-Weibull gradient noise*. We propose two novel pairwise ℓ_1 uniform model stability tools for the generalization analysis of pairwise SGD to prevent the sampling barrier for stability analysis. After giving the relationship between stability and generalization, we establish the generalization bounds for non-convex pairwise SGD. Even under the general non-convex setting, our result is comparable with previous works for convex pairwise SGD (Lei, Liu, and Ying 2021; Lei, Ledent, and Kloft 2020; Yang et al. 2021). Subsequently, the refined bounds are derived by introducing the heavy-tailed sub-Weibull gradient noise assumption, where the standard requirement of the bounded gradient is removed.
- *Learning guarantees with gradient dominance condition*. Sharper bounds for generalization error and excess risk are provided for the non-convex pairwise SGD under an additional gradient dominance condition. The current analysis extends the previous one for pointwise SGD with sub-Weibull tails (Li and Liu 2022) to the more complicated pairwise setting, which shows the competitive convergence rates and some positive dependencies on heavy tails for optimization. Finally, we develop our analysis to the corresponding minibatch case and give the first-ever-known stability-based learning guarantees which are consistent with empirical observations (Dekel et al. 2012; Woodworth, Patel, and Srebro 2020).

Related Work

Analysis of pairwise SGD via algorithmic stability. Algorithmic stability has gained much attention in statistical learning theory due to its attractive properties, i.e., the independence to hypothesis function space and wide applicability (Lei and Ying 2020; Hardt, Recht, and Singer 2016). This analysis technique is applied to investigate theoretical foundations of pairwise SGD (Lei, Liu, and Ying 2021; Lei, Ledent, and Kloft 2020; Shen et al. 2019; Yang et al. 2021). For the convex pairwise SGD, Shen et al. (2019) established the bounds of the expected optimization error and excess risk after illustrating the trade-off between the stability and optimization error. Moreover, some systematic studies are provided in Lei, Liu, and Ying (2021); Lei, Ledent, and Kloft (2020); Yang et al. (2021) to cover more general cases (i.e., without the bounded loss assumption or smoothness assumption). For the non-convex pairwise SGD, (Lei, Liu, and Ying 2021) investigated the stability and generalization of pairwise SGD under the gradient dominance condition, while the derived bounds are not tight enough. Therefore, it is necessary to further explore the learning guarantees of non-convex pairwise SGD from the perspective of algorithmic stability. Please see *Appendix G* for the outlines of algorithmic stability.

Analysis of SGD with heavy-tailed gradient noise. The heavy-tailed performance of SGD has been studied extensively, see e.g., Simsekli, Sagun, and Gürbüzbalaban (2019); Nguyen et al. (2019); Hodgkinson and Mahoney (2021). In a seminal paper, Vladimirova et al. (2019) found that the Bayesian neural network presents a heavier-tailed unit distribution than Gaussian prior (Lee et al. 2018) while deepening the model. After that, several works (Simsekli et al. 2019; Simsekli, Sagun, and Gürbüzbalaban 2019; Panigrahi et al. 2019) verified that SGD also has heavier-tailed distribution than sub-Gaussian distribution. It is demonstrated in Nguyen et al. (2019); Hodgkinson and Mahoney (2021) that the generalization ability of SGD may suffer from the heavy-tailed gradient noise, but it is also demonstrated that when the tail is heavy but not too heavy, heavy tails might help generalization (Raj et al. 2023a,b). Based on uniform convergence analysis, the high probability guarantees for non-convex pointwise SGD are stated in Madden, Dall'Anese, and Becker (2020); Li and Liu (2022) under heavy-tailed gradient noise assumption. However, as far as we know, there are no stability-based learning guarantees for pairwise SGD with heavy tails. In this paper, we aim to make an effort to fill this theoretical gap.

Preliminaries

Problem setup

According to an unknown probability measure ρ defined over a sample space $\mathcal{Z}$, we draw each sample $z_i (1 \leq i \leq n)$

independently and get the training set $S := \{z_1, ..., z_n\} \in \mathcal{Z}^n$. The goal of pairwise learning is to find a data-driven predictor such that the population risk (the risk on the expectation of the whole sample space $\mathcal{Z}$)

$$F(w) := \mathbb{E}_{z,\tilde{z}}[f(w; z, \tilde{z})] \tag{1}$$

is as small as possible, where $f(w; z, \tilde{z}) : \mathcal{W} \times \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_+$ is a non-negative loss function, w is the model parameter belonging to the hypothesis function space $\mathcal{W}$, $z, \tilde{z}$ are independently distributed as ρ, and $\mathbb{E}_{z,\tilde{z}}$ denotes the expectation with respect to (w.r.t.) samples z and $\tilde{z}$. Due to the inaccessibility of $F(w)$, we often formulate pairwise learning algorithms by minimizing the empirical risk

$$F_S(w) := \frac{1}{n(n-1)} \sum_{i,j \in [n], i \neq j} f(w; z_i, z_j), \tag{2}$$

where $[n] := \{1, ..., n\}$. Let

$$w(S) \in \arg\min_{w \in \mathcal{W}} F_S(w), \;\; w^* \in \arg\min_{w \in \mathcal{W}} F(w), \tag{3}$$

where $F(w), F_S(w)$ are defined in (1) and (2), respectively. For feasibility, let $A(S, \xi)$ be the model parameter trained by algorithm $A : \mathcal{Z}^n \to \mathcal{W}$ with random parameter $\xi \in \mathcal{R}$ on dataset S.

Definition 1 (SGD for Pairwise Learning)**.** *For $t \in \mathbb{N}$, let $\{w_t\}_{t=1}^T$ be an update sequence of model parameters with the initial state $w_1 = 0$, and let $\{\eta_t\}_{t=1}^T$ be a stepsize sequence. Denote $\nabla f(w_t; z_{i_t}, z_{j_t})$ as the gradient of the loss function $f(w_t; z_{i_t}, z_{j_t})$ w.r.t. the first argument w_t, where (z_{i_t}, z_{j_t}) is a sample pair selected to update model parameters in the t-th iteration, and (i_t, j_t) is independently drawn from $\{(i, j) : i, j \in [n], i \neq j\}$. Then, the pairwise SGD is updated by*

$$w_{t+1}(\xi) = w_t - \eta_t \nabla f(w_t; z_{i_t}, z_{j_t}). \tag{4}$$

For pairwise SGD (Definition 1), the random parameter ξ is the sample index set $\{(i_1, j_1)..., (i_T, j_T)\}$ during the T iterations, where we let the trained model be $A(S, \xi) = w_T$. Since $\mathbb{E}[F_S(w(S)) - F(w^*)] \leq 0$, the excess risk of w_T can be decomposed by

$$\begin{aligned}
&\mathbb{E}[F(w_T) - F(w^*)] \\
=&\mathbb{E}[F(w_T) - F_S(w_T) + F_S(w_T) - F_S(w(S))] \\
\leq&|\mathbb{E}[F(w_T) - F_S(w_T) + F_S(w_T) - F_S(w(S))]| \\
\leq&|\mathbb{E}[F(w_T) - F_S(w_T)]| + |\mathbb{E}[F_S(w_T) - F_S(w(S))]|,
\end{aligned} \tag{5}$$

where $\mathbb{E}[\cdot]$ denotes the expectation w.r.t. all randomness, i.e., S and ξ. Usually, we call the first term $|\mathbb{E}[F(w_T) - F_S(w_T)]|$ as the generalization error and the second term $|\mathbb{E}[F_S(w_T) - F_S(w(S))]|$ as the optimization error.

Quantitative Relationships between Stability and Generalization

In this paper, we analyze the theoretical generalization performance of heavy-tailed non-convex pairwise SGD via pairwise ℓ_1 uniform model stability measuring the model parameter sensitivity to a small perturbation of S, which is different from the ones concerning the sensitivity of loss function value, e.g., the uniform stability (Hardt, Recht, and Singer 2016; Bousquet and Elisseeff 2002; Shalev-Shwartz et al. 2010) and the on-average stability (Lei, Ledent, and Kloft 2020; Kuzborskij and Lampert 2018; Lei and Ying 2021).

Definition 2. *Let $S = \{z_i\}_{i=1}^n, S' = \{z_i'\}_{i=1}^n$ be drawn independently from ρ. Define that, for $\forall i, j \in [n], i \neq j$,*

$$S_{i,j} = \{z_1, ..., z_{i-1}, z_i', z_{i+1}, ..., z_{j-1}, z_j', z_{j+1}, ..., z_n\}.$$

Denote $\|\cdot\|$ as the Euclidean norm. A pairwise learning algorithm A is ℓ_1 uniform model ϵ-stable if

$$\mathbb{E}_{S,S',\xi}\left[\|A(S_{i,j}, \xi) - A(S, \xi)\|\right] \leq \epsilon.$$

A pairwise learning algorithm A is ℓ_1 uniform model ϵ-stable with expectation w.r.t. ξ if

$$\mathbb{E}_{\xi}\left[\|A(S_{i,j}, \xi) - A(S, \xi)\|\right] \leq \epsilon, \;\; \forall S, S' \in \mathcal{Z}^n.$$

Some stability tools for pointwise learning have been extended to the case of pairwise learning. For example, Shen et al. (2019), Yang et al. (2021) and Lei, Liu, and Ying (2021) provided the definitions of uniform stability (6), uniform model stability (7) and on-average model stability (8) for pairwise learning, respectively, whose definitions are listed as follows:

$$\sup_{\substack{z,\tilde{z} \in \mathcal{Z}, \\ S,S' \in \mathcal{Z}^n}} \mathbb{E}_{\xi}[|f(A(S, \xi); z, \tilde{z}) - f(A(S_i, \xi); z, \tilde{z})|] \leq \epsilon, \tag{6}$$

$$\sup_{z,\tilde{z} \in \mathcal{Z}, S,S' \in \mathcal{Z}^n} \mathbb{E}_{\xi}[\|A(S, \xi) - A(S_i, \xi)\|] \leq \epsilon, \tag{7}$$

$$\mathbb{E}_{S,S',\xi}\left[\frac{1}{n}\sum_{i \in [n]} \|A(S_i, \xi) - A(S, \xi)\|\right] \leq \epsilon, \tag{8}$$

where $S_i = \{z_1, ..., z_{i-1}, z_i', z_{i+1}, ..., z_n\}, \forall i \in [n]$. Considering that the pairwise SGD involves dependent $\mathcal{O}\left(n^2\right)$ terms resulting in the sampling barrier for stability analysis, Definition 2 nails down the novel pairwise ℓ_1 uniform model stability and ℓ_1 uniform model stability with expectation w.r.t. ξ inspired by the stability definitions of Lei, Ledent, and Kloft (2020); Yang et al. (2021); Lei (2023). Our analysis treats a sample pair $(z_i, z_j), i \neq j$ as a single sample to avoid extremely complex sampling situations in stability analysis.

Assumption 1. *(a) For any $z, \tilde{z} \in \mathcal{Z}$, $w, w' \in \mathcal{W}$ and $L > 0$, a differentiable loss function $f(w; z, z')$ is L-Lipschitz continuous w.r.t the first argument w if $\|\nabla f(w; z, \tilde{z})\| \leq L$, which means that $|f(w; z, \tilde{z}) - f(w'; z, \tilde{z})| \leq L\|w - w'\|$.*

(b) For any $z, \tilde{z} \in \mathcal{Z}$, $w, w' \in \mathcal{W}$ and $\beta > 0$, a differentiable loss function $f(w; z, z')$ is β-smooth w.r.t the first argument w if $\|\nabla f(w; z, \tilde{z}) - \nabla f(w'; z, \tilde{z})\| \leq \beta\|w - w'\|$.

Some previous work assumed that the gradient and the loss function itself are both Lipschitz (Hardt, Recht, and Singer 2016) and gave some examples, e.g., ranking and metric learning with the logistic loss and the hinge loss (Lei and Ying 2020). These two assumptions are used extensively in statistical learning theory. However, the Lipschitz continuity assumption may not hold in some learning environments (Lei and Ying 2020). In these cases, many stability-based generalization bounds under this assumption don't match the algorithmic deployment in real applications. As for smoothness, it is assumed throughout the whole paper.

Definition 3 ((Vladimirova et al. 2020)). *We say X is a sub-Weibull random variable if the moment generating function (MGF) $\mathbb{E}\left[\exp\left((|X|/K)^{\frac{1}{\theta}}\right)\right] \leq 2$ holds for some positive parameters K and $\theta \geq 1/2$, and denote it as $X \sim subW(\theta, K)$.*

The sub-Weibull random variable X becomes the sub-Gaussian as $\theta = 1/2$ (Vershynin 2018) or the sub-Exponential distribution as $\theta = 1$ (Vladimirova et al. 2020). We consider the pairwise SGD with heavy tails and let $\theta > 1/2$ in the rest of this paper. For ease of understanding, some necessary preliminaries of sub-Weibull distribution are provided in *Appendix B*.

This work mainly studies the learning guarantees of pairwise SGD with a type of heavy-tailed gradient noise (sub-Weibull). In this case, we can obtain some specific upper bounds of gradient (Theorems 4, 5). Therefore, it is unnecessary to make the Lipschitz continuity assumption. Some prior work (Vladimirova et al. 2020; Li and Liu 2022) showed that these upper bounds are small.

Assumption 2 (Sub-Weibull Gradient Noise). *For the t-th iteration of (4), we assume $\|\nabla f(w_t; z_{i_t}, z_{j_t}) - \nabla F_S(w_t)\| \sim subW(\theta, K)$ with $\theta > 1/2, K > 0$, i.e.,*

$$\mathbb{E}_{i_t, j_t}\left[\exp\left(\left(\frac{\|\nabla f(w_t; z_{i_t}, z_{j_t}) - \nabla F_S(w_t)\|}{K}\right)^{\frac{1}{\theta}}\right)\right] \leq 2.$$

Recently, rich works have shown that SGD and its variants exhibit heavier noise than sub-Gaussian (Simsekli et al. 2019; Simsekli, Sagun, and Gürbüzbalaban 2019; Madden, Dall'Anese, and Becker 2020; Panigrahi et al. 2019; Zhang et al. 2020a; Wang et al. 2021). Hence, it is natural to consider Assumption 2 here for the pairwise SGD with heavy tails. In our analysis, the gradient noise assumption provides some refined bounds of gradient noise (Lemma 3 in *Appendix C.1*) which are key to bridging the connection between pairwise ℓ_1 uniform model stability and generalization error (Theorem 2) and stating stability bounds (Theorems 4, 5) without the bounded gradient assumption.

Based on the pairwise ℓ_1 uniform model stability, Theorems 1 and 2 give two generalization upper bounds for pairwise SGD under the Lipschitz condition (Assumption 1 (a)) and the heavy-tailed sub-Weibull gradient noise condition (Assumption 2) respectively.

Theorem 1. *Let S, S' and $S_{i,j}$ be constructed as Definition 2. Assume that pairwise learning algorithm A, associated with L-Lipschitz continuous loss function, is ℓ_1 uniform model ϵ-stable. Then,*

$$\mathbb{E}[|F_S(A(S,\xi)) - F(A(S,\xi))|] \leq L\epsilon.$$

Theorem 1 shows the generalization upper bound with expectation w.r.t. S, ξ can be controlled by the pairwise ℓ_1 uniform model stability bound even for the general non-convex pairwise SGD. Theorem 1 is consistent with the related results of stability and generalization for pointwise learning (Theorem 2 in Lei and Ying (2020)) and pairwise learning (Theorem 1 in Lei, Liu, and Ying (2021)), where the slight difference is induced by the divergence among stability definitions. However, the Lipschitz continuity condition is necessary for our proof framework under this case.

Our next generalization result is a new guarantee which is just averaged over the random algorithmic parameter ξ and conditioned on the dataset S under the milder condition, heavy-tailed gradient noise.

Theorem 2. *Let S, S' and $S_{i,j}$ be constructed as Definition 2. Assume that gradient-based pairwise learning algorithm A, associated with the loss function whose gradient noise obeys $subW(\theta, K)$, is ℓ_1 uniform model ϵ-stable with expectation w.r.t. ξ, and the norm of the gradient $\mathbb{E}_\xi[\|\nabla f(w_t; z_{i_t}, z_{j_t})\|]$ for the loss function f is upper bounded by $L(\theta)$. Let constant $M > 0$ and, for all $z, \tilde{z} \in \mathcal{Z}, \xi \in \mathcal{R}$, $f(A(S,\xi); z, \tilde{z}) \leq M$. Then, for any $\delta \in (0, 1/e)$, we have*

$$\begin{aligned}&\mathbb{E}_\xi[|F(A(S,\xi)) - F_S(A(S,\xi))|] \\ \leq& 4L(\theta)\epsilon + e\Big(12\sqrt{2}M(n-1)^{-\frac{1}{2}}\sqrt{\log(e/\delta)} \\ &+ 48\sqrt{6}L(\theta)\epsilon\lceil\log_2(n-1)\rceil\log(e/\delta)\Big)\end{aligned}$$

with probability at least $1 - \delta$.

Theorem 2 verifies the generalization bound via pairwise ℓ_1 uniform model stability with expectation w.r.t. ξ enjoys some attractive properties, e.g., independence of the Lipschitz continuity assumption. Specifically, the heavy-tailed sub-Weibull gradient noise assumption enjoys the bounded second-order gradient property. Therefore, its first-order gradient is also bounded and the bound $L(\theta)$ is mild which is validated in Theorems 4 and 5.

Main Results

In this section, we develop the learning guarantees of pairwise SGD under various conditions. In the first two cases, only generalization guarantees are provided to reveal the dependence of generalization on the parameter of heavy-tailed distribution. In the last two cases, we additionally investigate optimization guarantees to further analyze the corresponding heavy-tailed dependence. We summarize our bounds and some related results in Tables 1, 2 and 4 (*Appendix F*). To improve readability, we also make some discussions about the dependencies of our results in *Appendix F.1*, and the detailed comparisons with other results for pairwise learning and pointwise learning from several aspects including analysis tools, assumptions and algorithms in *Appendix F.2*.

Reference	Assumptions			Tool	Stability Bound
	L	μ	θ		
Shen et al. (2019) (Thm. 3.5)	$\surd$	$\times$	$\times$	Uniform	$\mathcal{O}\left((\beta n)^{-1}L^{\frac{2}{\beta c+1}}T^{\frac{\beta c}{\beta c+1}}\right)$
Lei, Liu, and Ying (2021) (Thm. 15)	$\surd$	$\surd$	$\times$	Uniform	$\mathcal{O}\left((\beta n)^{-1}L^{2}T^{\frac{\beta c}{\beta c+1}}\right)$
Ours (Thm. 3)	$\surd$	$\times$	$\times$	Uniform model	$\mathcal{O}\left((\beta n)^{-1}LT^{\frac{1}{2}}\log T\right)$
Ours (Thm. 4)	$\times$	$\times$	$\surd$	Uniform model	$\mathcal{O}\left((\beta n)^{-1}\sqrt{g(2\theta)}\log^{\theta}(1/\delta)T^{\frac{1}{2}}(\log T)^{\frac{3}{2}}\right)$
Ours (Thm. 5)	$\times$	$\surd$	$\surd$	Uniform model	$\mathcal{O}\left((\beta n)^{-1}\sqrt{g(2\theta)}\log^{\theta}(1/\delta)T^{\frac{1}{4}}(\log T)^{\frac{3}{2}}\right)$

Table 1: Summary of stability bounds for pairwise SGD with non-convex loss functions (Thm.-Theorem; $\surd$-has such a property; $\times$-hasn't such a property; L, μ, θ-the parameters of Lipschitz continuity, PL condition and sub-Weibull distribution; c-a non-negative constant). See *Appendix G* for details of stability tools. Note that, considering the smoothness assumption is used in every work, we don't mention it in this table.

General Non-convex Pairwise SGD

Now we state the quantitative characterization of the pairwise ℓ_1 uniform model stability for the general non-convex pairwise SGD.

Theorem 3 (Stability of SGD: Lipschitz Case)**.** *Given S, S' and $S_{i,j}$ in Definition 2, let $\{w_t\}_{t=1}^T$ and $\{w_t'\}_{t=1}^T$ be produced by (4) on S and $S_{i,j}$ respectively, where $\eta_t = \eta_1 t^{-1}, \eta_1 \leq (2\beta)^{-1}$, and let the parameters $A(S,\xi) = w_T$ and $A(S_{i,j},\xi) = w_T'$ after T iterations. Under Assumption 1, there holds $\mathbb{E}\left[\|w_T - w_T'\|\right] \leq \mathcal{O}\Big((\beta n)^{-1}LT^{\frac{1}{2}}\log T\Big)$.*

Theorem 3 illustrates the pairwise ℓ_1 uniform model stability bound $\mathcal{O}\left((\beta n)^{-1}LT^{\frac{1}{2}}\log T\right)$ for non-convex pairwise SGD when the loss function is Lipschitz continuous and smooth. Shen et al. (2019) provided a uniform stability bound $\mathcal{O}\left((\beta n)^{-1}L^{\frac{2}{\beta c+1}}T^{\frac{\beta c}{\beta c+1}}\right)$, where the constant $c > 0$. In practice, Theorem 3 is also a general result $\mathcal{O}\left(cn^{-1}LT^{\beta c}\log T\right)$. When Shen et al. (2019) makes the same setting $c = 1/\beta$ as our setting, their result is $\beta^{-\frac{1}{2}}$-times larger than Theorem 3. Besides, Theorem 3 shows that the smaller the step size, the better the stability bound, which is different from Shen et al. (2019).

Non-convex Pairwise SGD without Lipschitz Condition

In this section, the refined bounds of stability and generalization error are given by employing the heavy-tailed gradient noise assumption to remove the Lipschitz continuity assumption.

Theorem 4 (Stability of SGD: Sub-Weibull Case)**.** *Given S, S' and $S_{i,j}$ described in Definition 2, let $\{w_t\}_{t=1}^T$ and $\{w_t'\}_{t=1}^T$ be produced by (4) on S and $S_{i,j}$ respectively, where $\eta_t = \eta_1 t^{-1}, \eta_1 \leq (2\beta)^{-1}$, and let the parameters $A(S,\xi) = w_T$ and $A(S_{i,j},\xi) = w_T'$ after T iterations. Under Assumptions 1 (b) and 2, for all $\delta \in (0,1)$, there hold*

$$\mathbb{E}_\xi[\|\nabla f(w_t; z_{i_t}, z_{j_t})\|] \leq \mathcal{O}\left(\sqrt{g(2\theta)\log^{2\theta}(1/\delta)\log T}\right)$$

and

$$\begin{aligned}&\mathbb{E}_\xi\left[\|w_T - w_T'\|\right]\\ \leq&\mathcal{O}\left((\beta n)^{-1}\sqrt{g(2\theta)}\log^{\theta}(1/\delta)T^{\frac{1}{2}}\log^{\frac{3}{2}}T\right)\end{aligned}$$

with probability at least $1-\delta$, where $g(\theta) = (4e)^\theta$ for $\theta \leq 1$ and $g(\theta) = 2(2e\theta)^\theta$ for $\theta \geq 1$.

Theorem 4 assures the pairwise ℓ_1 uniform model stability upper bound with the order $\mathcal{O}\left((\beta n)^{-1}\sqrt{g(2\theta)}\log^{\theta}(1/\delta)T^{\frac{1}{2}}\log^{\frac{3}{2}}T\right)$ with expectation w.r.t. ξ, where the heavy-tailed gradient noise assumption is employed to get rid of the bounded gradient assumption. Observe that, the dependence on T for the bound of Theorem 4 is just $\sqrt{\log T}$-times larger than Theorem 3 and the additional dependence on the heavy tail parameter θ is often bounded (Vladimirova et al. 2020). Thus, it is better than the dependence on the Lipschitz parameter L which is likely infinite for some learning environments. Due to the above reasons, the bound of Theorem 4 is tighter than the one of Theorem 3.

Theorem 3 in Lei, Liu, and Ying (2021) provided a ℓ_2 on-average model stability bound $\mathcal{O}\left(\left(\frac{1}{n}+\frac{T}{n^2}\right)\sum_{t=1}^{T}\eta_t^2\mathbb{E}[F_S(w_t)]\right)$ for pairwise SGD with convex loss functions, which involves $\sum_{t=1}^{T}\eta_t^2\mathbb{E}[F_S(w_t)]$. However, it is hard to ensure the summation of empirical risks is small enough. Therefore, the current result enjoys more adaptivity and flexibility since it is not affected by the quality of the initial empirical risks and nears optimum under the milder assumptions, i.e., non-convex loss and heavy-tailed gradient noise.

Non-convex Pairwise SGD with PL Condition

Inspired from Li and Liu (2022), we further investigate learning guarantees of non-convex pairwise SGD under the PL condition which assures that the estimator w satisfying $\|\nabla F_S(w)\| = 0$ is a global minimizer of empirical risk $F_S(w)$ (2).

Assumption 3 (Polyak-Lojasiewicz (PL) Condition)**.** *For any $w \in \mathcal{W}$ and $S \in \mathcal{Z}^n$, the empirical risk $F_S(w)$ (2) satisfies the PL condition with parameter $\mu > 0$ if* $\|\nabla F_S(w)\|^2 \geq 2\mu \left(F_S(w) - F_S(w(S))\right)$.

The PL condition, also called gradient dominance condition (Lei, Liu, and Ying 2021; Zhou, Liang, and Zhang 2022; Foster, Sekhari, and Sridharan 2018; Reddi et al. 2016), can be viewed as a mild control over the curvature of loss function and has been employed for the non-convex generalization analysis (Lei and Tang 2021; Li and Liu 2022; Lei and Ying 2021). This condition demonstrates that the lower bound of the quadratic of objective gradient is $2\mu \left(F_S(w) - F_S(w(S))\right)$ and will increase as the model parameter w is far away from the empirically optimal parameter $w(S)$ (Karimi, Nutini, and Schmidt 2016). Note that, the PL condition assures that any w satisfying $\|\nabla F_S(w)\| = 0$ is a global minimizer (Charles and Papailiopoulos 2018).

Theorem 5 (Stability of SGD: PL Case)**.** *Given S, S' and $S_{i,j}$ in Definition 2 and $w(S)$ in (3), let $\{w_t\}_{t=1}^T$ and $\{w'_t\}_{t=1}^T$ be produced by (4) on S and $S_{i,j}$ respectively, where $\eta_t = \eta_1 t^{-1}, \eta_1 \leq (4\beta)^{-1}, 1 - \mu\eta_1 \geq 0$. Take $a_1 = 1 - \prod_{i=1}^t \left(1 - \frac{1}{2}\mu\eta_i\right)$. Under Assumptions 1 (b), 2 and 3, for all $\delta \in (0,1)$, there hold $\mathbb{E}_\xi[\|\nabla f(w_t; z_{i_t}, z_{j_t})\|] \leq \mathcal{O}\left(\sqrt{g(2\theta)\log^{2\theta}(1/\delta)\log T}\right)$ and*

$$\mathbb{E}_\xi\left[\|w_T - w'_T\|\right] \leq \mathcal{O}\Bigg((\beta n)^{-1} T^{\frac{1}{4}} \log T \sqrt{a_1 \beta \mathbb{E}_\xi[F_S(w(S))] + g(2\theta)\log^{2\theta}(1/\delta)\log T}\Bigg)$$

with probability at least $1 - \delta$.

Compared with Theorem 4, the stability bound $\mathcal{O}\left((\beta n)^{-1} T^{\frac{1}{4}} \sqrt{g(2\theta)} \log^{\theta}(1/\delta) \log^{\frac{3}{2}} T\right)$ in Theorem 5 involves a different term $T^{\frac{1}{4}}$ which is better than $T^{\frac{1}{2}}$ when $\mathbb{E}_\xi[F_S(w(S))] = \mathcal{O}\left(n^{-1}\right)$. The reason why the dependence on T can be improved is that we employ the PL condition to make a slightly different analysis (see the proof in *Appendix D.4*) which allows a smaller learning rate to achieve a tighter stability bound. Specifically, for Theorem 4, we need to set a larger step size $\eta_t \leq (2\beta)^{-1}$ to remove the term $\frac{1}{2}\sum_{t'=1}^t \eta_{t'} ||\nabla F_S(w_{t'})||^2$, which is important for our analysis. For Theorem 5, we adopt a new strategy (related to PL condition) to get the upper bound of $F_S(w_t)$. To ensure the condition $\beta\eta_t^2 - \frac{1}{2}\eta_t \leq 0$ holds, we must set a smaller step size than $(2\beta)^{-1}$. Therefore, we select $\eta_t \leq (4\beta)^{-1}$, which leads to a result outperforming Theorem 4 by a factor of $T^{\frac{1}{4}}$.

Lei, Liu, and Ying (2021) provided a uniform stability bound $\mathcal{O}\left((\beta n)^{-1} L^2 T^{\frac{\beta c}{\beta c+1}}\right)$, where the constant $c = 1/\mu$ (μ is the parameter of PL condition). In general, μ is typically a very small value (Examples 1 and 2 in Lei and Ying (2021)) which leads to a large value of c. Thus, $T^{\frac{\beta c}{\beta c+1}}$ is closer to T than $T^{1/2}\log T$ of our bound. In other words, our bound is tighter than Lei, Liu, and Ying (2021).

Lei, Liu, and Ying (2021) have established the generalization bound $\mathcal{O}\left((\beta n)^{-1} L^2 T^{\beta/(\beta+\mu)}\right)$ for non-convex pairwise SGD under the gradient dominance condition. Different from the existing work that relies on uniform stability (Lei, Liu, and Ying 2021), we provides tighter generalization bound under weaker assumptions when the mild condition $\sqrt{g(2\theta)}\log^{\theta+1}(1/\delta)T^{\frac{1}{4}}\log^2 T \log_2 n \leq L^2 T^{\frac{\beta}{\beta+\mu}-\frac{1}{4}}$ holds.

Theorem 6 (Excess risk of SGD: PL Case)**.** *Given w^* in (3) and $\{w_t\}_{t=1}^T$ produced by (4) on S, where $\eta_t = \eta_1 t^{-1}, \eta_1 \leq (4\beta)^{-1}, 1 - \mu\eta_1 \geq 0$. Take $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$. Under Assumptions 1 (b), 2 and 3, for all $\delta \in (0, 1/e)$, there holds*

$$\begin{aligned}&\mathbb{E}_\xi[F(w_T) - F(w^*)] \\ \leq &\mathcal{O}\Bigg(T^{-2} + (\beta T)^{-1}\Gamma(2\theta+1) + \sqrt{\frac{\log(1/\delta)}{n}} \\ &+ (\beta n)^{-1}\sqrt{g(2\theta)}\log^{\theta+1}(1/\delta)T^{\frac{1}{4}}\log^2 T\log_2 n\Bigg)\end{aligned}$$

with probability at least $1 - \delta$.

With the help of the PL condition, we can guarantee that the algorithm can find a global minimizer (Karimi, Nutini, and Schmidt 2016; Foster, Sekhari, and Sridharan 2018; Lei, Liu, and Ying 2021; Yunyan Bai 2024). Therefore, we can use $\mathbb{E}_\xi[F_S(w_T) - F_S(w(S))]$ instead of $\mathbb{E}_\xi[\|\nabla F_S(w_T)\|]$ to measure the optimization performance of pairwise SGD in Definition 1. For the non-convex pairwise SGD, Theorem 6 provides the optimization error bound $\mathcal{O}\Big(T^{-2} + (\beta T)^{-1}\Gamma(2\theta+1)\Big)$ and the excess risk bound $\mathcal{O}\Big(T^{-2} + (\beta T)^{-1}\Gamma(2\theta+1) + \sqrt{n^{-1}\log(1/\delta)} + (\beta n)^{-1}\sqrt{g(2\theta)}\log^{\theta+1}(1/\delta)T^{\frac{1}{4}}\log^2 T\log_2 n\Big)$ by (5). The derived optimization error bound is comparable with $\mathcal{O}\Big(T^{-1}\Big)$ stated in Lemma D.1 (e) of *Appendix D* in (Lei, Liu, and Ying 2021). For excess risk, our bound is $\mathcal{O}\Big(\sqrt{n^{-1}\log(1/\delta)} + \beta^{-1}n^{-\frac{3}{4}}\sqrt{g(2\theta)}\log^{\theta+1}(1/\delta)\log^2 n\log_2 n\Big)$ as $T \asymp n$, which is comparable with the related results (Lei and Tang 2021; Li and Liu 2022) for the pointwise SGD and enjoys nice property, i.e., the independence of the dimension d. Moreover, our bound is also consistent with the excess risk bound $\mathcal{O}\left((\beta n)^{-1}L^2 T^{\frac{\beta}{\beta+\mu}} + T^{-1}\right)$ (Lei, Liu, and Ying 2021) which implies the convergence order $\mathcal{O}\left(n^{-\frac{\beta/\mu+1}{2\beta/\mu+1}}\right)$ as $T \asymp n^{\frac{\beta+\mu}{2\beta+\mu}}$. As shown in Table 2, our results fill the theoretical gap of stability-based excess risk analysis for the non-convex pairwise SGD with heavy tails, and guarantee a satisfactory convergence rate.

Non-convex Minibatch SGD

We further investigate the stability and generalization of non-convex pairwise minibatch SGD. To our surprise, this issue has not been studied in machine learning literature.

Reference	Assumptions			Excess risk bound
	β	μ	θ	
Madden, Dall'Anese, and Becker (2020) ♣▼ (Thm. 9)	√	√	×	$*\mathcal{O}\left(T^{-1}\log\left(1/\delta\right)\right)$
Lei and Tang (2021) ♣▼ (Thm. 7)	√	√	×	$*\mathcal{O}\left(n^{-1}\left(d+\log\left(1/\delta\right)\right)\log^2 n\log^2\left(1/\delta\right)\right)$
Li and Liu (2022) ♣▼ (Thm. 3.11)	√	√	√	$*\mathcal{O}\Big(n^{-1}\left(d+\log\left(1/\delta\right)\right)\log^{2\theta+1}\left(1/\delta\right)\log^{\frac{3(\theta-1)}{2}}\left(n/\theta\right)\log n\Big)$
Lei, Liu, and Ying (2021) ♠▲ (Thm. 15)	√	√	×	$\mathcal{O}\Big((\beta n)^{-1}L^2T^{\frac{\beta}{\beta+\mu}}+T^{-1}\Big)$
Ours ♠▲ (Thm. 6)	√	√	√	$\mathcal{O}\Big(\frac{\beta+T\Gamma(2\theta+1)}{\beta T^2}+\sqrt{\frac{\log(1/\delta)}{n}}+\frac{\sqrt{g(2\theta)}\log^{\theta+1}(1/\delta)T^{\frac{1}{4}}\log^2 T\log_2 n}{\beta n}\Big)$
Ours (Minibatch) ♠▲ (Thm. 7)	√	√	√	$\mathcal{O}\Big(\frac{\beta b+T\Gamma(2\theta+1)}{\beta bT^2}+\sqrt{\frac{\log(1/\delta)}{n}}+\frac{\Gamma(2\theta+1)T^{\frac{1}{4}}\log^2 T\log(1/\delta)\log_2 n}{\beta n}\Big)$

Table 2: Summary of excess risk bounds for non-convex SGD via uniform convergence approaches and stability analysis (Thm.-Theorem; ♣-uniform convergence; ♠-stability; ▼-pointwise learning; ▲-pairwise learning; β,μ,θ-the parameters of smoothness, PL condition and sub-Weibull distribution; √-has such a property; ×-hasn't such a property; d-dimension of hypothesis function space; $*$-high-probability bound; δ-some probability).

Definition 4 (Minibatch SGD for Pairwise Learning)**.** *For $\{w_t\}_{t=1}^T,\{\eta_t\}_{t=1}^T$ described in Definition 1 and the batch size b, denote $\nabla f(w_t;z_{i_{t,m}},z_{j_{t,m}})$ as the gradient of the loss function $f(w_t;z_{i_{t,m}},z_{j_{t,m}})$ w.r.t. the first argument w_t, where $(z_{i_{t,m}},z_{j_{t,m}}),m\in[b]$ is the m-th sample pair selected to update model parameters in the t-th iteration, and $(i_{t,m},j_{t,m})$ is independently drawn from $\{(i,j):i,j\in[n],i\neq j\}$. Then, the pairwise minibatch SGD is updated by*

$$w_{t+1}(\xi)=w_t-\frac{\eta_t}{b}\sum_{m=1}^{b}\nabla f(w_t;z_{i_{t,m}},z_{j_{t,m}}). \tag{9}$$

The pairwise minibatch SGD in Definition 4 reduces to the pairwise SGD in Definition 1 as $b=1$. Note that, when $b=n(n-1)$, Definition 4 is inconsistent with the full-batch SGD for the reason that $(z_{i_{t,m}},z_{j_{t,m}})$ is independently selected from all sample pairs $\{(z_i,z_j):z_i,z_j\in S,z_i\neq z_j\}$, which means that some certain sample pair can be selected more than once at each iteration.

Theorem 7 (Excess risk of Minibatch SGD: PL Case)**.** *Let $A(S,\xi)$ be produced by the pairwise minibatch SGD (9) on S, where $\eta_t=\eta_1t^{-1},\eta_1\leq(4\beta)^{-1},1-\mu\eta_1\geq 0$. Under Assumptions 1 (b), 2 and 3, for all $\delta\in(0,1/e)$, there holds*

$$\begin{aligned}&\mathbb{E}_\xi[F(w_T)-F(w^*)]\\ \leq&\mathcal{O}\Bigg(T^{-2}+(\beta bT)^{-1}\Gamma(2\theta+1)+\sqrt{\frac{\log(1/\delta)}{n}}\\ &+(\beta n)^{-1}\Gamma(2\theta+1)T^{\frac{1}{4}}\log^2 T\log_2 n\log(1/\delta)\Bigg)\end{aligned}$$

with probability at least $1-\delta$.

Theorem 7 provides the first optimal optimization error and stability-based excess risk bounds for non-convex pairwise minibatch SGD with heavy tails and the PL condition. Particularly, the detailed comparisons of our results are summarized in Table 4 of *Appendix F*.

Theorem 7 is consistent with many empirical observations (Cotter et al. 2011; Dekel et al. 2012; Li et al. 2014; Lin et al. 2020; Woodworth, Patel, and Srebro 2020). Specifically, the dependencies on the batch size b and the iteration number T for Theorem 7 are consistent with some prior observations, e.g., Figures 1, 2 in Cotter et al. (2011). Some previous work (Lei, Sun, and Liu 2023) provided some stability-based bounds (Theorems 2,3,5) for pointwise minibatch SGD which shows a similar negative dependence on b. Note that, the references we provide above are about pointwise learning. Considering the fact that pairwise SGD is regarded as a special pointwise SGD in our analysis, we use these references to validate our results.

Conclusions

This paper aims to investigate the impact of heavy-tailed gradient on the learning guarantees of non-convex pairwise SGD. We derive the first near-optimal bounds of generalization error and excess risk for the heavy-tailed non-convex pairwise SGD and minibatch SGD, respectively, where the algorithmic stability analysis technique is developed to overcome the obstacle induced by the complicated pairwise objective and the minibatch strategy. Our results verify the effect of the heavy-tailed gradient noise on removing the bounded gradient assumption, which brings some mild positive dependencies on the heavy-tailed strength θ.

Although our results achieve some satisfactory orders, we provide some dependencies on θ,T which may exist in practice. Therefore, in the future, some lower bounds are needed to combine with our upper bounds to ensure the tightness of our results. Besides, it is interesting to further investigate the stability and generalization of pairwise minibatch SGD obeying the distributions with heavier tails (such as α-stable distributions (Simsekli, Sagun, and Gürbüzbalaban 2019)).

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Nos. 12301651, 62306338, 62376104, 12426512, 12071166), the Fundamental Research Funds for the Central Universities of China (Nos. 2662024XXPY001, 2662023LXPY005), the Independent Innovation Research Project of China University of Petroleum (East China) (No. 23CX06033A) and the Key Research and Development of Hubei Province (No. 2023BBB119).

References

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

Bottou, L.; and Bousquet, O. 2007. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems (NIPS)*, 161–168.

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

Charles, Z.; and Papailiopoulos, D. S. 2018. Stability and generalization of learning algorithms that converge to global optima. In *International Conference on Machine Learning (ICML)*, volume 80, 744–753.

Clémençon, S.; Lugosi, G.; and Vayatis, N. 2008. Ranking and empirical minimization of U-statistics. *The Annals of Statistics*, 36(2): 844–874.

Cortes, C.; and Mohri, M. 2003. AUC optimization vs. error rate minimization. In *Advances in Neural Information Processing Systems (NIPS)*, 313–320.

Cotter, A.; Shamir, O.; Srebro, N.; and Sridharan, K. 2011. Better mini-batch algorithms via accelerated gradient methods. In *Advances in Neural Information Processing Systems (NIPS)*, 1647–1655.

Cutkosky, A.; and Mehta, H. 2021. High-probability bounds for non-convex stochastic optimization with heavy tails. In *Advances in Neural Information Processing Systems (NeurIPS)*, 4883–4895.

Dekel, O.; Gilad-Bachrach, R.; Shamir, O.; and Xiao, L. 2012. Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13: 165–202.

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

Gao, W.; Jin, R.; Zhu, S.; and Zhou, Z. 2013. One-pass AUC optimization. In *International Conference on Machine Learning (ICML)*, volume 28, 906–914.

Hardt, M.; Recht, B.; and Singer, Y. 2016. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning (ICML)*, volume 48, 1225–1234.

Hodgkinson, L.; and Mahoney, M. W. 2021. Multiplicative noise and heavy tails in stochastic optimization. In *International Conference on Machine Learning (ICML)*, volume 139, 4262–4274.

Jin, R.; Wang, S.; and Zhou, Y. 2009. Regularized distance metric learning: Theory and algorithm. In *Advances in Neural Information Processing Systems (NIPS)*, 862–870.

Karimi, H.; Nutini, J.; and Schmidt, M. 2016. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)*, volume 9851, 795–811.

Köppel, M.; Segner, A.; Wagener, M.; Pensel, L.; Karwath, A.; and Kramer, S. 2019. Pairwise learning to rank by neural networks revisited: Reconstruction, theoretical analysis and practical performance. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)*, volume 11908, 237–252.

Kuzborskij, I.; and Lampert, C. H. 2018. Data-dependent stability of stochastic gradient descent. In *International Conference on Machine Learning (ICML)*, volume 80, 2820–2829.

Lee, J.; Bahri, Y.; Novak, R.; Schoenholz, S. S.; Pennington, J.; and Sohl-Dickstein, J. 2018. Deep neural networks as Gaussian processes. In *International Conference on Learning Representations (ICLR)*.

Lei, Y. 2023. Stability and generalization of stochastic optimization with nonconvex and nonsmooth problems. In *Conference on Learning Theory (COLT)*, 191–227.

Lei, Y.; Hu, T.; and Tang, K. 2021. Generalization performance of multi-pass stochastic gradient descent with convex loss functions. *Journal of Machine Learning Research*, 22: 25:1–25:41.

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.; 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.; Sun, T.; and Liu, M. 2023. Stability and generalization for minibatch SGD and local SGD. arXiv:2310.01139.

Lei, Y.; and Tang, K. 2021. Learning rates for stochastic gradient descent with nonconvex objectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12): 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)*, volume 119, 5809–5819.

Lei, Y.; and Ying, Y. 2021. Sharper generalization bounds for learning with gradient-dominated objective functions. In *International Conference on Learning Representations (ICLR)*, 1–23.

Li, M.; Zhang, H.; Lei, Q.; Fan, Z.; Liu, J.; and Du, J. 2022. Pairwise contrastive learning network for action quality assessment. In *European Conference on Computer Vision (ECCV)*, volume 13664, 457–473.

Li, M.; Zhang, T.; Chen, Y.; and Smola, A. J. 2014. Efficient mini-batch training for stochastic optimization. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, 661–670.

Li, S.; and Liu, Y. 2022. High probability guarantees for nonconvex stochastic gradient descent with heavy tails. In *International Conference on Machine Learning (ICML)*, volume 162, 12931–12963.

Lin, T.; Stich, S. U.; Patel, K. K.; and Jaggi, M. 2020. Don't use large mini-batches, use local SGD. In *International Conference on Learning Representations (ICLR)*.

Madden, L.; Dall'Anese, E.; and Becker, S. 2020. High-probability convergence bounds for non-convex stochastic gradient descent. arXiv:2006.05610.

Mukherjee, S.; and Zhou, D. 2006. Learning coordinate covariances via gradients. *Journal of Machine Learning Research*, 7: 519–549.

Nguyen, T. H.; Simsekli, U.; Gürbüzbalaban, M.; and Richard, G. 2019. First exit time analysis of stochastic gradient descent under heavy-tailed Gradient Noise. In *Advances in Neural Information Processing Systems (NeurIPS)*, 273–283.

Panigrahi, A.; Somani, R.; Goyal, N.; and Netrapalli, P. 2019. Non-Gaussianity of stochastic gradient noise. arXiv:1910.09626.

Príncipe, J. C. 2010. *Information theoretic learning: Renyi's entropy and kernel perspectives*. Springer.

Raj, A.; Barsbey, M.; Gürbüzbalaban, M.; Zhu, L.; and Simsekli, U. 2023a. Algorithmic stability of heavy-tailed stochastic gradient descent on least squares. In *International Conference on Algorithmic Learning Theory (ALT)*, volume 201, 1292–1342.

Raj, A.; Zhu, L.; Gürbüzbalaban, M.; and Simsekli, U. 2023b. Algorithmic stability of heavy-tailed SGD with general loss functions. In *International Conference on Machine Learning, (ICML)*, volume 202, 28578–28597.

Reddi, S. J.; Hefny, A.; Sra, S.; Póczos, B.; and Smola, A. J. 2016. Stochastic variance reduction for nonconvex optimization. In *International Conference on Machine Learning (ICML)*, volume 48, 314–323.

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

Shen, W.; Yang, Z.; Ying, Y.; and Yuan, X. 2019. Stability and optimization error of stochastic gradient descent for pairwise learning. arXiv:1904.11316.

Simsekli, U.; Gürbüzbalaban, M.; Nguyen, T. H.; Richard, G.; and Sagun, L. 2019. On the heavy-tailed theory of stochastic gradient descent for deep neural networks. arXiv:1912.00018.

Simsekli, U.; Sagun, L.; and Gürbüzbalaban, M. 2019. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning (ICML)*, volume 97, 5827–5837.

Tang, H.; Liu, G.; Dai, S.; Ye, K.; Zhao, K.; Wang, W.; Yang, C.; He, L.; Leow, A. D.; Thompson, P. M.; Huang, H.; and Zhan, L. 2024. Interpretable spatio-temporal embedding for brain structural-effective network with ordinary differential equation. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, volume 15002, 227–237.

Vapnik, V. N. 1998. Statistical learning theory. *Encyclopedia of the Sciences of Learning*, 41(4): 3185–3185.

Vershynin, R. 2018. *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press.

Vladimirova, M.; Girard, S.; Nguyen, H.; and Arbel, J. 2020. Sub-Weibull distributions: Generalizing sub-Gaussian and sub-Exponential properties to heavier-tailed distributions. *Stat*, 9(1): e318:1–8.

Vladimirova, M.; Verbeek, J.; Mesejo, P.; and Arbel, J. 2019. Understanding priors in Bayesian neural networks at the unit level. In *International Conference on Machine Learning (ICML)*, volume 97, 6458–6467.

Wang, H.; Gürbüzbalaban, M.; Zhu, L.; Simsekli, U.; and Erdogdu, M. A. 2021. Convergence rates of stochastic gradient descent under infinite noise variance. In *Advances in Neural Information Processing Systems (NeurIPS)*, 18866–18877.

Wang, P.; Lei, Y.; Ying, Y.; and Zhang, H. 2022. Differentially private SGD with non-smooth losses. *Applied and Computational Harmonic Analysis*, 56: 306–336.

Woodworth, B. E.; Patel, K. K.; and Srebro, N. 2020. Minibatch vs local SGD for heterogeneous distributed learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Xing, E. P.; Ng, A. Y.; Jordan, M. I.; and Russell, S. 2002. Distance metric learning with application to clustering with side-information. In *Advances in Neural Information Processing Systems (NIPS)*, 505–512.

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.

Yunyan Bai, L. L., Yuxing Liu. 2024. On the complexity of finite-sum smooth optimization under the Polyak–Łojasiewicz condition. In *International Conference on Machine Learning (ICML)*, volume 235, 2392–2417.

Zhang, J.; He, T.; Sra, S.; and Jadbabaie, A. 2020a. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations (ICLR)*.

Zhang, J.; Karimireddy, S. P.; Veit, A.; Kim, S.; Reddi, S. J.; Kumar, S.; and Sra, S. 2020b. Why are adaptive methods good for attention models? In *Advances in Neural Information Processing Systems (NeurIPS)*.

Zhou, P.; Feng, J.; Ma, C.; Xiong, C.; Hoi, S. C.; and E, W. 2020. Towards theoretically understanding why SGD generalizes better than adam in deep learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Zhou, Y.; Liang, Y.; and Zhang, H. 2022. Understanding generalization error of SGD in nonconvex optimization. *Machine Learning*, 111(1): 345–375.