ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230418

High Probability Analysis for Non-Convex Stochastic Optimization with Clipping

Shaojie Li[a;b] **and Yong Liu** [a;b;*]

[a]Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China
[b]Beijing Key Laboratory of Big Data Management and Analysis Methods, Beijing, China

Abstract. Gradient clipping is a commonly used technique to stabilize the training process of neural networks. A growing body of studies has shown that gradient clipping is a promising technique for dealing with the heavy-tailed behavior that emerged in stochastic optimization as well. While gradient clipping is significant, its theoretical guarantees are scarce. Most theoretical guarantees only provide an in-expectation analysis and only focus on optimization performance. In this paper, we provide high probability analysis in the non-convex setting and derive the optimization bound and the generalization bound simultaneously for popular stochastic optimization algorithms with gradient clipping, including stochastic gradient descent and its variants of momentum and adaptive stepsizes. With the gradient clipping, we study a heavy-tailed assumption that the gradients only have bounded α-th moments for some $\alpha \in (1, 2]$, which is much weaker than the standard bounded second-moment assumption. Overall, our study provides a relatively complete picture for the theoretical guarantee of stochastic optimization algorithms with clipping.

1 Introduction

Stochastic optimization has played a crucial role in modern machine learning and data-driven optimization since many machine learning problems can be transformed into a stochastic optimization problem [5, 4]. The past decades have witnessed the prosperous development of stochastic optimization algorithms. For example, stochastic gradient descent (SGD) [51] has shown great success in the training of a large number of learning tasks [26, 24]. In practice, SGD works by querying an oracle iteratively to obtain unbiased gradient estimates built on one or several training examples in place of the exact gradient. Its simplicity in implementation and low memory requirements per iteration make it easier to scale into the big data era [30, 3].

Driven by the empirical success of SGD, a great deal of work has been done on design modifications to improve its performance in various ways. One popular modification is to use the adaptive stepsizes. [13, 40] propose the provably convergent adaptive gradient (AdaGrad) and demonstrate that the sparsity of the gradient suggests outperformance. Another popular modification of SGD is the momentum technique. Momentum uses a running average of the past gradient values [47, 43], and intuitively, adding momentum accelerates convergence by circumventing sharp curvatures and long ravines of the sub-level sets of the objective function [49]. These stochastic optimization algorithms have shown distinct advantages in different learning tasks [64, 61]. The superior empirical performance has attracted many researchers to investigate their guarantees and understand their theoretical properties.

Recently, a number of works have interestingly shown that stochastic optimization algorithms easily exhibit a heavy-tailed behavior [61, 53, 45, 52, 7, 21]. For example, [61] provide empirical study and show that large natural language processing models, e.g., Bert [56, 12], have heavy-tailed gradients. In this spirit, existing guarantees of assuming bounded variance or light sub-Gaussian tail seem to be inappropriate [61, 62, 9]. In particular, in practice the variance can be very large, possibly even infinite, but the α-th moment is bounded for some $\alpha \in (1, 2]$ [9, 62]. For a more realistic analysis, it is essential to investigate the theoretical guarantees of stochastic optimization algorithms under this heavy-tailed condition. However, the setup becomes complicated, which hinders the use of conventional convergence analysis techniques that rely on the existence of the second-order moment.

Gradient clipping is an effective tool for dealing with heavy-tailed random variables [9, 62]. The intuition behind this is that the clipped version of a heavy-tailed random variable will have much more benign properties when the clipping parameters are well chosen. Thus, gradient clipping is a promising technique for dealing with the heavy-tailed behavior in stochastic optimization. Additionally, gradient clipping can stabilize the gradient updates and thus stabilize the training process of stochastic optimization [39]. It is believed to effectively alleviate the gradient explosion problem without adding additional cost to the original update [59]. As such, it has been a common choice for many application domains of machine learning, especially the language processing tasks [46, 58].

Theoretically, some recent works have studied the optimization guarantee for stochastic optimization algorithms with gradient clipping [19, 59, 39, 60, 62, 61, 9]. However, these optimization guarantees are typically either provided for the convex optimization problems [19, 39] or derived in expectation [59, 39, 60, 62, 61]. Unfortunately, the expectation bound does not capture the behavior of stochastic optimization algorithms within one or several runs, which is relevant to the probabilistic property of stochastic optimization algorithms. Also, in real-world applications such as neural networks, since the training process can take hours or even days, algorithms are usually run only once, so it is important to obtain high probability guarantees [36, 23, 57, 9, 19].

Furthermore, to the best of our knowledge, existing learning guarantees of stochastic optimization algorithms with clipping are almost all derived from the optimization performance perspective. In

* Corresponding Author. Email: liuyonggsai@ruc.edu.cn.

Please check ArXiv or contact the authors for any appendices or supplementary material mentioned in the paper.

machine learning, our primary interest would be the generalization performance of the trained model on testing examples, which is quite different from the empirical performance on training examples [33, 44, 5]. To be specific, the optimization performance concerns how the learning algorithm minimizes the empirical risk, while generalization performance concerns how the predictive models learned from training samples behave on the testing samples. Thus, to investigate the learning guarantees of clipped stochastic optimization algorithms, it is necessary to consider both the optimization and generalization guarantees.

Motivated by the problems we discussed above, this paper considers three popular stochastic optimization algorithms, i.e., stochastic gradient descent (SGD), stochastic gradient descent with momentum (SGDM), and stochastic gradient descent with adaptive stepsizes (SGDAS), in the non-convex setting. We establish both the high probability optimization bound and the high probability generalization bound for their clipped version under the bounded α-th moment assumption. The results cover SGD and the well-known momentum technique and adaptive stepsizes and reveal the learning performance of the clipped stochastic optimization algorithms from both the perspective of convergence and generalization. In Table 1, we provide an intuitive display of the results this paper obtained.

This paper is organized as follows. We first review the related work in Section 2 and then introduce the preliminaries relevant to our discussion in Section 3. Section 4 presents the main results, where we derive a series of learning guarantees for stochastic optimization algorithms with clipping. In Section 5, we conclude this paper. Some Lemmas useful to our discussions and proofs are shown in Section 6, and the complete proofs are provided in the Appendix[1].

2 Related Work

High Probability Bounds. Most of the literature provides guarantees in expectation for stochastic optimization algorithms [23]. The high probability guarantees of SGD are mainly provided for the convex setting [28, 25, 48, 19, 23, 10, 11, 20, 27, 32, 37, 31, 14, 2]. As a comparison, high probability studies on the non-convex setting are scarce. Specifically, [16, 33, 38, 34] provide high probability bounds for non-convex SGD and [36, 64, 57, 29] for non-convex adaptive SGD. Unsatisfied, all these works assume the light sub-Gaussian tail or bounded variance. Very recently, motivated by recent research on the heavy-tailed phenomena in stochastic optimization, [9] give high probability bounds in the non-convex setting by assuming the bounded α-th moment, a heavy-tailed assumption allowing unbounded variance. We mention to readers here that in some literature [38, 19, 34], the "heavy-tailedness" refers to non-sub-Gaussianity. While in this paper, by stochastic gradient with heavy-tailed distribution, we mean such a stochastic gradient allows unbounded variance. Overall, high probability bounds for stochastic optimization algorithms under the heavy-tailed assumption allowing unbounded variance are scarce.

Gradient Clipping. Gradient clipping is a commonly used technique in the training process of neural networks [18, 41]. In [59, 19, 39, 60, 62, 61, 9, 34], the optimization guarantees of clipping are investigated. Specifically, [19] study convex SGD and consider the smoothness and bounded variance conditions. [39] then study convex SGD but the non-smooth case. [60] study non-convex SGD, using a relaxed smoothness condition and a stronger assumption than the bounded variance. [59] then provide improved convergence analysis of [60] with joint consideration of clipped gradient and clipped momentum. [62, 61] study non-convex SGD under the bounded α-th moment condition. Notably that the above works all focus on in-expectation optimization guarantees. Under the bounded α-th moment condition, [9] combine the gradient clipping and normalized gradient descent and derive the first high probability optimization guarantees for SGDM. The work [34] then provide the first high probability optimization guarantee for SGD with the sub-Weibull gradient noise. Therefore, from the related work, one can see that the high probability optimization analysis of non-convex stochastic optimization algorithms with clipping has not been thoroughly studied and is far from being understood. Even worse, there is almost no research on its generalization performance analysis. This paper makes an effort in this direction.

[1] https://arxiv.org/abs/2307.13680.

3 Preliminaries

3.1 Notations

Let P be a probability measure defined on a sample space $\mathcal{Z}$, many learning problems of machine learning can be cast into the following stochastic optimization problem with a hypothesis space indexed by $\mathcal{W} \subseteq \mathbb{R}^d$:

$$\min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w}) := \mathbb{E}_{z\sim P}[f(\mathbf{w}; z)],$$

where the objective $f : \mathcal{W} \times \mathcal{Z} \mapsto \mathbb{R}_+$ is possibly non-convex and $\mathbb{E}_{z\sim P}$ denotes the expectation with respect to (w.r.t.) the random variable z drawn form P. In machine learning, $F(\mathbf{w})$ is typically referred to as population risk [6].

For the above stochastic optimization problem, people want to learn a prediction model with a small population risk. However, $F(\mathbf{w})$ is typically not accessible since the underlying distribution P is unknown. In practice, we often sample a set of i.i.d. training data $S = \{z_1, ..., z_n\}$ from P and minimize the following empirical risk:

$$F_S(\mathbf{w}) := \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{w}; z_i).$$

Various stochastic optimization algorithms, e.g. SGD and its variants of momentum and adaptive stepsizes, have been proposed to optimize the empirical risk $F_S(\mathbf{w})$ and have shown their distinct advantages in different learning tasks [4, 16]. Perhaps SGD is the most popular stochastic optimization algorithm due to its simplicity in implementation, low computational complexity, and sound practical behavior. For this reason, we show the pseudocode of SGD in Algorithm 1. SGD iteratively moves models along the reverse direction of an unbiased gradient estimate $\nabla f(\mathbf{w}_t; z_{j_t})$, i.e.,

$$\mathbb{E}_{j_t}[\nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t)] = 0,$$

and the simplicity has made SGD become one of the workhorses behind many machine learning tasks [5, 4, 30, 33]. Its clipped version and other variants will be presented in Section 4.

We then introduce some notations used in this paper. Let $b = \sup_{z\in\mathcal{Z}} \|\nabla f(0; z)\|$, where $\nabla f(\cdot; z)$ denotes the gradient of f w.r.t. the first argument and $\|\cdot\|$ denotes the Euclidean norm. Let $B(\mathbf{0}, R) := \{\mathbf{w} \in \mathbb{R}^d : \|\mathbf{w} - \mathbf{0}\| \le R\}$ denote a ball with center $\mathbf{0} \in \mathbb{R}^d$ and radius R, denoted by B_R. We also denote $A \asymp B$ if there exists universal constants $C_1, C_2 > 0$ such that $C_1 A \le B \le C_2 A$. Standard order of magnitude notation such as $\mathcal{O}(\cdot)$ will be used.

Algorithm 1 SGD

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

1: **for** $t = 1, ..., T$ **do**
2: draw j_t from the uniform distribution over the set $\{j : j \in [n]\}$
3: update $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla f(\mathbf{w}_t; z_{j_t})$.
4: **end for**

3.2 Assumptions

We first present the assumption of smoothness.

Assumption 1. *Let the constant $L > 0$. A differentiable function $g : \mathcal{W} \mapsto \mathbb{R}$ is L-smooth if*

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

Remark 2. This assumption is necessary to have the convergence of the gradients to zero [36]. It is standard in the optimization and generalization literature, e.g. [14, 22, 50, 15, 9, 26], to mention but a few. In this paper, for the optimization guarantees, we just need the empirical risk F_S to be smooth, i.e., for any $\mathbf{w}, \mathbf{w}' \in \mathcal{W}$, there holds $\|\nabla F_S(\mathbf{w}) - \nabla F_S(\mathbf{w}')\| \leq L\|\mathbf{w} - \mathbf{w}'\|$. While for the generalization guarantees, we need the function f to be smooth, i.e., for any sample $z \in \mathcal{Z}$ and $\mathbf{w}, \mathbf{w}' \in \mathcal{W}$, there holds $\|\nabla f(\mathbf{w}; z) - \nabla f(\mathbf{w}'; z)\| \leq L\|\mathbf{w} - \mathbf{w}'\|$.

With the smoothness assumption, we have the useful "descent lemma" [42]:

$$g(\mathbf{w}) - g(\mathbf{w}') \leq \langle \mathbf{w} - \mathbf{w}', \nabla g(\mathbf{w}') \rangle + \frac{L}{2}\|\mathbf{w} - \mathbf{w}'\|^2.$$

We then show our assumption on the stochastic gradient.

Assumption 3. *There exists positive real numbers $\alpha \in (1, 2]$ and $G > 0$ such that for all $\mathbf{w}_t$,*

$$\mathbb{E}_{j_t}[\|\nabla f(\mathbf{w}_t; z_{j_t})\|^\alpha] \leq G^\alpha.$$

Remark 4. It is possible that the variance of $\nabla f(\mathbf{w}_t; z_{j_t})$ is unbounded while simultaneously satisfying Assumption 3 for $\alpha < 2$, e.g. the Pareto or α-stable Levy random variables, please refer to Section 2.1 in [62] for details. This assumption is thus much weaker than the standard bounded second moment assumption. It is shown that the unbounded variance strongly corrupts the optimization process and that previous convergence proofs for SGD fail [62]. Thus, it is essential to investigate the theoretical guarantees of stochastic optimization algorithms under this heavy-tailed condition. This paper uses gradient clipping to establish high probability guarantees for many popular stochastic optimization algorithms under this assumption.

4 Main Results

In this section, we present the main results of this paper. We first consider SGD with gradient clipping in Section 4.1, and then SGDM with joint consideration of gradient clipping and momentum clipping in Section 4.2. Further, we study AdaGrad with gradient clipping in Section 4.3.1 and study a more general template of adaptive algorithms in Section 4.3.2.

In the general nonconvex case, since obtaining the global minimum is NP-hard in general, we cannot guarantee that the algorithm can find a global minimizer. Therefore, we are interested in finding the ϵ-stationary point of first-order gradient for both the optimization guarantees and the generalization guarantees [16, 31, 36, 38, 61, 9].

Algorithm 2 SGD with Clipping

Input: initial point $\mathbf{w}_1 = 0$, step sizes $\{\eta_t\}_t$, dataset $S = \{z_1, ..., z_n\}$, and clipping parameter $\tau > 0$.

1: **for** $t = 1, ..., T$ **do**
2: draw j_t from the uniform distribution over the set $\{j : j \in [n]\}$
3: obtain $\nabla \bar{f}(\mathbf{w}_t; z_{j_t}) = \frac{\nabla f(\mathbf{w}_t; z_{j_t})}{\|\nabla f(\mathbf{w}_t; z_{j_t})\|} \min\{\tau, \|\nabla f(\mathbf{w}_t; z_{j_t})\|\}$
4: update $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla \bar{f}(\mathbf{w}_t; z_{j_t})$.
5: **end for**

4.1 SGD with Clipping

The pseudocode of SGD with clipping is shown in Algorithm 2. In each iterate, SGD moves models along the reverse direction of a clipped gradient $\nabla \bar{f}(\mathbf{w}_t; z_{j_t})$, which is a biased estimate. We first present the optimization guarantee and then the generalization guarantee for clipped SGD.

Theorem 5. *Suppose the empirical risk F_S satisfies Assumption 1 and suppose Assumption 3 holds. Let $\mathbf{w}_t$ be the iterate produced by Algorithm 2. Set $\eta_t = \eta = p\frac{1}{T^{\frac{\alpha}{3\alpha-2}}}$ and $\tau = qT^{\frac{1}{3\alpha-2}}$ for some positive constants p, q such that $q \leq T^{\frac{2\alpha-2}{\alpha(3\alpha-2)}}$ and $\eta \leq 1/(12L)$. Then for any $\delta \in (0, 1)$, with probability $1 - \delta$, we have*

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla F_S(\mathbf{w}_t)\|^2 = \mathcal{O}\left(\frac{1}{T^{\frac{2\alpha-2}{3\alpha-2}}} \log\frac{1}{\delta}\right).$$

Remark 6. Theorem 5 suggests that if the empirical risk is smooth and the stochastic gradient follows from the heavy-tailed assumption, the optimization guarantee of clipped SGD has a convergence rate of the order $\mathcal{O}(\log(\frac{1}{\delta})/T^{\frac{2\alpha-2}{3\alpha-2}})$. When $\alpha = 2$, it implies $\mathcal{O}(\log(\frac{1}{\delta})/T^{\frac{1}{2}})$. We now compare Theorem 5 with the related work of clipping. Theorem 3.1 in [19] provides a high probability convergence bound for clipped SGD under the smoothness, convexity, and bounded variance conditions. Theorem 8 in [60] provides an in-expectation analysis for non-convex clipped SGD under a relaxed smoothness condition and a stronger assumption than the bounded variance, i.e., $\|\nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t)\| \leq G$ holds for any $\mathbf{w}_t$ and z_{j_t} almost surely. The most relevant result to Theorem 5 is Theorem 2 in [62]. Theorem 5 provides a high-probability result for non-convex clipped SGD, matching the in-expectation convergence rate of Theorem 2 in [62] up to logarithmic factors. It has been shown in [9] and Theorem 6 of [61] that this rate is optimal. The benefit of the high probability bound is that it holds for any training data S drawn from P and over the randomness of the algorithm. To our best knowledge, Theorem 5 provides the first high probability optimization bound for clipped SGD under an unbounded variance assumption. We sketch the proof technique of Theorem 5. The proof begins with the "descent lemma" and some decompositions, resulting in Eq. (1) in the Appendix. Unlike the in-expectation analysis, the high probability analysis requires to construct some martingale difference sequences, e.g. $\sum_{t=1}^{T} L\eta^2(\|\nabla \bar{f}(\mathbf{w}_t; z_{j_t}) - \mathbb{E}_{j_t}\nabla \bar{f}(\mathbf{w}_t; z_{j_t})\|^2 - \mathbb{E}_{j_t}\|\nabla \bar{f}(\mathbf{w}_t; z_{j_t}) - \mathbb{E}_{j_t}\nabla \bar{f}(\mathbf{w}_t; z_{j_t})\|^2)$ and $-\sum_{t=1}^{T} \eta\langle \nabla \bar{f}(\mathbf{w}_t; z_{j_t}) - \mathbb{E}_{j_t}\nabla \bar{f}(\mathbf{w}_t; z_{j_t}), \nabla F_S(\mathbf{w}_t)\rangle$. Some concentration inequalities on martingales should be used to bound these terms. The key point lies in that the Auzan-Hoeffding inequality for martingales with bounded increments fails to give the optimal rate of Theorem 5, especially when dealing with $\sum_{t=1}^{T} L\eta^2(\|\nabla \bar{f}(\mathbf{w}_t; z_{j_t}) - \mathbb{E}_{j_t}\nabla \bar{f}(\mathbf{w}_t; z_{j_t})\|^2 - \mathbb{E}_{j_t}\|\nabla \bar{f}(\mathbf{w}_t; z_{j_t}) - \mathbb{E}_{j_t}\nabla \bar{f}(\mathbf{w}_t; z_{j_t})\|^2)$. For the purpose of

the optimal rate, one must consider the conditional variance and use the Bernstein-type concentration inequality (Lemma 19). Notably that for clipped SGD, its conditional variance should be carefully controlled. Other terms like the bias $\sum_{t=1}^{T} L\eta^2 \|\mathbb{E}_{j_t} \nabla \bar{f}(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t)\|^2$ and the variance $\sum_{t=1}^{T} L\eta^2 \mathbb{E}_{j_t} \|\nabla \bar{f}(\mathbf{w}_t; z_{j_t}) - \mathbb{E}_{j_t} \nabla \bar{f}(\mathbf{w}_t; z_{j_t})\|^2$ of the clipped stochastic gradient can be bounded by its boundedness and the heavy-tailed assumption. After getting the bound in Eq. (8) in the Appendix, carefully selecting the stepsize η and clipping parameter τ obtains the optimal rate of Theorem 5.

Theorem 7. *Suppose the function f satisfies Assumption 1 and suppose Assumption 3 holds. Let $\mathbf{w}_t$ be the iterate produced by Algorithm 2. Set $\eta_t = \eta = p\frac{1}{T^{\frac{\alpha}{3\alpha-2}}}$ and $\tau = qT^{\frac{1}{3\alpha-2}}$ for some positive constants p, q such that $q \leq T^{\frac{2\alpha-2}{\alpha(3\alpha-2)}}$ and $\eta \leq 1/(12L)$. Select $T \asymp (\frac{n}{d})^{\frac{3\alpha-2}{4\alpha-4}}$. Then for any $\delta \in (0,1)$, with probability $1-\delta$, we have*

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

Remark 8. Theorem 7 shows that if the function f is smooth and the stochastic gradient follows from the heavy-tailed assumption, the generalization guarantee of clipped SGD has a convergence rate of the order $\mathcal{O}((\frac{d}{n})^{\frac{1}{2}} \log\frac{1}{\delta})$ when the iterate number $T \asymp (\frac{n}{d})^{\frac{3\alpha-2}{4\alpha-4}}$. Lemma 4.3 in [22] provides an in-expectation analysis for clipped SGD by the lens of algorithmic stability [6]. For the generalization analysis of clipped stochastic optimization algorithms, we have not found other related results in the literature. The proof of Theorem 7 begins with a decomposition, resulting in Eq. (14) in the Appendix, where $\sum_{t=1}^{T} \|\nabla F_S(\mathbf{w}_t)\|^2$ corresponds to Theorem 5 and $T\|\nabla F(\mathbf{w}_T) - \nabla F_S(\mathbf{w}_T)\|^2$ can be bounded by the uniform convergence of gradients (Lemma 21). In using Lemma 21, we need to quantify the value of R_T, which reveals the space complexity induced by the iterate update of SGD. In this spirit, we need to give the bound of SGD's iterate $\max_{1\leq t\leq T} \|\mathbf{w}_t\|$, see Eq. (11) in the Appendix. We show that this term can be bounded by the bias of the clipped stochastic gradient, the empirical risk, and the Pinelis-Bernstein inequality for martingales difference sequences (Lemma 20). Again, the conditional variance should be carefully controlled to guarantee the convergence rate of Theorem 7 when using Lemma 20. One can see that $\frac{1}{T}\sum_{t=1}^{T} \|\nabla F_S(\mathbf{w}_t)\|^2$ is decreasing along the training process, while $\|\nabla F(\mathbf{w}_T) - \nabla F_S(\mathbf{w}_T)\|^2$ is increasing, which suggests the space complexity is keeping grow along the training process. Thus, Theorem 7 reveals that an implicit regularization can be achieved by tuning the number of passes to balance the optimization and generalization error for the clipped stochastic gradient descent.

4.2 SGDM with Clipping

The pseudocode of SGDM with clipping is shown in Algorithm 3. Algorithm 3 incorporates the momentum update, $\mathbf{m}_t = \gamma \mathbf{m}_{t-1} + (1-\gamma)\nabla \bar{f}(\mathbf{w}_t; z_{j_t})$, to SGD. We first give the optimization guarantee and then the generalization guarantee for clipped SGDM.

Theorem 9. *Suppose the empirical risk F_S satisfies Assumption 1 and suppose Assumption 3 holds. Let $\mathbf{w}_t$ be the iterate produced by Algorithm 3. Set $\tau_1 = \frac{pG}{(1-\gamma)^{1/\alpha}}$, $1-\gamma = \frac{s}{T^{\frac{\alpha}{3\alpha-2}}}$, $\eta_t = \eta = \frac{q}{T^{\frac{\alpha}{3\alpha-2}}}$, and $\tau_2 = \frac{r}{T^{\frac{\alpha-1}{3\alpha-2}}}$ for some positive constants p, s, q, r such that $1-\gamma \leq 1$. Then for any $\delta \in (0,1)$, with probability $1-\delta$, we have*

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla F_S(\mathbf{w}_t)\| = \mathcal{O}\left(\frac{1}{T^{\frac{\alpha-1}{3\alpha-2}}} \log\frac{T}{\delta}\right).$$

Algorithm 3 SGDM with Clipping

Input: initial point $\mathbf{w}_1 = 0$, $\mathbf{m}_0 = 0$, step sizes $\{\eta_t\}_t$, dataset $S = \{z_1, ..., z_n\}$, momentum parameter γ, and clipping parameters $\tau_1, \tau_2 > 0$.

1: **for** $t = 1, ..., T$ **do**
2: draw j_t from the uniform distribution over the set $\{j : j \in [n]\}$
3: obtain $\nabla \bar{f}(\mathbf{w}_t; z_{j_t}) = \frac{\nabla f(\mathbf{w}_t; z_{j_t})}{\|\nabla f(\mathbf{w}_t; z_{j_t})\|} \min\{\tau_1, \|\nabla f(\mathbf{w}_t; z_{j_t})\|\}$
4: update $\mathbf{m}_t = \gamma \mathbf{m}_{t-1} + (1-\gamma)\nabla \bar{f}(\mathbf{w}_t; z_{j_t})$
5: obtain $\bar{\mathbf{m}}_t = \frac{\mathbf{m}_t}{\|\mathbf{m}_t\|} \min\{\tau_2, \|\mathbf{m}_t\|\}$
6: update $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \bar{\mathbf{m}}_t$.
7: **end for**

Remark 10. Theorem 9 shows that the optimization guarantee of SGDM with gradient clipping and momentum clipping has a convergence rate of the order $\mathcal{O}(\log\frac{T}{\delta}/T^{\frac{\alpha-1}{3\alpha-2}})$. Note that according to Jensen's inequality, the bound in Theorem 5 implies that $\frac{1}{T}\sum_{t=1}^{T} \|\nabla F_S(\mathbf{w}_t)\| = ((\frac{1}{T}\sum_{t=1}^{T} \|\nabla F_S(\mathbf{w}_t)\|)^2)^{1/2} \leq (\frac{1}{T}\sum_{t=1}^{T} \|\nabla F_S(\mathbf{w}_t)\|^2)^{1/2} \leq \mathcal{O}(\log\frac{1}{\delta}/T^{\frac{\alpha-1}{3\alpha-2}})$. Thus, Theorem 9 presents a similar order bound to Theorem 5. An improvement of Theorem 9 is that its stepsize η does not depend on the smoothness parameter L, i.e., completely oblivious to the knowledge of smoothness. We now compare Theorem 9 with the related work of clipping. As we discussed in Section 2, [59] also study SGDM with both gradient clipping and momentum clipping. Their updates are $\mathbf{m}_{t+1} = \gamma \mathbf{m}_t + (1-\gamma)\nabla f(\mathbf{w}_t; z_{j_t})$ and then $\mathbf{w}_{t+1} = \mathbf{w}_t - [v\min(\eta, \frac{\tau}{\|\mathbf{m}_{t+1}\|})\mathbf{m}_{t+1} + (1-v)\min(\eta, \frac{\tau}{\|\nabla f(\mathbf{w}_t; z_{j_t})\|})\nabla f(\mathbf{w}_t; z_{j_t})]$, where $v \in [0,1]$ is an interpolation parameter. Theorem 3.2 in [59] provides an expected optimization bound under a relaxed smoothness and a stronger assumption than the bounded variance, i.e., $\|\nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t)\| \leq G$ holds for any $\mathbf{w}_t$ and z_{j_t} almost surely, where the latter assumption is restrictive, hindering the scope of application of their results. Another related work is [9]. Theorem 2 in [9] gives a high-probability bound under the same conditions to Theorem 9 by combining the gradient clipping, momentum, and normalized momentum. Their updates are $\mathbf{m}_t = \gamma \mathbf{m}_{t-1} + (1-\gamma)\nabla \bar{f}(\mathbf{w}_t; z_{j_t})$ and then $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \frac{\mathbf{m}_t}{\|\mathbf{m}_t\|}$. In Algorithm 3, we study the clipped version of momentum. Algorithm 3 is more similar to the framework proposed in [59], where both the gradient clipping and momentum clipping are all considered. The proof techniques between ours and [59, 9] are different. We now compare Theorem 9 with [9] considering the two works all focus on high probability bound. Due to $\|\frac{\mathbf{m}_t}{\|\mathbf{m}_t\|}\| = 1$, [9] show that $F_S(\mathbf{w}_{t+1}) - F_S(\mathbf{w}_t) \leq -\eta\langle \frac{\mathbf{m}_t}{\|\mathbf{m}_t\|}, \nabla F_S(\mathbf{w}_t)\rangle + \frac{L}{2}\eta^2 = -\eta\langle \frac{\mathbf{m}_t}{\|\mathbf{m}_t\|}, \nabla F_S(\mathbf{w}_t) - \mathbf{m}_t\rangle - \eta\|\mathbf{m}_t\| + \frac{L}{2}\eta^2 \leq \eta\|\frac{\mathbf{m}_t}{\|\mathbf{m}_t\|}\|\|\nabla F_S(\mathbf{w}_t) - \mathbf{m}_t\| - \eta\|\mathbf{m}_t - \nabla F_S(\mathbf{w}_t) + \nabla F_S(\mathbf{w}_t)\| + \frac{L}{2}\eta^2 \leq 2\eta\|\nabla F_S(\mathbf{w}_t) - \mathbf{m}_t\| - \eta\|\nabla F_S(\mathbf{w}_t)\| + \frac{L}{2}\eta^2$, which implies $\|\nabla F_S(\mathbf{w}_t)\| \leq 2\|\nabla F_S(\mathbf{w}_t) - \mathbf{m}_t\| + \frac{L}{2}\eta - \frac{(F_S(\mathbf{w}_{t+1}) - F_S(\mathbf{w}_t))}{\eta}$. [9] then use Freedman's inequality to bound the term $\|\nabla F_S(\mathbf{w}_t) - \mathbf{m}_t\|$. However, the clipped momentum doesn't have the property $\|\frac{\mathbf{m}_t}{\|\mathbf{m}_t\|}\| = 1$. In the proof of Theorem 9, we need to consider two cases, i.e., $\|\mathbf{m}_t\| \geq \tau_2$ and $\|\mathbf{m}_t\| < \tau_2$. In the former

case, we need to prove that $\|\nabla F_S(\mathbf{w}_t)\| \leq 3\frac{F_S(\mathbf{w}_t)-F_S(\mathbf{w}_{t+1})}{\eta\tau_2} + 4\|\mathbf{m}_t - \nabla F_S(\mathbf{w}_t)\| + \frac{3L}{2}\eta\tau_2$, and in the latter case, we need to prove that $\|\nabla F_S(\mathbf{w}_t)\|^2 \leq \frac{2(F_S(\mathbf{w}_t)-F_S(\mathbf{w}_{t+1}))}{\eta} + \|\nabla F_S(\mathbf{w}_t) - \mathbf{m}_t\|^2 + L\eta\tau_2^2$. We then use the Pinelis-Bernstein inequality for martingales difference sequences (Lemma 20) to bound the terms $\|\nabla F_S(\mathbf{w}_t)-\mathbf{m}_t\|$ and $\|\nabla F_S(\mathbf{w}_t)-\mathbf{m}_t\|^2$. Additionally, in practice, a more common application of the normalized momentum should be $\mathbf{w}_{t+1} = \mathbf{w}_t-\eta_t\frac{\mathbf{m}_t}{\|\mathbf{m}_t\|+\beta}$ with $\beta > 0$. However, this pattern of iterate update violates the property $\|\frac{\mathbf{m}_t}{\|\mathbf{m}_t\|}\| = 1$, which plays an essential role in the proof in [9]. It is unclear whether the proof techniques of [9] can guarantee the convergence for this more commonly used pattern of iterate update. The clear motivations of our study on momentum clipping include that the clipping doesn't have such an issue and is more common in practice, that Appendix C in [59] suggests that there are some practical issues that make normalized momentum less favorable than traditional clipping methods, and that [59] only provide in-expectation analysis. Considering the above analysis, we believe that Theorem 9 is an important result for stochastic optimization with clipping.

Theorem 11. *Suppose the function f satisfies Assumption 1 and suppose Assumption 3 holds. Let $\mathbf{w}_t$ be the iterate produced by Algorithm 3. Set $\tau_1 = \frac{pG}{(1-\gamma)^{1/\alpha}}$, $1-\gamma = \frac{s}{T^{\frac{\alpha}{3\alpha-2}}}$, $\eta_t = \eta = \frac{q}{T^{\frac{\alpha}{3\alpha-2}}}$, and $\tau_2 = \frac{r}{T^{\frac{\alpha-1}{3\alpha-2}}}$ for some positive constants p, s, q, r such that $1-\gamma \leq 1$. Select $T \asymp (\frac{n}{d})^{\frac{3\alpha-2}{4\alpha-4}}$. Then for any $\delta \in (0,1)$, with probability $1-\delta$, we have*

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\mathbf{w}_t)\| = \mathcal{O}\left(\left(\frac{d}{n}\right)^{\frac{1}{4}}\log\frac{n}{d\delta}\right).$$

Remark 12. According to Jensen's inequality, Theorem 11 shows a generalization bound of a similar order to Theorem 7. For the generalization analysis of clipped SGDM and even SGDM, we have not found related results in the literature. The analysis pattern of Theorem 11 follows Remark 8. Thus, Theorem 11 also reveals that an implicit regularization can be achieved by tuning the number of passes to balance the optimization and generalization error for the clipped stochastic gradient descent with momentum.

4.3 SGDAS with Clipping

After the momentum technique, this section studies SGD with the adaptive stepsizes. We first consider AdaGrad and then a more general form of adaptive accelerated algorithms, including AdaGrad and adaptive RSAG as specific examples.

4.3.1 AdaGrad

The pseudocode of AdaGrad with clipping is shown in Algorithm 4. Compared to the original AdaGrad [13, 40], in each iterate, Algorithm 4 uses a clipped gradient estimate $\nabla\bar{f}(\mathbf{w}_t; z_{j_t})$. We first give the optimization guarantee and then the generalization guarantee.

Theorem 13. *Suppose the empirical risk F_S satisfies Assumption 1 and suppose Assumption 3 holds. Assume that $F_S(\mathbf{w}) \leq M$ for all $\mathbf{w}$ for some M. Let $\mathbf{w}_t$ be the iterate produced by Algorithm 4. Set $\tau = pT^{\frac{1}{3\alpha-2}}$ for some positive constants p such that $p \leq T^{\frac{2\alpha-2}{\alpha(3\alpha-2)}}$. Then for any $\delta \in (0,1)$, with probability $1-\delta$, we have*

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F_S(\mathbf{w}_t)\|^2 = \mathcal{O}\left(\frac{1}{T^{\frac{2\alpha-2}{3\alpha-2}}}\log\frac{1}{\delta}\right).$$

Algorithm 4 AdaGrad with Clipping

Input: initial point $\mathbf{w}_1 = 0$, step sizes $\{\eta_t\}_t$, dataset $S = \{z_1, ..., z_n\}$, $G_0 > 0$, and $\tau > 0$.

1: **for** $t = 1, ..., T$ **do**
2: draw j_t from the uniform distribution over the set $\{j : j \in [n]\}$
3: obtain $\nabla\bar{f}(\mathbf{w}_t; z_{j_t}) = \frac{\nabla f(\mathbf{w}_t; z_{j_t})}{\|\nabla f(\mathbf{w}_t; z_{j_t})\|}\min\{\tau, \|\nabla f(\mathbf{w}_t; z_{j_t})\|\}$
4: obtain $\eta_t = \frac{1}{\sqrt{G_0^2+\sum_{k=1}^{t}\|\nabla\bar{f}(\mathbf{w}_t; z_{j_t})\|^2}}$
5: update $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t\nabla\bar{f}(\mathbf{w}_t; z_{j_t})$.
6: **end for**

Remark 14. Theorem 13 shows that if F_S is smooth and bounded and the stochastic gradient follows from the heavy-tailed assumption, the optimization guarantee of clipped AdaGrad has a convergence rate of the order $\mathcal{O}(\log\frac{1}{\delta}/T^{\frac{2\alpha-2}{3\alpha-2}})$. Theorem 13 requires F_S to be bounded additionally. This assumption also appears in Theorem 4 of [8] and Theorem 6 of [55] when they prove the convergence rate for adaptive algorithms. To our best knowledge, Theorem 13 provides the first optimization bound for AdaGrad with clipping. The proof technique of clipped AdaGrad is different from the clipped SGD and SGDM. With the "descent lemma" and some decompositions, we instead prove that $\sum_{t=1}^{T}\|\nabla F_S(\mathbf{w}_t)\|^2 \leq (2M+L)\sqrt{G_0^2+\sum_{t=1}^{T}\|\nabla\bar{f}(\mathbf{w}_t; z_{j_t})\|^2} - \sum_{t=1}^{T}\langle\nabla\bar{f}(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t), \nabla F_S(\mathbf{w}_t)\rangle$ for clipped AdaGrad. Then, we need to bound the terms $\sum_{t=1}^{T}\|\nabla\bar{f}(\mathbf{w}_t; z_{j_t})\|^2$ and $-\sum_{t=1}^{T}\langle\nabla\bar{f}(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t), \nabla F_S(\mathbf{w}_t)\rangle$ with the term $\sum_{t=1}^{T}\|\nabla F_S(\mathbf{w}_t)\|^2$, see Eqs. (21), (22) and (24) in the Appendix for details. To guarantee the optimal rate of Theorem 13, in bounding the two terms, we need to use the Bernstein-type concentration inequality (Lemma 19) since the Auzan-Hoeffding inequality for martingales with bounded increments leads to the sub-optimal rates. During this process, the conditional variance must be carefully considered. Finally, solving the quadratic inequality of $\sum_{t=1}^{T}\|\nabla F_S(\mathbf{w}_t)\|^2$, we get the optimization bound of Theorem 13. We now compare Theorem 13 with the results of clipped SGD and clipped SGDM (Theorem 5 and Theorem 9). An improvement of Theorem 13 is that compared to clipped SGD, the stepsize η_t of clipped AdaGrad does not depend on the smoothness parameter L and the parameter α of Assumption 3 and compared to clipped SGDM, the stepsize η_t of clipped AdaGrad does not depend on the parameter α of Assumption 3.

Theorem 15. *Suppose the function f satisfies Assumption 1 and suppose Assumption 3 holds. Assume that $F_S(\mathbf{w}) \leq M$ for all $\mathbf{w}$ for some M. Let $\mathbf{w}_t$ be the iterate produced by Algorithm 4. Set $\tau = pT^{\frac{1}{3\alpha-2}}$ for some positive constants p such that $p \leq T^{\frac{2\alpha-2}{\alpha(3\alpha-2)}}$. Select $T \asymp (\frac{n}{d})^{\frac{3\alpha-2}{4\alpha-4}}$. Then for any $\delta \in (0,1)$, with probability $1-\delta$, we have*

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\mathbf{w}_t)\|^2 = \mathcal{O}\Big(\Big(\frac{d}{n}\Big)^{\frac{2\alpha-2}{5\alpha-4}}\log(\frac{1}{\delta})\log\Big(1+(\frac{n}{d})^{\frac{3\alpha}{5\alpha-4}}\Big)\Big).$$

Remark 16. Theorem 15 shows that the generalization guarantee of clipped AdaGrad has a convergence rate of the order $\mathcal{O}((\frac{d}{n})^{\frac{2\alpha-2}{5\alpha-4}}\log(1/\delta)\log(1+(\frac{n}{d})^{\frac{3\alpha}{5\alpha-4}}))$ when the iterate number $T \asymp (\frac{n}{d})^{\frac{3\alpha-2}{4\alpha-4}}$. When $\alpha = 2$, Theorem 15 implies $\mathcal{O}((\frac{d}{n})^{\frac{1}{3}}\log(1/\delta)\log(1+\frac{n}{d}))$. For the generalization analysis of clipped AdaGrad and even AdaGrad, we have not found related results in the literature. The analysis pattern of Theorem 15 follows

Algorithm 5 Adaptive Accelerated Algorithms with Clipping

Input: initial point $\mathbf{w}_1 = \tilde{\mathbf{w}}_1$, $\alpha_t \in (0,1]$, step sizes $\{\eta_t\}_t$ and $\{\beta_t\}_t$, dataset $S = \{z_1, ..., z_n\}$, $G_0 > 0$, and $\tau > 0$.

1: **for** $t = 1, ..., T$ **do**
2: draw j_t from the uniform distribution over the set $\{j : j \in [n]\}$
3: obtain $\bar{\mathbf{w}}_t = \alpha_t \mathbf{w}_t + (1-\alpha_t)\tilde{\mathbf{w}}_t$
4: obtain $\nabla \bar{f}(\bar{\mathbf{w}}_t; z_{j_t}) = \frac{\nabla f(\bar{\mathbf{w}}_t; z_{j_t})}{\|\nabla f(\bar{\mathbf{w}}_t; z_{j_t})\|} \min\{\tau, \|\nabla f(\bar{\mathbf{w}}_t; z_{j_t})\|\}$
5: update $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla \bar{f}(\bar{\mathbf{w}}_t; z_{j_t})$
6: update $\tilde{\mathbf{w}}_{t+1} = \bar{\mathbf{w}}_t - \beta_t \nabla \bar{f}(\bar{\mathbf{w}}_t; z_{j_t})$.
7: **end for**

Remark 8 and also reveals the implicit regularization effect. Investigating whether the generalization bound of clipped AdaGrad can achieve the similar order to SGD or SGDM is an interesting open problem.

4.3.2 Adaptive Accelerated Algorithms

We then study a general form of adaptive accelerated algorithm, Algorithm 5, which corresponds to a clipped version of Algorithm 2 in [29]. We introduce some adaptive algorithms covered by Algorithm 5. Define $\lambda_t = \frac{1}{\sqrt{G_0^2 + \sum_{k=1}^t \|\nabla \bar{f}(\bar{\mathbf{w}}_t; z_{j_t})\|^2}}$. When $\eta_t = \beta_t = \lambda_t$, Algorithm 5 becomes the clipped AdaGrad. When $\eta_t = \lambda_t$ and $\beta_t = (1+\alpha_t)\eta_t$, where $\alpha_t = \frac{2}{t+1}$, Algorithm 5 becomes the clipped RSAG [17]. Note that for Algorithm 5, we are interested in the iterate $\bar{\mathbf{w}}_t$. Assumption 3 should be assumed on $\bar{\mathbf{w}}$, i.e., $\mathbb{E}_{j_t}[\|\nabla f(\bar{\mathbf{w}}_t; z_{j_t})\|^\alpha] \leq G^\alpha$. We present the optimization guarantee below.

Theorem 17. *Suppose the empirical risk F_S satisfies Assumption 1 and suppose Assumption 3 holds. Assume that $F_S(\mathbf{w}) \leq M$ for all $\mathbf{w}$ for some M. Let $\bar{\mathbf{w}}_t$ be the iterate produced by Algorithm 5. Set $\tau = pT^{\frac{1}{3\alpha-2}}$ for some positive constants p such that $p \leq T^{\frac{2\alpha-2}{\alpha(3\alpha-2)}}$. Then for any $\delta \in (0,1)$, with probability $1-\delta$, we have*

$$\frac{1}{T}\sum_{t=1}^T \|\nabla F_S(\bar{\mathbf{w}}_t)\|^2 = \mathcal{O}\left(\frac{1}{T^{\frac{2\alpha-2}{3\alpha-2}}} \log\frac{1}{\delta}\right).$$

Remark 18. Algorithm 5 corresponds to a specific reformulation of Nesterov's acceleration [17]. This reformulation was referred to as linear coupling in [1], which is a combination of mirror descent, SGD, and averaging. Theorem 17 shows a similar $\mathcal{O}(\log\frac{1}{\delta}/T^{\frac{2\alpha-2}{3\alpha-2}})$ rate to Theorem 13. When $\alpha = 2$, it implies $\mathcal{O}(\log(\frac{1}{\delta})/T^{\frac{1}{2}})$. In the related work, [29] provide a convergence rate of the order $\mathcal{O}(\log(1/\delta)/\sqrt{T})$ for Algorithm 5 without clipping by assuming the smoothness, Lipschitz continuity of F_S, bounded variance, and $\|\nabla f(\mathbf{w}_t; z_{j_t})\| \leq G$ holding for all $\mathbf{w}_t$ and z_{j_t} almost surely. By comparison, Theorem 17 gives the guarantee for a heavy-tailed assumption allowing unbounded variance, and the overall conditions are weaker than [29].

4.4 Summary of Results

We provide the results obtained in this paper and the high probability results of related work in the non-convex setting with gradient clipping in Table 1. Here, we provide some descriptions of Table 1. S means the smoothness, S-S means second-order smoothness, and α means Assumption 3. We say a function g is ρ-second-order smoothness if for every $\mathbf{w}_1, \mathbf{w}_2 \in \mathcal{W}$ and $\mathbf{y} \in \mathbb{R}^d$, there holds

$$\|(\nabla^2 g(\mathbf{w}_1) - \nabla^2 g(\mathbf{w}_2))\mathbf{y}\|^2 \leq \rho\|\mathbf{w}_1 - \mathbf{w}_2\|\|\mathbf{y}\|.$$

One can derive the convergence bound and generalization bound for Algorithm 2 with the second-order smoothness by incorporating our proof technique and the technique of [9]. We leave it to the interested readers. The difference between ours and [9] has been discussed in Remark 10.

Moreover, the comparison between our results and the results of related work (in-expectation analysis and high probability analysis) has been discussed in previous Remarks. We won't repeat it here and only provide an intuitive display of the related results here. One can see from Table 1 that we have provided a series of high probability convergence bounds and high probability generalization bounds for non-convex stochastic optimization with clipping that the related work does not involve.

5 Conclusions

This paper provides a high probability analysis for non-convex stochastic optimization with clipping. We establish learning guarantees for clipped SGD and its variants of momentum and adaptive stepsizes under a heavy-tailed assumption of the stochastic gradients. Our analysis involves joint consideration of optimization and generalization performance, which systematically demonstrates the learning guarantees of non-convex stochastic optimization with gradient clipping from the two perspectives, and covers many popular stochastic optimization algorithms. We believe our theoretical findings can provide deep insights into the theoretical properties of stochastic optimization with clipping.

6 Auxiliary Lemmas

The following Lemma 19 and Lemma 20 provide concentration inequalities for martingales.

Lemma 19 ([63]). *Let $z_1, ..., z_n$ be a sequence of randoms variables such that z_k may depend the previous variables $z_1, ..., z_{k-1}$ for all $k = 1, ..., n$. Consider a sequence of functionals $\xi_k(z_1, ..., z_k)$, $k = 1, ..., n$. Let $\sigma_n^2 = \sum_{k=1}^n \mathbb{E}_{z_k}[(\xi_k - \mathbb{E}_{z_k}[\xi_k])^2]$ be the conditional variance. Assume $|\xi_k - \mathbb{E}_{z_k}[\xi_k]| \leq b$ for each k. Let $\rho \in (0,1]$ and $\delta \in (0,1)$. With probability at least $1-\delta$ we have*

$$\sum_{k=1}^n \xi_k - \sum_{k=1}^n \mathbb{E}_{z_k}[\xi_k] \leq \frac{\rho\sigma_n^2}{b} + \frac{b\log\frac{1}{\delta}}{\rho}.$$

Lemma 20 ([54]). *Let $\{\xi_k\}_{k\in\mathbb{N}}$ be a martingale difference sequence in $\mathbb{R}^d$. Suppose that almost surely $\|\xi_k\| \leq D$ and $\sum_{k=1}^t \mathbb{E}[\|\xi_k\|^2 | \xi_1, ..., \xi_{k-1}] \leq \sigma_t^2$. Then, for any $0 < \delta < 1$, the following inequality holds with probability at least $1-\delta$*

$$\max_{1\leq j\leq t}\left\|\sum_{k=1}^j \xi_k\right\| \leq 2\left(\frac{D}{3} + \sigma_t\right)\log\frac{2}{\delta}.$$

The following Lemma 21 states the uniform convergence of the gradient, which will be used to derive the generalization bound of this paper.

Table 1. Summary of Results.

REF.	ALGORITHM	ASSUMPTION	MEASURE	GUARANTEE
[9]	SGDM	S, α	$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F_S(\mathbf{w}_t)\|$	$\mathcal{O}\Big(\frac{\log(T/\delta)}{T^{\frac{\alpha-1}{3\alpha-2}}}\Big)$
		S, α, S-S	$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F_S(\mathbf{w}_t)\|$	$\mathcal{O}\Big(\frac{\log(T/\delta)}{T^{\frac{2\alpha-2}{5\alpha-3}}}\Big)$
OURS	SGD	S, α	$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F_S(\mathbf{w}_t)\|^2$	$\mathcal{O}\Big(\frac{1}{T^{\frac{2\alpha-2}{3\alpha-2}}}\log\frac{1}{\delta}\Big)$
		S, α	$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\mathbf{w}_t)\|^2$	$\mathcal{O}\Big(\big(\frac{d}{n}\big)^{\frac{1}{2}}\log\frac{1}{\delta}\Big)$
	SGDM	S, α	$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F_S(\mathbf{w}_t)\|$	$\mathcal{O}\Big(\frac{1}{T^{\frac{\alpha-1}{3\alpha-2}}}\log\frac{T}{\delta}\Big)$
		S, α	$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\mathbf{w}_t)\|$	$\mathcal{O}\Big(\big(\frac{d}{n}\big)^{\frac{1}{4}}\log\frac{n}{d\delta}\Big)$
	ADAGRAD	S, α	$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F_S(\mathbf{w}_t)\|^2$	$\mathcal{O}\Big(\frac{1}{T^{\frac{2\alpha-2}{3\alpha-2}}}\log\frac{1}{\delta}\Big)$
		S, α	$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\mathbf{w}_t)\|^2$	$\mathcal{O}\Big(\big(\frac{d}{n}\big)^{\frac{2\alpha-2}{5\alpha-4}}\log\frac{1}{\delta}\log(1+(\frac{n}{d})^{\frac{3\alpha}{5\alpha-4}})\Big)$
	ALGORITHM 5	S, α	$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F_S(\bar{\mathbf{w}}_t)\|^2$	$\mathcal{O}\Big(\frac{1}{T^{\frac{2\alpha-2}{3\alpha-2}}}\log\frac{1}{\delta}\Big)$

Lemma 21 ([33]). *Let $\delta\in(0,1)$, $R>0$, and $S=\{z_1,...,z_n\}$ be a set of i.i.d. samples. Suppose the function f satisfies Assumption 1. Then with probability at least $1-\delta$ we have*

$$\sup_{\mathbf{w}\in B_R}\|\nabla F(\mathbf{w})-\nabla F_S(\mathbf{w})\|$$
$$\leq\frac{(LR+b)}{\sqrt{n}}\Big(2+2\sqrt{48e\sqrt{2}(\log 2+d\log(3e))}+\sqrt{2\log(\frac{1}{\delta})}\Big),$$

where e is the base of the natural logarithm.

The following Lemma 22 is from online learning and is often used in the study of adaptive algorithms [29, 57, 35, 36].

Lemma 22. *Let $a_1,...,a_n$ be a sequence of non-negative real numbers. Then, it holds that*

$$\sqrt{\sum_{i=1}^{n}a_i}\leq\sum_{i=1}^{n}\frac{a_i}{\sqrt{\sum_{k=1}^{i}a_k}}\leq 2\sqrt{\sum_{i=1}^{n}a_i},$$

and

$$\sum_{i=1}^{n}\frac{a_i}{\sum_{k=1}^{i}a_k}\leq 1+\log\Big(1+\sum_{i=1}^{n}a_i\Big).$$

Acknowledgements

We thank the anonymous reviewers for their valuable and constructive suggestions and comments. This work is supported by the National Natural Science Foundation of China (NO.62076234); the National Key Research and Development Project (No.2022YFB2703102); the Beijing Natural Science Foundation (No.4222029); the "Intelligent Social Governance Interdisciplinary Platform, Major Innovation & Planning Interdisciplinary Platform for the "Double-First Class" Initiative, Renmin University of China"; the Beijing Outstanding Young Scientist Program (NO.BJJWZYJH012019100020098); the Public Computing Cloud, Renmin University of China; the Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China (NO.2021030199); the Huawei-Renmin University joint program on Information Retrieval; the Unicom Innovation Ecological Cooperation Plan; and the CCF-Huawei Populus Grove Fund.

References

[1] Zeyuan Allen-Zhu and Lorenzo Orecchia, 'Linear coupling: An ultimate unification of gradient and mirror descent', *arXiv preprint arXiv:1407.1537*, (2014).

[2] Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar, 'Stability of stochastic gradient descent on nonsmooth convex losses', in *Advances in Neural Information Processing Systems*, pp. 4381–4391, (2020).

[3] Léon Bottou, 'Large-scale machine learning with stochastic gradient descent', in *Proceedings of COMPSTAT'2010*, 177–186, (2010).

[4] Léon Bottou, Frank E Curtis, and Jorge Nocedal, 'Optimization methods for large-scale machine learning', *Siam Review*, **60**(2), 223–311, (2018).

[5] Olivier Bousquet and Léon Bottou, 'The tradeoffs of large scale learning', in *Advances in Neural Information Processing Systems*, pp. 161–168, (2007).

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

[7] Alexander Camuto, Xiaoyu Wang, Lingjiong Zhu, Chris Holmes, Mert Gürbüzbalaban, and Umut Şimşekli, 'Asymmetric heavy tails and implicit bias in gaussian noise injections', in *International Conference on Machine Learning*, (2021).

[8] Ashok Cutkosky and Harsh Mehta, 'Momentum improves normalized sgd', in *International Conference on Machine Learning*, pp. 2260–2268, (2020).

[9] Ashok Cutkosky and Harsh Mehta, 'High-probability bounds for non-convex stochastic optimization with heavy tails', in *Advances in Neural Information Processing Systems*, (2021).

[10] Damek Davis and Dmitriy Drusvyatskiy, 'High probability guarantees for stochastic convex optimization', in *Conference on Learning Theory*, pp. 1411–1427, (2020).

[11] Damek Davis, Dmitriy Drusvyatskiy, Lin Xiao, and Junyu Zhang, 'From low probability to high confidence in stochastic convex optimization.', *Journal of Machine Learning Research*, **22**, 49–1, (2021).

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, 'Bert: Pre-training of deep bidirectional transformers for language understanding', *arXiv preprint arXiv:1810.04805*, (2018).

[13] John Duchi, Elad Hazan, and Yoram Singer, 'Adaptive subgradient methods for online learning and stochastic optimization.', *Journal of machine learning research*, **12**(7), (2011).

[14] Vitaly Feldman and Jan Vondrak, 'High probability generalization bounds for uniformly stable algorithms with nearly optimal rate', in *Conference on Learning Theory*, pp. 1270–1279, (2019).

[15] Dylan J. Foster, Ayush Sekhari, and Karthik Sridharan, 'Uniform convergence of gradients for non-convex learning and optimization', in *Advances in Neural Information Processing Systems*, pp. 8745–8756, (2018).

[16] Saeed Ghadimi and Guanghui Lan, 'Stochastic first- and zeroth-order

methods for nonconvex stochastic programming', *Siam Journal on Optimization*, **23**(4), 2341–2368, (2013).
[17] Saeed Ghadimi and Guanghui Lan, 'Accelerated gradient methods for nonconvex nonlinear and stochastic programming', *Mathematical Programming*, **156**(1), 59–99, (2016).
[18] Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep learning*, MIT press, 2016.
[19] Eduard Gorbunov, Marina Danilova, and Alexander Gasnikov, 'Stochastic optimization with heavy-tailed noise via accelerated gradient clipping', in *Advances in Neural Information Processing Systems*, pp. 15042–15053, (2020).
[20] Eduard Gorbunov, Marina Danilova, Innokentiy Shibaev, Pavel Dvurechensky, and Alexander Gasnikov, 'Near-optimal high probability complexity bounds for non-smooth stochastic optimization with heavy-tailed noise', *arXiv preprint arXiv:2106.05958*, (2021).
[21] Mert Gurbuzbalaban, Umut Simsekli, and Lingjiong Zhu, 'The heavy-tail phenomenon in sgd', in *International Conference on Machine Learning*, pp. 3964–3975, (2021).
[22] Moritz Hardt, Benjamin Recht, and Yoram Singer, 'Train faster, generalize better: stability of stochastic gradient descent', in *International Conference on Machine Learning*, pp. 1225–1234, (2016).
[23] Nicholas JA Harvey, Christopher Liaw, Yaniv Plan, and Sikander Randhawa, 'Tight analyses for non-smooth stochastic gradient descent', in *Conference on Learning Theory*, pp. 1579–1613, (2019).
[24] Elad Hazan et al., 'Introduction to online convex optimization', *Foundations and Trends® in Optimization*, **2**(3-4), 157–325, (2016).
[25] Elad Hazan and Satyen Kale, 'Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization', *Journal of Machine Learning Research*, **15**(1), 2489–2512, (2014).
[26] Prateek Jain and Purushottam Kar, 'Non-convex optimization for machine learning', *Foundations and Trends® in Machine Learning*, **10**(3-4), 142–336, (2017).
[27] Prateek Jain, Dheeraj Nagaraj, and Praneeth Netrapalli, 'Making the last iterate of sgd information theoretically optimal', in *Conference on Learning Theory*, pp. 1752–1755, (2019).
[28] Sham M Kakade and Ambuj Tewari, 'On the generalization ability of online strongly convex programming algorithms', in *Advances in Neural Information Processing Systems*, pp. 801–808, (2009).
[29] Ali Kavis, Kfir Yehuda Levy, and Volkan Cevher, 'High probability bounds for a class of nonconvex algorithms with adagrad stepsize', in *International Conference on Learning Representations*, (2022).
[30] Guanghui Lan, *First-order and Stochastic Optimization Methods for Machine Learning*, Springer Nature, 2020.
[31] Yunwen Lei, Ting Hu, and Ke Tang, 'Generalization performance of multi-pass stochastic gradient descent with convex loss functions.', *Journal of Machine Learning Research*, **22**, 25–1, (2021).
[32] Yunwen Lei and Ke Tang, 'Stochastic composite mirror descent: Optimal bounds with high probabilities', in *Advances in Neural Information Processing Systems*, pp. 1519–1529, (2018).
[33] Yunwen Lei and Ke Tang, 'Learning rates for stochastic gradient descent with nonconvex objectives', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (2021).
[34] Shaojie Li and Yong Liu, 'High probability guarantees for nonconvex stochastic gradient descent with heavy tails', in *International Conference on Machine Learning*, pp. 12931–12963, (2022).
[35] Xiaoyu Li and Francesco Orabona, 'On the convergence of stochastic gradient descent with adaptive stepsizes', in *International Conference on Artificial Intelligence and Statistics*, pp. 983–992, (2019).
[36] Xiaoyu Li and Francesco Orabona, 'A high probability analysis of adaptive sgd with momentum', in *Workshop on Beyond First Order Methods in ML Systems at ICML*, (2020).
[37] Ben London, 'A pac-bayesian analysis of randomized learning with application to stochastic gradient descent', in *Advances in Neural Information Processing Systems*, pp. 2931–2940, (2017).
[38] Liam Madden, Emiliano Dall'Anese, and Stephen Becker, 'High-probability convergence bounds for non-convex stochastic gradient descent', *arXiv preprint arXiv:2006.05610v4*, (2021).
[39] Vien V Mai and Mikael Johansson, 'Stability and convergence of stochastic gradient clipping: Beyond lipschitz continuity and smoothness', in *International Conference on Machine Learning*, pp. 7325–7335, (2021).
[40] H Brendan McMahan and Matthew Streeter, 'Adaptive bound optimization for online convex optimization', *arXiv preprint arXiv:1002.4908*, (2010).
[41] Aditya Krishna Menon, Ankit Singh Rawat, Sashank J Reddi, and Sanjiv Kumar, 'Can gradient clipping mitigate label noise?', in *International Conference on Learning Representations*, (2019).
[42] Iu. E. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, 2014.
[43] Yurii E Nesterov, 'A method for solving the convex programming problem with convergence rate $o(1/k^2)$', in *Dokl. akad. nauk Sssr*, volume 269, pp. 543–547, (1983).
[44] Behnam Neyshabur, Srinadh Bhojanapalli, David Mcallester, and Nati Srebro, 'Exploring generalization in deep learning', in *Advances in Neural Information Processing Systems*, pp. 5947–5956, (2017).
[45] Abhishek Panigrahi, Raghav Somani, Navin Goyal, and Praneeth Netrapalli, 'Non-gaussianity of stochastic gradient noise', *arXiv preprint arXiv:1910.09626*, (2019).
[46] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio, 'On the difficulty of training recurrent neural networks', in *International conference on machine learning*, pp. 1310–1318, (2013).
[47] Boris T Polyak, 'Some methods of speeding up the convergence of iteration methods', *Ussr computational mathematics and mathematical physics*, **4**(5), 1–17, (1964).
[48] Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan, 'Making gradient descent optimal for strongly convex stochastic optimization', in *International Conference on Machine Learning*, pp. 1571–1578, (2012).
[49] Ali Ramezani-Kebrya, Ashish Khisti, and Ben Liang, 'On the generalization of stochastic gradient descent with momentum', *arXiv preprint arXiv:2102.13653*, (2021).
[50] Sashank J. Reddi, Ahmed Hefny, Suvrit Sra, Barnabás Póczós, and Alex Smola, 'Stochastic variance reduction for nonconvex optimization', in *International Conference on Machine Learning*, pp. 314–323, (2016).
[51] Herbert Robbins and Sutton Monro, 'A stochastic approximation method', *The annals of mathematical statistics*, 400–407, (1951).
[52] Umut Şimşekli, Mert Gürbüzbalaban, Thanh Huy Nguyen, Gaël Richard, and Levent Sagun, 'On the heavy-tailed theory of stochastic gradient descent for deep neural networks', *arXiv preprint arXiv:1912.00018*, (2019).
[53] Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban, 'A tail-index analysis of stochastic gradient noise in deep neural networks', in *International Conference on Machine Learning*, pp. 5827–5837, (2019).
[54] Pierre Tarres and Yuan Yao, 'Online learning as stochastic approximation of regularization paths: Optimality and almost-sure convergence', *IEEE Transactions on Information Theory*, **60**(9), 5716–5735, (2014).
[55] Hoang Tran and Ashok Cutkosky, 'Better sgd using second-order momentum', *arXiv preprint arXiv:2103.03265*, (2021).
[56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, 'Attention is all you need', in *Advances in neural information processing systems*, pp. 5998–6008, (2017).
[57] Rachel Ward, Xiaoxia Wu, and Leon Bottou, 'Adagrad stepsizes: Sharp convergence over nonconvex landscapes', in *International Conference on Machine Learning*, pp. 6677–6686, (2019).
[58] Yang You, Igor Gitman, and Boris Ginsburg, 'Scaling sgd batch size to $32k$ for imagenet training', *arXiv preprint arXiv:1708.03888*, (2017).
[59] Bohang Zhang, Jikai Jin, Cong Fang, and Liwei Wang, 'Improved analysis of clipping algorithms for non-convex optimization', *arXiv preprint arXiv:2010.02519*, (2020).
[60] Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie, 'Why gradient clipping accelerates training: A theoretical justification for adaptivity', in *International Conference on Learning Representations*, (2019).
[61] Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, and Suvrit Sra, 'Why are adaptive methods good for attention models?', in *Advances in Neural Information Processing Systems*, (2020).
[62] Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank J Reddi, Sanjiv Kumar, and Suvrit Sra, 'Why adam beats sgd for attention models', (2019).
[63] Tong Zhang, 'Data dependent concentration bounds for sequential prediction algorithms', in *Conference on Learning Theory*, pp. 173–187, (2005).
[64] Dongruo Zhou, Jinghui Chen, Yuan Cao, Yiqi Tang, Ziyan Yang, and Quanquan Gu, 'On the convergence of adaptive gradient methods for nonconvex optimization', *arXiv preprint arXiv:1808.05671*, (2018).