High-Probability Bound for Non-Smooth Non-Convex Stochastic Optimization with Heavy Tails

Langqi Liu [1 2] Yibo Wang [1 2] Lijun Zhang [1 2]

Abstract

Recently, Cutkosky et al. introduce the online-to-non-convex framework, which utilizes online learning methods to solve non-smooth non-convex optimization problems, and achieves an $\mathcal{O}(\epsilon^{-3}\delta^{-1})$ gradient complexity for finding (δ, ϵ)-stationary points. However, their results rely on the bounded variance assumption of stochastic gradients and only hold in expectation. To address these limitations, we investigate the case that stochastic gradients obey heavy-tailed distributions with finite $\mathfrak{p}$-th moments for some $\mathfrak{p} \in (1, 2]$, and propose a novel algorithm which is able to identify a (δ, ϵ)-stationary point with high probability, after consuming $\tilde{\mathcal{O}}(\epsilon^{-\frac{2\mathfrak{p}-1}{\mathfrak{p}-1}}\delta^{-1})$ stochastic gradients. The key idea is first incorporating the gradient clipping technique into the online-to-non-convex framework to produce a sequence of points, the averaged gradient norms of which is no greater than ϵ. Then, we propose a validation method to select one (δ, ϵ)-stationary point among the candidates. When gradient distributions have bounded variance, i.e., $\mathfrak{p} = 2$, our result turns into $\tilde{\mathcal{O}}(\epsilon^{-3}\delta^{-1})$, which improves the existing $\tilde{\mathcal{O}}(\epsilon^{-4}\delta^{-1})$ high-probability bound. When the objective is smooth, our algorithm can also find an ϵ-stationary point with $\tilde{\mathcal{O}}(\epsilon^{-\frac{3\mathfrak{p}-2}{\mathfrak{p}-1}})$ gradient queries.

1. Introduction

Non-convex optimization holds a critical position within machine learning because the training process of many machine learning models, especially deep neural networks, can be formulated as solving non-convex optimization problems. With the trend towards building larger and more powerful machine learning models, there is a growing demand for efficient methods of non-convex optimization, which has aroused extensive research interest (Ghadimi & Lan, 2013; Kingma & Ba, 2015; Carmon et al., 2017; Goyal et al., 2017; Levy et al., 2021; Qiu et al., 2022; Jiang et al., 2022).

The majority of prior research in non-convex optimization relies on the smoothness assumption of the objective function, under which we aim to find an ϵ-stationary point, i.e., a point $\mathbf{x}$ satisfying $\|\nabla F(\mathbf{x})\| \leq \epsilon$ (Allen-Zhu, 2018; Tripuraneni et al., 2018; Fang et al., 2018; Zhou et al., 2018; Cutkosky & Orabona, 2019; Cutkosky & Mehta, 2021; Liu et al., 2022; Faw et al., 2022; Nguyen et al., 2023). However, such an assumption often fails to hold in numerous real-world applications. For instance, the objectives of many neural networks are non-smooth due to the usage of the ReLU activation function. As a result, several studies have been conducted to explore the theoretical guarantees of optimizing non-smooth non-convex objectives (Benaïm et al., 2005; Majewski et al., 2018; Davis et al., 2020).

The analysis of non-smooth non-convex optimization differs significantly from that of smooth non-convex optimization. As demonstrated by Zhang et al. (2020c), no algorithm can find an ϵ-stationary point of a non-smooth non-convex objective within finite time. To provide an alternative, they introduce the notion of (δ, ϵ)-stationary points and prove its tractability for non-smooth non-convex objectives. Instead of evaluating the gradient at a precise point $\mathbf{x}$, this notion describes whether there exists a subset within the ball centered at $\mathbf{x}$ with radius δ, so that by randomly drawing points from this subset, the norm of the expected gradient is smaller than ϵ. Compared to the definition of ϵ-stationary points, this notion is more general since it recovers the previous one when $\delta = 0$. Then, the notion of (δ, ϵ)-stationarity has attained widespread acceptance in subsequent research (Davis et al., 2022; Kornowski & Shamir, 2022; Tian & So, 2022; Lin et al., 2022; Kornowski & Shamir, 2023; Jordan et al., 2023). Previous studies (Zhang et al., 2020c; Tian et al., 2022) have proposed algorithms that utilize $\tilde{\mathcal{O}}(\epsilon^{-4}\delta^{-1})$ gradient queries to find a (δ, ϵ)-stationary point with high probability. Recently, Cutkosky et al. (2023) introduce the online-to-non-convex framework, which transforms non-smooth

[1] National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China [2] School of Artificial Intelligence, Nanjing University, Nanjing, China. Correspondence to: Lijun Zhang <zhanglj@lamda.nju.edu.cn>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

non-convex optimization to an online learning problem, and improve the query complexity to $\mathcal{O}(\epsilon^{-3}\delta^{-1})$, but their results only hold in expectation.

Moreover, prior studies on non-smooth non-convex optimization (Zhang et al., 2020c; Tian et al., 2022; Cutkosky et al., 2023) typically assume that gradient distributions have bounded variance. However, empirical observations suggest that such an assumption is too idealistic. For example, Zhang et al. (2020b) report that the gradient distributions exhibit heavy-tailed characteristics during the training of BERT (Vaswani et al., 2017). Here, heavy tails mean that the distributions have bounded $\mathfrak{p}$-th moments for some $\mathfrak{p} \in (1, 2]$. The assumption of heavy tails is more general than that of bounded variance because it holds for a broader class of distributions, while bounded variance only corresponds to the special case of $\mathfrak{p} = 2$. An abundant of prior research explores learning and optimization problems in the presence of heavy tails (Audibert & Catoni, 2010; Hsu & Sabato, 2014; Zhang & Zhou, 2018; Lu et al., 2019; Lugosi & Mendelson, 2019; Gürbüzbalaban et al., 2021; Xue et al., 2023). In recent years, several studies have investigated non-convex stochastic optimization with heavy-tailed gradients (Cutkosky & Mehta, 2021; Sadiev et al., 2023; Nguyen et al., 2023), but their methods are limited to smooth non-convex objectives.

To overcome the limitations of previous research, we propose a novel algorithm with high-probability guarantee for non-smooth non-convex stochastic optimization under heavy-tailed gradients. Inspired by Cutkosky et al. (2023), we incorporate the gradient clipping technique into the online-to-non-convex framework to handle heavy-tailed gradients. Theoretical analysis shows that our modification yields a high-probability upper bound for the averaged gradient norms across a series of candidate points. Furthermore, we propose a validation method based on sampling and gradient clipping to efficiently estimate the gradient norm of a specific candidate. By repeatedly validating candidate points, we can explicitly select a genuine (δ, ϵ)-stationary point among candidates. We summarize our contributions as follows:

- We propose an algorithm with the complexity of $\tilde{\mathcal{O}}(\epsilon^{-\frac{2\mathfrak{p}-1}{\mathfrak{p}-1}}\delta^{-1})$ gradient queries for finding a (δ, ϵ)-stationary point with high probability. To our knowledge, this is the first work that targets non-smooth non-convex stochastic optimization with heavy-tailed gradients.
- When the stochastic gradients have bounded variance, i.e., $\mathfrak{p} = 2$, the query complexity of our algorithm reduces to $\tilde{\mathcal{O}}(\epsilon^{-3}\delta^{-1})$, improving the existing high-probability result, i.e., $\tilde{\mathcal{O}}(\epsilon^{-4}\delta^{-1})$ (Zhang et al., 2020c; Tian et al., 2022).
- When the objective is smooth, our algorithm is able to identify an ϵ-stationary with $\tilde{\mathcal{O}}(\epsilon^{-\frac{3\mathfrak{p}-2}{\mathfrak{p}-1}})$ gradient queries, nearly matching previous studies (Zhang et al., 2020b; Cutkosky & Mehta, 2021; Nguyen et al., 2023).

2. Related Work

In this section, we briefly review recent studies on non-convex stochastic optimization, including smooth and non-smooth settings.

2.1. Smooth Non-Convex Optimization

For smooth non-convex objectives, Ghadimi & Lan (2013) prove that stochastic gradient descent (SGD) finds a point $\mathbf{x}$ satisfying $\mathbb{E}[\|\nabla F(\mathbf{x})\|] \leq \epsilon$ with $\mathcal{O}(\epsilon^{-4})$ gradient queries, when gradients have bounded variance. They also prove that combining SGD with a post-optimization phase achieves a query complexity of $\tilde{\mathcal{O}}(\epsilon^{-4} + \epsilon^{-2}q^{-1})$ for ensuring $\|\nabla F(\mathbf{x})\| \leq \epsilon$ with probability $1 - q$. The aforementioned theoretical guarantees have (nearly) matched the lower bound, as Arjevani et al. (2023) prove that $\Omega(\epsilon^{-4})$ queries are necessary to find an ϵ-stationary point of smooth non-convex objectives. If additional properties (e.g., second-order smoothness or mean-squared smoothness) are assumed, it is possible to achieve lower complexity (Allen-Zhu, 2018; Fang et al., 2018; Levy et al., 2021). When the distributions of stochastic gradients are heavy-tailed, Zhang et al. (2020b) establish the lower bound of $\Omega(\epsilon^{-\frac{3\mathfrak{p}-2}{\mathfrak{p}-1}})$ queries for identifying an ϵ-stationary point in expectation and propose a matching algorithm. Later, Cutkosky & Mehta (2021) achieve a high-probability bound of $\tilde{\mathcal{O}}(\epsilon^{-\frac{3\mathfrak{p}-2}{\mathfrak{p}-1}})$. Under the assumption of bounded variance, i.e., $\mathfrak{p} = 2$, these results recover previous bounds. In recent studies, Sadiev et al. (2023) and Nguyen et al. (2023) also examine the setting of heavy-tailed noise with a similar metric and weaker constraints.

2.2. Non-Smooth Non-Convex Optimization

The optimization of non-smooth non-convex objectives has undergone rapid development in recent years. Zhang et al. (2020c) demonstrate that ϵ-stationary points are intractable in the non-smooth non-convex setting and introduce the alternative notion of (δ, ϵ)-stationarity. In the stochastic finite variance setting, they propose an algorithm that achieves $\tilde{\mathcal{O}}(\epsilon^{-4}\delta^{-1})$ query complexity in finding a (δ, ϵ)-stationary point with high probability. They also consider the setting of deterministic gradients and achieve an $\tilde{\mathcal{O}}(\epsilon^{-3}\delta^{-1})$ bound. One limitation of their study is the use of a directional gradient oracle that requires the subgradients to satisfy a certain linear equation, which is impractical. To address this drawback, Tian et al. (2022) adopt the standard stochastic gradient oracle and propose algorithms with the same complexity of $\tilde{\mathcal{O}}(\epsilon^{-4}\delta^{-1})$. However, the algorithms of Zhang et al.

(2020c) and Tian et al. (2022) only promise the existence of one (δ, ϵ)-stationary point within the output set, without explicitly selecting a specific point. Recently, Cutkosky et al. (2023) introduce the online-to-non-convex framework, which transforms the problem of minimizing non-smooth non-convex objectives to the minimization of shifting regret over linear losses. Their algorithm enjoys a query complexity of $\mathcal{O}(\epsilon^{-3}\delta^{-1})$ for finding a (δ, ϵ)-stationary point in expectation, which is proved to be optimal. We observe that all existing studies assume the variance of stochastic gradients to be bounded, which means their analyses do not apply to heavy-tailed gradients. Additionally, the in-expectation guarantee is not satisfactory, as the algorithm may perform poorly in one individual run.

3. Main Results

In this section, we first present necessary preliminaries, including basic assumptions and definitions, then introduce our motivation and algorithms.

3.1. Preliminaries

We aim to optimize a non-convex and *possibly non-smooth* function $F : \mathcal{H} \mapsto \mathbb{R}$, where $\mathcal{H}$ denotes a real Hilbert space (e.g., $\mathbb{R}^d$). If not otherwise specified, we use $\|\cdot\|$ to denote the ℓ_2-norm. We assume that $F^\star \triangleq \inf_{\mathbf{x}} F(\mathbf{x}) > -\infty$ and F is differentiable.

In this paper, we focus on designing first-order optimization strategies. Our algorithm accesses the information about F through a stochastic gradient oracle (Tian et al., 2022).

Definition 3.1. Given a differentiable function F and a query point $\mathbf{x} \in \mathcal{H}$, the oracle $\mathrm{GRAD} : \mathcal{H} \times \mathcal{Z} \mapsto \mathcal{H}$ samples an i.i.d. random variable $\mathbf{z} \in \mathcal{Z}$ and returns $\mathrm{GRAD}(\mathbf{x}, \mathbf{z}) \in \mathcal{H}$ which satisfies

$$\mathbb{E}_{\mathbf{z}}\left[\mathrm{GRAD}\left(\mathbf{x}, \mathbf{z}\right)\right] = \nabla F\left(\mathbf{x}\right).$$

Following previous studies (Cutkosky & Mehta, 2021; Cutkosky et al., 2023), we introduce three basic assumptions.

Assumption 3.2. The gradient of the objective function F is bounded by G, i.e.,

$$\max_{\mathbf{x} \in \mathcal{H}} \|\nabla F\left(\mathbf{x}\right)\| \leq G.$$

Assumption 3.3. The $\mathfrak{p}$-th moment of the possibly heavy-tailed stochastic gradient $\mathbf{g} = \mathrm{GRAD}\left(\mathbf{x}, \mathbf{z}\right)$ is bounded by $\sigma^{\mathfrak{p}}$ for some $\mathfrak{p} \in (1, 2]$, i.e.,

$$\mathbb{E}_{\mathbf{z}}\left[\|\mathbf{g} - \mathbb{E}_{\mathbf{z}}\left[\mathbf{g}\right]\|^{\mathfrak{p}}\right] \leq \sigma^{\mathfrak{p}},\ \forall \mathbf{x} \in \mathcal{H}.$$

Assumption 3.4. The objective function F is *well-behaved*, i.e., for any $\mathbf{x}, \mathbf{y} \in \mathcal{H}$

$$F\left(\mathbf{y}\right) - F\left(\mathbf{x}\right) = \int_0^1 \left\langle \nabla F\left(\mathbf{x} + t\left(\mathbf{y} - \mathbf{x}\right)\right), \mathbf{y} - \mathbf{x}\right\rangle \mathrm{d}t.$$

Remark 1. Since we have assumed that F is differentiable, Assumption 3.4 holds true due to the Fundamental Theorem of Calculus. To handle objectives that are not everywhere differentiable, one can apply the randomized smoothing technique (Duchi et al., 2012; Yousefian et al., 2012; Nesterov & Spokoiny, 2017) to avoid non-differentiable points. The processed objective exhibits milder properties: it is differentiable and well-behaved, meanwhile its value is close to the original one at any point.

Following Cutkosky et al. (2023), we introduce the definition of (δ, ϵ)-stationarity.

Definition 3.5. A point $\mathbf{x}$ is a (δ, ϵ)-stationary point of a differentiable function F if there is a finite set $S \subset \mathbb{B}(\mathbf{x}, \delta)$, where $\mathbb{B}(\mathbf{x}, \delta)$ denotes a ball centered at $\mathbf{x}$ with radius δ, such that for $\mathbf{y}$ selected uniformly at random from S, $\mathbb{E}[\mathbf{y}] = \mathbf{x}$ and $\|\mathbb{E}[\nabla F(\mathbf{y})]\| \leq \epsilon$.

Remark 2. This definition of a (δ, ϵ)-stationary point is essentially the same as used in Zhang et al. (2020c); Davis et al. (2020); Tian et al. (2022), yet it is slightly more stringent. Specifically, Definition 3.5 requires the explicit selection of a finite support set S and the unbiasedness in calculating the norm of expected gradient. These subtle differences provide convenience for validation: given a support set S', once it is feasible to compute a high-precision approximation of $\|\mathbb{E}_{\mathbf{y} \in S'}[\nabla F(\mathbf{y})]\|$ from stochastic gradients, we are able to validate the corresponding solution.

As a counterpart of Definition 3.5, we also introduce the following notation of $\|\nabla F(\mathbf{x})\|_\delta$.

Definition 3.6. Given a point $\mathbf{x}$, a number $\delta > 0$ and a differentiable function F, define

$$\|\nabla F\left(\mathbf{x}\right)\|_\delta \triangleq \inf_{S \subset \mathbb{B}(\mathbf{x}, \delta), \frac{1}{|S|}\sum_{\mathbf{y} \in S} \mathbf{y} = \mathbf{x}} \left\| \frac{1}{|S|} \sum_{\mathbf{y} \in S} \nabla F\left(\mathbf{y}\right) \right\|.$$

With Definition 3.6, we can use the inequality $\|\nabla F(\mathbf{x})\|_\delta \leq \epsilon$ to concisely demonstrate that $\mathbf{x}$ is a (δ, ϵ)-stationary point.

3.2. Motivation

Based on the online-to-non-convex framework (Cutkosky et al., 2023), we endeavor to develop an algorithm that achieves a high-probability bound for non-smooth non-convex stochastic optimization with heavy tails. To reach our goal, we first identify the bottleneck in the previous analysis.

Our general intuition is that, within the online-to-non-convex framework, the linear losses fed to the online learning algorithm are defined as

$$\ell_n\left(\mathbf{x}\right) = \left\langle \mathbf{g}_n, \mathbf{x}\right\rangle,\ \forall n \in [N],$$

Algorithm 1 Candidate Generation Algorithm

```
Input: Initial point $\mathbf{x}_0$, number of restarts $K$, round length $T$, clipping parameter $\tau$, and domain radius $D$
Set $M = KT$, $\eta = D/\tau$, $\Delta_1 = \mathbf{0}$
for $n = 1 \cdots M$ do
    Update $\mathbf{x}_n$ according to (1)
    Generate $s_n$ and compute $\mathbf{w}_n$ according to (2)
    Sample $\mathbf{z}_n$ and generate $\mathbf{g}_n = \mathrm{GRAD}(\mathbf{w}_n, \mathbf{z}_n)$
    Compute $\hat{\mathbf{g}}_n$ according to (3)
    Update $\Delta_{n+1}$ according to (4)
end for
Set $\mathbf{w}_t^k = \mathbf{w}_{(k-1)T+t}$ for $k \in [K]$ and $t \in [T]$
Set $\bar{\mathbf{w}}^k$ according to (5)
Return $\mathbf{w}_1, \cdots, \mathbf{w}_M$ and $\bar{\mathbf{w}}^1, \cdots, \bar{\mathbf{w}}^K$
```

where $\ell_n(\cdot)$ is the linear loss function in the n-th round, $\mathbf{g}_n$ is the corresponding stochastic gradient, and $[N]$ denotes $\{1, 2, \cdots, N\}$. The challenge of achieving a high-probability theoretical guarantee for heavy-tailed gradients stems from the uncontrollable $\mathbf{g}_n$: we can not even bound $\sum_{n=1}^{M} \|\mathbf{g}_n\|^2$ in expectation under heavy tails, causing the regret of the online learning algorithm to be out of control.

To address the above challenge, we employ the gradient clipping technique, which introduces a truncated value $\hat{\mathbf{g}}_n$ that has favorable properties while generally remaining close to $\mathbf{g}_n$. This technique is commonly used in training deep learning models to mitigate the gradient explosion problem (Mikolov, 2012; Goodfellow et al., 2016). We demonstrate that by constructing the linear losses with clipped gradients, the shifting regret of the online learning algorithm can be effectively controlled with high probability. Meanwhile, we prove that the deviation caused by clipping operations is of the same order as the regret. Moreover, the gradient clipping technique also plays an important role in our validation method, helping attain high precision.

3.3. Algorithms

We follow the online-to-non-convex framework (Cutkosky et al., 2023) to generate a sequence of candidate points $\{\bar{\mathbf{w}}^k\}$. Specifically, we update the iterate $\mathbf{x}_n$ with a learned step Δ_n, i.e.,

$$\mathbf{x}_n = \mathbf{x}_{n-1} + \Delta_n. \tag{1}$$

Meanwhile, we randomly select a point $\mathbf{w}_n$ between $\mathbf{x}_{n-1}$ and $\mathbf{x}_n$, i.e.,

$$\mathbf{w}_n = \mathbf{x}_{n-1} + s_n \Delta_n, \tag{2}$$

where $s_n \in [0, 1]$ is a uniformly distributed random variable. Next, we query the stochastic gradient oracle to access the gradient $\mathbf{g}_n$ at point $\mathbf{w}_n$. The reason for introducing $\mathbf{w}_n$ originates from its connection with the decrease in the objective value between two consecutive iterations, as

$$\begin{aligned} F(\mathbf{x}_n) - F(\mathbf{x}_{n-1}) &= \int_0^1 \langle \nabla F(\mathbf{x}_{n-1} + s\Delta_n), \Delta_n \rangle \, \mathrm{d}s \\ =&\mathbb{E}_{s_n} [\nabla F(\mathbf{x}_{n-1} + s_n \Delta_n)] = \mathbb{E}_{\mathbf{w}_n} [\nabla F(\mathbf{w}_n)] \\ =&\langle \mathbb{E}_{\mathbf{w}_n, \mathbf{z}_n} [\mathbf{g}_n], \Delta_n \rangle \end{aligned}$$

holds for the well-behaved objective F (c.f. Lemma 4.1), where $\mathbf{z}_n$ is the i.i.d. random variable sampled within the gradient oracle. Later, we compute the clipped gradient with a clipping parameter $\tau > 0$, i.e.,

$$\hat{\mathbf{g}}_n = \frac{\mathbf{g}_n}{\|\mathbf{g}_n\|} \min(\tau, \|\mathbf{g}_n\|) \tag{3}$$

to protect our algorithm from extraordinarily large stochastic gradients. The computed $\hat{\mathbf{g}}_n$ is used to update Δ_n, which follows the strategy of the standard online gradient descent (OGD) (Zinkevich, 2003) with the fixed step size η, i.e.,

$$\Delta_{n+1} = \Pi_{\mathbb{B}(\mathbf{0}, D)} [\Delta_n - \eta \hat{\mathbf{g}}_n], \tag{4}$$

where $\Pi_{\mathbb{B}(\mathbf{0}, D)} [\mathbf{x}]$ denotes projecting $\mathbf{x}$ to the nearest point measured by ℓ_2-norm within $\mathbb{B}(\mathbf{0}, D)$. After M iterations, we partition the elements of $\{\mathbf{w}_n\}$ into groups of size T in sequential order, and compute the average of each group to produce K centers, i.e.,

$$\bar{\mathbf{w}}^k = \frac{1}{T} \sum_{t=1}^{T} \mathbf{w}_t^k, \ \forall k \in [K]. \tag{5}$$

We return the candidate point $\bar{\mathbf{w}}^k$ as well as its corresponding support set $\{\mathbf{w}_1^k, \cdots, \mathbf{w}_T^k\}$ for each $k \in [K]$. The complete procedure is summarized in Algorithm 1.

The difference between our algorithm and that of Cutkosky et al. (2023, Algorithm 1) is updating Δ_n with the clipped gradient $\hat{\mathbf{g}}_n$ rather than the original one $\mathbf{g}_n$, which allows us to establish an upper bound for $F(\mathbf{x}_M) - F(\mathbf{x}_0) = \sum_{n=1}^{M} \langle \mathbb{E}_{\mathbf{w}_n, \mathbf{z}_n} [\mathbf{g}_n], \Delta_n \rangle$ with high probability. To be precise, we conduct the following decomposition:

$$\begin{aligned} &\sum_{n=1}^{M} \langle \mathbb{E}_{\mathbf{w}_n, \mathbf{z}_n} [\mathbf{g}_n], \Delta_n \rangle \\ =&\sum_{n=1}^{M} \langle \mathbb{E}_{\mathbf{z}_n} [\mathbf{g}_n], \mathbf{u}_n \rangle + \sum_{n=1}^{M} \langle \mathbb{E}_{\mathbf{w}_n, \mathbf{z}_n} [\mathbf{g}_n] - \mathbb{E}_{\mathbf{z}_n} [\mathbf{g}_n], \Delta_n \rangle \\ &+ \sum_{n=1}^{M} \langle \hat{\mathbf{g}}_n, \Delta_n - \mathbf{u}_n \rangle + \sum_{n=1}^{M} \langle \mathbb{E}_{\mathbf{z}_n} [\mathbf{g}_n] - \hat{\mathbf{g}}_n, \Delta_n - \mathbf{u}_n \rangle, \end{aligned}$$

where $\{\mathbf{u}_n\}$ is an arbitrary sequence of points. The first term of the decomposition can be negative if $\mathbf{u}_n$ is chosen according to $\mathbb{E}_{\mathbf{z}_n} [\mathbf{g}_n]$. The second term is the summation of a martingale difference sequence and can be bounded by

Hoeffding-Azuma inequality. The third and fourth terms correspond to the regret of the online learning algorithm and the deviation of $\hat{\mathbf{g}}_n$ from $\mathbb{E}_{\mathbf{z}_n}[\mathbf{g}_n]$, respectively. Compared with the analysis of Cutkosky et al. (2023, Theorem 8), we substitute $\mathbf{g}_n$ in the third and fourth terms with $\hat{\mathbf{g}}_n$, so that our analysis can benefit from the benign properties of $\hat{\mathbf{g}}_n$.

By properly setting the parameters of Algorithm 1, we prove that $\bar{\mathbf{w}}^1, \cdots, \bar{\mathbf{w}}^K$ are good candidates for (δ, ϵ)-stationary points on average, as demonstrated in the following theorem.

Theorem 3.7. *With the notation in Algorithm 1, set the clipping parameter $\tau = T^{1/\mathfrak{p}} \left(\sigma^p + G^{\mathfrak{p}}\right)^{1/\mathfrak{p}} \log(2K/q)^{-1/\mathfrak{p}}$. Under Assumptions 3.2, 3.3 and 3.4, with probability at least $1-q$, we have*

$$\begin{aligned}&\frac{1}{K}\sum_{k=1}^{K}\left\|\frac{1}{T}\sum_{t=1}^{T}\nabla F\left(\mathbf{w}_t^k\right)\right\| \\ \leq &\frac{F\left(\mathbf{x}_0\right)-F^\star}{DM} + \frac{AKT^{1/\mathfrak{p}}}{M} + \frac{B}{\sqrt{M}},\end{aligned}$$

for any $D > 0$, where $A = \mathcal{O}(\log(K/q)^{(\mathfrak{p}-1)/\mathfrak{p}})$, $B = \mathcal{O}(\sqrt{\log(1/q)})$.

Remark 3. We emphasize that Theorem 3.7 is the most crucial step in achieving the high-probability bound under heavy-tailed stochastic gradients. Specifically, we prove that $\frac{1}{K}\sum_{k=1}^{K}\|\frac{1}{T}\sum_{t=1}^{T}\nabla F(\mathbf{w}_t^k)\|$ keeps decreasing as K and T grow. Since $\bar{\mathbf{w}}^k$ is the center of $\mathbf{w}_1^k, \cdots, \mathbf{w}_T^k$, $\|\frac{1}{T}\sum_{t=1}^{T}\nabla F(\mathbf{w}_t^k)\|$ describes the norm of average gradient across the supporting set of $\bar{\mathbf{w}}^k$ and is closely related to the definition of (δ, ϵ)-stationarity.

Then, we adjust T with respect to the query budget N and set D to yield the following corollary.

Corollary 3.8. *Suppose we have a budget of N gradient queries. Set $T = \min(\lceil(\frac{AN\delta}{F(\mathbf{x}_0)-F^\star})^{\mathfrak{p}/(2\mathfrak{p}-1)}\rceil, \frac{N}{2})$, $K = \lfloor\frac{N}{T}\rfloor$ and $D = \delta/T$ for an arbitrary $\delta > 0$. Under Assumptions 3.2, 3.3 and 3.4, with probability at least $1-q$, Algorithm 1 ensures*

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^{K}\left\|\nabla F\left(\bar{\mathbf{w}}^k\right)\right\|_\delta \leq &\frac{2\left(F\left(\mathbf{x}_0\right)-F^\star\right)}{\delta N} + \frac{2B}{\sqrt{N}} \\ &+ \max\left(\frac{4A^{\frac{\mathfrak{p}}{2\mathfrak{p}-1}}\left(F\left(\mathbf{x}_0\right)-F^\star\right)^{\frac{\mathfrak{p}-1}{2\mathfrak{p}-1}}}{(\delta N)^{\frac{\mathfrak{p}-1}{2\mathfrak{p}-1}}}, \frac{4A}{N^{\frac{\mathfrak{p}-1}{\mathfrak{p}}}}\right),\end{aligned}$$

where A, B are given in Theorem 3.7.

Remark 4. When N is sufficiently large, Corollary 3.8 implies that

$$\frac{1}{K}\sum_{k=1}^{K}\left\|\nabla F\left(\bar{\mathbf{w}}^k\right)\right\|_\delta \leq \mathcal{O}\left(\left(\frac{\log\left(K/q\right)}{\delta N}\right)^{\frac{\mathfrak{p}-1}{2\mathfrak{p}-1}}\right).$$

Algorithm 2 Approximation Algorithm with Adjustable Precision

Input: A given point set $S = \{\mathbf{w}_1, \cdots, \mathbf{w}_T\}$, total number of rounds R, and clipping parameter τ

```
for $r = 1$ to $R$ do
    Sample $\mathbf{w}_r$ from $S$ uniformly at random
    Sample $\mathbf{z}_r$ and generate $\mathbf{g}_r = \mathrm{GRAD}\left(\mathbf{w}_r, \mathbf{z}_r\right)$
    Compute $\tilde{\mathbf{g}}_r$ according to (6)
end for
Return $\mathbf{g}_{\text{est}} = \frac{1}{R}\sum_{r=1}^{R}\tilde{\mathbf{g}}_r$
```

The above convergence rate indicates that setting $N = \tilde{\mathcal{O}}(\epsilon^{-\frac{2\mathfrak{p}-1}{\mathfrak{p}-1}}\delta^{-1}\log(1/q))$ ensures $\frac{1}{K}\sum_{k=1}^{K}\|\nabla F(\bar{\mathbf{w}}^k)\|_\delta \leq \epsilon$ with high probability. This guarantee reflects the average quality of candidate points $\{\bar{\mathbf{w}}^k\}$, but in practice, we are usually required to ultimately output one valid (δ, ϵ)-stationary point among candidates, i.e., identifying a k satisfying $\|\nabla F(\bar{\mathbf{w}}^k)\|_\delta \leq \epsilon$. Therefore, we need an algorithm that approximates $\|\frac{1}{T}\sum_{t=1}^{T}\nabla F(\mathbf{w}_t^k)\|$ for a specific $k \in [K]$, which establishes a upper bound for $\|\nabla F(\bar{\mathbf{w}}^k)\|_\delta$, in order to perform validation.

Aiming at efficiently estimating the true value of the averaged gradient over a specific set $S = \{\mathbf{w}_1, \cdots, \mathbf{w}_T\}$, we also utilize the gradient clipping technique to alleviate the impact of heavy tails. In the r-th round, we randomly select a point $\mathbf{w}_r$ from S, access its stochastic gradient $\mathbf{g}_r$, and calculate the clipped gradient $\tilde{\mathbf{g}}_r$ by

$$\tilde{\mathbf{g}}_r = \frac{\mathbf{g}_r}{\|\mathbf{g}_r\|}\min\left(\tau, \|\mathbf{g}_r\|\right). \tag{6}$$

After R rounds, we compute $\mathbf{g}_{\text{est}} = \frac{1}{R}\sum_{r=1}^{R}\tilde{\mathbf{g}}_r$ to estimate $\frac{1}{T}\sum_{t=1}^{T}\nabla F(\mathbf{w}_t)$. We summarize the procedure in Algorithm 2 and provide the following guarantee for bounding the approximation error.

Theorem 3.9. *Set the clipping parameter $\tau = R^{1/\mathfrak{p}}\left(\sigma^{\mathfrak{p}} + G^{\mathfrak{p}}\right)^{1/\mathfrak{p}}\log(1/q)^{-1/\mathfrak{p}}$. Under Assumptions 3.2 and 3.3, with probability at least $1-q$, Algorithm 2 ensures*

$$\left\|\frac{1}{R}\sum_{r=1}^{R}\tilde{\mathbf{g}}_r - \frac{1}{T}\sum_{t=1}^{T}\nabla F\left(\mathbf{w}_t\right)\right\| = \mathcal{O}\left(\left(\frac{\log\left(1/q\right)}{R}\right)^{\frac{\mathfrak{p}-1}{\mathfrak{p}}}\right).$$

Remark 5. According to Theorem 3.9, we can set $R = \mathcal{O}(\epsilon^{-\frac{\mathfrak{p}}{\mathfrak{p}-1}}\log(1/q))$ to ensure that the approxiation error is smaller than ϵ. We notice that under the setting of bounded variance, a similar method to our Algorithm 2 exists in the literature (Kornowski & Shamir, 2023, Algorithm 3), albeit with a different implementation and analysis. At a high level, they use the average of original gradients $\frac{1}{R}\sum_{r=1}^{R}\mathbf{g}_r$ to estimate $\frac{1}{T}\sum_{t=1}^{T}\nabla F(\mathbf{w}_t)$. Since original gradients do not have bounded norms, they employ the standard Markov's

Algorithm 3 Non-Smooth Non-Convex Optimization for Heavy-Tailed Stochastic Gradients

```
Input: Initial point $\mathbf{x}_0$, query budget $N$, and validation round length $R$
Call Algorithm 1 with $\mathbf{x}_0$ and $N$, other parameters are set according to Theorem 3.7 and Corollary 3.8
Receive $\mathbf{w}_1, \cdots, \mathbf{w}_M$ and $\bar{\mathbf{w}}^1, \cdots, \bar{\mathbf{w}}^K$
Set $\mathbf{w}_{\text{out}} = null$, $\epsilon_{\text{thres}} = 5\epsilon/6$
while $\mathbf{w}_{\text{out}} = null$ do
    Select $k \in [K]$ uniformly at random
    Call Algorithm 2 with $\{\mathbf{w}_1^k, \cdots, \mathbf{w}_T^k\}$ and $R$, the other parameter is set according to Theorem 3.9
    Receive $\mathbf{g}_{\text{est}}$
    if $\|\mathbf{g}_{\text{est}}\| \leq \epsilon_{\text{thres}}$ then
        Set $\mathbf{w}_{\text{out}} = \bar{\mathbf{w}}^k$
    end if
end while
Return $\mathbf{w}_{\text{out}}$
```

inequality, and their algorithm requires $\mathcal{O}(\epsilon^{-2}\log(1/q)/q)$ gradient queries to achieve a precision of ϵ with probability $1-q$. In comparison, our Algorithm 2 only requires $\mathcal{O}(\epsilon^{-2}\log(1/q))$ queries under the same setting, eliminating the undesirable $1/q$ factor. Moreover, our guarantee naturally extends to the setting of heavy-tailed stochastic gradients.

By assembling Algorithms 1 and 2, we devise an algorithm that outputs a (δ, ϵ)-stationary point with high probability. We start by running Algorithm 1 to generate $\mathbf{w}_1, \cdots, \mathbf{w}_M$ and $\bar{\mathbf{w}}^1, \cdots, \bar{\mathbf{w}}^K$, which yields the high-probability guarantee of $\frac{1}{K}\sum_{k=1}^K \|\nabla F(\bar{\mathbf{w}}^k)\|_\delta \leq \epsilon/3$. Then, we choose $k \in [K]$ randomly and verify whether the current $\bar{\mathbf{w}}^k$ is indeed a (δ, ϵ)-stationary point. To perform validation, we set the acceptance threshold to be $\epsilon_{\text{thres}} = 5\epsilon/6$ and control the precision of Algorithm 2 to be $\epsilon/6$, so that we are likely to filter out any invalid $\bar{\mathbf{w}}^k$ satisfying $\|\nabla F(\bar{\mathbf{w}}^k)\|_\delta > \epsilon$. We summarize the procedure in Algorithm 3 and provide the following theorem.

Theorem 3.10. *Set* $N = \tilde{\mathcal{O}}(\epsilon^{-\frac{2\mathfrak{p}-1}{\mathfrak{p}-1}}\delta^{-1}\log(1/q))$ *and* $R = \mathcal{O}(\epsilon^{-\frac{\mathfrak{p}}{\mathfrak{p}-1}}\log(1/q))$. *Under Assumptions 3.2, 3.3 and 3.4, with probability at least* $1-q$, *Algorithm 3 guarantees to output a* (δ, ϵ)*-stationary point with* $\tilde{\mathcal{O}}(\epsilon^{-\frac{2\mathfrak{p}-1}{\mathfrak{p}-1}}\delta^{-1}\log(1/q) + \epsilon^{-\frac{\mathfrak{p}}{\mathfrak{p}-1}}\log(1/q)^2)$ *queries of stochastic gradient oracle.*

Remark 6. When the variance of stochastic gradients is bounded by σ^2, existing algorithms require $\tilde{\mathcal{O}}\left(\frac{(\sigma^2+G^2)^{3/2}(F(\mathbf{x}_0)-F^\star)\log(1/q)}{\epsilon^4\delta}\right)$ gradient queries to identify a (δ, ϵ)-stationary point with probability $1-q$ (Zhang et al., 2020c; Tian et al., 2022). This setting corresponds to the special case of $\mathfrak{p} = 2$, where Theorem 3.10 implies a query complexity of $\tilde{\mathcal{O}}\left(\frac{(\sigma^2+G^2)(F(\mathbf{x}_0)-F^\star)\log(1/q)}{\epsilon^3\delta}\right)$. In comparison, our bound has a smaller dependence not only on ϵ, but also on parameters σ and G.

When the objective is smooth, Algorithm 3 can also adapt to identify ϵ-stationary points by exploiting the smoothness property (Nesterov, 2004).

Assumption 3.11. The objective function F is H-smooth, i.e., for any $\mathbf{x}, \mathbf{y} \in \mathcal{H}$

$$\|\nabla F(\mathbf{x}) - \nabla F(\mathbf{y})\| \leq H\|\mathbf{x} - \mathbf{y}\|.$$

Using the strategy of converting guarantees for non-smooth objectives to smooth objectives (Cutkosky et al., 2023, Section 5), Algorithm 3 is equipped with the following property.

Corollary 3.12. *Further set* $\delta = \epsilon/H$. *Under Assumptions 3.2, 3.3, 3.4 and 3.11, with probability at least* $1-q$, *Algorithm 3 guarantees to output an* ϵ*-stationary point with* $\tilde{\mathcal{O}}(\epsilon^{-\frac{3\mathfrak{p}-2}{\mathfrak{p}-1}}\log(1/q) + \epsilon^{-\frac{\mathfrak{p}}{\mathfrak{p}-1}}\log(1/q)^2)$ *queries of stochastic gradient oracle.*

Remark 7. Corollary 3.12 suggests that our high-probability bound matches (up to logarithmic factors) the $\Omega(\epsilon^{-\frac{3\mathfrak{p}-2}{\mathfrak{p}-1}})$ in-expectation lower bound (Zhang et al., 2020b).

4. Theoretical Analysis

In this section, we only provide the detailed proofs of Theorems 3.7 and 3.9. The omitted proofs can be found in Appendix B.

4.1. Proof of Theorem 3.7

In the beginning, we introduce the following two lemmas that will be used later.

Lemma 4.1. *With the notation in Algorithm 1, under Assumption 3.4, we have*

$$F(\mathbf{x}_M) - F(\mathbf{x}_0) = \sum_{n=1}^M \langle \mathbb{E}_{\mathbf{w}_n, \mathbf{z}_n}[\mathbf{g}_n], \Delta_n\rangle. \tag{7}$$

Lemma 4.2. *With the notation in Algorithm 1, under Assumptions 3.2 and 3.4, we have*

$$\|\mathbb{E}_{\mathbf{w}_n, \mathbf{z}_n}[\mathbf{g}_n]\| \leq G. \tag{8}$$

Remark 8. We acknowledge that Lemma 4.1 is an intermediate result of Cutkosky et al. (2023, Theorem 7). However, their subsequent decomposition

$$\sum_{n=1}^M \langle \mathbb{E}_{\mathbf{w}_n, \mathbf{z}_n}[\mathbf{g}_n], \Delta_n\rangle = \sum_{n=1}^M \langle \mathbf{g}_n, \Delta_n - \mathbf{u}_n\rangle$$
$$+ \sum_{n=1}^M \langle \mathbb{E}_{\mathbf{w}_n, \mathbf{z}_n}[\mathbf{g}_n] - \mathbf{g}_n, \Delta_n\rangle + \sum_{n=1}^M \langle \mathbf{g}_n, \mathbf{u}_n\rangle$$

for the in-expectation result does not apply to our case, where $\{\mathbf{u}_n\}$ is an arbitrary sequence of points within $\mathbb{B}(\mathbf{0}, D)$. When obtaining an in-expectation guarantee, each term of the decomposition can be simply bounded by its expected value. In contrast, we must bound each term with high probability and then apply the union bound. One might conjecture that we can directly utilize concentration inequalities to conduct analysis, which is a common approach for obtaining high-probability guarantees. However, this approach is not effective in this setting, since $\|\mathbf{g}_n\|$ is unbounded and we can only apply weak concentration inequalities, which consequently yields suboptimal results. To make the situation worse, in the case of heavy-tailed gradient distributions, even the in-expectation result by Cutkosky et al. (2023) fails to hold. Specifically, the assumptions of $\mathrm{Var}(\mathbf{g}_n) \le \sigma^2$ and $\mathbb{E}[\|\mathbf{g}_n\|^2] \le G^2$ required by their analysis are not met.

To tackle this challenge, we introduce the gradient clipping technique, which throws out the outliers and ensures the following nice properties of clipped gradients for any $\tau > 0$:

$$\left\|\mathbb{E}_{\mathbf{z}_n}\left[\hat{\mathbf{g}}_n\right] - \mathbb{E}_{\mathbf{z}_n}\left[\mathbf{g}_n\right]\right\| \le \frac{2^{\mathfrak{p}-1}\left(\sigma^{\mathfrak{p}} + G^{\mathfrak{p}}\right)}{\tau^{\mathfrak{p}-1}}, \tag{9}$$

$$\mathbb{E}_{\mathbf{z}_n}\left[\left\|\hat{\mathbf{g}}_n\right\|^2\right] \le 2^{\mathfrak{p}-1}\tau^{2-\mathfrak{p}}\left(\sigma^{\mathfrak{p}} + G^{\mathfrak{p}}\right). \tag{10}$$

The proof is provided in Appendix A.

By exploiting Lemma 4.1, we have

$$\begin{aligned}
F\left(\mathbf{x}_M\right) - F\left(\mathbf{x}_0\right) &\overset{(7)}{=} \sum_{n=1}^{M}\left\langle \mathbb{E}_{\mathbf{w}_n,\mathbf{z}_n}\left[\mathbf{g}_n\right], \Delta_n\right\rangle \\
&= \underbrace{\sum_{n=1}^{M}\left\langle \hat{\mathbf{g}}_n, \Delta_n - \mathbf{u}_n\right\rangle}_{\texttt{term (a)}} + \underbrace{\sum_{n=1}^{M}\left\langle \mathbb{E}_{\mathbf{z}_n}\left[\mathbf{g}_n\right], \mathbf{u}_n\right\rangle}_{\texttt{term (b)}} \\
&\quad + \underbrace{\sum_{n=1}^{M}\left\langle \mathbb{E}_{\mathbf{w}_n,\mathbf{z}_n}\left[\mathbf{g}_n\right] - \mathbb{E}_{\mathbf{z}_n}\left[\mathbf{g}_n\right], \Delta_n\right\rangle}_{\texttt{term (c)}} \\
&\quad + \underbrace{\sum_{n=1}^{M}\left\langle \mathbb{E}_{\mathbf{z}_n}\left[\mathbf{g}_n\right] - \mathbb{E}_{\mathbf{z}_n}\left[\hat{\mathbf{g}}_n\right], \Delta_n - \mathbf{u}_n\right\rangle}_{\texttt{term (d)}} \\
&\quad + \underbrace{\sum_{n=1}^{M}\left\langle \mathbb{E}_{\mathbf{z}_n}\left[\hat{\mathbf{g}}_n\right] - \hat{\mathbf{g}}_n, \Delta_n - \mathbf{u}_n\right\rangle}_{\texttt{term (e)}},
\end{aligned}$$

where $\{\mathbf{u}_n\}$ is an arbitrary sequence of points whose norms are bounded by D. Specifically, we take $\mathbf{u}_{(k-1)T+1} = \cdots = \mathbf{u}_{kT} = \mathbf{u}^k = D\frac{\sum_{t=1}^{T}\nabla F(\mathbf{w}_t^k)}{\|\sum_{t=1}^{T}\nabla F(\mathbf{w}_t^k)\|}$ for $\forall k \in [K]$ throughout the analysis of this theorem. In the following, we analyze each term separately.

For $\texttt{term (a)}$, our specification of $\{\mathbf{u}_n\}$ implies

$$\texttt{term (a)} = \sum_{k=1}^{K}\sum_{t=1}^{T}\left\langle \hat{\mathbf{g}}_t^k, \Delta_t^k - \mathbf{u}^k\right\rangle,$$

where $\hat{\mathbf{g}}_t^k = \hat{\mathbf{g}}_{(k-1)T+t}$ and $\Delta_t^k = \Delta_{(k-1)T+t}$. We can interpret $\texttt{term (a)}$ as optimizing the online linear functions $\ell_n\left(\mathbf{x}\right) = \left\langle \hat{\mathbf{g}}_n, \mathbf{x}\right\rangle$ with K restarts. In this sense, $\texttt{term (a)}$ directly corresponds to the shifting regret of the online learning algorithm.

The standard OGD-style update strategy of Δ_n with the fixed step size η has the regret guarantee of

$$\sum_{t=1}^{T}\ell_t\left(\mathbf{w}_t\right) - \ell_t\left(\mathbf{u}\right) \le \frac{1}{2\eta}\left\|\mathbf{u} - \mathbf{x}_1\right\|^2 + \frac{\eta}{2}\sum_{t=1}^{T}\left\|\nabla \ell_t\left(\mathbf{w}_t\right)\right\|^2$$

for arbitrary sequence of convex functions ℓ_t, domain bounded $\{\mathbf{w}_t\}$ and $\mathbf{u}$, according to Orabona (2019, Theorem 2.13). Applying this result to Algorithm 1, we derive

$$\sum_{t=1}^{T}\left\langle \hat{\mathbf{g}}_t^k, \Delta_t^k - \mathbf{u}^k\right\rangle \le \frac{2}{\eta}D^2 + \frac{\eta}{2}\sum_{t=1}^{T}\left\|\hat{\mathbf{g}}_t^k\right\|^2 \tag{11}$$

for T rounds within the k-th restart.

Note that the properties of clipped gradient ensure $\|\hat{\mathbf{g}}_t^k\| \le \tau$ and $\mathbb{E}_{\mathbf{z}_n}[\|\hat{\mathbf{g}}_t^k\|^2] \le 2^{\mathfrak{p}-1}\tau^{2-\mathfrak{p}}\left(\sigma^{\mathfrak{p}} + G^{\mathfrak{p}}\right)$. We can further bound $\sum_{t=1}^{T}\|\hat{\mathbf{g}}_t^k\|^2$ using concentration inequality for sub-exponential sequences (c.f. Lemma D.1). From Lemma D.1, with probability at least $1 - q/\left(2K\right)$, we have

$$\begin{aligned}
\sum_{t=1}^{T}\left\|\hat{\mathbf{g}}_t^k\right\|^2 \le& 6\tau^{2-\mathfrak{p}}T\left(\sigma^{\mathfrak{p}} + G^{\mathfrak{p}}\right) + \frac{5}{3}\tau^2\log\frac{2K}{q} \\
=& \frac{23}{3}\tau^2\log\frac{2K}{q}.
\end{aligned} \tag{12}$$

We combine (11), (12) and $\eta = D/\tau$ to bound the regret of each restart, and then apply union bound for K restarts. With probability at least $1 - q/2$, we have

$$\texttt{term (a)} \le DK\tau\left(2 + \frac{23}{6}\log\frac{2K}{q}\right).$$

For $\texttt{term (b)}$, our specification of $\{\mathbf{u}_n\}$ and $\{\mathbf{u}^k\}$ directly implies

$$\texttt{term (b)} = -DT\sum_{k=1}^{K}\left\|\frac{1}{T}\sum_{t=1}^{T}\nabla F\left(\mathbf{w}_t^k\right)\right\|.$$

For $\texttt{term (c)}$, letting $V_n = \left\langle \mathbb{E}_{\mathbf{w}_n,\mathbf{z}_n}\left[\mathbf{g}_n\right] - \mathbb{E}_{\mathbf{z}_n}\left[\mathbf{g}_n\right], \mathbf{g}_n\right\rangle$, we find that $\{V_n\}$ forms a martingale difference sequence. Thus, $\texttt{term (c)}$ can be bounded using the Hoeffding-Azuma inequality for martingales (c.f. Lemma D.2).

Before applying the concentration inequality, we need to bound V_n:

$$V_n \leq \left\|\mathbb{E}_{\mathbf{w}_n,\mathbf{z}_n}\left[\mathbf{g}_n\right] - \mathbb{E}_{\mathbf{z}_n}\left[\mathbf{g}_n\right]\right\| \left\|\Delta_n\right\| \leq 2DG,$$

where the last step follows from $\|\mathbf{u}_n\| \leq D$, $\|\mathbb{E}_{\mathbf{z}_n}[\mathbf{g}_n]\| \leq G$ and Lemma 4.2. From Lemma D.2, with probability at least $1 - q/6$, we have

$$\texttt{term}\,(\texttt{c}) = \sum_{n=1}^{M} V_n \leq 4DG\sqrt{\frac{M}{2}\log\frac{6}{q}}.$$

For $\texttt{term}\,(\texttt{d})$, we utilize the property (9) of clipped gradients to bound this term, that is

$$\begin{aligned}\texttt{term}\,(\texttt{d}) &= \sum_{n=1}^{M} \left\langle \mathbb{E}_{\mathbf{z}_n}\left[\mathbf{g}_n\right] - \mathbb{E}_{\mathbf{z}_n}\left[\hat{\mathbf{g}}_n\right], \Delta_n - \mathbf{u}_n \right\rangle \\ &\leq \sum_{n=1}^{M} \left\|\mathbb{E}_{\mathbf{z}_n}\left[\mathbf{g}_n\right] - \mathbb{E}_{\mathbf{z}_n}\left[\hat{\mathbf{g}}_n\right]\right\| \left\|\Delta_n - \mathbf{u}_n\right\| \\ &\overset{(9)}{\leq} DKT \cdot \frac{2^{\mathfrak{p}}\left(\sigma^{\mathfrak{p}} + G^{\mathfrak{p}}\right)}{\tau^{\mathfrak{p}-1}}.\end{aligned}$$

For $\texttt{term}\,(\texttt{e})$, we can bound this term by utilizing the property (10) of clipped gradients. Let $D_n = \langle \mathbb{E}_{\mathbf{z}_n}[\hat{\mathbf{g}}_n] - \hat{\mathbf{g}}_n, \Delta_n - \mathbf{u}_n \rangle$, then $\{D_n\}$ is a martingale difference sequence. We make use of Freedman's inequality for martingales (c.f. Lemma D.3).

In order to apply Lemma D.3, we respectively bound the following two terms

$$D_n \leq \left\|\mathbb{E}_{\mathbf{z}_n}\left[\hat{\mathbf{g}}_n\right] - \hat{\mathbf{g}}_n\right\| \left\|\Delta_n - \mathbf{u}_n\right\| \leq 4\tau D,$$

$$\begin{aligned}\mathbb{E}_{\mathbf{z}_n}\left[D_n^2\right] \leq& \mathbb{E}_{\mathbf{z}_n}\left[\left\|\mathbb{E}_{\mathbf{z}_n}\left[\hat{\mathbf{g}}_n\right] - \hat{\mathbf{g}}_n\right\|^2 \left\|\Delta_n - \mathbf{u}_n\right\|^2\right] \\ &\overset{(10)}{\leq} D^2 2^{\mathfrak{p}+1}\tau^{2-\mathfrak{p}}\left(\sigma^{\mathfrak{p}} + G^{\mathfrak{p}}\right),\end{aligned}$$

where we utilize $\|\Delta_n\| \leq D$, $\|\mathbf{u}_n\| \leq D$ and $\|\hat{\mathbf{g}}_n\| \leq \tau$. From Lemma D.3, with probability at least $1 - q/6$, we have

$$\texttt{term}\,(\texttt{e}) \leq \frac{8D\tau}{3}\log\frac{6}{p} + D\sqrt{2^{\mathfrak{p}+2}\left(\sigma^{\mathfrak{p}} + G^{\mathfrak{p}}\right)M\tau^{2-\mathfrak{p}}\log\frac{6}{q}}.$$

We apply the union bound to combine the high-probability guarantee of each term in the decomposition, so that

$$\begin{aligned}&F\left(\mathbf{x}_M\right) - F\left(\mathbf{x}_0\right) \\ \leq& DK\tau\left(2 + \frac{23}{6}\log\frac{2K}{q}\right) - DT\sum_{k=1}^{K}\left\|\frac{1}{T}\sum_{t=1}^{T}\nabla F\left(\mathbf{w}_t^k\right)\right\| \\ &+ 4DG\sqrt{\frac{M}{2}\log\frac{6}{q}} + DKT \cdot \frac{2^{\mathfrak{p}}\left(\sigma^{\mathfrak{p}} + G^{\mathfrak{p}}\right)}{\tau^{\mathfrak{p}-1}} \\ &+ \frac{8D\tau}{3}\log\frac{6}{q} + D\sqrt{2^{\mathfrak{p}+2}\left(\sigma^{\mathfrak{p}} + G^{\mathfrak{p}}\right)M\tau^{2-\mathfrak{p}}\log\frac{6}{q}}\end{aligned}$$

holds with probability at least $1 - q$.

Finally, we rearrange the above inequality and divide both sides by DM, and then replace all τ by $T^{1/\mathfrak{p}}(\sigma^{\mathfrak{p}} + G^{\mathfrak{p}})^{1/\mathfrak{p}}\log(2K/q)^{-1/\mathfrak{p}}$ to finish the proof. The precise values of A and B are given by

$$A = \left(\sigma^{\mathfrak{p}} + G^{\mathfrak{p}}\right)^{1/\mathfrak{p}}\log\left(2K/q\right)^{-1/\mathfrak{p}}\left(8 + \frac{29}{2}\log\left(2K/q\right)\right),$$

$$B = \sqrt{8\log\left(6/q\right)}.$$

4.2. Proof of Theorem 3.9

As $\mathbb{E}_{\mathbf{w}_r,\mathbf{z}_r}[\mathbf{g}_r] = \frac{1}{T}\sum_{t=1}^{T}\nabla F(\mathbf{w}_t)$ holds true due to our choice of $\mathbf{w}_r$, we can decompose the approximation error of Algorithm 2 as

$$\begin{aligned}\left\|\frac{1}{R}\sum_{r=1}^{R}\tilde{\mathbf{g}}_r - \frac{1}{T}\sum_{t=1}^{T}\nabla F\left(\mathbf{w}_t\right)\right\| \leq& \underbrace{\frac{1}{R}\left\|\sum_{r=1}^{R}\tilde{\mathbf{g}}_r - \mathbb{E}_{\mathbf{w}_r,\mathbf{z}_r}\left[\tilde{\mathbf{g}}_r\right]\right\|}_{\texttt{term}\,(\texttt{f})} \\ &+ \underbrace{\frac{1}{R}\sum_{r=1}^{R}\left\|\mathbb{E}_{\mathbf{w}_r,\mathbf{z}_r}\left[\tilde{\mathbf{g}}_r\right] - \mathbb{E}_{\mathbf{w}_r,\mathbf{z}_r}\left[\mathbf{g}_r\right]\right\|}_{\texttt{term}\,(\texttt{g})}.\end{aligned}$$

For $\texttt{term}\,(\texttt{f})$, we let $X_r = \tilde{\mathbf{g}}_r - \mathbb{E}_{\mathbf{w}_r,\mathbf{z}_r}[\tilde{\mathbf{g}}_r]$ and find that $\{X_r\}$ forms a vector-valued martingale difference sequence. Then we bound $\|X_r\|$ and $\mathbb{E}_{\mathbf{w}_r,\mathbf{z}_r}[\|X_r\|^2]$ in order to apply concentration inequality for vector-valued martingales (c.f. Lemma D.4):

$$\left\|X_r\right\| \leq \left\|\tilde{\mathbf{g}}_r\right\| + \left\|\mathbb{E}_{\mathbf{w}_r,\mathbf{z}_r}\left[\tilde{\mathbf{g}}_r\right]\right\| \leq 2\tau,$$

$$\mathbb{E}_{\mathbf{w}_r,\mathbf{z}_r}\left[\left\|X_r\right\|^2\right] \leq \mathbb{E}_{\mathbf{w}_r,\mathbf{z}_r}\left[\left\|\tilde{\mathbf{g}}_r\right\|^2\right] \overset{(10)}{\leq} 2\tau^2R^{-1}\log\left(1/q\right).$$

Apply Lemma D.4 by taking $X_r = \tilde{\mathbf{g}}_r - \mathbb{E}_{\mathbf{w}_r,\mathbf{z}_r}[\tilde{\mathbf{g}}_r]$, $b_r = 2\tau$, $\sigma_r^2 = 2\tau^2R^{-1}\log(1/q)$, $q = q$ and $\nu = 2\tau$ to have

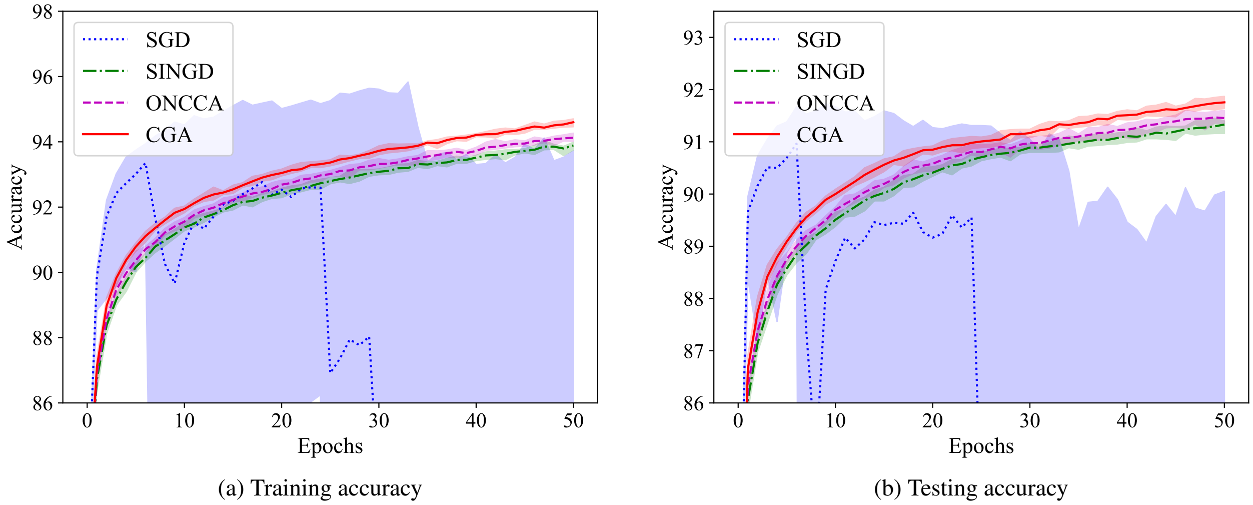
$$\begin{aligned}\left\|\sum_{r=1}^{R}X_r\right\| \leq& 5\tau\sqrt{2\log\frac{1}{q}\log\frac{64}{q}} + 46\tau\log\left(\frac{224}{q}\cdot 3^2\right) \\ \leq& \tau\left(\frac{5}{2}\log 64 + 46\log 2016 + \frac{107}{2}\log\frac{1}{q}\right).\end{aligned}$$

Thus with probability at least $1 - q$, we have

$$\texttt{term}\,(\texttt{f}) \leq \frac{\tau}{R}\left(\frac{5}{2}\log 64 + 46\log 2016 + \frac{107}{2}\log\frac{1}{q}\right).$$

For $\texttt{term}\,(\texttt{g})$, we expand the expectation on $\mathbf{w}_r$, then bound this part using property (9) of clipped gradients:

$$\begin{aligned}&\texttt{term}\,(\texttt{g}) \\ \leq& \frac{1}{TR}\sum_{t=1}^{T}\sum_{r=1}^{R}\left\|\mathbb{E}_{\mathbf{z}_r}[\tilde{\mathrm{G}}\left(\mathbf{w}_t,\mathbf{z}_r\right)] - \mathbb{E}_{\mathbf{z}_r}\left[\mathrm{GRAD}\left(\mathbf{w}_t,\mathbf{z}_r\right)\right]\right\| \\ \overset{(9)}{\leq}& 2\tau R^{-1}\log\left(1/q\right),\end{aligned}$$



(a) Training accuracy

(b) Testing accuracy

Figure 1. Accuracy of different methods versus the number of epochs.

where $\tilde{\mathrm{G}}\left(\mathbf{w}_t, \mathbf{z}_r\right)$ denotes applying gradient clipping to $\mathrm{GRAD}\left(\mathbf{w}_t, \mathbf{z}_r\right)$ with parameter τ.

We apply the union bound and substitute τ with $R^{1/\mathfrak{p}}(\sigma^{\mathfrak{p}} + G^{\mathfrak{p}})^{1/\mathfrak{p}} \log(1/q)^{-1/\mathfrak{p}}$ to finish the proof.

5. Preliminary Experiment

We conduct a preliminary experiment to evaluate our proposed algorithm, with details provided in Appendix C. We train the ResNet18 (He et al., 2016) model on the CIFAR10 (Krizhevsky & Hinton, 2009) dataset, which consists of a training set of 50k images and a testing set of 10k images from 10 classes. As we are investigating the heavy-tailed distributions of stochastic gradients, we manually add heavy-tailed noise to the gradients. We compare the performance of CGA (Candidate Generation Algorithm, our Algorithm 1) with three benchmarks: SGD with momentum, SINGD (Stochastic-INGD, Algorithm 2 of Zhang et al., 2020c) and ONCCA (Online-to-Non-Convex Conversion Algorithm, Algorithm 1 of Cutkosky et al., 2023). For each algorithm, we choose the output of the last iteration for evaluation.

From the results presented in Figure 1, We observe that SGD performs poorly and does not converge due to the heavy-tailed noise. Meanwhile, SINGD, ONCCA and CGA make steady progress towards higher accuracy. Although SINGD and ONCCA do not explicitly address heavy-tailed gradients, the normalization steps in their update strategies seem to mitigate the effects. Our CGA further outperforms SINGD and ONCCA on both the training and testing sets in terms of accuracy, which demonstrates the effectiveness of applying gradient clipping to the stochastic gradients.

6. Conclusion and Future Work

In this paper, we propose a novel algorithm for non-smooth non-convex stochastic optimization that supports heavy-tailed gradients. As far as we know, it is the first attempt to investigate such setting. We integrate the gradient clipping technique into the online-to-non-convex framework to produce candidate points, and validate a solution among candidates by computing the approximation of the expected gradient norm. Our algorithm provides a high-probability bound of consuming $\tilde{\mathcal{O}}(\epsilon^{-\frac{2\mathfrak{p}-1}{\mathfrak{p}-1}}\delta^{-1})$ gradient queries for finding a (δ, ϵ)-stationary point. When the gradient distributions have bounded variance, our complexity becomes $\tilde{\mathcal{O}}(\epsilon^{-3}\delta^{-1})$, better than the existing $\tilde{\mathcal{O}}(\epsilon^{-4}\delta^{-1})$ high-probability bound. For smooth objectives, our algorithm identifies ϵ-stationary points with a query complexity of $\tilde{\mathcal{O}}(\epsilon^{-\frac{3\mathfrak{p}-2}{\mathfrak{p}-1}})$, nearly matching the lower bound.

For future work, since our algorithm requires the knowledge of problem parameters σ and G, it is of interest to explore the possibility for any parameter-free extensions, for instance, removing the dependence of the clipping parameter on either σ (Liu & Zhou, 2023) or G (Sadiev et al., 2023). Another possible direction is designing a time-varying clipping parameter, which may extend the theoretical guarantee to an unknown time horizon and enable more refined analysis (Nguyen et al., 2023).

Acknowledgements

This work was partially supported by NSFC (62122037, U23A20382), and the Collaborative Innovation Center of Novel Software Technology and Industrialization.

Impact Statement

This paper is mainly theoretical and discusses algorithms for optimization. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Allen-Zhu, Z. Natasha 2: Faster non-convex optimization than SGD. In *Advances in Neural Information Processing Systems 31*, pp. 2680–2691, 2018.

Arjevani, Y., Carmon, Y., Duchi, J. C., Foster, D. J., Srebro, N., and Woodworth, B. E. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199 (1):165–214, 2023.

Audibert, J.-Y. and Catoni, O. Robust linear least squares regression. *Annals of Statistics*, 39:2766–2794, 2010.

Benaïm, M., Hofbauer, J., and Sorin, S. Stochastic approximations and differential inclusions. *SIAM Journal on Control and Optimization*, 44(1):328–348, 2005.

Carmon, Y., Duchi, J. C., Hinder, O., and Sidford, A. "Convex until proven guilty": Dimension-free acceleration of gradient descent on non-convex functions. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 654–663, 2017.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, learning, and games*. Cambridge University Press, 2006.

Cutkosky, A. and Mehta, H. High-probability bounds for non-convex stochastic optimization with heavy tails. In *Advances in Neural Information Processing Systems 34*, pp. 4883–4895, 2021.

Cutkosky, A. and Orabona, F. Momentum-based variance reduction in non-convex SGD. In *Advances in Neural Information Processing Systems 32*, pp. 15210–15219, 2019.

Cutkosky, A., Mehta, H., and Orabona, F. Optimal stochastic non-smooth non-convex optimization through online-to-non-convex conversion. *Proceedings of the 40th International Conference on Machine Learning*, pp. 6643–6670, 2023.

Davis, D., Drusvyatskiy, D., Kakade, S. M., and Lee, J. D. Stochastic subgradient method converges on tame functions. *Foundations of Computational Mathematics*, 20(1): 119–154, 2020.

Davis, D., Drusvyatskiy, D., Lee, Y. T., Padmanabhan, S., and Ye, G. A gradient sampling method with complexity guarantees for lipschitz functions in high and low dimensions. In *Advances in Neural Information Processing Systems 35*, pp. 6692–6703, 2022.

Duchi, J. C., Bartlett, P. L., and Wainwright, M. J. Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization*, 22(2):674–701, 2012.

Fang, C., Li, C. J., Lin, Z., and Zhang, T. SPIDER: near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems 31*, pp. 687–697, 2018.

Faw, M., Tziotis, I., Caramanis, C., Mokhtari, A., Shakkottai, S., and Ward, R. A. The power of adaptivity in SGD: self-tuning step sizes with unbounded gradients and affine variance. In *Proceedings of the 35th Conference on Learning Theory*, pp. 313–355, 2022.

Ghadimi, S. and Lan, G. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Goodfellow, I. J., Bengio, Y., and Courville, A. C. *Deep Learning*. MIT Press, 2016.

Goyal, P., Dollár, P., Girshick, R. B., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., and He, K. Accurate, large minibatch SGD: training imagenet in 1 hour. *ArXiv e-prints*, arXiv:1706.02677, 2017.

Gürbüzbalaban, M., Simsekli, U., and Zhu, L. The heavy-tail phenomenon in SGD. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 3964–3975, 2021.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the 29th IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Hsu, D. J. and Sabato, S. Heavy-tailed regression with a generalized median-of-means. In *Proceedings of the 31th International Conference on Machine Learning*, pp. 37–45, 2014.

Jiang, W., Li, G., Wang, Y., Zhang, L., and Yang, T. Multi-block-single-probe variance reduced estimator for coupled compositional optimization. In *Advances in Neural Information Processing Systems 35*, pp. 32499–32511, 2022.

Jordan, M. I., Kornowski, G., Lin, T., Shamir, O., and Zampetakis, M. Deterministic nonsmooth nonconvex optimization. In *the 36th Conference on Learning Theory*, volume 195, pp. 4570–4597, 2023.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Kornowski, G. and Shamir, O. On the complexity of finding small subgradients in nonsmooth optimization. *ArXiv e-prints*, arXiv:2209.10346, 2022.

Kornowski, G. and Shamir, O. An algorithm with optimal dimension-dependence for zero-order nonsmooth nonconvex stochastic optimization. *ArXiv e-prints*, arXiv:2307.04504, 2023.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Levy, K. Y., Kavis, A., and Cevher, V. STORM+: fully adaptive SGD with recursive momentum for nonconvex optimization. In *Advances in Neural Information Processing Systems 34*, pp. 20571–20582, 2021.

Lin, T., Zheng, Z., and Jordan, M. I. Gradient-free methods for deterministic and stochastic nonsmooth nonconvex optimization. In *Advances in Neural Information Processing Systems 35*, pp. 26160–26175, 2022.

Liu, Z. and Zhou, Z. Stochastic nonsmooth convex optimization with heavy-tailed noises: High-probability bound, in-expectation rate and initial distance adaptation. *ArXiv e-prints*, arXiv:2303.12277, 2023.

Liu, Z., Nguyen, T. D., Nguyen, T. H., Ene, A., and Nguyen, H. L. META-STORM: generalized fully-adaptive variance reduced SGD for unbounded functions. *ArXiv e-prints*, arXiv:2209.14853, 2022.

Lu, S., Wang, G., Hu, Y., and Zhang, L. Optimal algorithms for Lipschitz bandits with heavy-tailed rewards. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 4154–4163, 2019.

Lugosi, G. and Mendelson, S. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19:1145–1190, 2019.

Majewski, S., Miasojedow, B., and Moulines, E. Analysis of nonsmooth stochastic approximation: the differential inclusion approach. *ArXiv e-prints*, arXiv:1805.01916, 2018.

Mikolov, T. *Statistical language models based on neural networks*. PhD thesis, Brno University of Technology, 2012.

Nesterov, Y. *Introductory lectures on convex optimization: a basic course*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, 2004.

Nesterov, Y. E. and Spokoiny, V. G. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

Nguyen, T. D., Nguyen, T. H., Ene, A., and Nguyen, H. Improved convergence in high probability of clipped gradient methods with heavy tailed noise. In *Advances in Neural Information Processing Systems 36*, pp. 24191–24222, 2023.

Orabona, F. A modern introduction to online learning. *ArXiv e-prints*, arXiv:1912.13213 (v5), 2019.

Qiu, Z., Hu, Q., Zhong, Y., Zhang, L., and Yang, T. Large-scale stochastic optimization of NDCG surrogates for deep learning with provable convergence. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 18122–18152, 2022.

Sadiev, A., Danilova, M., Gorbunov, E., Horváth, S., Gidel, G., Dvurechensky, P., Gasnikov, A., and Richtárik, P. High-probability bounds for stochastic optimization and variational inequalities: the case of unbounded variance. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 29563–29648, 2023.

Tian, L. and So, A. M.-C. No dimension-free deterministic algorithm computes approximate stationarities of lipschitzians. *ArXiv e-prints*, arXiv:2210.06907, 2022.

Tian, L., Zhou, K., and So, A. M. On the finite-time complexity and practical computation of approximate stationarity concepts of lipschitz functions. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 21360–21379, 2022.

Tripuraneni, N., Stern, M., Jin, C., Regier, J., and Jordan, M. I. Stochastic cubic regularization for fast nonconvex optimization. In *Advances in Neural Information Processing Systems 31*, pp. 2904–2913, 2018.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pp. 5998–6008, 2017.

Xue, B., Wang, Y., Wan, Y., Yi, J., and Zhang, L. Efficient algorithms for generalized linear bandits with heavy-tailed rewards. In *Advances in Neural Information Processing Systems 36*, pp. 70880–70891, 2023.

Yousefian, F., Nedic, A., and Shanbhag, U. V. On stochastic gradient and subgradient methods with adaptive steplength sequences. *Automatica*, 48(1):56–67, 2012.

Zhang, J. and Cutkosky, A. Parameter-free regret in high probability with heavy tails. In *Advances in Neural Information Processing Systems 35*, pp. 8000–8012, 2022.

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

Zhang, J., Karimireddy, S. P., Veit, A., Kim, S., Reddi, S. J., Kumar, S., and Sra, S. Why are adaptive methods good for attention models? In *Advances in Neural Information Processing Systems 33*, pp. 15383–15393, 2020b.

Zhang, J., Lin, H., Jegelka, S., Sra, S., and Jadbabaie, A. Complexity of finding stationary points of nonconvex nonsmooth functions. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 11173–11182, 2020c.

Zhang, L. and Zhou, Z. ℓ_1-regression with heavy-tailed distributions. In *Advances in Neural Information Processing Systems 31*, pp. 1084–1094, 2018.

Zhou, D., Xu, P., and Gu, Q. Stochastic nested variance reduced gradient descent for nonconvex optimization. In *Advances in Neural Information Processing Systems 31*, pp. 3925–3936, 2018.

Zinkevich, M. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning*, pp. 928–936, 2003.

A. Properties of Clipped Gradients

These properties are first introduced by Zhang et al. (2020b) and similarly analyzed by Zhang & Cutkosky (2022). We reproduce the following results from Zhang & Cutkosky (2022, Lemma 15) for completeness.

Lemma A.1. *Suppose $\mathbf{g}_t$ is a heavy-tailed random vector, $\|\mathbb{E}\left[\mathbf{g}_t\right]\| \leq G$, $\mathbb{E}[\|\mathbf{g}_t - \mathbb{E}[\mathbf{g}_t]\|^{\mathfrak{p}}] \leq \sigma^{\mathfrak{p}}$ for some $\mathfrak{p} \in (1, 2]$ and $\sigma < \infty$. Define truncated gradient $\hat{\mathbf{g}}_t$ with a positive clipping parameter τ:*

$$\hat{\mathbf{g}}_t = \frac{\mathbf{g}_t}{\|\mathbf{g}_t\|} \min\left(\tau, \|\mathbf{g}_t\|\right),$$

then we have:

$$\begin{aligned}
\|\mathbb{E}\left[\hat{\mathbf{g}}_t\right] - \mathbb{E}\left[\mathbf{g}_t\right]\| &\leq \frac{2^{\mathfrak{p}-1}\left(\sigma^{\mathfrak{p}} + G^{\mathfrak{p}}\right)}{\tau^{\mathfrak{p}-1}}, \\
\mathbb{E}\left[\|\hat{\mathbf{g}}_t\|^2\right] &\leq 2^{\mathfrak{p}-1}\tau^{2-\mathfrak{p}}\left(\sigma^{\mathfrak{p}} + G^{\mathfrak{p}}\right).
\end{aligned}$$

First we bound the bias by

$$\begin{aligned}
\|\mathbb{E}\left[\hat{\mathbf{g}}_t\right] - \mathbb{E}\left[\mathbf{g}_t\right]\| = &\|\mathbb{E}\left[\hat{\mathbf{g}}_t - \mathbf{g}_t\right]\| \\
\leq &\mathbb{E}\left[\|\hat{\mathbf{g}}_t - \mathbf{g}_t\|\right] \\
\leq &\mathbb{E}\left[\|\mathbf{g}_t\| \mathbf{1}\left[\|\mathbf{g}_t\| \geq \tau\right]\right] \\
\leq &\mathbb{E}[\|\mathbf{g}_t\|^{\mathfrak{p}} / \tau^{\mathfrak{p}-1}] \\
\leq &\mathbb{E}[(\|\mathbf{g}_t - \mathbb{E}\left[\mathbf{g}_t\right]\| + \|\mathbb{E}\left[\mathbf{g}_t\right]\|)^{\mathfrak{p}} / \tau^{\mathfrak{p}-1}] \\
= &\frac{2^{\mathfrak{p}}}{\tau^{\mathfrak{p}-1}}\mathbb{E}[(\frac{1}{2}\|\mathbf{g}_t - \mathbb{E}\left[\mathbf{g}_t\right]\| + \frac{1}{2}\|\mathbb{E}\left[\mathbf{g}_t\right]\|)^{\mathfrak{p}}] \\
\leq &\frac{2^{\mathfrak{p}-1}}{\tau^{\mathfrak{p}-1}}\left(\mathbb{E}\left[\|\mathbf{g}_t - \mathbb{E}\left[\mathbf{g}_t\right]\|^{\mathfrak{p}}\right] + \mathbb{E}\left[\|\mathbf{g}_t\|^{\mathfrak{p}}\right]\right) \\
\leq &\frac{2^{\mathfrak{p}-1}\left(\sigma^{\mathfrak{p}} + G^{\mathfrak{p}}\right)}{\tau^{\mathfrak{p}-1}},
\end{aligned}$$

where the first and the second last inequality follow from convexity of $\|\cdot\|$ and $(\cdot)^{\mathfrak{p}}$ accordingly and linearity of expectation. Then we continue to bound the variance with similar algebra and have

$$\begin{aligned}
\mathbb{E}\left[\|\hat{\mathbf{g}}_t\|^2\right] \leq &\mathbb{E}\left[\|\mathbf{g}_t\|^{\mathfrak{p}}\tau^{2-\mathfrak{p}}\right] \\
\leq &\mathbb{E}\left[(\|\mathbf{g}_t - \mathbb{E}\left[\mathbf{g}_t\right]\| + \|\mathbb{E}\left[\mathbf{g}_t\right]\|)^{\mathfrak{p}}\tau^{2-\mathfrak{p}}\right] \\
= &2^{\mathfrak{p}}\tau^{2-\mathfrak{p}}\mathbb{E}[(\frac{1}{2}\|\mathbf{g}_t - \mathbb{E}\left[\mathbf{g}_t\right]\| + \frac{1}{2}\|\mathbb{E}\left[\mathbf{g}_t\right]\|)^{\mathfrak{p}}] \\
\leq &2^{\mathfrak{p}-1}\tau^{2-\mathfrak{p}}\left(\mathbb{E}\left[\|\mathbf{g}_t - \mathbb{E}\left[\mathbf{g}_t\right]\|^{\mathfrak{p}}\right] + \mathbb{E}\left[\|\mathbf{g}_t\|^{\mathfrak{p}}\right]\right) \\
\leq &2^{\mathfrak{p}-1}\tau^{2-\mathfrak{p}}\left(\sigma^{\mathfrak{p}} + G^{\mathfrak{p}}\right).
\end{aligned}$$

B. Omitted Proofs in Sections 3 and 4

B.1. Proof for Corollary 3.8

Corollary 3.8. *Suppose we have a budget of N gradient queries. Set $T = \min(\lceil(\frac{AN\delta}{F(\mathbf{x}_0)-F^\star})^{\mathfrak{p}/(2\mathfrak{p}-1)}\rceil, \frac{N}{2})$, $K = \lfloor\frac{N}{T}\rfloor$ and $D = \delta/T$ for an arbitrary $\delta > 0$. Under Assumptions 3.2, 3.3 and 3.4, with probability at least $1 - q$, Algorithm 1 ensures*

$$\begin{aligned}
\frac{1}{K}\sum_{k=1}^{K}\left\|\nabla F\left(\bar{\mathbf{w}}^k\right)\right\|_\delta \leq &\frac{2\left(F\left(\mathbf{x}_0\right) - F^\star\right)}{\delta N} + \frac{2B}{\sqrt{N}} \\
&+ \max\left(\frac{4A^{\frac{\mathfrak{p}}{2\mathfrak{p}-1}}\left(F\left(\mathbf{x}_0\right) - F^\star\right)^{\frac{\mathfrak{p}-1}{2\mathfrak{p}-1}}}{\left(\delta N\right)^{\frac{\mathfrak{p}-1}{2\mathfrak{p}-1}}}, \frac{4A}{N^{\frac{\mathfrak{p}-1}{\mathfrak{p}}}}\right),
\end{aligned}$$

where A, B are given in Theorem 3.7.

Proof. Recall that $\|\Delta_n\| \leq D$, we can then bound the distance between $\bar{\mathbf{w}}^k = \frac{1}{T}\sum_{t=1}^{T} \mathbf{w}_t^k$ and $\mathbf{w}_1^k, \cdots, \mathbf{w}_T^k$ for a fixed $k \in [K]$. For $1 < t \leq T$, with update rule $\mathbf{x}_n = \mathbf{x}_{n-1} + \Delta_n$ in Line 5 and $\mathbf{w}_n = \mathbf{x}_{n-1} + s_n \Delta_n$ in Line 7 of Algorithm 1, we have

$$\begin{aligned}
\left\|\mathbf{w}_t^k - \mathbf{w}_1^k\right\| &= \left\|\left(\mathbf{x}_{t-1}^k + s_t^k \Delta_t^k\right) - \left(\mathbf{x}_1^k + \left(s_1^k - 1\right)\Delta_1^k\right)\right\| \\
&\leq \left\|\sum_{i=2}^{t-1} \Delta_i^k\right\| + \left\|\left(s_t^k - 1\right)\Delta_t^k\right\| + \left\|\left(s_1^k - 1\right)\Delta_1^k\right\| \\
&\leq \sum_{i=2}^{t-1} \left\|\Delta_i^k\right\| + 2D \leq tD.
\end{aligned}$$

Thus, the distance between $\bar{\mathbf{w}}^k$ and $\mathbf{w}_t^k$ for $\forall t \in [T]$ is bounded by

$$\left\|\bar{\mathbf{w}}^k - \mathbf{w}_t^k\right\| = \left\|\frac{1}{T}\sum_{i=1}^{T} \mathbf{w}_i^k - \mathbf{w}_t^k\right\| = \frac{1}{T}\sum_{i=1}^{T} \left\|\mathbf{w}_i^k - \mathbf{w}_t^k\right\| \leq TD = \delta.$$

We relate Theorem 3.7 to our setting of D, K, T to obtain

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\left\|\frac{1}{T}\sum_{t=1}^{T} \nabla F\left(\mathbf{w}_t^k\right)\right\| \\
\leq& \frac{2\left(F\left(\mathbf{x}_0\right) - F^\star\right)}{DN} + \frac{2AKT^{1/\mathfrak{p}}}{N} + \frac{2B}{\sqrt{N}} \\
\leq& \frac{2T\left(F\left(\mathbf{x}_0\right) - F^\star\right)}{\delta N} + \frac{2A}{T^{\frac{\mathfrak{p}-1}{\mathfrak{p}}}} + \frac{2B}{\sqrt{N}} \\
\leq& \max\left(\frac{4A^{\frac{\mathfrak{p}}{2\mathfrak{p}-1}}\left(F\left(\mathbf{x}_0\right) - F^\star\right)^{\frac{\mathfrak{p}-1}{2\mathfrak{p}-1}}}{(\delta N)^{\frac{\mathfrak{p}-1}{2\mathfrak{p}-1}}}, \frac{4A}{N^{\frac{\mathfrak{p}-1}{\mathfrak{p}}}}\right) + \frac{2\left(F\left(\mathbf{x}_0\right) - F^\star\right)}{\delta N} + \frac{2B}{\sqrt{N}},
\end{aligned} \tag{13}$$

with probability at least $1 - q$, where the first step follows from $N < 2KT = 2M$, the second step follows from $D = \delta / T$ and $N \geq KT$, the last step follows from our choice of T.

Note that we have $\bar{\mathbf{w}}^k = \frac{1}{T}\sum_{t=1}^{T} \mathbf{w}_t^k$ for $\forall k \in [K]$, then we can relate (13) to Definition 3.6. We have a support set $S' = \{\mathbf{w}_1^k, \cdots, \mathbf{w}_T^k\}$ whose center is $\bar{\mathbf{w}}^k$ and satisfying $S' \subset \mathbb{B}\left(\bar{\mathbf{w}}^k, \delta\right)$, thus it establishes an upper bound for $\|\nabla F(\bar{\mathbf{w}}^k)\|_\delta$, that is

$$\left\|\nabla F\left(\bar{\mathbf{w}}^k\right)\right\|_\delta \triangleq \inf_{S \subset \mathbb{B}(\bar{\mathbf{w}}^k, \delta), \frac{1}{|S|}\sum\limits_{\mathbf{y}\in S} \mathbf{y} = \bar{\mathbf{w}}^k} \left\|\frac{1}{|S|}\sum_{\mathbf{y}\in S} \nabla F\left(\mathbf{y}\right)\right\| \leq \left\|\frac{1}{T}\sum_{t=1}^{T} \nabla F\left(\mathbf{w}_t^k\right)\right\|.$$

Taking the average over K restarts to have $\frac{1}{K}\sum_{k=1}^{K} \|\nabla F(\bar{\mathbf{w}}^k)\|_\delta \leq \frac{1}{K}\sum_{k=1}^{K} \|\frac{1}{T}\sum_{t=1}^{T} \nabla F(\mathbf{w}_t^k)\|$, we finally combine it with (13) to finish the proof. □

B.2. Proof for Theorem 3.10

Theorem 3.10. *Set $N = \tilde{\mathcal{O}}(\epsilon^{-\frac{2\mathfrak{p}-1}{\mathfrak{p}-1}}\delta^{-1}\log(1/q))$ and $R = \mathcal{O}(\epsilon^{-\frac{\mathfrak{p}}{\mathfrak{p}-1}}\log(1/q))$. Under Assumptions 3.2, 3.3 and 3.4, with probability at least $1 - q$, Algorithm 3 guarantees to output a (δ, ϵ)-stationary point with $\tilde{\mathcal{O}}(\epsilon^{-\frac{2\mathfrak{p}-1}{\mathfrak{p}-1}}\delta^{-1}\log(1/q) + \epsilon^{-\frac{\mathfrak{p}}{\mathfrak{p}-1}}\log(1/q)^2)$ queries of stochastic gradient oracle.*

Proof. We apply Corollary 3.8 with the specification of taking $\epsilon_{\text{gen}} = \epsilon/3$ and $q_{\text{gen}} = q/3$. Then, with a budget of $N = \tilde{\mathcal{O}}({\epsilon_{\text{gen}}}^{-\frac{2\mathfrak{p}-1}{\mathfrak{p}-1}}\delta^{-1}\log(1/q_{\text{gen}}))$ gradient queries, we ensure $\frac{1}{K}\sum_{k=1}^{K}\|\frac{1}{T}\sum_{t=1}^{T}\nabla F(\mathbf{w}_t^k)\| \leq \epsilon/3$ with high probability. Meanwhile, since norms are non-negative, at least half of $\|\frac{1}{T}\sum_{t=1}^{T}\nabla F(\mathbf{w}_t^k)\|$ are no greater than $2\epsilon/3$. Therefore, by selecting $k \in [K]$ at random, with probability $1 - q/3$ we will meet a k that satisfies $\|\frac{1}{T}\sum_{t=1}^{T} \nabla F(\mathbf{w}_t^k)\| \leq 2\epsilon/3$ in $\log(3/q)$ rounds.

We continue to analyze the query complexity of each validation algorithm call. We apply Theorem 3.9 by setting the precision of approximation to $\epsilon_{\text{val}} = \epsilon/6$ and $q_{\text{val}} = q/(3\log(1/q))$. Thus, each call of Algorithm 2 requires $R = \mathcal{O}({\epsilon_{\text{val}}}^{-\frac{\mathtt{p}}{\mathtt{p}-1}} \log(1/q_{\text{val}}))$ gradient queries. Under this setting, any $\|\frac{1}{T}\sum_{t=1}^{T} \nabla F(\mathbf{w}_t^k)\| > \epsilon$ will be rejected by our threshold of $5\epsilon/6$ with probability $1 - q_{\text{val}}$, which guarantees the validity of the accepted solution. From another perspective, any $\|\frac{1}{T}\sum_{t=1}^{T} \nabla F(\mathbf{w}_t^k)\| \leq 2\epsilon/3$ will be accepted with probability $1 - q_{\text{val}}$, meaning that a valid solution is likely to be accepted within $\mathcal{O}(\log(1/q))$ rounds.

Since we call Algorithm 1 for one time and call Algorithm 2 for $\mathcal{O}(\log(1/q))$ times, we require $\tilde{\mathcal{O}}(\epsilon^{-\frac{3\mathtt{p}-2}{\mathtt{p}-1}} \log(1/q) + \epsilon^{-\frac{\mathtt{p}}{\mathtt{p}-1}} \log(1/q)^2)$ gradient queries in total. □

B.3. Proof for Corollary 3.12

Corollary 3.12. *Further set $\delta = \epsilon/H$. Under Assumptions 3.2, 3.3, 3.4 and 3.11, with probability at least $1 - q$, Algorithm 3 guarantees to output an ϵ-stationary point with $\tilde{\mathcal{O}}(\epsilon^{-\frac{3\mathtt{p}-2}{\mathtt{p}-1}} \log(1/q) + \epsilon^{-\frac{\mathtt{p}}{\mathtt{p}-1}} \log(1/q)^2)$ queries of stochastic gradient oracle.*

Proof. When the objective function F is smooth, we will see that with minor modification, the (δ, ϵ)-stationary point we identified converts to $(0, \epsilon')$-stationary point for some appropriate ϵ'. The conversion is based on the following lemma (Cutkosky et al., 2023, Proposition 14).

Lemma B.1. *Suppose that F is H-smooth (that is, ∇F is H-Lipschitz) and $\mathbf{x}$ also satisfies $\|\nabla F(\mathbf{x})\|_\delta \leq \epsilon$. Then, $\|\nabla F(\mathbf{x})\| \leq \epsilon + H\delta$.*

From Lemma B.1, by setting $\delta = \epsilon/H$, we have $\|\nabla F(\mathbf{x})\| \leq 2\epsilon$ under Assumption 3.11, indicating that the (δ, ϵ)-stationary point $\mathbf{x}$ is also a 2ϵ-stationary point. Thus, the number of required queries changes from $\tilde{\mathcal{O}}(\epsilon^{-\frac{2\mathtt{p}-1}{\mathtt{p}-1}} \delta^{-1} \log(1/q) + \epsilon^{-\frac{\mathtt{p}}{\mathtt{p}-1}} \log(1/q)^2)$ to $\tilde{\mathcal{O}}(\epsilon^{-\frac{3\mathtt{p}-2}{\mathtt{p}-1}} \log(1/q) + \epsilon^{-\frac{\mathtt{p}}{\mathtt{p}-1}} \log(1/q)^2)$ by our specification of δ. □

B.4. Proof for Lemma 4.1

This lemma is originally introduced by Zhang et al. (2020a) and is referred to as the random scaling trick. We reproduce the proof for completeness.

Lemma 4.1. *With the notation in Algorithm 1, under Assumption 3.4, we have*

$$F(\mathbf{x}_M) - F(\mathbf{x}_0) = \sum_{n=1}^{M} \langle \mathbb{E}_{\mathbf{w}_n, \mathbf{z}_n}[\mathbf{g}_n], \Delta_n \rangle. \tag{7}$$

Proof. Under Assumption 3.4, we have

$$\begin{aligned} F(\mathbf{x}_M) - F(\mathbf{x}_0) &= \sum_{n=1}^{M} F(\mathbf{x}_n) - F(\mathbf{x}_{n-1}) \\ &= \sum_{n=1}^{M} \int_0^1 \langle \nabla F(\mathbf{x}_{n-1} + s(\mathbf{x}_n - \mathbf{x}_{n-1})), \mathbf{x}_n - \mathbf{x}_{n-1} \rangle \, \mathrm{d}s \qquad (14) \\ &= \sum_{n=1}^{M} \int_0^1 \langle \nabla F(\mathbf{x}_{n-1} + s\Delta_n), \Delta_n \rangle \, \mathrm{d}s. \end{aligned}$$

Meanwhile, we can simplify the above equation with the following observation:

$$\mathbb{E}_{\mathbf{w}_n, \mathbf{z}_n}[\mathbf{g}_n] = \mathbb{E}_{\mathbf{w}_n}[\mathbb{E}_{\mathbf{z}_n}[\mathrm{GRAD}(\mathbf{w}_n, \mathbf{z}_n)]] = \mathbb{E}_{s_n}[\nabla F(\mathbf{x}_{n-1} + s_n\Delta_n)] = \int_0^1 \nabla F(\mathbf{x}_{n-1} + s\Delta_n) \, \mathrm{d}s, \tag{15}$$

where the first equality holds as $\mathbf{w}_n$ and $\mathbf{z}_n$ are independent. Combining (14) and (15) to finish the proof. □

B.5. Proof for Lemma 4.2

Lemma 4.2. *With the notation in Algorithm 1, under Assumptions 3.2 and 3.4, we have*

$$\left\|\mathbb{E}_{\mathbf{w}_n,\mathbf{z}_n}\left[\mathbf{g}_n\right]\right\| \le G. \tag{8}$$

Proof. Under Assumption 3.2, we further bound the norm of (15) to have

$$\left\|\mathbb{E}_{\mathbf{w}_n,\mathbf{z}_n}\left[\mathbf{g}_n\right]\right\| \overset{(15)}{=} \left\|\int_0^1 \nabla F\left(\mathbf{x}_{n-1}+s\Delta_n\right)\mathrm{d}s\right\| \le \int_0^1 \left\|\nabla F\left(\mathbf{x}_{n-1}+s\Delta_n\right)\right\|\mathrm{d}s \le G.$$

□

C. Experimental Setting

In this part, we provide experimental details of Section 5.

To our knowledge, only Zhang et al. (2020c) has conducted experiment in non-smooth non-convex stochastic optimization. Therefore, we adopt their setup to design our experiment.

As we are investigating the heavy-tailed distributions of stochastic gradients, we manually add heavy-tailed noise to the gradients. The noise is added in the following manner. First, we concatenate the computed gradients into a vector $\mathbf{g}_{\text{raw}}$. Then, we generate a vector $\mathbf{v}$ of the same size, where each element follows the continuous uniform distribution over the interval $[-1,1]$. Next, we generate a variable X following the Pareto distribution described by shape $\alpha = 1.5$ and scale $x_m = 4$, which implies $\mathbb{E}[X^{\mathfrak{p}}] = \frac{\alpha x_m{}^{\mathfrak{p}}}{\alpha-\mathfrak{p}}$ for $\mathfrak{p} < \alpha$. Finally, we combine $\mathbf{g}_{\text{raw}}$ with the generated noise to compute the noisy gradient as $\mathbf{g}_{\text{noisy}} = \mathbf{g}_{\text{raw}} + \mathbf{v}X/\|\mathbf{v}\|$, and feed $\mathbf{g}_{\text{noisy}}$ as the gradient accessed through the oracle to algorithms. Under this setting, we have $\mathbb{E}[\|\mathbf{g}_{\text{noisy}} - \mathbf{g}_{\text{raw}}\|^{\mathfrak{p}}] = \mathbb{E}[X^{\mathfrak{p}}] = \frac{\alpha x_m{}^{\mathfrak{p}}}{\alpha-\mathfrak{p}}$.

We compare the performance of our algorithm with three benchmarks. The specific settings are as follows:

- SGD with momentum. We set the learning rate as 0.01 and momentum as 0.9.
- SINGD (Stochastic-INGD, Algorithm 2 of Zhang et al., 2020c). We use $\beta = 0.9$, $p = 1$, $q = 10$, and multiply the learning rate by an additional factor of 0.1. This setting follows from their experiment.
- ONCCA (Online-to-Non-Convex Conversion Algorithm, Algorithm 1 of Cutkosky et al., 2023). We use $D = 2.5\times10^{-2}$ and $\eta = 2.5 \times 10^{-3}$.
- CGA (Candidate Generation Algorithm, our Algorithm 1). We use $D = 2.5 \times 10^{-2}$, $\tau = 10$ and $\eta = 2.5 \times 10^{-3}$.

All four algorithms are equipped with a weight decay parameter of 5×10^{-4}.

We run these four algorithms with the same pre-trained model, which achieves 81.96% accuracy on the training set and 81.10% on the testing set. We report their performance over the next 50 epochs, with the accuracy averaged over 20 individual runs.

D. Some Useful Concentration Inequalities

We make use of several concentration inequalities in Section 4. Here we provide these inequalities for reference.

The first is bounded squared sum for sub-exponential sequences (Zhang & Cutkosky, 2022, Lemma 24).

Lemma D.1. *Suppose $\{X_t, \mathcal{F}_t\}$ is a sequence with $|X_t| \le b$ and $\mathbb{E}[X_t^2|\mathcal{F}_t] \le \sigma^2$ almost surely for some fixed σ, b. Then with probability at least $1-q$ we have*

$$\sum_{t=1}^{T} X_t^2 \le \frac{3\sigma^2}{2}T + \frac{5}{3}b^2\log\frac{1}{q}.$$

The second is Hoeffding-Azuma inequality for martingale difference sequences stated below (Cesa-Bianchi & Lugosi, 2006).

Lemma D.2. *Let $V_1, V_2, \cdots$ be a martingale difference sequence with respect to some sequence $X_1, X_2, \cdots$ such that $V_i \in [A_i, A_i + c_i]$ for some random variable A_i, measurable with respect to $X_1, \cdots, X_{i-1}$ and a positive constant c_i. Then* *for any $t > 0$,*

$$\Pr\left[\sum_{i=1}^{n} V_i > t\right] \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n} c_i^2}\right).$$

When we specify $|V_i| \leq b$, Lemma D.2 is equivalent to guarantee

$$\sum_{i=1}^{n} V_i \leq 2b\sqrt{\frac{n}{2}\log\frac{1}{q}}$$

with probability at least $1 - q$.

The third is Freedman's inequality for martingales (Cutkosky & Mehta, 2021, Lemma 11).

Lemma D.3. *Let $D_1, D_2, \cdots$ be a martingale difference sequence adapted to a filtration $\mathcal{F}_1, \mathcal{F}_2, \cdots$ such that $D_i \leq R$ almost surely for all i. Let $\mathbb{E}_i$ indicate expectation conditioned on $\mathcal{F}_i$. Suppose further that for all i with probability 1,*

$$\mathbb{E}_{t-1}\left[D_i^2\right] \leq \sigma_i^2.$$

Then with probability at least $1 - q$, for all n we have

$$\sum_{i=1}^{n} D_i \leq \frac{2R}{3}\log\frac{1}{p} + \sqrt{2\sum_{i=1}^{n}\sigma_i^2 \log\frac{1}{q}}.$$

The fourth is vector-valued martingale difference sequence concentration inequality (Zhang & Cutkosky, 2022, Theorem 19).

Lemma D.4. *Suppose that $\{X_t, \mathcal{F}_t\}$ is a vector-valued martingale difference sequence such that $\mathbb{E}[\|X_t\|^2|\mathcal{F}_{t-1}] \leq \sigma_t^2$ and $\|X_t\| \leq b_t$ almost everywhere for some sequence $\{\sigma_t, b_t\}$ such that σ_t, b_t is $\mathcal{F}_{t-1}$-measurable. Let $\nu \geq 0$ be an arbitrary constant. Then with probability at least $1 - q$, for all t we have*

$$\begin{aligned}\left\|\sum_{i=1}^{t} X_i\right\| \leq & 5\sqrt{\sum_{i=1}^{t}\sigma_i^2 \log\left(\frac{16}{q}\left[\log\left(\left[\sqrt{\sum_{i=1}^{t}\sigma_i^2/\nu^2}\right]_1\right) + 2\right]^2\right)} \\ & + 23\max\left(\nu, \max_{i\leq t} b_i\right)\log\left(\frac{224}{q}\left[\log\left(\frac{2\max\left(\nu, \max_{i\leq t} b_i\right)}{\nu}\right) + 2\right]^2\right)\end{aligned}$$

where $[x]_1 = \max(1, x)$.