Revisiting Gradient Normalization and Clipping for Nonconvex SGD under Heavy-Tailed Noise: Necessity, Sufficiency, and Acceleration

Tao Sun   SUNTAO.SALTFISH@OUTLOOK.COM
College of Computer Science and Technology
National University of Defense Technology
Changsha, Hunan, China

Xinwang Liu[*]   XINWANGLIU@NUDT.EDU.CN
College of Computer Science and Technology
National University of Defense Technology
Changsha, Hunan, China

Kun Yuan[*]   KUNYUAN@PKU.EDU.CN
Center for Machine Learning Research (CMLR)
Peking University
Beijing, China

Editor: Quanquan Gu

Abstract

Gradient clipping has long been considered essential for ensuring the convergence of Stochastic Gradient Descent (SGD) in the presence of heavy-tailed gradient noise. In this paper, we revisit this belief and explore whether gradient normalization can serve as an effective alternative or complement. We prove that, under individual smoothness assumptions, gradient normalization alone is sufficient to guarantee convergence of the nonconvex SGD. Moreover, when combined with clipping, it yields far better rates of convergence under more challenging noise distributions. We provide a unifying theory describing normalization-only, clipping-only, and combined approaches. Moving forward, we investigate existing variance-reduced algorithms, establishing that, in such a setting, normalization alone is sufficient for convergence. Finally, we present an accelerated variant that under second-order smoothness improves convergence. Our results provide theoretical insights and practical guidance for using normalization and clipping in nonconvex optimization with heavy-tailed noise.

Keywords: Gradient Normalization, Gradient Clipping, Heavy-tailed Noise, Nonconvex SGD, Nonconvex Variance Reduction

1. Introduction

Stochastic Gradient Descent (SGD) (Robbins and Monro, 1951), a widely used method, serves as the primary optimization algorithm for addressing fundamental optimization problems in machine learning and statistics

$$\min_{\boldsymbol{w} \in \mathbb{R}^d} f(\boldsymbol{w}) := \mathbb{E}_{\xi \sim \mathcal{D}} f(\boldsymbol{w}; \xi), \tag{1}$$

---

[*]. Corresponding authors.

where $\mathcal{D}$ denotes the probability distribution over the statistical sample space Ξ. The convergence properties of SGD have been a significant area of research in machine learning, with many previous studies making substantial contributions to this topic (Ghadimi and Lan, 2013; Cutkosky and Orabona, 2019; Arjevani et al., 2023; Johnson and Zhang, 2013). Early analyses often assumed that the noise caused by the stochastic gradient, denoted as $\boldsymbol{g}^t$, in each iteration could be uniformly variance-bounded, i.e., $\mathbb{E}\|\boldsymbol{g}^t - \nabla f(\boldsymbol{w}^t)\|^2 \leq \sigma^2$ for some $\sigma > 0$ and all $t \in \mathbb{Z}^+$. However, recent studies (Zhang et al., 2020; Nguyen et al., 2019; Simsekli et al., 2019) have shown that this assumption of uniformly bounded variance is unrealistic for training neural networks, particularly in the context of language models (Kunstner et al., 2024). To address this, the concept of heavy-tailed noise is mathematically defined and utilized to analyze the convergence of SGD (Zhang et al., 2020). Below, we revisit the assumption related to heavy-tailed noise[1].

Assumption 1 *There exists a constant $\sigma > 0$ and $p \in (1, 2]$ such that*

$$\sup_{\boldsymbol{w}\in\mathbb{R}^d} \{\mathbb{E}_{\xi\sim\mathcal{D}}\|\nabla f(\boldsymbol{w};\xi) - \nabla f(\boldsymbol{w})\|^p\} \leq \sigma^p.$$

This assumption is significantly weaker than the standard bounded variance assumption. Specifically, when $p = 2$, this assumption reduces to the standard bounded variance assumption. At first glance, based solely on its mathematical form, Assumption 1 might seem to be merely a minor relaxation of the standard bounded variance assumption, leading one to believe that it would have minimal impact on the analysis and performance of SGD. However, this assumption introduces a significant shift in the convergence properties of SGD. Specifically, in (Zhang et al., 2020), the authors provide an example illustrating that when $1 < p < 2$, SGD fails to converge, highlighting the critical implications of this seemingly modest change.

1.1 Main Approaches to Handling Heavy-Tailed Noise

To ensure the convergence of SGD in heavy-tailed noise, (Zhang et al., 2020) introduced a clipping technique, which they proved to be effective for handling such noise. The resulting method, SGDC, effectively reduces the impact of extreme gradient values, thereby stabilizing the training process. The scheme of SGD with clipping is presented in Algorithm 1, where the clipping is defined as $\mathrm{Clip}_h(\boldsymbol{w}) := \min\left\{1, \frac{h}{\|\boldsymbol{w}\|}\right\}\boldsymbol{w}$ with $\boldsymbol{w} \in \mathbb{R}^d$.

Algorithm 1 SGD with Clipping (SGDC) (Zhang et al., 2020)

Require: parameters $\gamma > 0$, $h > 0$
    Initialization: $\boldsymbol{w}^0$
    for $t = 0, 1, 2, \ldots$
      step 1: Samples unbiased stochastic gradient $\boldsymbol{g}^t$
      step 2: Updates $\boldsymbol{w}^{t+1} = \boldsymbol{w}^t - \gamma\mathrm{Clip}_h(\boldsymbol{g}^t)$
    end for
---

1. Indeed, there is an alternative definition of heavy-tailed noise. In (Gorbunov et al., 2020), heavy-tailed noise is described as satisfying the uniformly bounded variance condition but not exhibiting sub-Gaussian behavior. This definition is, in fact, stricter than the assumption of bounded variance noise and even more so compared to Assumption 1. Our paper refers to the heavy-tailed noise as Assumption 1.

Building on SGDC, (Cutkosky and Mehta, 2021) integrated gradient normalization with clipping, further enhancing the robustness of SGD against heavy-tailed noise in stochastic optimization. The details of this combined approach are presented in Algorithm 2.

Algorithm 2 Normalized SGD with Clipping (NSGDC) (Cutkosky and Mehta, 2021)

Require: parameters $\gamma > 0$, $0 \leq \theta < 1$, $h > 0$

    Initialization: $\boldsymbol{w}^0 = \boldsymbol{w}^1$, $\boldsymbol{m}^0 = \boldsymbol{0}$

    for $t = 1, 2, \ldots$

      step 1: Samples unbiased stochastic gradient $\boldsymbol{g}^t$

      step 2: Computes $\boldsymbol{m}^t = \theta \boldsymbol{m}^{t-1} + (1 - \theta)\text{Clip}_h(\boldsymbol{g}^t)$

      step 3: Updates $\boldsymbol{w}^{t+1} = \boldsymbol{w}^t - \gamma \frac{\boldsymbol{m}^t}{\|\boldsymbol{m}^t\|}$

    end for

It has been proven that NSGDC converges under a very general smoothness assumption, i.e., the global gradient ∇f is Lipschitz. Additionally, when the gradient of the individual function $f(\cdot; \xi)$ is Lipschitz, a variance-reduced variant of NSGDC is established as discussed by Liu et al. (2023), which is presented by Algorithm 3.

Algorithm 3 Variance-Reduced NSGDC (NSGCD-VR) (Liu et al., 2023)

Require: parameters $\gamma > 0$, $0 \leq \theta < 1$, $h > 0$

    Initialization: $\boldsymbol{w}^0 = \boldsymbol{w}^1$, $\boldsymbol{m}^0 = \boldsymbol{0}$

    for $t = 1, 2, \ldots$

      step 1: Sample the data $\xi^t \sim \mathcal{D}$ and $\boldsymbol{g}^t = \nabla f(\boldsymbol{w}^t; \xi^t)$

      step 2: $\boldsymbol{m}^t = \theta \boldsymbol{m}^{t-1} + (1 - \theta)\text{Clip}_h(\boldsymbol{g}^t) + \theta \nabla f(\boldsymbol{w}^t; \xi^t) - \theta \nabla f(\boldsymbol{w}^{t-1}; \xi^t)$

      step 3: $\boldsymbol{w}^{t+1} = \boldsymbol{w}^t - \gamma \frac{\boldsymbol{m}^t}{\|\boldsymbol{m}^t\|}$

    end for

1.2 Fundamental Open Questions and Main Results

As discussed in Section 1.1, current methods for addressing heavy-tailed noise predominantly employ either gradient clipping or a combination of gradient normalization and clipping. The ubiquitous utilization of gradient clipping in these approaches raises the following fundamental question:

Q1. *Is gradient clipping genuinely indispensable for handling heavy-tailed noise? Can gradient normalization alone ensure SGD convergence with heavy-tailed noise?*

This paper investigates the convergence properties of gradient normalization, both with and without variance reduction, in the presence of heavy-tailed noise. Our analysis demonstrates that gradient normalization alone suffices to guarantee SGD convergence under heavy-tailed noise, thereby addressing Q1. However, this result raises a further question: given that either gradient normalization or gradient clipping alone can ensure SGD convergence under heavy-tailed noise, is it necessary to combine these techniques in algorithmic development? Specifically,

Q2. *Does combining gradient normalization with clipping offer additional advantages over using either technique alone?*

This paper provides the affirmative answer to this question without any mini-batch size requirement. We demonstrate that combining gradient normalization with clipping (NS-GDC, Algorithm 2) substantially improves upon the convergence properties of SGD with gradient clipping alone (SGDC, Algorithm 1) (Zhang et al., 2020) under heavy-tailed noise. Our analysis reveals two key advantages. First, NSGDC eliminates the logarithmic factors that affect the convergence rate in SGDC (Zhang et al., 2020), achieving an optimal convergence rate that matches the theoretical lower bound. Second, we establish that NSGDC significantly outperforms SGDC, with the performance gap growing arbitrarily large as the gradient noise σ^2 approaches zero. Moreover, our analysis proves tight, as it successfully recovers the convergence rate in the deterministic setting where $\sigma^2 = 0$. We prove that NSGDC achieves convergence rates comparable to those of SGD with gradient normalization alone, while relying on global Lipschitz assumptions. Notably, both methods enjoy convergence guarantees without any requirement on the mini-batch size. This property is particularly appealing from a generalization perspective, as large mini-batches have been shown to negatively affect generalization performance (Keskar et al., 2017). These findings together comprehensively address Q2. All these convergence as mentioned above results are summarized in Table 1.

The final open question pertains to momentum acceleration in SGD. While extensive research on accelerated SGD with bounded gradient noise exists, studies on accelerated SGD under heavy-tailed noise remain limited. Given this context, a natural question arises:

Q3. *Can NSGDC achieve accelerated convergence in the presence of heavy-tailed noise?*

This paper introduces an accelerated variant of NSGDC that employs two momentum terms. Under a second-order Lipschitz condition, we prove that this new algorithm converges significantly faster than the standard NSGDC, thus addressing Q3.

Clipping: necessary or beneficial? The above questions aim to clarify a fundamental issue in optimization with heavy-tailed noise: is gradient clipping truly necessary for convergence, or merely helpful for acceleration? This work provides a unified view. We do not argue that clipping should always be removed; rather, we identify the regimes in which it is unnecessary, and others where it provably improves convergence rates. Our results span multiple algorithmic variants – both with and without clipping – and aim to characterize its theoretical necessity and utility in a principled manner.

1.3 Contributions

Our contributions can be summarized as follows.

- We investigate the convergence of gradient normalization in the presence of heavy-tailed noise. Our analysis shows that gradient normalization alone ensures SGD convergence in this setting, addressing Q1.

- We present a refined convergence analysis for NSGDC and NSGDC-VR, showing that combining gradient normalization with gradient clipping offers additional advantages beyond employing either technique alone, addressing Q2. Previously, the best-known convergence rates for NSGDC and NSGDC-VR from (Liu et al., 2023) matched those

Table 1: A comprehensive comparison between our results and existing convergence rates for SGD under heavy-tailed noise, including (Liu et al., 2023; Cutkosky and Mehta, 2021; Nguyen et al., 2023; Zhang et al., 2020). "BG" indicates the assumption of bounded gradient with constant $G > 0$ (G is always much larger than σ), "GL" means the global gradient Lipschitz, and "IL" is the individual Lipschitz (Assumption 2). Our convergence rate is characterized by $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{w}^t)\|$, the convergence rate in (Zhang et al., 2020) is characterized by $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{w}^t)\|^2$, while others are characterized by $\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(\boldsymbol{w}^t)\|$ with probability $1 - \delta$. Our rate significantly outperforms those of (Cutkosky and Mehta, 2021; Liu et al., 2023) when $\sigma \neq 0$. Additionally, for very small values of σ, our rates show substantial improvement over (Liu et al., 2023; Cutkosky and Mehta, 2021; Nguyen et al., 2023; Zhang et al., 2020), particularly for $1 \leq p < 2$, with the advantage being most pronounced as p approaches 1. Our convergence rates are established in expectation, leading to improvements over previous high-probability results, which exhibit a gap in the dependence on σ even when $p = 2$.

Algorithms	Papers	Rate	Assumptions
SGDC	Zhang et al. (2020)	$\mathcal{O}\left(\frac{1}{T^{\frac{p-1}{3p-2}}} + \frac{1}{T^{\frac{2p-p^2}{3p-2}}\,\sigma^{\frac{2p^2}{3p-2}}}\right)$	GL
SGDC	Nguyen et al. (2023)	$\mathcal{O}\left(\frac{\sigma\ln\frac{1}{\delta}}{T^{\frac{p-1}{3p-2}}} + \frac{\ln\frac{1}{\delta}}{T^{\frac{p-1}{3p-2}}}\right)$	GL
NSGD	Theorem 2	$\mathcal{O}\left(\frac{\sigma^{\frac{2p-2}{3p-2}}}{T^{\frac{p-1}{3p-2}}} + \frac{1}{T^{1/2}}\right)$	IL
NSGDC	Cutkosky and Mehta (2021)	$\mathcal{O}\left(\frac{G\ln\frac{T}{\delta}}{T^{\frac{p-1}{3p-2}}}\right)$	BG, GL
NSGDC	Liu et al. (2023)	$\mathcal{O}\left(\max\left\{\frac{\sigma\ln\frac{T}{\delta}}{T^{\frac{p-1}{3p-2}}}, \frac{1}{T^{\frac{p-1}{3p-2}}}\right\}\right)$	GL
NSGDC	Theorem 3	$\mathcal{O}\left(\frac{\sigma^{\frac{p}{3p-2}}}{T^{\frac{p-1}{3p-2}}} + \frac{1}{T^{1/2}}\right)$	GL

of SGDC (Zhang et al., 2020), failing to justify any theoretical benefits of combining normalization and clipping.

- We propose an enhanced NSGDC variant incorporating two momentum terms, achieving substantially faster convergence rates than the standard NSGDC algorithm, addressing Q3.

1.4 More Related Works

SGD under Heavy-tailed Noise. In (Zhang et al., 2020), the authors first establish the convergence of SGD with gradient clipping under heavy-tailed noise (Assumption 1). They demonstrate that the convergence rate, in expectation, for $\frac{\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{w}^t)\|}{T}$ is $\mathcal{O}(T^{-\frac{p-1}{3p-2}})$ in the nonconvex case. This rate matches the lower bound for nonconvex stochastic optimization, also proven in the same work for cases where the global function $\nabla f(\cdot)$ is Lipschitz. (Sadiev et al., 2023) investigates the application of stochastic algorithms to solve variational inequalities under heavy-tailed noise, a generalization that extends the applicability

of these methods beyond traditional stochastic optimization problems. In (Cutkosky and Mehta, 2021), the authors prove a high-probability convergence rate of $\mathcal{O}\left(\frac{\ln T}{T^{\frac{p-1}{3p-2}}}\right)$ when using a combination of gradient normalization and clipping, with the logarithmic factor being eliminated in (Liu et al., 2023). Further improvements are presented in (Puchkin et al., 2024), where the authors assume a specific structure for the heavy-tailed noise and achieve enhanced results. Additionally, paper (Vural et al., 2022) examines the convergence rate of the Stochastic Mirror Descent algorithm in stochastic convex optimization with heavy-tailed noise, providing lower bounds that emphasize the algorithm's efficiency without the need for gradient clipping. In the latter work (Nguyen et al., 2023), the authors offer an improved convergence rate for SGD with clipping under heavy-tailed noise for convex cases. In (Sun et al., 2025), the authors prove that SGD with gradient normalization and momentum still converges for a class of unbounded variance noise even under delayed updates and weakened smoothness assumptions.

Nonconvex Variance Reduction. Variance reduction techniques have gained significant attention as effective strategies for improving the efficiency of stochastic optimization, particularly in convex finite-sum problems (Johnson and Zhang, 2013; Shalev-Shwartz and Zhang, 2013; Defazio et al., 2014). Building on this foundation, researchers have extended these methods to nonconvex smooth optimization. Among the most influential works, Zhou et al. (2020) proposed a nested variance-reduced stochastic gradient algorithm that introduces multi-level reference point construction to control gradient variance more tightly. This design achieves improved complexity guarantees for reaching both first- and second-order stationary points and provides a unified framework for analyzing variance reduction in nonconvex settings. In parallel, Fang et al. (2018) introduced the another variance reduction algorithm, which also targets variance reduction for nonconvex finite-sum problems and achieves near-optimal sample complexity. Complementing these algorithmic developments, Zhou and Gu (2019) established lower bounds for smooth nonconvex finite-sum optimization, showing that many first-order variance-reduction methods are close to optimal up to logarithmic factors. These results provide both a theoretical benchmark and practical guidance for future algorithm design. Subsequent works (Zhou et al., 2018b; Cutkosky and Orabona, 2019; Tran-Dinh et al., 2019; Zhou and Gu, 2020; Li et al., 2021) have explored simpler and more robust variants, extending the applicability of variance reduction to broader settings. More recently, Liu et al. (2023) extended variance-reduction techniques to optimization under heavy-tailed noise, addressing the limitations of standard assumptions in stochastic gradient methods.

SGD under Additional Second-Order Smoothness. In stochastic settings, nonconvex accelerated algorithms that rely on both first- and second-order smoothness typically achieve a convergence rate of $\mathcal{O}(T^{-\frac{2}{7}})$ (Tripuraneni et al., 2018; Allen-Zhu, 2018; Fang et al., 2019; Zhou et al., 2018a; Sun et al., 2023a; Cutkosky and Mehta, 2020b). When compared to nonconvex variance reduction algorithms, nonconvex accelerated algorithms follow a different path in terms of their smoothness requirements. While variance reduction methods primarily focus on reducing stochastic noise in gradients to improve convergence, nonconvex accelerated algorithms take advantage of higher-order smoothness. Specifically, these algorithms do not demand increased first-order smoothness, such as the Lipschitz continu-

ity of the stochastic gradient. Instead, they rely more heavily on second-order smoothness properties, i.e., the Lipschitz continuity of the Hessian, to ensure acceleration.

Two concurrent works. We note that two concurrent studies independently investigate gradient normalization alone under heavy-tailed noise (Hübler et al., 2024; Liu and Zhou, 2024), both of which appeared on arXiv around the same time as our paper[2]. In (Hübler et al., 2024), the authors show that gradient normalization with mini-batches guarantees the convergence of SGD under heavy-tailed noise, achieving the optimal lower bound on sample complexity. They also provide high-probability bounds for the convergence rate. In work (Liu and Zhou, 2024), the authors incorporate momentum into mini-batch gradient normalization for SGD under heavy-tailed noise and establish convergence rates in expectation. Both concurrent works operate under the GL assumption. This paper treats both regimes: our IL-based results focus on the effect of normalization-only and variance reduction algorithms under heavy-tailed noise, while our GL-based results (Theorem 3) combine normalization with clipping and compete head-on with Liu and Zhou (2024); Hübler et al. (2024).

Our work differs in several aspects from these two studies. First, we do not focus solely on gradient normalization for heavy-tailed noise; instead, our paper also considers several other algorithmic variants. Second, our method imposes no constraint on the mini-batch size, which naturally leads to a different proof technique. The minimal requirement on mini-batch size is particularly beneficial for generalization, as large mini-batches have been shown to degrade generalization performance (Keskar et al., 2017). Third, our findings highlight the benefit of jointly applying gradient normalization and clipping.

Notation: This paper uses $\mathbb{E}[\cdot]$ to denote the expectation with respect to the underlying probability space. We use $\|\cdot\|$ to denote the L_2-norm of a vector and $\|\cdot\|_{\mathrm{op}}$ to denote the spectral norm of a matrix. The minimum value of the function f is denoted as $\min f$. Given two sequences $\{a_t\}$ and $\{b_t\}$, we write $a_t = \mathcal{O}(b_t)$ if there exists a positive constant $0 < C < +\infty$ such that $a_t \leq Cb_t$, and write $a_t = \Theta(b_t)$ if $a_t = \mathcal{O}(b_t)$ and $b_t = \mathcal{O}(a_t)$. Given an event $\mathcal{E}$, we define $1_{\mathcal{E}}$ as the indicator function that takes the value 1 if $\mathcal{E}$ happens and 0 otherwise.

2. Assumptions and Technical Novelty

2.1 Other Necessary Assumptions

Besides the heavy-tailed noise, we need other basic assumptions utilized for the analysis in our paper. The first assumption pertains to first-order smoothness and unbiased noise, commonly employed within the stochastic community.

Assumption 2' *There exists constant $L > 0$, for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, function $f(\cdot)$ obeys*

$$\|\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\| \leq L\|\boldsymbol{y} - \boldsymbol{x}\|.$$

We assume stochastic gradient $\boldsymbol{g}^t$ satisfies

$$\mathbb{E}\boldsymbol{g}^t = \nabla f(\boldsymbol{w}^t).$$

---

2. (Hübler et al., 2024) was posted on October 17, 2024; (Liu and Zhou, 2024) was posted on December 27, 2024; and the first version of our paper was posted on October 21, 2024.

The function value of f satisfies $\inf f > -\infty$.

Assumption 2' is indeed the global Lipschitz, which is quite general as it only requires the gradient of the objective function f to be Lipschitz. In many cases, a stronger assumption can be applied, where the gradient of the objective function $f(\cdot; \xi)$ is Lipschitz (referred to as "*individual Lipschitzness*" in this paper).

Assumption 2 *There exists constant $L > 0$, for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, function $f(\cdot; \xi)$ obeys*

$$\|\nabla f(\boldsymbol{y}; \xi) - \nabla f(\boldsymbol{x}; \xi)\| \le L\|\boldsymbol{y} - \boldsymbol{x}\|.$$

We assume stochastic gradient $\nabla f(\cdot; \xi)$ satisfies

$$\mathbb{E}_{\xi \sim \mathcal{D}}[\nabla f(\boldsymbol{w}; \xi)] = \nabla f(\boldsymbol{w}),$$

for any $\boldsymbol{w} \in \mathbb{R}^d$. In addition, we assume that the stochastic gradient is bounded at the initialization point w_0 as

$$B := \sup_{\xi \sim \mathcal{D}}\{\|\nabla f(\boldsymbol{w}^0; \xi)\|\} < +\infty, \ a.s. \tag{2}$$

for some $B > 0$. The function value of $f(\cdot; \xi)$ satisfies $\inf f(\boldsymbol{w}; \xi) > -\infty$.

Compared to Assumption 2', Assumption 2 employs an additional requirement with respect to the boundedness of the initial point $\boldsymbol{w}^0$, i.e., (2). We provide justification for the reasonability of this condition. The Lipschitz continuity of $\nabla f(\cdot; \xi)$ indicates

$$\|\nabla f(\boldsymbol{w}^0; \xi)\|^2 \le 2L[f(\boldsymbol{w}^0; \xi) - \min_{\boldsymbol{w}} f(\boldsymbol{w}; \xi)] \le 2L[f(\boldsymbol{w}^0; \xi) - \inf f(\boldsymbol{w}; \xi)].$$

If the measurement of the set

$$\boldsymbol{V} := \{\xi \sim \mathcal{D} \mid \|\nabla f(\boldsymbol{w}^0; \xi)\| = +\infty\}$$

is nonzero. Hence, we can get

$$+\infty = \int_{\xi \in \boldsymbol{V}} \|\nabla f(\boldsymbol{w}^0; \xi)\|^2 m(d\xi) \le \mathbb{E}\|\nabla f(\boldsymbol{w}^0; \xi)\|^2 \le 2L\mathbb{E}[f(\boldsymbol{w}^0; \xi) - \inf f(\boldsymbol{w}; \xi)],$$

indicating that initialization point $\boldsymbol{w}^0$ satisfies $f(\boldsymbol{w}^0) = \mathbb{E}[f(\boldsymbol{w}^0; \xi)] = +\infty$. Hence, the set $\boldsymbol{V}$ must have measure zero, and we use condition (2).

In mathematical formulations, the concept of individual Lipschitz continuity closely resembles that of global Lipschitz continuity. However, considering individual Lipschitzness, we can derive improved lower bounds for stochastic nonconvex optimization problems (Carmon et al., 2020). The improved lower bound suggests that faster algorithms can be developed for scenarios characterized by this type of smoothness. Indeed, the nonconvex variance reduction method has been specifically designed to take advantage of this smoothness (Cutkosky and Orabona, 2019; Liu et al., 2023).

2.2 A Concise Explanation of How We Achieve Improved Results

This section concisely explains why we can derive improved results for both NSGDC and NSGDC-VR. Previous studies have used complicated high-probability analysis methods, such as (Cutkosky and Mehta, 2021; Liu et al., 2023; Nguyen et al., 2023). The difficulty arises from the fact that it is challenging to efficiently bound $\mathbb{E}\|\text{Clip}_h(\boldsymbol{g}^t) - \nabla f(\boldsymbol{w}^t)\|^2$, which is a crucial term in these analyses. Indeed, the gradient normalization technique used in NSGDC and NSGDC-VR provides a useful bound as

$$\|\nabla f(\boldsymbol{w}^t)\| \leq \|\nabla f(\boldsymbol{w}^0)\| + L\gamma T.$$

This inequality allows us to set $h \geq 2(\|\nabla f(\boldsymbol{w}^0)\| + L\gamma T)$, ensuring that $\|\nabla f(\boldsymbol{w}^t)\| \leq \frac{h}{2}$ for all iterations. By leveraging this bound, we simplify the analysis significantly. According to [Lemma 5, (Liu et al., 2023)], we can establish the following bounds:

$$\mathbb{E}\|\text{Clip}_h(\boldsymbol{g}^t) - \mathbb{E}\text{Clip}_h(\boldsymbol{g}^t)\|^2 \leq 10\sigma^p h^{2-p}, \|\mathbb{E}\text{Clip}_h(\boldsymbol{g}^t) - \nabla f(\boldsymbol{w}^t)\|^2 \leq 4\sigma^{2p}h^{-2(p-1)}.$$

We provide further details on these bounds in Lemma 7. Using these bounds, we can then present the improved results.

For NSGD and NSGD-VR, the analysis focuses on bounding $\mathbb{E}\|\boldsymbol{g}^t - \nabla f(\boldsymbol{w}^t)\|^2$. By leveraging individual Lipschitz smoothness and gradient normalization, we can establish the following inequality

$$\|\boldsymbol{g}^t - \nabla f(\boldsymbol{w}^t)\| \leq 2(B + L\gamma T),$$

which leads to the bound

$$\mathbb{E}\|\boldsymbol{g}^t - \nabla f(\boldsymbol{w}^t)\|^2 \leq [2(B + L\gamma T)]^{2-p}\mathbb{E}\|\boldsymbol{g}^t - \nabla f(\boldsymbol{w}^t)\|^p \leq [2(B + L\gamma T)]^{2-p}\sigma^p.$$

Further details of this bound are provided in Lemma 8. These bounds significantly simplify the proof process by reducing the complexity of the analysis.

3. Gradient Normalization Alone Guarantees Convergence of SGD under Heavy-Tailed Noise

We present the scheme of NSGD in Algorithm 4. Compared with Algorithm 2, Algorithm 4 can be regarded as setting $h = +\infty$ in Algorithm 2. This adjustment eliminates the clipping hyperparameter in Algorithm 2, leading to simpler analysis and more straightforward practical implementations.

Algorithm 4 Normalized Stochastic Gradient Descent (NSGD) (Cutkosky and Mehta, 2020a)

Require: parameters $\gamma > 0$, $0 \leq \theta < 1$
    Initialization: $\boldsymbol{w}^0 = \boldsymbol{w}^1$, $\boldsymbol{m}^0 = \boldsymbol{0}$
    for $t = 1, 2, \ldots$
    step 1: Sample the data $\xi^t \sim \mathcal{D}$ and $\boldsymbol{m}^t = \theta\boldsymbol{m}^{t-1} + (1-\theta)\nabla f(\boldsymbol{w}^t; \xi^t)$
    step 2: $\boldsymbol{w}^{t+1} = \boldsymbol{w}^t - \gamma\frac{\boldsymbol{m}^t}{\|\boldsymbol{m}^t\|}$
    end for

Similar to the NSGDC-VR, we can also introduce the variance reduction scheme of NSGD in Algorithm 5, which, like its counterpart, sets the hyperparameter h to $+\infty$ in Algorithm 3. This modification effectively simplifies the algorithm while maintaining its variance reduction capabilities.

Algorithm 5 Normalized Stochastic Gradient Descent with Variance Reduction (NSGD-VR)

Require: parameters $\gamma > 0$, $0 \le \theta < 1$
    Initialization: $\boldsymbol{w}^0 = \boldsymbol{w}^1$, $\boldsymbol{m}^0 = \boldsymbol{0}$
    for $t = 1, 2, \ldots$
      step 1: Sample the data $\xi^t \sim \mathcal{D}$ and $\boldsymbol{m}^t = \theta \boldsymbol{m}^{t-1} + \nabla f(\boldsymbol{w}^t; \xi^t) - \theta \nabla f(\boldsymbol{w}^{t-1}; \xi^t)$
      step 2: $\boldsymbol{w}^{t+1} = \boldsymbol{w}^t - \gamma \frac{\boldsymbol{m}^t}{\|\boldsymbol{m}^t\|}$
    end for

3.1 NSGD under Heavy-tailed Noise

We present the convergence rate of NSGD under the heavy-tailed noise in the following theorem.

Theorem 1 *Let $(\boldsymbol{w}^t)_{t \ge 0}$ be generated by Algorithm 4. Suppose Assumptions 1-2 hold. If $1 - \theta = \min\left\{\frac{\max\{(L\Delta)^{\frac{1}{2}}, 1\}}{\sigma^{\frac{4p-4}{3p-2}} T^{\frac{p}{3p-2}}}, 1\right\}$, $\gamma = \sqrt{\frac{\Delta}{L}} \frac{\sqrt{1-\theta}}{\sqrt{T}}$ with $\Delta := f(\boldsymbol{w}^0) - \min f$, then it holds that*

$$
\frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \|\nabla f(\boldsymbol{w}^t)\| \le \frac{4(L\Delta)^{\frac{1}{4}} \sigma^{\frac{2p-2}{3p-2}}}{T^{\frac{p-1}{3p-2}}} + \frac{2\|\nabla f(\boldsymbol{w}^0)\| \sigma^{\frac{4p-4}{3p-2}}}{T^{\frac{2p-2}{3p-2}}}
$$

$$
+ \frac{4(L\Delta)^{\frac{6-3p}{8}} \left(\frac{B}{L}+1\right)^{\frac{2-p}{2}} \sigma^{\frac{5p^2-12p+8}{6p-4}}}{T^{\frac{(p-2)^2}{2(3p-2)} + \frac{p-1}{3p-2}}} + \frac{4(L\Delta)^{\frac{2-p}{4}} \left(\frac{B}{L}+1\right)^{\frac{2-p}{2}} \sigma^{\frac{5p^2-12p+8}{6p-4}}}{T^{\frac{(p-2)^2}{2(3p-2)} + \frac{p-1}{3p-2}}}
$$

$$
+ \frac{4(L\Delta)^{\frac{2-p}{4}} \left(\frac{B}{L}+1\right)^{\frac{2-p}{2}} \sigma^{p/8}}{T^{\frac{1}{4}}} + \frac{4\sqrt{L\Delta}}{T^{\frac{1}{2}}} + \frac{2\|\nabla f(\boldsymbol{w}^0)\|}{T}.
$$

When $\sigma = 0$, the convergence rate presented above matches that of GD, whereas when $p = 2$, it reduces to the convergence rate of standard SGD as $\mathcal{O}\left(\frac{\sqrt{\sigma}}{T^{\frac{1}{4}}}\right)$. When $\sigma > 0$ is large, the right-hand side of the convergence bound can further be simplified as

$$
\mathcal{O}\left(\frac{(L\Delta)^{\frac{1}{4}} \sigma^{\frac{2p-2}{3p-2}}}{T^{\frac{p-1}{3p-2}}} + \frac{\sigma^{\frac{p}{8}}}{T^{\frac{1}{4}}} + \frac{1}{T^{1/2}}\right) = \mathcal{O}\left(\frac{(L\Delta)^{\frac{1}{4}} \sigma^{\frac{2p-2}{3p-2}}}{T^{\frac{p-1}{3p-2}}}\right).
$$

When $\sigma = 0$, the convergence rate presented above matches that of GD, whereas when $p = 2$, it reduces to the convergence rate of standard SGD.

In Theorem 1, we show that gradient clipping is unnecessary to guarantee convergence for heavy-tailed noise, which contrasts with the approach in (Cutkosky and Mehta, 2021; Liu et al., 2023) where gradient clipping is used alongside gradient normalization. Removing the

gradient clipping eliminates the need to tune the clipping hyperparameter, simplifying the algorithm. The convergence rate we achieve is as fast as that of NSGDC, as demonstrated in Theorem 3 presented in the next section. However, it is essential to note that Theorem 1 relies on a stronger Lipschitz assumption, specifically individual Lipschitzness rather than global Lipschitzness. This assumption is crucial because it allows us to bound $\|\nabla f(\boldsymbol{w}^t; \xi^t)\|$ without relying on gradient clipping in NSGD. Although individual Lipschitzness is considered a stronger condition than global Lipschitzness, it is highly relevant and natural in the context of learning. This is because the smoothness of the expected objective function typically originates from the loss function $f(\cdot; \xi)$ evaluated at each individual sample.

Indeed, the concurrent works on NSGD with heavy-tailed noise (Liu and Zhou, 2024; Hübler et al., 2024) both establish the convergence of NSGD under heavy-tailed noise and global Lipschitz continuity, achieving the same optimal sampling complexity as our method. Notably, Hübler et al. (2024) presented a stronger result by providing high-probability convergence guarantees, in contrast to our expectation-based analysis. However, Hübler et al. (2024) impose specific requirements on the mini-batch size used in the algorithm. In contrast, our convergence guarantees do not rely on any mini-batch assumption. This distinction is particularly relevant in practice, as large mini-batch sizes are known to often degrade generalization performance. Compared with (Liu and Zhou, 2024), our proven convergence rate demonstrates improved dependence on the noise level, particularly in terms of the power of σ, especially when σ is large and p is close to 1. Specifically, our result achieves a rate of $\mathcal{O}\left(\frac{\sigma^{\frac{2p-2}{3p-2}}}{T^{\frac{p-1}{3p-2}}}\right)$, whereas the convergence rate reported in (Liu and Zhou, 2024) is $\mathcal{O}\left(\frac{\sigma^{\frac{p}{3p-2}}}{T^{\frac{p-1}{3p-2}}}\right)$.

3.2 NSGD-VR under Heavy-tailed Noise

We present the convergence for the variance reduction scheme of NSGD under the heavy-tailed noise.

Theorem 2 *Let $(\boldsymbol{w}^t)_{t\geq 0}$ be generated by Algorithm 5. Suppose Assumptions 1-2 hold. If $1 - \theta = \min\left\{\frac{1}{\sigma^{\frac{p}{2p-1}} T^{\frac{p}{2p-1}}}, 1\right\}, \gamma = \frac{\sqrt[4]{1-\theta}}{L\sqrt{T}}$, then it holds that*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{w}^t)\| \leq \frac{2\|\nabla f(\boldsymbol{w}^0)\|\sigma^{\frac{p}{2p-1}} + 4\left(\frac{B}{L}+1\right)^{\frac{2-p}{2}}\sigma^{\frac{p}{2p-1}} + 4\left(\frac{B}{L}+1\right)^{\frac{2-p}{2}}\sigma^{\frac{p}{4(2p-1)}}}{T^{\frac{p-1}{2p-1}}}$$

$$+ \frac{[L(f(\boldsymbol{w}^0) - \min f) + 5]\sigma^{\frac{p}{4(2p-1)}}}{T^{\frac{p-1}{2p-1}}} + \frac{L(f(\boldsymbol{w}^0) - \min f) + 4\left(\frac{B}{L}+1\right)^{\frac{2-p}{2}} + 5}{\sqrt{T}} + \frac{2\|\nabla f(\boldsymbol{w}^0)\|}{T}.$$

When $\sigma > 0$ is large, we can get a simplified rate for NSGD-VR under heavy-tailed noise as

$$\mathcal{O}\left(\frac{\sigma^{\frac{p}{2p-1}}}{T^{\frac{p-1}{2p-1}}} + \frac{1}{T^{1/2}}\right).$$

When $p = 2$, i.e., in the traditional bounded variance case, the convergence rate of NSGD-VR for $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{w}^t)\|$ is $\mathcal{O}\left(\frac{1}{T^{\frac{1}{3}}}\right)$ as $\sigma \neq 0$, which matches the speed of STORM

Table 2: A comparison between Theorem 2 and existing convergence rates (Liu et al., 2023). Our convergence rate is characterized by $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{w}^t)\|$, whereas (Liu et al., 2023) adopts $\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(\boldsymbol{w}^t)\|$ with a confidence level of at least $1-\delta$.

Paper	Rate under $\sigma \neq 0$	Rate under $\sigma = 0$
Liu et al. (2023)	$\mathcal{O}\left(\dfrac{\sigma \ln T}{T^{\frac{p-1}{3p-2}}}\right)$	N/A
Our result	$\mathcal{O}\left(\dfrac{\sigma^{\frac{p}{2p-1}}}{T^{\frac{p-1}{3p-2}}}\right)$	$\mathcal{O}\left(\dfrac{1}{T^{1/2}}\right)$

(Cutkosky and Orabona, 2019) but without requiring the bounded gradient assumption. According to (Arjevani et al., 2023), the lower bound for nonconvex first-order stochastic methods in individually smooth optimization is also $\mathcal{O}\left(\frac{1}{T^{\frac{1}{3}}}\right)$. Hence, NSGD-VR achieves the lower bound when $p = 2$. The convergence rate of NSGD-VR is $\mathcal{O}\left(\frac{\sigma^{\frac{p}{2p-1}}}{T^{\frac{p-1}{2p-1}}}\right)$, which is better than that of (Liu et al., 2023) as $\mathcal{O}\left(\frac{\sigma \ln T}{T^{\frac{p-1}{2p-1}}}\right)$. Our results also demonstrate that gradient clipping is unnecessary for heavy-tailed noise to achieve nonconvex variance reduction. Furthermore, compared to (Liu et al., 2023), our findings fix the case when $\sigma = 0$, which is not addressed with clipping and normalization. When $\sigma = 0$, the rate is $\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$, matching the speed of gradient descent for nonconvex optimization. The comparison is described by Table 2.

Indeed, as shown in Theorem 2, the convergence rate of NSGD-VR is in the same order as that of NSGDC-VR, as established in Theorem 4 presented in the next section. Both algorithms achieve this rate under the same assumptions, indicating that the clipping procedure is entirely redundant for variance reduction under heavy-tailed noise.

4. Improved Results of Gradient Normalization with Clipping under Heavy-tailed Noise

This section presents improved convergence rates for the gradient normalization with clipping methods. Building on the same assumptions in the most recent paper (Liu et al., 2023), we demonstrate that our approaches achieve superior convergence rates.

4.1 Improved Convergence Rate of NSGDC

The improved convergence rate of NSGDC is presented by Theorem 3.

Theorem 3 *Let $(\boldsymbol{w}^t)_{t\geq 0}$ be generated by Algorithm 2. Suppose Assumptions 1-2' hold. If* $1 - \theta = \min\left\{\frac{\max\left\{(L\Delta)^{\frac{p}{3p-2}},1\right\}}{\max\left\{\sigma^{\frac{4p-4}{3p-2}},\sigma^{\frac{2p}{3p-2}}\right\}T^{\frac{p}{3p-2}}},1\right\}$, $\gamma = (1_{\{L\Delta\geq 1\}}\sqrt{\frac{\Delta}{L}} + 1_{\{L\Delta<1\}}\frac{1}{L})\frac{\sqrt{1-\theta}}{\sqrt{T}}$ *with* $\Delta :=$ $f(\boldsymbol{w}^0) - \min f$, $h = 2(\gamma L T^{\frac{2p}{3p-2}} \cdot \max\{1, \sigma^{\frac{3p}{3p-2}} \cdot 1_{\{\sigma^2 T \geq L\Delta\}}\} + \|\nabla f(\boldsymbol{w}^0)\|)$, *then it holds that*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{w}^t)\| \leq \frac{4(L\Delta)^{\frac{p-1}{3p-2}}\sigma^{\frac{p}{3p-2}} + 23\sigma^{\frac{p}{3p-2}} + 3\sigma^{\frac{2p-2}{3p-2}} + 16\sigma^{\frac{p-1}{3p-2}} + 16\sigma^{\frac{p}{8}}}{T^{\frac{p-1}{3p-2}}}$$

$$+ \frac{4(L\Delta)^{\frac{p-1}{3p-2}}\sigma^{\frac{2p-2}{3p-2}} + 16(L\Delta)^{\frac{p}{3p-2}}\sigma^{\frac{p-1}{3p-2}} + 16(L\Delta)^{\frac{6p-2p^2-2}{6p-4}}\sigma^{\frac{p^2}{6p-4}} + 16(L\Delta)^{\frac{1}{4}}\sigma^{\frac{p}{8}}}{T^{\frac{p-1}{3p-2}}}$$

$$+ \frac{16[(L\Delta)^{\frac{p}{6p-4}} + 1]\cdot\|\nabla f(\boldsymbol{w}^0)\|^{\frac{2-p}{2}}\sigma^{\frac{3p^2-6p+4}{2(3p-2)}}}{T^{\frac{p}{6p-4}}} + \frac{2\|\nabla f(\boldsymbol{w}^0)\|\sigma^{\frac{4p-4}{3p-2}} + 2\|\nabla f(\boldsymbol{w}^0)\|\sigma^{\frac{2p}{3p-2}}}{T^{\frac{2p-2}{3p-2}}}$$

$$+ \frac{4\sqrt{L\Delta} + 16\|\nabla f(\boldsymbol{w}^0)\|^{\frac{2-p}{2}} + 7}{T^{\frac{1}{2}}} + \frac{2\|\nabla f(\boldsymbol{w}^0)\|}{T}.$$

When $\sigma = 0$, the convergence rate described above aligns with that of GD. In contrast, when $p = 2$, it simplifies to the convergence rate of standard SGD as $\mathcal{O}\left(\frac{\sqrt{\sigma}}{T^{\frac{1}{4}}}\right)$. As σ is large, we can get a simplified rate for NSGDC under the heavy-tailed noise as

$$\mathcal{O}\left(\frac{(L\Delta)^{\frac{p-1}{3p-2}}\sigma^{\frac{p}{3p-2}}}{T^{\frac{p-1}{3p-2}}} + \frac{\sigma^{\frac{2p}{3p-2}}}{T^{\frac{2p-2}{3p-2}}} + \frac{1}{T^{1/2}}\right) = \mathcal{O}\left(\frac{(L\Delta)^{\frac{p-1}{3p-2}}\sigma^{\frac{p}{3p-2}}}{T^{\frac{p-1}{3p-2}}}\right).$$

The presented bound exhibits the same dependence on $(L\Delta)^{\frac{p-1}{3p-2}}$ as established in (Liu and Zhou, 2024). We also revise Theorem 1 to align the dependence on $L\Delta$.

We present a discussion of our proven results compared to several closely related works (Cutkosky and Mehta, 2021; Liu et al., 2023; Zhang et al., 2020; Nguyen et al., 2023).

First, we establish convergence rates under either weaker or comparable assumptions. Specifically, when $\sigma \neq 0$, our dominant rate achieves $\mathcal{O}\left(\frac{1}{T^{\frac{p-1}{3p-2}}}\right)$ in expectation, without relying on the uniformly bounded gradient assumption. In comparison, (Cutkosky and Mehta, 2021) and (Liu et al., 2023) obtain high-probability convergence rates of the form $\mathcal{O}\left(\frac{\ln\frac{T}{\delta}}{T^{\frac{p-1}{3p-2}}}\right)$ with probability at least $1 - \delta$ for a small $\delta > 0$. Similarly, (Nguyen et al., 2023) establishes high-probability guarantees under heavy-tailed noise, incurring additional logarithmic dependence on δ. Note that high-probability bounds are generally stronger than expectation bounds, and the additional logarithmic factors are standard consequences of union bound arguments across iterations. In contrast, our result is characterized in expectation, and thus naturally avoids logarithmic terms such as $\ln T$, providing a clean dependence on T without affecting the probabilistic level of guarantee. Accordingly, our expectation-based convergence result and existing high-probability results address different aspects of algorithmic behavior, and should be interpreted within their respective probabilistic frameworks.

Second, compared to all four works mentioned above (Liu et al., 2023; Cutkosky and Mehta, 2021; Nguyen et al., 2023; Zhang et al., 2020), our results explicitly characterize the subordinate rate. Note that when $\sigma = 0$ (i.e., the deterministic case for any $p \in (1, 2]$), the rate is $\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$ for $\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(\boldsymbol{w}^t)\|$, matching the convergence rate of gradient descent.

In contrast, none of the previous works recover the deterministic results when $\sigma = 0$. We list the detailed comparison in Table 1. The result also explains the necessity of using gradient normalization in clipping SGD when σ is tiny.

Third, when the noise level σ is large, we derive a convergence rate of $\mathcal{O}\left(\frac{\sigma^{\frac{p}{3p-2}}}{T^{\frac{p-1}{3p-2}}}\right)$, which improves upon the $\widetilde{\mathcal{O}}\left(\frac{\sigma}{T^{\frac{p-1}{3p-2}}}\right)$ result in high-probability form shown in (Liu et al., 2023), since the exponent $\frac{p}{3p-2} < 1$ for all $p \in (1, 2)$. This demonstrates a substantial gain in noise robustness offered by our approach. Indeed, the rate matches the lower bound under heavy-tailed noise with respect to both σ and $1/T$ (Liu and Zhou, 2024).

Compared with concurrent work on NSGD (Hübler et al., 2024; Liu and Zhou, 2024), our analysis of NSGDC underscores the benefits of combining gradient clipping with normalization. Notably, relative to Hübler et al. (2024), NSGDC eliminates the need for the mini-batch assumption and achieves improved dependence on the noise level σ in terms of gradient computation cost. Furthermore, while Liu and Zhou (2024) establish a convergence rate of $\mathcal{O}\left(\frac{\sigma^{\frac{p}{3p-2}}}{T^{\frac{p-1}{3p-2}}} + \frac{\sigma}{T^{\frac{p-1}{2p-1}}}\right)$ for NSGD, our results show that NSGDC achieves $\mathcal{O}\left(\frac{\sigma^{\frac{p}{3p-2}}}{T^{\frac{p-1}{3p-2}}} + \frac{\sigma^{\frac{2p}{3p-2}}}{T^{\frac{2p-2}{3p-2}}}\right)$, offering a more favorable subordinate term.

4.2 Improved Convergence Rate of NSGDC-VR

We also present the improved convergence result for NSGDC-VR in Theorem 4.

Theorem 4 *Let $(\boldsymbol{w}^t)_{t\geq 0}$ be generated by Algorithm 3. Suppose Assumptions 1-2 hold. If $1 - \theta = \min\left\{\frac{1}{\max\left\{\sigma^{\frac{p}{2p-1}}, \sigma^{\frac{2p-2}{2p-1}}\right\}T^{\frac{p}{2p-1}}}, 1\right\}$, $\gamma = \frac{\sqrt[4]{1-\theta}}{L\sqrt{T}}$, $h = 2(\gamma L T^{\frac{5p+2}{8p-4}} \cdot \max\{1, \sigma^{\frac{9p}{8p-4}}\} + \|\nabla f(\boldsymbol{w}^0)\|)$, then it holds that*

$$
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{w}^t)\| \leq \frac{20\sigma^{\frac{p}{2p-1}} + 2\|\nabla f(\boldsymbol{w}^0)\|(\sigma^{\frac{p}{2p-1}} + \sigma^{\frac{2p-2}{2p-1}}) + \sigma^{\frac{p}{16}}}{T^{\frac{p-1}{2p-1}}}
$$
$$
+ \frac{[L(f(\boldsymbol{w}^0) - \min f) + 5](\sigma^{\frac{p-1}{4p-2}} + \sigma^{\frac{p}{8p-4}})}{T^{\frac{3p-2}{8p-4}}} + \frac{16\|\nabla f(\boldsymbol{w}^0)\|^{\frac{2-p}{2}}(\sigma^{\frac{p^2-p}{2p-1}} + \sigma^{\frac{2p^2-5p+2}{4p-2}})}{T^{\frac{p}{4p-2}}}
$$
$$
+ \frac{16\sigma^{\frac{5p^2-9p+6}{8p-4}}}{T^{\frac{5p-1}{8p-4}}} + \frac{L(f(\boldsymbol{w}^0) - \min f) + 16\|\nabla f(\boldsymbol{w}^0)\|^{\frac{2-p}{2}} + 9}{T^{\frac{1}{2}}} + \frac{2\|\nabla f(\boldsymbol{w}^0)\|}{T}.
$$

When $\sigma = 0$, the convergence rate proved above corresponds to that of GD. On the other hand, if $p = 2$, it simplifies to the convergence rate of standard SGD, expressed as $\mathcal{O}\left(\frac{\sqrt{\sigma}}{T^{\frac{1}{4}}}\right)$. When σ becomes large, we derive a simplified rate for NSGDC-VR as

$$
\mathcal{O}\left(\frac{\sigma^{\frac{p}{2p-1}}}{T^{\frac{p-1}{2p-1}}} + \frac{1}{T^{1/2}}\right).
$$

The rate in Theorem 4 is on par with that of Theorem 2. We stress the "improved result" refers to the comparison with (Liu et al., 2023). More precisely, unlike previous results

14

on variance reduction algorithms under heavy-tailed noise (Liu et al., 2023), our work successfully characterizes the subordinate term, which was not addressed in earlier studies. When $\sigma \neq 0$, our method achieves a convergence rate as fast as that established in (Liu et al., 2023) as $\mathcal{O}\Big(\frac{1}{T^{\frac{p-1}{2p-1}}}\Big)$. Moreover, when $\sigma = 0$, our result matches the convergence rate of gradient descent in the deterministic case, a scenario not covered in prior research. When σ is large, our rate $\mathcal{O}\Big(\frac{\sigma^{\frac{p}{2p-1}}}{T^{\frac{p-1}{2p-1}}}\Big)$ is better than that in (Liu et al., 2023), which is $\mathcal{O}\Big(\frac{\sigma}{T^{\frac{p-1}{2p-1}}}\Big)$, since $\frac{p}{2p-1} \leq 1$ as $p \in (1, 2]$.

5. Accelerations with Second-order Smoothness under Heavy-tailed Noise

This section explores how second-order smoothness enhances the convergence rate of SGD under heavy-tailed noise. We present an approach inspired by the accelerated algorithm introduced in (Cutkosky and Mehta, 2020b).

Algorithm 6 Accelerated Normalized Stochastic Gradient Descent with Clipping under Second-Order Smoothness (A-NSGDC)

Require: parameters $\gamma > 0$, $0 \leq \theta < 1$, $\zeta \geq 0$
    Initialization: $\boldsymbol{w}^0 = \boldsymbol{w}^1$, $\boldsymbol{m}^0 = \boldsymbol{0}$
    for $t = 1, 2, \ldots$
    step 1 $\boldsymbol{v}^t = \boldsymbol{w}^t + \zeta(\boldsymbol{w}^t - \boldsymbol{w}^{t-1})$
    step 2: Sample unbiased stochastic gradient $\mathbb{E}\boldsymbol{g}^t = \nabla f(\boldsymbol{v}^t)$
    step 3: Updates $\boldsymbol{m}^t = \theta\boldsymbol{m}^{t-1} + (1 - \theta)\mathrm{Clip}_h(\boldsymbol{g}^t)$
    step 4: $\boldsymbol{w}^{t+1} = \boldsymbol{w}^t - \gamma\frac{\boldsymbol{m}^t}{\|\boldsymbol{m}^t\|}$
    end for

The acceleration of A-NSGDC is proved under the second-order Lipschitz continuity.

Assumption 3 *There exists $H > 0$ such that the Hessian matrix of function f satisfies $\|\nabla^2 f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{y})\|_{\mathrm{op}} \leq H\|\boldsymbol{x} - \boldsymbol{y}\|$ for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$.*

In the stochastic setting, the second-order Lipschitz continuity is commonly employed for analyzing accelerated nonconvex optimization, as demonstrated in (Tripuraneni et al., 2018; Allen-Zhu, 2018; Fang et al., 2019; Zhou et al., 2018a; Sun et al., 2023b; Cutkosky and Mehta, 2020b; Sun et al., 2023a).

In the following, we present the accelerated convergence rate for A-NSGDC under second-order smoothness with heavy-tailed noise.

Theorem 5 *Let $(\boldsymbol{w}^t)_{t \geq 0}$ be generated by Algorithm 6. Suppose Assumptions 1, 2', and 3 hold. If $1 - \theta = \min\Big\{\frac{1}{\max\big\{\sigma^{\frac{6}{7}}, \sigma^{\frac{3}{7}}\big\}T^{\frac{2p}{4p-1}}}, 1\Big\}$, $\gamma = \frac{(1-\theta)^{\frac{2}{3}}}{LT^{\frac{2p+3}{12p-3}}}\big(\frac{1_{\theta=0}}{T^{\frac{8p-9}{24p-6}}} + 1_{\theta\neq0}\big)$, $h = 2(\gamma LT^{\frac{2p+3}{4p-1}} \cdot$*

$\max\{1, \sigma^{\frac{13p-10}{7p-7}}\} + \|\nabla f(\boldsymbol{w}^0)\|)$, $\zeta = \frac{\theta}{1-\theta}$, *then it holds that*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{w}^t)\| \leq \frac{L(\sigma^{\frac{4}{7}}+\sigma^{\frac{2}{7}})(f(\boldsymbol{w}^1)-\min f) + 20(\sigma^{\frac{4}{7}}+\sigma^{\frac{2}{7}}) + 16\sigma^{\frac{p}{16}}}{T^{\frac{2p-2}{4p-1}}}$$

$$+ \frac{16\|\nabla f(\boldsymbol{w}^0)\|^{\frac{2-p}{2}}(\sigma^{\frac{4}{7}}+\sigma^{\frac{2}{7}})}{T^{\frac{p}{4p-1}}} + \frac{2\|\nabla f(\boldsymbol{w}^0)\|(\sigma^{\frac{6}{7}}+\sigma^{\frac{3}{7}})}{T^{\frac{2p-1}{4p-1}}} + \frac{H(\sigma^{\frac{4}{7}}+\sigma^{\frac{2}{7}})}{L^2 T^{\frac{2}{4p-1}}} + \frac{\sigma^{\frac{4}{7}}+\sigma^{\frac{1}{28}}}{2T^{\frac{2p+1}{4p-1}}}$$

$$+ \frac{L(f(\boldsymbol{w}^1)-\min f) + 5 + 16\|\nabla f(\boldsymbol{w}^0)\|^{\frac{2-p}{2}}}{T^{1/2}} + \frac{2\|\nabla f(\boldsymbol{w}^0)\|}{T}.$$

In the case where $\sigma = 0$, the convergence rate established above reduces to that of gradient descent (GD). Conversely, when $p = 2$, it recovers the standard SGD rate, given by $\mathcal{O}\left(\frac{\sqrt{\sigma}}{T^{1/4}}\right)$. Furthermore, as σ increases significantly, we can further simplify the rate as

$$\mathcal{O}\left(\frac{\sigma^{\frac{4}{7}}}{T^{\frac{2p-2}{4p-1}}} + \frac{1}{T^{1/2}}\right).$$

To our knowledge, Theorem 5 provides the first convergence result for an accelerated non-convex first-order method in the presence of heavy-tailed noise under second-order Lipschitzness. The theorem is proven under a broad smoothness assumption, specifically that the function f is Lipschitz continuous. When $\sigma \neq 0$, Theorem 5 demonstrates a convergence rate of $\mathcal{O}\left(\frac{1}{T^{\frac{2p-2}{4p-1}}}\right)$, which is faster than the rate of NSGDC. In particular, when $p = 2$, the rate is $\mathcal{O}\left(\frac{1}{T^{\frac{2}{7}}}\right)$, aligning with results from (Carmon et al., 2018; Agarwal et al., 2017; Tripuraneni et al., 2018; Allen-Zhu, 2018; Fang et al., 2019; Zhou et al., 2018a; Sun et al., 2023b; Cutkosky and Mehta, 2020b; Sun et al., 2023a). Furthermore, in the deterministic case where $\sigma = 0$, the convergence rate improves to $\mathcal{O}\left(\frac{1}{T^{\frac{1}{2}}}\right)$, matching that of gradient descent.

By setting $h = +\infty$ in A-NSGDC, we obtain the accelerated version of NSGD under second-order Lipschitz continuity. The same convergence rate as A-NSGDC can be derived following a similar proof strategy (see details in Appendix G). However, we do not further present the result of the accelerated NSGD under second-order Lipschitzness in the main text because NSGD is proved under individual Lipschitz continuity, where variance reduction techniques are more effective. The convergence rate of NSGD-VR is $\mathcal{O}\left(\frac{1}{T^{\frac{p-1}{2p-1}}}\right)$, which is faster than the $\mathcal{O}\left(\frac{1}{T^{\frac{2p-2}{4p-1}}}\right)$ rate achieved under second-order Lipschitzness. Therefore, applying the accelerated scheme is unnecessary, even in the presence of second-order Lipschitz continuity.

The structure of our proof for Theorem 5 is related to that of (Cutkosky and Mehta, 2021), but analyzing the algorithm in expectation provides a complementary perspective. In particular, the expectation-based framework allows for a simpler analysis and, more importantly, makes the role of the hyperparameters clearer. This transparency means that hyperparameter choices in the accelerated scheme are more flexible and interpretable, as opposed to high-probability analysis in (Cutkosky and Mehta, 2021), where the admissible

parameter range is quite constrained and technically involved. Therefore, Theorem 5 should be viewed as complementing the high-probability result of (Cutkosky and Mehta, 2021): our main contribution lies in demystifying parameter tuning and providing a simplified framework, whereas (Cutkosky and Mehta, 2021) provides the high-probability guarantees.

A natural question that arises is whether we can integrate the accelerated scheme with variance reduction techniques to further enhance the performance of algorithms under the assumption of second-order smoothness. This combination would ideally allow us to achieve faster convergence. Unfortunately, despite our efforts, we have not yet succeeded in developing such a combined scheme. The main technical challenge stems from a fundamental incompatibility between the accelerated scheme and the variance reduction process. The accelerated scheme approximates the stochastic gradient using a second-order Taylor expansion, which inherently alters the structure of the stochastic gradient. This alteration disrupts the conditions required for effective variance reduction, making it difficult to harmonize these two approaches within a single framework.

6. Concluding Remarks

This work presents significant theoretical advances for NSGDC and NSGDC-VR in addressing nonconvex optimization problems under heavy-tailed noise. Beyond proving improved convergence results, we showed that gradient normalization alone can ensure convergence under heavy-tailed noise without clipping, broadening the applicability of normalization techniques. By leveraging second-order smoothness, we also developed accelerated algorithms that enhance performance in nonconvex settings. These contributions provide new directions for optimizing stochastic algorithms and reinforce the utility of gradient normalization in complicated, noise-affected environments.

Our results can be extended to a broader class of gradient normalizations as defined by (Sun et al., 2023a), outlined below.

Definition 6 *We define $\mathbb{G}$ as a general normalization if, for any $\boldsymbol{x} \in \mathbb{R}^d$, it satisfies*

$$\langle \boldsymbol{x}, \mathbb{G}(\boldsymbol{x}) \rangle \geq l\|\boldsymbol{x}\|_\diamond, \ \|\mathbb{G}(\boldsymbol{x})\| \leq U,$$

where $l, U > 0$, and $\|\cdot\|_\diamond$ denotes a specific norm on $\mathbb{R}^d$.

From this definition, we observe that the sign operator and normalization are particular cases of general normalization. The results established in this paper can thus be directly extended to this broader setting by substituting $\frac{\boldsymbol{m}^t}{\|\boldsymbol{m}^t\|}$ with $\mathbb{G}(\boldsymbol{m}^t)$ in all relevant algorithms. The proofs remain nearly identical to that presented in the appendix.

Acknowledgements

Appendix A. Technical Lemmas

Lemma 7 *Let $(\boldsymbol{w}^t)_{t \geq 1}$ be generated by Algorithms 2, 3 and 6, if Assumptions 1 and 2' hold and $h \geq 2(\|\nabla f(\boldsymbol{w}^0)\| + L\gamma T)$, we have*

$$\begin{aligned}
\mathbb{E}\|\mathrm{Clip}_h(\boldsymbol{g}^t) - \mathbb{E}\mathrm{Clip}_h(\boldsymbol{g}^t)\|^2 &\leq 10h^{2-p}\sigma^p, \\
\|\mathbb{E}\mathrm{Clip}_h(\boldsymbol{g}^t) - \nabla f(\boldsymbol{w}^t)\| &\leq 2\sigma^p h^{-(p-1)}.
\end{aligned} \tag{3}$$

Lemma 8 *Let $(\boldsymbol{w}^t)_{t \geq 1}$ be generated by Algorithms 4, 5 and 7, if Assumptions 1 and 2 hold, it follows that*

$$\mathbb{E}_{\xi^t \sim \mathcal{D}}\|\nabla f(\boldsymbol{w}^t; \xi^t) - \nabla f(\boldsymbol{w}^t)\|^2 \leq 4(B + L\gamma T)^{2-p}\sigma^p.$$

Lemma 9 *(Cutkosky and Mehta, 2020b; Zhao et al., 2021) Let $\boldsymbol{w}^\dagger, \boldsymbol{m} \in \mathbb{R}^d$ be arbitrary vectors, and*

$$\boldsymbol{w}^\ddagger = \boldsymbol{w}^\dagger - \gamma\frac{\boldsymbol{m}}{\|\boldsymbol{m}\|}, \tag{4}$$

and $\boldsymbol{\epsilon} := \boldsymbol{m} - \nabla f(\boldsymbol{w}^\dagger)$. When the gradient is Lipschitz, we have

$$f(\boldsymbol{w}^\ddagger) - f(\boldsymbol{w}^\dagger) \leq -\gamma\|\nabla f(\boldsymbol{w}^\dagger)\| + 2\gamma\|\boldsymbol{m} - \nabla f(\boldsymbol{w}^\dagger)\| + \frac{L\gamma^2}{2}.$$

Appendix B. Proof of Theorem 1

We designate the following variables as

$$\boldsymbol{\epsilon}^t := \boldsymbol{m}^t - \nabla f(\boldsymbol{w}^t), \boldsymbol{\delta}^t := \nabla f(\boldsymbol{w}^t; \xi^t) - \nabla f(\boldsymbol{w}^t), \boldsymbol{s}^t := \nabla f(\boldsymbol{w}^{t-1}) - \nabla f(\boldsymbol{w}^t). \tag{5}$$

Based on the scheme of momentum updating in the NSGD,

$$\begin{aligned}
\boldsymbol{m}^t &= \theta\boldsymbol{m}^{t-1} + (1 - \theta)\nabla f(\boldsymbol{w}^t; \xi^t) \\
&= \theta\Big[\boldsymbol{\epsilon}^{t-1} + \nabla f(\boldsymbol{w}^{t-1})\Big] + (1 - \theta)\Big[\boldsymbol{\delta}^t + \nabla f(\boldsymbol{w}^t)\Big].
\end{aligned}$$

The relation above then gives us

$$\boldsymbol{\epsilon}^t = \boldsymbol{m}^t - \nabla f(\boldsymbol{w}^t) = \theta\boldsymbol{\epsilon}^{t-1} + \theta\boldsymbol{s}^t + (1 - \theta)\boldsymbol{\delta}^t.$$

By employing mathematical induction, we can proceed to derive

$$\boldsymbol{\epsilon}^t = \theta^t\boldsymbol{\epsilon}^0 + \sum_{k=1}^t \theta^{t-k}\boldsymbol{s}^k + (1 - \theta)\sum_{k=1}^t \theta^{t-k}\boldsymbol{\delta}^k.$$

By taking the norms of both sides of the inequality and considering the expectation,

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{\epsilon}^t\| &\leq \sum_{k=1}^t \theta^{t-k}\mathbb{E}\|\boldsymbol{s}^k\| + (1 - \theta)\mathbb{E}\Big\|\sum_{k=1}^t \theta^{t-k}\boldsymbol{\delta}^k\Big\| + \theta^t\|\boldsymbol{\epsilon}^0\| \\
&\leq \frac{L\gamma}{1 - \theta} + (1 - \theta)\mathbb{E}\Big\|\sum_{k=1}^t \theta^{t-k}\boldsymbol{\delta}^k\Big\| + \theta^t\|\boldsymbol{\epsilon}^0\|,
\end{aligned} \tag{6}$$

where we used the gradient Lipschitz property of function f as

$$\mathbb{E}\|\boldsymbol{s}^k\| = \mathbb{E}\|\nabla f(\boldsymbol{w}^k) - \nabla f(\boldsymbol{w}^{k-1})\| \leq \mathbb{E}\left[L\|\boldsymbol{w}^k - \boldsymbol{w}^{k-1}\|\right] \leq L\gamma.$$

With the definition and Lemma 8, we can get

$$\mathbb{E}\Big\|\sum_{k=1}^t \theta^{t-k}\boldsymbol{\delta}^k\Big\| \leq \sqrt{\mathbb{E}\Big\|\sum_{k=1}^t \theta^{t-k}\boldsymbol{\delta}^k\Big\|^2} = \sqrt{\sum_{k=1}^t \theta^{2t-2k}\mathbb{E}\|\boldsymbol{\delta}^k\|^2}$$

$$\leq \frac{\sqrt{4(B+L\gamma t)^{2-p}\sigma^p}}{\sqrt{1-\theta^2}} \leq \frac{2(B+L\gamma t)^{\frac{2-p}{2}}\sigma^{p/2}}{\sqrt{1-\theta}}.$$

The bound of the right side of (6) then reduces to

$$\mathbb{E}\|\boldsymbol{\epsilon}^t\| \leq \frac{L\gamma}{1-\theta} + 2\sqrt{1-\theta}(B+L\gamma t)^{\frac{2-p}{2}}\sigma^{p/2} + \theta^t\|\boldsymbol{\epsilon}^0\|$$
$$\leq \frac{L\gamma}{1-\theta} + 2\sqrt{1-\theta}(B+L\gamma T)^{\frac{2-p}{2}}\sigma^{p/2} + \theta^t\|\boldsymbol{\epsilon}^0\|. \tag{7}$$

Summing the inequality from $t = 1$ to T, we then get

$$\sum_{t=1}^T \mathbb{E}\|\boldsymbol{\epsilon}^t\|/T \leq \frac{L\gamma}{1-\theta} + \frac{\|\nabla f(\boldsymbol{w}^0)\|}{(1-\theta)T} + 2\sqrt{1-\theta}(B+L\gamma T)^{\frac{2-p}{2}}\sigma^{p/2}. \tag{8}$$

Applying Lemma 9 with $\boldsymbol{w}^\dagger$ replaced by $\boldsymbol{w}^t$ and $\boldsymbol{m}$ replaced by $\boldsymbol{m}^t$, and considering expectations,

$$\mathbb{E}f(\boldsymbol{w}^{t+1}) - \mathbb{E}f(\boldsymbol{w}^t) \leq -\gamma\mathbb{E}\|\nabla f(\boldsymbol{w}^t)\| + 2\gamma\mathbb{E}\|\boldsymbol{\epsilon}^t\| + \frac{L\gamma^2}{2}.$$

Adding up the recursion from $t = 1$ to T, we obtain

$$\frac{1}{T}\sum_{t=1}^T \mathbb{E}\|\nabla f(\boldsymbol{w}^t)\| \leq \frac{f(\boldsymbol{w}^1) - \min f}{\gamma T} + 2\sum_{t=1}^T \mathbb{E}\|\boldsymbol{\epsilon}^t\|/T + L\gamma/2. \tag{9}$$

Employing the bound (8) and (9), we further get

$$\frac{1}{T}\sum_{t=1}^T \mathbb{E}\|\nabla f(\boldsymbol{w}^t)\|$$
$$\leq \frac{f(\boldsymbol{w}^1) - \min f}{\gamma T} + L\gamma/2 + \frac{2L\gamma}{1-\theta} + \frac{2\|\nabla f(\boldsymbol{w}^0)\|}{(1-\theta)T} + 4\sqrt{1-\theta}(B+L\gamma T)^{\frac{2-p}{2}}\sigma^{p/2}$$
$$\leq \frac{f(\boldsymbol{w}^1) - \min f}{\gamma T} + \frac{3L\gamma}{1-\theta} + \frac{2\|\nabla f(\boldsymbol{w}^0)\|}{(1-\theta)T} + 4\sqrt{1-\theta}(B+L\gamma T)^{\frac{2-p}{2}}\sigma^{p/2} \tag{10}$$
$$\leq \frac{f(\boldsymbol{w}^1) - \min f}{\gamma T} + \frac{3L\gamma}{1-\theta} + \frac{2\|\nabla f(\boldsymbol{w}^0)\|}{(1-\theta)T} + 4(B+L)^{\frac{2-p}{2}}\sqrt{1-\theta}(\gamma T)^{\frac{2-p}{2}}\sigma^{p/2},$$

where we used $L\gamma/2 \le \frac{L\gamma}{1-\theta}$ and $(B + L\gamma T) \le (B + L)\gamma T$ because $\gamma T \ge 1$ (Note that we have $\frac{f(\boldsymbol{w}^1) - \min f}{\gamma T}$ in the bound, $\gamma T \to +\infty$ to ensures the convergence).

Case I: $L\Delta \ge 1$. 1) If $\sigma^{\frac{4p-4}{3p-2}} T^{\frac{p}{3p-2}} \ge \sqrt{L\Delta}$, we have

$$1 - \theta = \frac{\sqrt{L\Delta}}{\sigma^{\frac{4p-4}{3p-2}} T^{\frac{p}{3p-2}}}, \gamma = \frac{\sqrt{\Delta}^{\frac{3}{4}}}{\sqrt[4]{L}\sigma^{\frac{2p-2}{3p-2}} T^{\frac{2p-1}{3p-2}}}.$$

Based on (10) and the fact $\boldsymbol{w}^0 = \boldsymbol{w}^1$, we have

$$
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{w}^t)\| \le \frac{4(L\Delta)^{\frac{1}{4}}\sigma^{\frac{2p-2}{3p-2}}}{T^{\frac{p-1}{3p-2}}} + \frac{2\|\nabla f(\boldsymbol{w}^0)\|\sigma^{\frac{4p-4}{3p-2}}}{T^{\frac{2p-2}{3p-2}}}
$$
$$
+ 4(L\Delta)^{\frac{6-3p}{8}}\left(\frac{B}{L}+1\right)^{\frac{2-p}{2}}\sigma^{\frac{5p^2-12p+8}{6p-4}}\frac{1}{T^{\frac{p}{2(3p-2)} - \frac{(p-1)(2-p)}{2(3p-2)}}} \tag{11}
$$
$$
\le \frac{4(L\Delta)^{\frac{1}{4}}\sigma^{\frac{2p-2}{3p-2}}}{T^{\frac{p-1}{3p-2}}} + \frac{2\|\nabla f(\boldsymbol{w}^0)\|\sigma^{\frac{4p-4}{3p-2}}}{T^{\frac{2p-2}{3p-2}}} + \frac{4(L\Delta)^{\frac{6-3p}{8}}\left(\frac{B}{L}+1\right)^{\frac{2-p}{2}}\sigma^{\frac{5p^2-12p+8}{6p-4}}}{T^{\frac{(p-2)^2}{2(3p-2)} + \frac{p-1}{3p-2}}},
$$

where we used $1 < p \le 2$.

2) If $\sigma^{\frac{4p-4}{3p-2}} T^{\frac{p}{3p-2}} \le 1$, we have

$$1 - \theta = 1, \gamma = \frac{\sqrt{\Delta}}{\sqrt{L}T^{\frac{1}{2}}}.$$

The upper bound (10) can be rebounded as

$$
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{w}^t)\|
$$
$$
\le \frac{4\sqrt{L\Delta}}{T^{\frac{1}{2}}} + \frac{2\|\nabla f(\boldsymbol{w}^0)\|}{T} + 4(L\Delta)^{\frac{2-p}{4}}\left(\frac{B}{L}+1\right)^{\frac{2-p}{2}}T^{\frac{2-p}{4}}\sigma^{p/2}
$$
$$
\le \frac{4\sqrt{L\Delta}}{T^{\frac{1}{2}}} + \frac{2\|\nabla f(\boldsymbol{w}^0)\|}{T} + 4(L\Delta)^{\frac{2-p}{4}}\left(\frac{B}{L}+1\right)^{\frac{2-p}{2}}\frac{\sigma^{p/8}}{T^{\frac{3p^2}{32p-32} - \frac{2-p}{4}}}.
$$

Notice that $\frac{3p^2}{32p-32} - \frac{2-p}{4} \ge \frac{1}{4}$, when $1 < p \le 2$. Thus, we are led to

$$
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{w}^t)\| \le \frac{4(L\Delta)^{\frac{2-p}{4}}\left(\frac{B}{L}+1\right)^{\frac{2-p}{2}}\sigma^{p/8}}{T^{\frac{1}{4}}} + \frac{4\sqrt{L\Delta}}{T^{\frac{1}{2}}} + \frac{2\|\nabla f(\boldsymbol{w}^0)\|}{T}. \tag{12}
$$

Taking maximum of the both bounds in (11) and (12), we then get the convergence as

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{w}^t)\| \leq \frac{4(L\Delta)^{\frac{1}{4}}\sigma^{\frac{2p-2}{3p-2}}}{T^{\frac{p-1}{3p-2}}} + \frac{2\|\nabla f(\boldsymbol{w}^0)\|\sigma^{\frac{4p-4}{3p-2}}}{T^{\frac{2p-2}{3p-2}}}$$

$$+ \frac{4(L\Delta)^{\frac{6-3p}{8}}\left(\frac{B}{L}+1\right)^{\frac{2-p}{2}}\sigma^{\frac{5p^2-12p+8}{6p-4}}}{T^{\frac{(p-2)^2}{2(3p-2)}+\frac{p-1}{3p-2}}}$$

$$+ \frac{4(L\Delta)^{\frac{2-p}{4}}\left(\frac{B}{L}+1\right)^{\frac{2-p}{2}}\sigma^{p/8}}{T^{\frac{1}{4}}} + \frac{4\sqrt{L\Delta}}{T^{\frac{1}{2}}} + \frac{2\|\nabla f(\boldsymbol{w}^0)\|}{T}.$$

Case II: $L\Delta < 1$. Similar to Case I, we can get

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{w}^t)\| \leq \frac{4\sqrt{L\Delta}\sigma^{\frac{2p-2}{3p-2}}}{T^{\frac{p-1}{3p-2}}} + \frac{2\|\nabla f(\boldsymbol{w}^0)\|\sigma^{\frac{4p-4}{3p-2}}}{T^{\frac{2p-2}{3p-2}}}$$

$$+ \frac{4(L\Delta)^{\frac{2-p}{4}}\left(\frac{B}{L}+1\right)^{\frac{2-p}{2}}\sigma^{\frac{5p^2-12p+8}{6p-4}}}{T^{\frac{(p-2)^2}{2(3p-2)}+\frac{p-1}{3p-2}}}$$

$$+ \frac{4(L\Delta)^{\frac{2-p}{4}}\left(\frac{B}{L}+1\right)^{\frac{2-p}{2}}\sigma^{p/8}}{T^{\frac{1}{4}}} + \frac{4\sqrt{L\Delta}}{T^{\frac{1}{2}}} + \frac{2\|\nabla f(\boldsymbol{w}^0)\|}{T}$$

$$\leq \frac{4(L\Delta)^{\frac{1}{4}}\sigma^{\frac{2p-2}{3p-2}}}{T^{\frac{p-1}{3p-2}}} + \frac{2\|\nabla f(\boldsymbol{w}^0)\|\sigma^{\frac{4p-4}{3p-2}}}{T^{\frac{2p-2}{3p-2}}}$$

$$+ \frac{4(L\Delta)^{\frac{2-p}{4}}\left(\frac{B}{L}+1\right)^{\frac{2-p}{2}}\sigma^{\frac{5p^2-12p+8}{6p-4}}}{T^{\frac{(p-2)^2}{2(3p-2)}+\frac{p-1}{3p-2}}}$$

$$+ \frac{4(L\Delta)^{\frac{2-p}{4}}\left(\frac{B}{L}+1\right)^{\frac{2-p}{2}}\sigma^{p/8}}{T^{\frac{1}{4}}} + \frac{4\sqrt{L\Delta}}{T^{\frac{1}{2}}} + \frac{2\|\nabla f(\boldsymbol{w}^0)\|}{T}$$

as $L\Delta < 1$.

Combining Case I and Case II, we then proved the final result.

Appendix C. Proof of Theorem 2

For Algorithm 5, we use different shorthand notation

$$\boldsymbol{\epsilon}^t := \boldsymbol{m}^t - \nabla f(\boldsymbol{w}^t), \boldsymbol{\delta}^t := \nabla f(\boldsymbol{w}^t; \xi^t) - \nabla f(\boldsymbol{w}^t),$$
$$\hat{\boldsymbol{s}}^t := \nabla f(\boldsymbol{w}^{t-1}) - \nabla f(\boldsymbol{w}^t) + \nabla f(\boldsymbol{w}^t; \xi^t) - \nabla f(\boldsymbol{w}^{t-1}; \xi^t). \tag{13}$$

The scheme of momentum updating in the NSGD-VR indicates

$$\boldsymbol{m}^t = \theta\boldsymbol{m}^{t-1} + \nabla f(\boldsymbol{w}^t; \xi^t) - \theta\nabla f(\boldsymbol{w}^{t-1}; \xi^t)$$
$$= \theta\left[\boldsymbol{\epsilon}^{t-1} + \nabla f(\boldsymbol{w}^{t-1})\right] + (1-\theta)\left[\boldsymbol{\delta}^t + \nabla f(\boldsymbol{w}^t)\right] + \theta\left[\nabla f(\boldsymbol{w}^t; \xi^t) - \nabla f(\boldsymbol{w}^{t-1}; \xi^t)\right].$$

Thus, we can derive

$$\boldsymbol{\epsilon}^t = \boldsymbol{m}^t - \nabla f(\boldsymbol{w}^t) = \theta\boldsymbol{\epsilon}^{t-1} + \theta\hat{\boldsymbol{s}}^t + (1-\theta)\boldsymbol{\delta}^t.$$

The relation above then gives us

$$\boldsymbol{\epsilon}^t = \theta^t\boldsymbol{\epsilon}^0 + \sum_{k=1}^t \theta^{t-k}\hat{\boldsymbol{s}}^k + (1-\theta)\sum_{k=1}^t \theta^{t-k}\boldsymbol{\delta}^k.$$

Taking norms of both sides and using the triangle inequality, it follows

$$\|\boldsymbol{\epsilon}^t\| \leq \theta^t\|\boldsymbol{\epsilon}^0\| + \|\sum_{k=1}^t \theta^{t-k}\hat{\boldsymbol{s}}^k\| + (1-\theta)\|\sum_{k=1}^t \theta^{t-k}\boldsymbol{\delta}^k\|.$$

Taking the expectations, we can further have

$$\mathbb{E}\|\boldsymbol{\epsilon}^t\| \leq \theta^t\mathbb{E}\|\boldsymbol{\epsilon}^0\| + \mathbb{E}\|\sum_{k=1}^t \theta^{t-k}\hat{\boldsymbol{s}}^k\| + (1-\theta)\mathbb{E}\|\sum_{k=1}^t \theta^{t-k}\boldsymbol{\delta}^k\|.$$

The upper bound of $\|\sum_{k=1}^t \theta^{t-k}\hat{\boldsymbol{s}}^k\|$ can be derived as follows

$$\mathbb{E}\|\sum_{k=1}^t \theta^{t-k}\hat{\boldsymbol{s}}^k\| \leq \sqrt{\mathbb{E}\|\sum_{k=1}^t \theta^{t-k}\hat{\boldsymbol{s}}^k\|^2} = \sqrt{\sum_{k=1}^t \theta^{2t-2k}\mathbb{E}\|\hat{\boldsymbol{s}}^k\|^2}$$

$$\leq \sqrt{\frac{4L^2\gamma^2}{1-\theta^2}} \leq \frac{2L\gamma}{\sqrt{1-\theta}},$$

where we used $\mathbb{E}\hat{\boldsymbol{s}}^k = \boldsymbol{0}$ and

$$\mathbb{E}\|\hat{\boldsymbol{s}}^k\|^2 \leq 2\mathbb{E}\|\nabla f(\boldsymbol{w}^{t-1}) - \nabla f(\boldsymbol{w}^t)\|^2 + 2\|\nabla f(\boldsymbol{w}^t;\xi^t) - \nabla f(\boldsymbol{w}^{t-1};\xi^t)\|^2$$

$$\leq 4L^2\mathbb{E}\|\boldsymbol{w}^t - \boldsymbol{w}^{t-1}\|^2 \leq 4L^2\gamma^2.$$

The bound of $\mathbb{E}\left\|\sum_{k=1}^t \theta^{t-k}\boldsymbol{\delta}^k\right\|$ follows the same of that in Theorem 1. Therefore, we are then led to the following result

$$\mathbb{E}\|\boldsymbol{\epsilon}^t\| \leq \frac{2L\gamma}{\sqrt{1-\theta}} + 2\sqrt{1-\theta}(B + L\gamma T)^{\frac{2-p}{2}}\sigma^{p/2} + \theta^t\|\boldsymbol{\epsilon}^0\| \tag{14}$$

as $1 \leq t \leq T$. Adding up the inequality from $t = 1$ to T, we obtain

$$\sum_{t=1}^T \mathbb{E}\|\boldsymbol{\epsilon}^t\|/T \leq \frac{2L\gamma}{\sqrt{1-\theta}} + \frac{\|\nabla f(\boldsymbol{w}^0)\|}{(1-\theta)T} + 2\sqrt{1-\theta}(B + L\gamma T)^{\frac{2-p}{2}}\sigma^{p/2}. \tag{15}$$

Noticing that (9) still holds in NSGD-VR, we then have

$$
\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{w}^t)\| \\
\leq \frac{f(\boldsymbol{w}^1)-\min f}{\gamma T}+L\gamma/2+\frac{4L\gamma}{\sqrt{1-\theta}}+\frac{2\|\nabla f(\boldsymbol{w}^0)\|}{(1-\theta)T}+4\sqrt{1-\theta}(B+L\gamma T)^{\frac{2-p}{2}}\sigma^{p/2} \\
\leq \frac{f(\boldsymbol{w}^1)-\min f}{\gamma T}+\frac{5L\gamma}{\sqrt{1-\theta}}+\frac{2\|\nabla f(\boldsymbol{w}^0)\|}{(1-\theta)T}+4\sqrt{1-\theta}(B+L\gamma T)^{\frac{2-p}{2}}\sigma^{p/2} \\
\leq \frac{f(\boldsymbol{w}^1)-\min f}{\gamma T}+\frac{5L\gamma}{\sqrt{1-\theta}}+\frac{2\|\nabla f(\boldsymbol{w}^0)\|}{(1-\theta)T}+4(B+L)^{\frac{2-p}{2}}\sqrt{1-\theta}(\gamma T)^{\frac{2-p}{2}}\sigma^{p/2},
\end{aligned}
\tag{16}
$$

where we used $L\gamma/2 \leq \frac{L\gamma}{\sqrt{1-\theta}}$ and $(B+L\gamma T)\leq(B+L)\gamma T$.

1). If $\sigma^{\frac{p}{2p-1}}T^{\frac{p}{2p-1}}\geq 1$, it follows

$$
1-\theta=\frac{1}{\sigma^{\frac{p}{2p-1}}T^{\frac{p}{2p-1}}},\gamma=\frac{1}{L\sigma^{\frac{p}{4(2p-1)}}T^{\frac{5p-2}{4(2p-1)}}}.
$$

We then simplify the inequality (16) as

$$
\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{w}^t)\| \\
\leq \frac{f(\boldsymbol{w}^1)-\min f}{\gamma T}+\frac{5L\gamma}{\sqrt{1-\theta}}+\frac{2\|\nabla f(\boldsymbol{w}^0)\|}{(1-\theta)T}+4(B+L)^{\frac{2-p}{2}}\sqrt{1-\theta}(\gamma T)^{\frac{2-p}{2}}\sigma^{p/2} \\
\leq \frac{L(f(\boldsymbol{w}^1)-\min f)\sigma^{\frac{p}{4(2p-1)}}+5\sigma^{\frac{p}{4(2p-1)}}}{T^{\frac{3p-2}{4(2p-1)}}}+\frac{2\|\nabla f(\boldsymbol{w}^0)\|\sigma^{\frac{p}{2p-1}}}{T^{\frac{p-1}{2p-1}}}+\frac{4\left(\frac{B}{L}+1\right)^{\frac{2-p}{2}}\sigma^{\frac{9p^2-10p}{8(2p-1)}}}{T^{\frac{3p^2-4p+4}{8(2p-1)}}} \\
= \frac{L(f(\boldsymbol{w}^1)-\min f)\sigma^{\frac{p}{4(2p-1)}}+5\sigma^{\frac{p}{4(2p-1)}}}{T^{\frac{3p-2}{4(2p-1)}}}+\frac{2\|\nabla f(\boldsymbol{w}^0)\|\sigma^{\frac{p}{2p-1}}}{T^{\frac{p-1}{2p-1}}}+\frac{4\left(\frac{B}{L}+1\right)^{\frac{2-p}{2}}\sigma^{\frac{9p^2-10p}{8(2p-1)}}}{T^{\frac{p-1}{2p-1}+\frac{3(p-2)^2}{8(2p-1)}}} \\
\leq \frac{L(f(\boldsymbol{w}^1)-\min f)\sigma^{\frac{p}{4(2p-1)}}+5\sigma^{\frac{p}{4(2p-1)}}}{T^{\frac{3p-2}{4(2p-1)}}}+\frac{2\|\nabla f(\boldsymbol{w}^0)\|\sigma^{\frac{p}{2p-1}}}{T^{\frac{p-1}{2p-1}}} \\
+\frac{4\left(\frac{B}{L}+1\right)^{\frac{2-p}{2}}\sigma^{\frac{12p^2-22p+12}{8(2p-1)}}}{T^{\frac{p-1}{2p-1}}},
\end{aligned}
\tag{17}
$$

23

due to $\frac{1}{T} \leq \sigma$. Based on the facts above and $\boldsymbol{w}^0 = \boldsymbol{w}^1$,

$$
\frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \|\nabla f(\boldsymbol{w}^t)\|
$$

$$
\leq \frac{2\|\nabla f(\boldsymbol{w}^0)\|\sigma^{\frac{p}{2p-1}} + 4\left(\frac{B}{L}+1\right)^{\frac{2-p}{2}} \sigma^{\frac{12p^2-22p+12}{8(2p-1)}}}{T^{\frac{p-1}{2p-1}}} + \frac{[L(f(\boldsymbol{w}^1) - \min f) + 5]\sigma^{\frac{p}{4(2p-1)}}}{T^{\frac{3p-2}{4(2p-1)}}} \quad (18)
$$

$$
\leq \frac{2\|\nabla f(\boldsymbol{w}^0)\|\sigma^{\frac{p}{2p-1}} + 4\left(\frac{B}{L}+1\right)^{\frac{2-p}{2}} \sigma^{\frac{12p^2-22p+12}{8(2p-1)}} + [L(f(\boldsymbol{w}^1) - \min f) + 5]\sigma^{\frac{p}{4(2p-1)}}}{T^{\frac{p-1}{2p-1}}}
$$

because $\frac{3p-2}{4(2p-1)} \geq \frac{p-1}{2p-1}$.

2). If $\sigma^{\frac{p}{2p-1}} T^{\frac{p}{2p-1}} \leq 1$, it follows

$$
1 - \theta = 1, \gamma = \frac{1}{LT^{1/2}}.
$$

We can rebound (16) as

$$
\frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \|\nabla f(\boldsymbol{w}^t)\| \leq \frac{L(f(\boldsymbol{w}^1) - \min f) + 5}{\sqrt{T}} + \frac{2\|\nabla f(\boldsymbol{w}^0)\|}{T} + \frac{4\left(\frac{B}{L}+1\right)^{\frac{2-p}{2}}}{T^{\frac{p^2}{2p-1} - \frac{2-p}{4}}}. \quad (19)
$$

It is easy to check that

$$
\frac{p^2}{2p-1} - \frac{2-p}{4} \geq \frac{1}{2}
$$

as $p \in (1, 2]$. The bound in (19) can be rebounded as

$$
\frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \|\nabla f(\boldsymbol{w}^t)\| \leq \frac{L(f(\boldsymbol{w}^1) - \min f) + 4\left(\frac{B}{L}+1\right)^{\frac{2-p}{2}} + 5}{\sqrt{T}} + \frac{2\|\nabla f(\boldsymbol{w}^0)\|}{T}
$$

$$
= \frac{L(f(\boldsymbol{w}^0) - \min f) + 4\left(\frac{B}{L}+1\right)^{\frac{2-p}{2}} + 5}{\sqrt{T}} + \frac{2\|\nabla f(\boldsymbol{w}^0)\|}{T}. \quad (20)
$$

Summing both bounds in (18) and (20), the convergence rate is proved as

$$
\frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \|\nabla f(\boldsymbol{w}^t)\|
$$

$$
\leq \frac{2\|\nabla f(\boldsymbol{w}^0)\|\sigma^{\frac{p}{2p-1}} + 4\left(\frac{B}{L}+1\right)^{\frac{2-p}{2}} \sigma^{\frac{12p^2-22p+12}{8(2p-1)}} + [L(f(\boldsymbol{w}^0) - \min f) + 5]\sigma^{\frac{p}{4(2p-1)}}}{T^{\frac{p-1}{2p-1}}}
$$

$$
+ \frac{L(f(\boldsymbol{w}^0) - \min f) + 4\left(\frac{B}{L}+1\right)^{\frac{2-p}{2}} + 5}{\sqrt{T}} + \frac{2\|\nabla f(\boldsymbol{w}^0)\|}{T}.
$$

24

Note the following inequality always holds

$$\sigma^y \leq \sigma^a + \sigma^x, \text{ as } 0 \leq x \leq y \leq a. \tag{21}$$

This holds because if $\sigma \geq 1$, then $\sigma^y \leq \sigma^a$; otherwise, $\sigma^y \leq \sigma^x$. By combining these cases, the inequality is established. Due to that $\frac{p}{4(2p-1)} \leq \frac{12p^2 - 22p + 12}{8(2p-1)} \leq \frac{p}{2p-1}$ when $1 \leq p \leq 2$, using (21), we then proved the result.

Appendix D. Proof of Theorem 3

This proof denotes the following shorthand notation

$$
\begin{aligned}
\boldsymbol{\epsilon}^t &:= \boldsymbol{m}^t - \nabla f(\boldsymbol{w}^t), \boldsymbol{\delta}^t := \mathrm{Clip}_h(\boldsymbol{g}^t) - \mathbb{E}\mathrm{Clip}_h(\boldsymbol{g}^t), \\
\boldsymbol{c}^t &:= \mathbb{E}\mathrm{Clip}_h(\boldsymbol{g}^t) - \nabla f(\boldsymbol{w}^t), \boldsymbol{s}^t := \nabla f(\boldsymbol{w}^{t-1}) - \nabla f(\boldsymbol{w}^t).
\end{aligned}
\tag{22}
$$

In NSGDC, we have

$$
\begin{aligned}
\boldsymbol{m}^t &= \theta \boldsymbol{m}^{t-1} + (1-\theta)\mathrm{Clip}_h(\boldsymbol{g}^t) \\
&= \theta\Big[\boldsymbol{\epsilon}^{t-1} + \nabla f(\boldsymbol{w}^{t-1})\Big] + (1-\theta)\Big[\boldsymbol{\delta}^t + \boldsymbol{c}^t + \nabla f(\boldsymbol{w}^t)\Big].
\end{aligned}
$$

The inequality then yields

$$\boldsymbol{\epsilon}^t = \boldsymbol{m}^t - \nabla f(\boldsymbol{w}^t) = \theta\boldsymbol{\epsilon}^{t-1} + \theta\boldsymbol{s}^t + (1-\theta)\boldsymbol{\delta}^t + (1-\theta)\boldsymbol{c}^t.$$

Using mathematical induction, we can derive

$$\boldsymbol{\epsilon}^t = \theta^t\boldsymbol{\epsilon}^0 + \sum_{k=1}^t \theta^{t-k}\boldsymbol{s}^k + (1-\theta)\sum_{k=1}^t \theta^{t-k}\boldsymbol{\delta}^k + (1-\theta)\sum_{k=1}^t \theta^{t-k}\boldsymbol{c}^k.$$

By taking the norms and expectations of both sides, we have

$$
\begin{aligned}
\mathbb{E}\|\boldsymbol{\epsilon}^t\| &\leq \sum_{k=1}^t \theta^{t-k}\mathbb{E}\|\boldsymbol{s}^k\| + (1-\theta)\sum_{k=1}^t \theta^{t-k}\mathbb{E}\|\boldsymbol{c}^k\| + (1-\theta)\mathbb{E}\Big\|\sum_{k=1}^t \theta^{t-k}\boldsymbol{\delta}^k\Big\| + \theta^t\|\boldsymbol{\epsilon}^0\| \\
&\leq \frac{L\gamma}{1-\theta} + 2\sigma^p h^{-(p-1)} + (1-\theta)\mathbb{E}\Big\|\sum_{k=1}^t \theta^{t-k}\boldsymbol{\delta}^k\Big\| + \theta^t\|\boldsymbol{\epsilon}^0\|,
\end{aligned}
\tag{23}
$$

where we used the gradient Lipschitz property of function f as

$$\mathbb{E}\|\boldsymbol{s}^k\| = \mathbb{E}\|\nabla f(\boldsymbol{w}^k) - \nabla f(\boldsymbol{w}^{k-1})\| \leq \mathbb{E}\Big[L\|\boldsymbol{w}^k - \boldsymbol{w}^{k-1}\|\Big] \leq L\gamma,$$

and Lemma 7 (note that $h \geq 2(\gamma LT + \|\nabla f(\boldsymbol{w}^0)\|)$). Now, we turn to the bound of $\mathbb{E}\Big\|\sum_{k=1}^t \theta^{t-k}\boldsymbol{\delta}^k\Big\|$. With the Cauchy's inequality, we can get

$$
\begin{aligned}
\mathbb{E}\Big\|\sum_{k=1}^t \theta^{t-k}\boldsymbol{\delta}^k\Big\| &\leq \sqrt{\mathbb{E}\Big\|\sum_{k=1}^t \theta^{t-k}\boldsymbol{\delta}^k\Big\|^2} = \sqrt{\sum_{k=1}^t \theta^{2t-2k}\mathbb{E}\|\boldsymbol{\delta}^k\|^2} \\
&\leq \frac{\sqrt{10h^{2-p}\sigma^p}}{\sqrt{1-\theta^2}} \leq \frac{4h^{\frac{2-p}{2}}\sigma^{p/2}}{\sqrt{1-\theta}}.
\end{aligned}
$$

25

Recalling inequality (23), we are led to

$$\mathbb{E}\|\boldsymbol{\epsilon}^t\| \leq \frac{L\gamma}{1-\theta} + 2\sigma^p h^{-(p-1)} + 4\sqrt{1-\theta}h^{\frac{2-p}{2}}\sigma^{p/2} + \theta^t\|\boldsymbol{\epsilon}^0\|,$$

which further indicates

$$\sum_{t=1}^{T}\mathbb{E}\|\boldsymbol{\epsilon}^t\|/T \leq \frac{L\gamma}{1-\theta} + 2\sigma^p h^{-(p-1)} + 4\sqrt{1-\theta}h^{\frac{2-p}{2}}\sigma^{p/2} + \frac{\|\nabla f(\boldsymbol{w}^0)\|}{(1-\theta)T}. \qquad (24)$$

Noticing that (9) still holds, further with (24), we can get the following bound

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{w}^t)\|$$

$$\leq \frac{f(\boldsymbol{w}^1) - \min f}{\gamma T} + L\gamma/2 + \frac{2L\gamma}{1-\theta} + \frac{2\|\nabla f(\boldsymbol{w}^0)\|}{(1-\theta)T} + 4\sigma^p h^{-(p-1)} + 8\sqrt{1-\theta}h^{\frac{2-p}{2}}\sigma^{p/2} \quad (25)$$

$$\leq \frac{f(\boldsymbol{w}^1) - \min f}{\gamma T} + \frac{3L\gamma}{1-\theta} + \frac{2\|\nabla f(\boldsymbol{w}^0)\|}{(1-\theta)T} + 4\sigma^p h^{-(p-1)} + 8\sqrt{1-\theta}h^{\frac{2-p}{2}}\sigma^{p/2},$$

where we used that $\frac{1}{2} \leq \frac{1}{1-\theta}$.

Case I: $L\Delta \geq 1$. We used a shorthand notation as $\zeta := (L\Delta)^{\frac{p}{3p-2}} \geq 1$.
1) If $\sigma < 1$ and $\sigma^{\frac{4p-4}{3p-2}}T^{\frac{p}{3p-2}} \geq \zeta$, we have

$$1-\theta = \frac{\zeta}{\sigma^{\frac{4p-4}{3p-2}}T^{\frac{p}{3p-2}}}, \gamma = \frac{\sqrt{\zeta\Delta}}{\sqrt{L}\sigma^{\frac{2p-2}{3p-2}}T^{\frac{2p-1}{3p-2}}}, h = \frac{2\sqrt{\zeta}\sqrt{L\Delta}T^{\frac{1}{3p-2}}}{\sigma^{\frac{2p-2}{3p-2}}} + 2\|\nabla f(\boldsymbol{w}^0)\|.$$

Direct calculations indicate $\sqrt{\zeta}\sqrt{L\Delta} = \zeta^{\frac{2p-1}{p}} \geq 1$. Based on (25), with the fact $0 \leq \sigma \leq 1$ and $\boldsymbol{w}^0 = \boldsymbol{w}^1$, we have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{w}^t)\| \leq \frac{(L\Delta)^{\frac{p-1}{3p-2}}\sigma^{\frac{2p-2}{3p-2}} + 3(L\Delta)^{\frac{p-1}{3p-2}}\sigma^{\frac{2p-2}{3p-2}} + 4\sigma^{\frac{p}{3p-2}}}{T^{\frac{p-1}{3p-2}}}$$

$$+ 16\sqrt{1-\theta}\left(\frac{\zeta^{\frac{2p-1}{p}}T^{\frac{1}{3p-2}}}{\sigma^{\frac{2p-2}{3p-2}}} + \|\nabla f(\boldsymbol{w}^0)\|\right)^{\frac{2-p}{2}}\sigma^{\frac{p}{2}} + \frac{2\|\nabla f(\boldsymbol{w}^0)\|\sigma^{\frac{4p-4}{3p-2}}}{T^{\frac{2p-2}{3p-2}}}$$

$$\leq \frac{4(L\Delta)^{\frac{p-1}{3p-2}}\sigma^{\frac{2p-2}{3p-2}} + 4\sigma^{\frac{p}{3p-2}} + 16\zeta^{\frac{6p-2p^2-2}{2p}}\sigma^{\frac{5p^2-12p+8}{2(3p-2)}}}{T^{\frac{p-1}{3p-2}}}$$

$$+ \frac{16\sqrt{\zeta}\|\nabla f(\boldsymbol{w}^0)\|^{\frac{2-p}{2}}\sigma^{\frac{3p^2-6p+4}{2(3p-2)}}}{T^{\frac{p}{6p-4}}} + \frac{2\|\nabla f(\boldsymbol{w}^0)\|\sigma^{\frac{4p-4}{3p-2}}}{T^{\frac{2p-2}{3p-2}}} \qquad (26)$$

$$\leq \frac{4(L\Delta)^{\frac{p-1}{3p-2}}\sigma^{\frac{2p-2}{3p-2}} + 4\sigma^{\frac{p}{3p-2}} + 16(L\Delta)^{\frac{p}{3p-2}}\sigma^{\frac{p-1}{3p-2}}}{T^{\frac{p-1}{3p-2}}}$$

$$+ \frac{16(L\Delta)^{\frac{p}{6p-4}}\|\nabla f(\boldsymbol{w}^0)\|^{\frac{2-p}{2}}\sigma^{\frac{3p^2-6p+4}{2(3p-2)}}}{T^{\frac{p}{6p-4}}} + \frac{2\|\nabla f(\boldsymbol{w}^0)\|\sigma^{\frac{4p-4}{3p-2}}}{T^{\frac{2p-2}{3p-2}}},$$

where we used that $(x+y)^\alpha \le x^\alpha + y^\alpha$ as $0 < \alpha \le 1$ and $x, y > 0^3$, and

$$\frac{p-1}{3p-2} \le \frac{5p^2 - 12p + 8}{6p-4}, \frac{6p - 2p^2 - 2}{6p-4} \le \frac{p}{3p-2}$$

as $1 \le p \le 2$.

2) If $\sigma < 1$ and $\sigma^{\frac{4p-4}{3p-2}} T^{\frac{p}{3p-2}} \le \zeta$, we have

$$1 - \theta = 1, \gamma = \sqrt{\frac{\Delta}{L}} \cdot \frac{1}{T^{\frac{1}{2}}}, h = 2(\sqrt{L\Delta} T^{\frac{p+2}{6p-4}} + \|\nabla f(\boldsymbol{w}^0)\|).$$

The upper bound (10) can be rebounded as

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\|\nabla f(\boldsymbol{w}^t)\|$$

$$\le \frac{4\sqrt{L\Delta}}{T^{\frac{1}{2}}} + \frac{2\|\nabla f(\boldsymbol{w}^0)\|}{T} + 4\sigma^p h^{-(p-1)} + 8h^{\frac{2-p}{2}} \sigma^{p/2}$$

$$\le \frac{4\sqrt{L\Delta}}{T^{\frac{1}{2}}} + \frac{2\|\nabla f(\boldsymbol{w}^0)\|}{T} + 4T^{-\frac{p^2}{4p-4} - \frac{(p-1)(p+2)}{6p-4}}$$

$$+ 16(L\Delta)^{\frac{1}{4}} T^{-\frac{3p^2}{32p-32} + \frac{(2-p)(p+2)}{12p-8}} \sigma^{\frac{p}{8}} + 16\|\nabla f(\boldsymbol{w}^0)\|^{\frac{2-p}{2}} T^{-\frac{p^2}{8p-8}}.$$

Noting that $\frac{p^2}{4p-4} + \frac{(p-1)(p+2)}{6p-4} \ge \frac{1}{2}$, $\frac{3p^2}{32p-32} - \frac{(2-p)(p+2)}{12p-8} \ge \frac{p-1}{3p-2}$, and $\frac{p^2}{8p-8} \ge \frac{1}{2}$ as $1 < p \le 2$, we then get the following bound

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\|\nabla f(\boldsymbol{w}^t)\| \le \frac{16(L\Delta)^{\frac{1}{4}} \sigma^{\frac{p}{8}}}{T^{\frac{p-1}{3p-2}}} + \frac{4\sqrt{L\Delta} + 16\|\nabla f(\boldsymbol{w}^0)\|^{\frac{2-p}{2}} + 4}{T^{\frac{1}{2}}} + \frac{2\|\nabla f(\boldsymbol{w}^0)\|}{T}. \quad (27)$$

3) Consider the subcase $\sigma \ge 1$ and $\sigma^{\frac{2p}{3p-2}} T^{\frac{p}{3p-2}} \ge \zeta$. Hence, we have

$$1 - \theta = \frac{\zeta}{\sigma^{\frac{2p}{3p-2}} T^{\frac{p}{3p-2}}}, \gamma = \sqrt{\frac{\zeta\Delta}{L}} \frac{1}{\sigma^{\frac{p}{3p-2}} T^{\frac{2p-1}{3p-2}}}, h = 2\sqrt{\zeta}\sqrt{L\Delta} T^{\frac{1}{3p-2}} \sigma^{\frac{2p}{3p-2}} + 2\|\nabla f(\boldsymbol{w}^0)\|.$$

---

3. We present a straightforward proof for this inequality. Without loss of generality, we can assume $x \ge y$. We can rewrite the inequality as $\left(\frac{x}{y} + 1\right)^\alpha \le \left(\frac{x}{y}\right)^\alpha + 1$. By letting $a = \frac{x}{y}$, it suffices to prove $(a+1)^\alpha - a^\alpha \le 1$ for $a \ge 1$. It is easy to see that $(a+1)^\alpha - a^\alpha = \alpha(a+z)^{-\alpha}$ with some $0 \le z \le 1$. We then proved the result by noting $a \ge 1$ and $0 < \alpha \le 1$.

Using (25), we then get

$$
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{w}^t)\| \leq \frac{4(L\Delta)^{\frac{p-1}{3p-2}}\sigma^{\frac{p}{3p-2}}+4\sigma^{\frac{p}{3p-2}}}{T^{\frac{p-1}{3p-2}}}
$$

$$
+16\sqrt{1-\theta}\Big(\sqrt{\zeta}\sqrt{L\Delta}T^{\frac{1}{3p-2}}\sigma^{\frac{2p}{3p-2}}+\|\nabla f(\boldsymbol{w}^0)\|\Big)^{\frac{2-p}{2}}\sigma^{\frac{p}{2}}+\frac{2\|\nabla f(\boldsymbol{w}^0)\|\sigma^{\frac{2p}{3p-2}}}{T^{\frac{2p-2}{3p-2}}}
$$

$$
\leq \frac{4(L\Delta)^{\frac{p-1}{3p-2}}\sigma^{\frac{p}{3p-2}}+4\sigma^{\frac{p}{3p-2}}+16(L\Delta)^{\frac{6p-2p^2-2}{6p-4}}\sigma^{\frac{p^2}{6p-4}}}{T^{\frac{p-1}{3p-2}}}
$$

$$
+\frac{16\sqrt{\zeta}\|\nabla f(\boldsymbol{w}^0)\|^{\frac{2-p}{2}}\sigma^{\frac{3p^2-4p}{2(3p-2)}}}{T^{\frac{p}{6p-4}}}+\frac{2\|\nabla f(\boldsymbol{w}^0)\|\sigma^{\frac{2p}{3p-2}}}{T^{\frac{2p-2}{3p-2}}}
$$

$$
\leq \frac{4(L\Delta)^{\frac{p-1}{3p-2}}\sigma^{\frac{p}{3p-2}}+4\sigma^{\frac{p}{3p-2}}+16(L\Delta)^{\frac{6p-2p^2-2}{6p-4}}\sigma^{\frac{p^2}{6p-4}}}{T^{\frac{p-1}{3p-2}}}+\frac{16\sqrt{\zeta}\|\nabla f(\boldsymbol{w}^0)\|^{\frac{2-p}{2}}\sigma^{\frac{3p^2-4p}{2(3p-2)}}}{T^{\frac{p}{6p-4}}}
$$

$$
+\frac{2\|\nabla f(\boldsymbol{w}^0)\|\sigma^{\frac{2p}{3p-2}}}{T^{\frac{2p-2}{3p-2}}}
$$

$$
\leq \frac{4(L\Delta)^{\frac{p-1}{3p-2}}\sigma^{\frac{p}{3p-2}}+4\sigma^{\frac{p}{3p-2}}+16(L\Delta)^{\frac{6p-2p^2-2}{6p-4}}\sigma^{\frac{p^2}{6p-4}}}{T^{\frac{p-1}{3p-2}}}
$$

$$
+\frac{16(L\Delta)^{\frac{p}{6p-4}}\|\nabla f(\boldsymbol{w}^0)\|^{\frac{2-p}{2}}\sigma^{\frac{3p^2-6p+4}{2(3p-2)}}}{T^{\frac{p}{6p-4}}}+\frac{2\|\nabla f(\boldsymbol{w}^0)\|\sigma^{\frac{2p}{3p-2}}}{T^{\frac{2p-2}{3p-2}}},
$$

(28)

where we used $\frac{3p^2-4p}{2(3p-2)}\leq\frac{3p^2-6p+4}{2(3p-2)}$ as $p\in(1,2]$.

4) If $\sigma\geq 1$ and $\sigma^{\frac{2p}{3p-2}}T^{\frac{p}{3p-2}}<\zeta$, we have

$$
1-\theta=1,\gamma=\sqrt{\frac{\Delta}{L}}\cdot\frac{1}{T^{\frac{1}{2}}},h=2(\sqrt{L\Delta}T^{\frac{p+2}{6p-4}}+\|\nabla f(\boldsymbol{w}^0)\|).
$$

This setting is identical to subcase 2) and the proof is similar.

Combing the bounds (26), (27) and (28), we then complete the upper bound as $L\Delta\geq 1$ as

$$
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{w}^t)\| \leq \frac{4(L\Delta)^{\frac{p-1}{3p-2}}\sigma^{\frac{p}{3p-2}}+4\sigma^{\frac{p}{3p-2}}}{T^{\frac{p-1}{3p-2}}}
$$

$$
+\frac{4(L\Delta)^{\frac{p-1}{3p-2}}\sigma^{\frac{2p-2}{3p-2}}+16(L\Delta)^{\frac{p}{3p-2}}\sigma^{\frac{p-1}{3p-2}}+16(L\Delta)^{\frac{6p-2p^2-2}{6p-4}}\sigma^{\frac{p^2}{6p-4}}+16(L\Delta)^{\frac{1}{4}}\sigma^{\frac{p}{8}}}{T^{\frac{p-1}{3p-2}}}
$$

$$
+\frac{16(L\Delta)^{\frac{p}{6p-4}}\|\nabla f(\boldsymbol{w}^0)\|^{\frac{2-p}{2}}\sigma^{\frac{3p^2-6p+4}{2(3p-2)}}}{T^{\frac{p}{6p-4}}}+\frac{2\|\nabla f(\boldsymbol{w}^0)\|\sigma^{\frac{4p-4}{3p-2}}+2\|\nabla f(\boldsymbol{w}^0)\|\sigma^{\frac{2p}{3p-2}}}{T^{\frac{2p-2}{3p-2}}}
$$

$$
+\frac{4\sqrt{L\Delta}+16\|\nabla f(\boldsymbol{w}^0)\|^{\frac{2-p}{2}}+4}{T^{\frac{1}{2}}}+\frac{2\|\nabla f(\boldsymbol{w}^0)\|}{T}.
$$

28

Case II: $L\Delta < 1$. Following the same process in Case I, we can get

$$
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{w}^t)\| \leq \frac{L\Delta\sigma^{\frac{p}{3p-2}} + 23\sigma^{\frac{p}{3p-2}}}{T^{\frac{p-1}{3p-2}}} + \frac{L\Delta\sigma^{\frac{2p-2}{3p-2}} + 3\sigma^{\frac{2p-2}{3p-2}} + 16\sigma^{\frac{p-1}{3p-2}} + 16\sigma^{\frac{p}{8}}}{T^{\frac{p-1}{3p-2}}}
$$

$$
+ \frac{16\|\nabla f(\boldsymbol{w}^0)\|^{\frac{2-p}{2}}\sigma^{\frac{3p^2-6p+4}{2(3p-2)}}}{T^{\frac{p}{6p-4}}} + \frac{2\|\nabla f(\boldsymbol{w}^0)\|\sigma^{\frac{4p-4}{3p-2}} + 2\|\nabla f(\boldsymbol{w}^0)\|\sigma^{\frac{2p}{3p-2}}}{T^{\frac{2p-2}{3p-2}}}
$$

$$
+ \frac{L\Delta + 16\|\nabla f(\boldsymbol{w}^0)\|^{\frac{2-p}{2}} + 7}{T^{\frac{1}{2}}} + \frac{2\|\nabla f(\boldsymbol{w}^0)\|}{T}
$$

$$
\leq \frac{(L\Delta)^{\frac{p-1}{3p-2}}\sigma^{\frac{p}{3p-2}} + 23\sigma^{\frac{p}{3p-2}}}{T^{\frac{p-1}{3p-2}}} + \frac{(L\Delta)^{\frac{p-1}{3p-2}}\sigma^{\frac{2p-2}{3p-2}} + 3\sigma^{\frac{2p-2}{3p-2}} + 16\sigma^{\frac{p-1}{3p-2}} + 16\sigma^{\frac{p}{8}}}{T^{\frac{p-1}{3p-2}}}
$$

$$
+ \frac{16\|\nabla f(\boldsymbol{w}^0)\|^{\frac{2-p}{2}}\sigma^{\frac{3p^2-6p+4}{2(3p-2)}}}{T^{\frac{p}{6p-4}}} + \frac{2\|\nabla f(\boldsymbol{w}^0)\|\sigma^{\frac{4p-4}{3p-2}} + 2\|\nabla f(\boldsymbol{w}^0)\|\sigma^{\frac{2p}{3p-2}}}{T^{\frac{2p-2}{3p-2}}}
$$

$$
+ \frac{\sqrt{L\Delta} + 16\|\nabla f(\boldsymbol{w}^0)\|^{\frac{2-p}{2}} + 7}{T^{\frac{1}{2}}} + \frac{2\|\nabla f(\boldsymbol{w}^0)\|}{T}
$$

as $L\Delta < 1$.

Combing the upper bounds in Case I and Case II together, we can get the final result.

Appendix E. Proof of Theorem 4

This section employs the following shorthand notation

$$
\begin{aligned}
\boldsymbol{\epsilon}^t &:= \boldsymbol{m}^t - \nabla f(\boldsymbol{w}^t), \boldsymbol{\delta}^t := \mathrm{Clip}_h(\nabla f(\boldsymbol{w}^t; \xi^t)) - \mathbb{E}\mathrm{Clip}_h(\nabla f(\boldsymbol{w}^t; \xi^t)), \\
\boldsymbol{c}^t &:= \mathbb{E}\mathrm{Clip}_h(\nabla f(\boldsymbol{w}^t; \xi^t)) - \nabla f(\boldsymbol{w}^t), \\
\hat{\boldsymbol{s}}^t &:= \nabla f(\boldsymbol{w}^{t-1}) - \nabla f(\boldsymbol{w}^t) + \nabla f(\boldsymbol{w}^t; \xi^t) - \nabla f(\boldsymbol{w}^{t-1}; \xi^t).
\end{aligned}
\tag{29}
$$

Like the proof of NSGDC, we have

$$
\boldsymbol{\epsilon}^t = \theta^t \boldsymbol{\epsilon}^0 + \sum_{k=1}^{t}\theta^{t-k}\hat{\boldsymbol{s}}^k + (1-\theta)\sum_{k=1}^{t}\theta^{t-k}\boldsymbol{\delta}^k + (1-\theta)\sum_{k=1}^{t}\theta^{t-k}\boldsymbol{c}^k,
$$

which then yields

$$
\mathbb{E}\|\boldsymbol{\epsilon}^t\| \leq \mathbb{E}\left\|\sum_{k=1}^{t}\theta^{t-k}\hat{\boldsymbol{s}}^k\right\| + (1-\theta)\sum_{k=1}^{t}\theta^{t-k}\mathbb{E}\|\boldsymbol{c}^k\| + (1-\theta)\mathbb{E}\left\|\sum_{k=1}^{t}\theta^{t-k}\boldsymbol{\delta}^k\right\| + \theta^t\|\boldsymbol{\epsilon}^0\|. \tag{30}
$$

The bound of $\mathbb{E}\left\|\sum_{k=1}^{t}\theta^{t-k}\boldsymbol{\delta}^k\right\|$ can follow the same as that in the last proof of NSGDC. Noticing that $\mathbb{E}\hat{\boldsymbol{s}}^k = \boldsymbol{0}$ and

$$
\begin{aligned}
\mathbb{E}\|\hat{\boldsymbol{s}}^k\|^2 &\leq 2\mathbb{E}\|\nabla f(\boldsymbol{w}^{k-1}) - \nabla f(\boldsymbol{w}^k)\|^2 + 2\|\nabla f(\boldsymbol{w}^k; \xi^k) - \nabla f(\boldsymbol{w}^{k-1}; \xi^k)\|^2 \\
&\leq 4L^2\mathbb{E}\|\boldsymbol{w}^k - \boldsymbol{w}^{k-1}\|^2 \leq 4L^2\gamma^2,
\end{aligned}
$$

we can get

$$
\mathbb{E}\|\sum_{k=1}^{t}\theta^{t-k}\hat{\boldsymbol{s}}^{k}\| \leq \sqrt{\mathbb{E}\|\sum_{k=1}^{t}\theta^{t-k}\hat{\boldsymbol{s}}^{k}\|^{2}} = \sqrt{\sum_{k=1}^{t}\theta^{2t-2k}\mathbb{E}\|\hat{\boldsymbol{s}}^{k}\|^{2}}
$$
$$
\leq \sqrt{\frac{4L^{2}\gamma^{2}}{1-\theta^{2}}} \leq \frac{2L\gamma}{\sqrt{1-\theta}},
$$

Based on (30), we then get

$$
\mathbb{E}\|\boldsymbol{\epsilon}^{t}\| \leq \frac{2L\gamma}{\sqrt{1-\theta}} + 2\sigma^{p}h^{-(p-1)} + 4\sqrt{1-\theta}h^{\frac{2-p}{2}}\sigma^{p/2} + \theta^{t}\|\boldsymbol{\epsilon}^{0}\|,
$$

which further indicates

$$
\sum_{t=1}^{T}\mathbb{E}\|\boldsymbol{\epsilon}^{t}\|/T \leq \frac{2L\gamma}{\sqrt{1-\theta}} + 2\sigma^{p}h^{-(p-1)} + 4\sqrt{1-\theta}h^{\frac{2-p}{2}}\sigma^{p/2} + \frac{\|\nabla f(\boldsymbol{w}^{0})\|}{(1-\theta)T}. \tag{31}
$$

Note that (9) still holds, it then follows

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{w}^{t})\| \\
&\leq \frac{f(\boldsymbol{w}^{1})-\min f}{\gamma T} + L\gamma/2 + \frac{4L\gamma}{\sqrt{1-\theta}} + \frac{2\|\nabla f(\boldsymbol{w}^{0})\|}{(1-\theta)T} \\
&\quad + 4\sigma^{p}h^{-(p-1)} + 8\sqrt{1-\theta}h^{\frac{2-p}{2}}\sigma^{p/2} \\
&\leq \frac{f(\boldsymbol{w}^{1})-\min f}{\gamma T} + \frac{5L\gamma}{\sqrt{1-\theta}} + \frac{2\|\nabla f(\boldsymbol{w}^{0})\|}{(1-\theta)T} + 4\sigma^{p}h^{-(p-1)} + 8\sqrt{1-\theta}h^{\frac{2-p}{2}}\sigma^{p/2}.
\end{aligned} \tag{32}
$$

1). If $\sigma \leq 1$ and $\sigma^{\frac{2p-2}{2p-1}}T^{\frac{p}{2p-1}} \geq 1$, we have

$$
1-\theta = \frac{1}{\sigma^{\frac{2p-2}{2p-1}}T^{\frac{p}{2p-1}}}, \gamma = \frac{1}{L\sigma^{\frac{p-1}{4p-2}}T^{\frac{5p-2}{8p-4}}}, h = \frac{2T^{\frac{1}{2p-1}}}{\sigma^{\frac{p-1}{4p-2}}} + 2\|\nabla f(\boldsymbol{w}^{0})\|.
$$

The upper bound of (32) can be reformulated as

$$
\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{w}^{t})\| &\leq \frac{[L(f(\boldsymbol{w}^{1})-\min f)+5]\sigma^{\frac{p-1}{4p-2}}}{T^{\frac{3p-2}{8p-4}}} + \frac{2\|\nabla f(\boldsymbol{w}^{0})\|\sigma^{\frac{2p-2}{2p-1}}}{T^{\frac{p-1}{2p-1}}} + \frac{4\sigma^{p+\frac{(p-1)^{2}}{4p-2}}}{T^{\frac{p-1}{2p-1}}} \\
&\quad + \frac{16\sigma^{\frac{5p^{2}-9p+6}{8p-4}}}{T^{\frac{5p-1}{8p-4}}} + \frac{16\|\nabla f(\boldsymbol{w}^{0})\|^{\frac{2-p}{2}}\sigma^{\frac{2p^{2}-5p+2}{4p-2}}}{T^{\frac{p}{4p-2}}} \\
&\leq \frac{[L(f(\boldsymbol{w}^{1})-\min f)+5]\sigma^{\frac{p-1}{4p-2}}}{T^{\frac{3p-2}{8p-4}}} + \frac{2\|\nabla f(\boldsymbol{w}^{0})\|\sigma^{\frac{2p-2}{2p-1}}}{T^{\frac{p-1}{2p-1}}} + \frac{4\sigma^{\frac{p}{2p-1}}}{T^{\frac{p-1}{2p-1}}} \\
&\quad + \frac{16\sigma^{\frac{5p^{2}-9p+6}{8p-4}}}{T^{\frac{5p-1}{8p-4}}} + \frac{16\|\nabla f(\boldsymbol{w}^{0})\|^{\frac{2-p}{2}}\sigma^{\frac{2p^{2}-5p+2}{4p-2}}}{T^{\frac{p}{4p-2}}},
\end{aligned} \tag{33}
$$

where we used $p + \frac{(p-1)^2}{4p-2} \geq \frac{p}{2p-1}$ as $p \in (1,2]$ and $0 \leq \sigma \leq 1$.

2). If $\sigma \leq 1$ and $\sigma^{\frac{2p-2}{2p-1}} T^{\frac{p}{2p-1}} \leq 1$, we have

$$1 - \theta = 1, \gamma = \frac{1}{L\sqrt{T}}, h = 2T^{\frac{p+4}{8p-4}} + 2\|\nabla f(\boldsymbol{w}^0)\|.$$

Based on these hyperparameters,

$$
\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{w}^t)\| &\leq \frac{L(f(\boldsymbol{w}^1) - \min f) + 5}{T^{\frac{1}{2}}} + \frac{2\|\nabla f(\boldsymbol{w}^0)\|}{T} + \frac{4}{T^{\frac{p^2}{2p-2} + \frac{(p-1)(p+4)}{8p-4}}} \\
&\quad + \frac{16\sigma^{\frac{p}{16}}}{T^{\frac{7p^2}{32p-32} - \frac{(2-p)(p+4)}{16p-8}}} + \frac{16\|\nabla f(\boldsymbol{w}^0)\|^{\frac{2-p}{2}}}{T^{\frac{p^2}{4p-4}}} \\
&\leq \frac{\sigma^{\frac{p}{16}}}{T^{\frac{p-1}{2p-1}}} + \frac{L(f(\boldsymbol{w}^1) - \min f) + 16\|\nabla f(\boldsymbol{w}^0)\|^{\frac{2-p}{2}} + 9}{T^{\frac{1}{2}}} + \frac{2\|\nabla f(\boldsymbol{w}^0)\|}{T},
\end{aligned}
\tag{34}
$$

because $\frac{p^2}{2p-2} + \frac{(p-1)(p+4)}{8p-4} \geq \frac{1}{2}$, $\frac{p^2}{4p-4} \geq \frac{1}{2}$, and $\frac{7p^2}{32p-32} - \frac{(2-p)(p+4)}{16p-8} \geq \frac{p-1}{2p-1}$ as $1 < p \leq 2$.

3). If $\sigma \geq 1$, we have $\sigma^{\frac{p}{2p-1}} T^{\frac{p}{2p-1}} \geq 1$, it then follows

$$1 - \theta = \frac{1}{\sigma^{\frac{p}{2p-1}} T^{\frac{p}{2p-1}}}, \gamma = \frac{1}{L\sigma^{\frac{p}{8p-4}} T^{\frac{5p-2}{8p-4}}}, h = 2T^{\frac{1}{2p-1}}\sigma^{\frac{2p}{2p-1}} + 2\|\nabla f(\boldsymbol{w}^0)\|.$$

We can rebound (32) as

$$
\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{w}^t)\| &\leq \frac{[L(f(\boldsymbol{w}^1) - \min f) + 5]\sigma^{\frac{p}{8p-4}}}{T^{\frac{3p-2}{8p-4}}} + \frac{2\|\nabla f(\boldsymbol{w}^0)\|\sigma^{\frac{p}{2p-1}}}{T^{\frac{p-1}{2p-1}}} + \frac{4\sigma^{\frac{p}{2p-1}}}{T^{\frac{p-1}{2p-1}}} \\
&\quad + \frac{16\sigma^{\frac{p}{2p-1}}}{T^{\frac{p-1}{2p-1}}} + \frac{16\|\nabla f(\boldsymbol{w}^0)\|^{\frac{2-p}{2}}\sigma^{\frac{p^2-p}{2p-1}}}{T^{\frac{p}{4p-2}}}.
\end{aligned}
\tag{35}
$$

The proof is completed by combining the bounds from equations (33), (34) and (34), along with the condition that $\boldsymbol{w}^0 = \boldsymbol{w}^1$.

Appendix F. Proof of Theorem 5

This section recruits the following notation

$$
\begin{aligned}
\boldsymbol{\epsilon}^t &:= \boldsymbol{m}^t - \nabla f(\boldsymbol{w}^t), \boldsymbol{\delta}^t := \mathrm{Clip}_h(\boldsymbol{g}^t) - \mathbb{E}\mathrm{Clip}_h(\boldsymbol{g}^t), \\
\boldsymbol{c}^t &:= \mathbb{E}\mathrm{Clip}_h(\boldsymbol{g}^t) - \nabla f(\boldsymbol{v}^t), \ \mathbf{D}(\boldsymbol{x}, \boldsymbol{y}) := \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}) - [\nabla^2 f(\boldsymbol{x})](\boldsymbol{x} - \boldsymbol{y}).
\end{aligned}
$$

If Assumption 3 holds, for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}$, the two-point operator satisfies

$$
\begin{aligned}
\|\mathbf{D}(\boldsymbol{x}, \boldsymbol{y})\| &= \|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}) - [\nabla^2 f(\boldsymbol{x})](\boldsymbol{x} - \boldsymbol{y})\| \\
&= \left\| \int_{s=0}^1 [\nabla^2 f(\boldsymbol{y} + (\boldsymbol{x} - \boldsymbol{y})s) - \nabla^2 f(\boldsymbol{x})](\boldsymbol{x} - \boldsymbol{y})ds \right\| \\
&\leq \int_{s=0}^1 \|\nabla^2 f(\boldsymbol{y} + (\boldsymbol{x} - \boldsymbol{y})s) - \nabla^2 f(\boldsymbol{x})\|_{\mathrm{op}} \|\boldsymbol{x} - \boldsymbol{y}\| ds \\
&\leq H \int_{s=0}^1 \|\boldsymbol{x} - \boldsymbol{y}\|^2 ds = \frac{H}{2} \|\boldsymbol{x} - \boldsymbol{y}\|^2.
\end{aligned}
$$

The scheme of the algorithm gives us

$$
\begin{aligned}
\boldsymbol{m}^t &= \theta \boldsymbol{m}^{t-1} + (1 - \theta)\mathrm{Clip}_h(\boldsymbol{g}^t) = \theta(\boldsymbol{\epsilon}^{t-1} + \nabla f(\boldsymbol{w}^{t-1})) + (1 - \theta)(\boldsymbol{\delta}^t + \boldsymbol{c}^t + \nabla f(\boldsymbol{v}^t)) \\
&= \theta(\boldsymbol{\epsilon}^{t-1} + \nabla f(\boldsymbol{w}^t) + [\nabla^2 f(\boldsymbol{w}^t)](\boldsymbol{w}^{t-1} - \boldsymbol{w}^t) + \mathbf{D}(\boldsymbol{w}^{t-1}, \boldsymbol{w}^t)) \\
&\quad + (1 - \theta)(\boldsymbol{\delta}^t + \boldsymbol{c}^t + \nabla f(\boldsymbol{w}^t) + [\nabla^2 f(\boldsymbol{w}^t)](\boldsymbol{v}^t - \boldsymbol{w}^t) + \mathbf{D}(\boldsymbol{v}^t, \boldsymbol{w}^t)) \\
&= \theta(\boldsymbol{\epsilon}^{t-1} + \nabla f(\boldsymbol{x}^t) + \mathbf{D}(\boldsymbol{w}^{t-1}, \boldsymbol{w}^t)) + (1 - \theta)(\boldsymbol{\delta}^t + \boldsymbol{c}^t + \nabla f(\boldsymbol{w}^t) + \mathbf{D}(\boldsymbol{v}^t, \boldsymbol{w}^t)),
\end{aligned}
$$

where we use the fact $\boldsymbol{v}^t - \boldsymbol{w}^t = \frac{\theta}{1-\theta}(\boldsymbol{w}^t - \boldsymbol{w}^{t-1})$. Based on the relation above, we further get

$$
\boldsymbol{\epsilon}^t = \boldsymbol{m}^t - \nabla f(\boldsymbol{x}^t) = \theta \boldsymbol{\epsilon}^{t-1} + \theta \mathbf{H}(\boldsymbol{w}^{t-1}, \boldsymbol{w}^t) + (1 - \theta)\mathbf{H}(\boldsymbol{v}^t, \boldsymbol{w}^t) + (1 - \theta)(\boldsymbol{\delta}^t + \boldsymbol{c}^t).
$$

Recursion of iteration above from 0 to t, we have

$$
\boldsymbol{\epsilon}^t = \theta \sum_{i=1}^{t-1} \theta^{t-i} \mathbf{H}(\boldsymbol{w}^{i-1}, \boldsymbol{w}^i) + (1 - \theta) \sum_{i=1}^{t-1} \theta^{t-i} \mathbf{H}(\boldsymbol{v}^i, \boldsymbol{w}^i) + (1 - \theta) \sum_{i=1}^{t-1} \theta^{t-i}(\boldsymbol{\delta}^i + \boldsymbol{c}^i) + \theta^t \boldsymbol{\epsilon}^0.
$$

Considering the norms and expectations of both sides of the equation above,

$$
\begin{aligned}
\mathbb{E}\|\boldsymbol{\epsilon}^t\| &\leq \theta \sum_{i=1}^{t-1} \theta^{t-i} \mathbb{E}\|\mathbf{H}(\boldsymbol{w}^{i-1}, \boldsymbol{w}^i)\| + (1 - \theta) \sum_{i=1}^{t-1} \theta^{t-i} \mathbb{E}\|\mathbf{H}(\boldsymbol{v}^i, \boldsymbol{w}^i)\| \\
&\quad + \mathbb{E}\|(1 - \theta) \sum_{i=1}^{t-1} \theta^{t-i}(\boldsymbol{\delta}^i + \boldsymbol{c}^i)\| + \theta^t \|\boldsymbol{\epsilon}^0\| \\
&\leq \frac{H}{2} \theta \sum_{i=1}^{t-1} \theta^{t-i} \mathbb{E}\|\boldsymbol{w}^{i-1} - \boldsymbol{w}^i\|^2 + \frac{H}{2}(1 - \theta) \sum_{i=1}^{t-1} \theta^{t-i} \mathbb{E}\|\boldsymbol{v}^i - \boldsymbol{w}^i\|^2 \\
&\quad + \mathbb{E}\|(1 - \theta) \sum_{i=1}^{t-1} \theta^{t-i}(\boldsymbol{\delta}^i + \boldsymbol{c}^i)\| + \theta^t \|\boldsymbol{\epsilon}^0\| \\
&\leq \frac{H}{2} \frac{\theta}{1 - \theta} \sum_{i=1}^{t-1} \theta^{t-i} \mathbb{E}\|\boldsymbol{w}^{i-1} - \boldsymbol{w}^i\|^2 + \mathbb{E}\|(1 - \theta) \sum_{i=1}^{t-1} \theta^{t-i}(\boldsymbol{\delta}^i + \boldsymbol{c}^i)\| + \theta^t \|\boldsymbol{\epsilon}^0\| \\
&\leq \frac{H}{2} \frac{\theta}{(1 - \theta)^2} \gamma^2 + \mathbb{E}\|(1 - \theta) \sum_{i=1}^{t-1} \theta^{t-i} \boldsymbol{\delta}^i\| + \mathbb{E}\|(1 - \theta) \sum_{i=1}^{t-1} \theta^{t-i} \boldsymbol{c}^i\| + \theta^t \|\boldsymbol{\epsilon}^0\|,
\end{aligned}
$$

where we used the gradient normalization scheme

$$\mathbb{E}\|\boldsymbol{w}^{i-1} - \boldsymbol{w}^i\|^2 \leq \gamma^2.$$

The bounds for $\mathbb{E}\|(1 - \theta) \sum_{i=1}^{t-1} \theta^{t-i} \boldsymbol{\delta}^i\|$ and $\mathbb{E}\|(1 - \theta) \sum_{i=1}^{t-1} \theta^{t-i} \boldsymbol{c}^i\|$ are analogous to that presented in Theorem 3. We then get

$$\mathbb{E}\|\boldsymbol{\epsilon}^t\| \leq \frac{H}{2} \frac{\theta}{(1 - \theta)^2} \gamma^2 + 2\sigma^p h^{-(p-1)} + 4\sqrt{1 - \theta} h^{\frac{2-p}{2}} \sigma^{p/2} + \theta^t \|\boldsymbol{\epsilon}^0\|.$$

Summing the inequality from $t = 1$ to T,

$$\sum_{t=1}^{T} \mathbb{E}\|\boldsymbol{\epsilon}^t\|/T \leq \frac{H}{2} \frac{\theta}{(1 - \theta)^2} \gamma^2 + 2\sigma^p h^{-(p-1)} + 4\sqrt{1 - \theta} h^{\frac{2-p}{2}} \sigma^{p/2} + \frac{\|\boldsymbol{\epsilon}^0\|}{(1 - \theta)T}.$$

Employing Lemma 9 with $\boldsymbol{w}^\dagger \to \boldsymbol{w}^t$ and $\boldsymbol{m} \to \boldsymbol{w}^t$,

$$f(\boldsymbol{w}^{t+1}) - f(\boldsymbol{w}^t) \leq -\gamma\|\nabla f(\boldsymbol{w}^t)\| + 2\gamma\|\boldsymbol{\epsilon}^t\| + \frac{L\gamma^2}{2}. \tag{36}$$

Summing the inequality from $t = 1$ to T,

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\|\nabla f(\boldsymbol{w}^t)\| \leq \frac{f(\boldsymbol{w}^1) - \min f}{\gamma T} + \frac{H\theta}{(1 - \theta)^2} \gamma^2$$
$$+ 4\sigma^p h^{-(p-1)} + 8\sqrt{1 - \theta} h^{\frac{2-p}{2}} \sigma^{p/2} + \frac{2\|\boldsymbol{\epsilon}^0\|}{(1 - \theta)T} + \frac{L\gamma}{2}. \tag{37}$$

1) As $0 \leq \sigma \leq 1$ and $\sigma^{3/7} T^{\frac{2p}{4p-1}} > 1$, and then

$$1 - \theta = \frac{1}{\sigma^{\frac{3}{7}} T^{\frac{2p}{4p-1}}}, \quad \gamma = \frac{1}{L\sigma^{\frac{2}{7}} T^{\frac{2p+1}{4p-1}}}, h = 2\Big(\frac{T^{\frac{2}{4p-1}}}{\sigma^{\frac{2}{7}}} + \|\nabla f(\boldsymbol{w}^0)\|\Big).$$

We then have the following rate as

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\|\nabla f(\boldsymbol{w}^t)\| \leq \frac{L\sigma^{\frac{2}{7}}(f(\boldsymbol{w}^1) - \min f)}{T^{\frac{2p-2}{4p-1}}} + \frac{H\sigma^{\frac{2}{7}}}{L^2 T^{\frac{2}{4p-1}}}$$
$$+ \frac{4\sigma^{p+\frac{2(p-1)}{7}}}{T^{\frac{2p-2}{4p-1}}} + 16\sqrt{1 - \theta}\Big(\frac{T^{\frac{2}{4p-1}}}{\sigma^{\frac{2}{7}}} + \|\nabla f(\boldsymbol{w}^0)\|\Big)^{\frac{2-p}{2}} \sigma^{p/2} + \frac{2\|\boldsymbol{\epsilon}^0\|\sigma^{\frac{3}{7}}}{T^{\frac{2p-1}{4p-1}}} + \frac{1}{2\sigma^{\frac{2}{7}} T^{\frac{2p+1}{4p-1}}}$$
$$\leq \frac{L\sigma^{\frac{2}{7}}(f(\boldsymbol{w}^1) - \min f)}{T^{\frac{2p-2}{4p-1}}} + \frac{H\sigma^{\frac{2}{7}}}{L^2 T^{\frac{2}{4p-1}}} + \frac{4\sigma^p}{T^{\frac{2p-2}{4p-1}}} + \frac{2\|\boldsymbol{\epsilon}^0\|\sigma^{\frac{3}{7}}}{T^{\frac{2p-1}{4p-1}}} + \frac{\sigma^{\frac{9-4p}{14p}}}{2T^{\frac{2p-2}{4p-1}}}$$
$$+ \frac{16\sigma^{\frac{9p-7}{14}}}{T^{\frac{2p-2}{4p-1}}} + \frac{16\|\nabla f(\boldsymbol{w}^0)\|^{\frac{2-p}{2}} \sigma^{\frac{7p-3}{14}}}{T^{\frac{p}{4p-1}}}. \tag{38}$$

Due to $0 \leq \sigma \leq 1$ and $1 < p \leq 2$, we have

$$\sigma^{p+\frac{2(p-1)}{7}} \leq \sigma^{\frac{2}{7}}, \sigma^{\frac{9p-7}{14}} \leq \sigma^{\frac{1}{7}}, \sigma^{\frac{9-4p}{14p}} \leq \sigma^{\frac{1}{28}}.$$

Thus, we are then led to

$$
\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{w}^t)\| &\leq \frac{L\sigma^{\frac{2}{7}}(f(\boldsymbol{w}^1)-\min f)}{T^{\frac{2p-2}{4p-1}}} + \frac{H\sigma^{\frac{2}{7}}}{L^2 T^{\frac{2}{4p-1}}} + \frac{20\sigma^{\frac{2}{7}}}{T^{\frac{2p-2}{4p-1}}} \\
&\quad + \frac{2\|\boldsymbol{\epsilon}^0\|\sigma^{\frac{3}{7}}}{T^{\frac{2p-1}{4p-1}}} + \frac{16\|\nabla f(\boldsymbol{w}^0)\|^{\frac{2-p}{2}}\sigma^{\frac{2}{7}}}{T^{\frac{p}{4p-1}}} + \frac{\sigma^{\frac{1}{28}}}{2T^{\frac{2p+1}{4p-1}}}.
\end{aligned}
\tag{39}
$$

2) When $0 \leq \sigma \leq 1$ and $\sigma^{3/7}T^{\frac{2p}{4p-1}} \leq 1$, we have

$$
1-\theta = 1, \ \gamma = \frac{1}{LT^{\frac{1}{2}}}, h = 2(T^{\frac{7}{8p-2}} + \|\nabla f(\boldsymbol{w}^0)\|).
$$

Recalling (37), we then get

$$
\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{w}^t)\| &\leq \frac{L(f(\boldsymbol{w}^1)-\min f)+1}{T^{1/2}} + 4T^{-\frac{4p^2+7p-7}{8p-2}} \\
&\quad + \frac{16\sigma^{\frac{p}{16}}}{T^{\frac{7(p+2)(p-1)}{16p-4}}} + 16\|\nabla f(\boldsymbol{w}^0)\|^{\frac{2-p}{2}}T^{-\frac{2p^2}{4p-1}} + \frac{2\|\boldsymbol{\epsilon}^0\|}{T} \\
&\leq \frac{L(f(\boldsymbol{w}^1)-\min f)+5+16\|\nabla f(\boldsymbol{w}^0)\|^{\frac{2-p}{2}}}{T^{1/2}} + \frac{16\sigma^{\frac{p}{16}}}{T^{\frac{2p-2}{4p-1}}} + \frac{2\|\boldsymbol{\epsilon}^0\|}{T}.
\end{aligned}
\tag{40}
$$

3) As $\sigma \geq 1$, we have $\sigma^{6/7}T^{\frac{2p}{4p-1}} > 1$ due to $T > 1$, and then

$$
1-\theta = \frac{1}{\sigma^{\frac{6}{7}}T^{\frac{2p}{4p-1}}}, \ \gamma = \frac{1}{L\sigma^{\frac{4}{7}}T^{\frac{2p+1}{4p-1}}}, h = 2\Big(T^{\frac{2}{4p-1}} \cdot \sigma^{\frac{7p-4}{7p-7}} + \|\nabla f(\boldsymbol{w}^0)\|\Big).
$$

We then have the following rate as

$$
\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{w}^t)\| &\leq \frac{L\sigma^{\frac{4}{7}}(f(\boldsymbol{w}^1)-\min f)}{T^{\frac{2p-2}{4p-1}}} + \frac{H\sigma^{\frac{4}{7}}}{L^2 T^{\frac{2}{4p-1}}} \\
&\quad + \frac{4\sigma^{\frac{4}{7}}}{T^{\frac{2p-2}{4p-1}}} + 16\sqrt{1-\theta}\Big(T^{\frac{2}{4p-1}} \cdot \sigma^{\frac{7p-4}{7p-7}} + \|\nabla f(\boldsymbol{w}^0)\|\Big)^{\frac{2-p}{2}}\sigma^{p/2} + \frac{2\|\boldsymbol{\epsilon}^0\|\sigma^{\frac{6}{7}}}{T^{\frac{2p-1}{4p-1}}} + \frac{1}{2\sigma^{\frac{4}{7}}T^{\frac{2p+1}{4p-1}}} \\
&\leq \frac{L\sigma^{\frac{4}{7}}(f(\boldsymbol{w}^1)-\min f)}{T^{\frac{2p-2}{4p-1}}} + \frac{H\sigma^{\frac{4}{7}}}{L^2 T^{\frac{2}{4p-1}}} + \frac{4\sigma^{\frac{4}{7}}}{T^{\frac{2p-2}{4p-1}}} + \frac{2\|\boldsymbol{\epsilon}^0\|\sigma^{\frac{6}{7}}}{T^{\frac{2p-1}{4p-1}}} + \frac{1}{2\sigma^{\frac{4}{7}}T^{\frac{2p+1}{4p-1}}} \\
&\quad + \frac{16\sigma^{\frac{p}{2}-\frac{3}{7}-\frac{(2-p)(7p-4)}{14p-14}}}{T^{\frac{2p-2}{4p-1}}} + \frac{16\|\nabla f(\boldsymbol{w}^0)\|^{\frac{2-p}{2}}\sigma^{\frac{p}{2}-\frac{3}{7}}}{T^{\frac{p}{4p-1}}}.
\end{aligned}
\tag{41}
$$

Using the fact $\sigma \geq 1$ and $1 < p \leq 2$, we have

$$
\frac{1}{\sigma^{\frac{4}{7}}} \leq \sigma^{\frac{4}{7}}, \sigma^{\frac{p}{2}-\frac{3}{7}-\frac{(2-p)(7p-4)}{14p-14}} \leq \sigma^{\frac{4}{7}}, \sigma^{\frac{p}{2}-\frac{3}{7}} \leq \sigma^{\frac{4}{7}}.
$$

34

Thus, we are then led to

$$
\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{w}^t)\| \leq{} & \frac{L\sigma^{\frac{4}{7}}(f(\boldsymbol{w}^1)-\min f)}{T^{\frac{2p-2}{4p-1}}}+\frac{H\sigma^{\frac{4}{7}}}{L^2 T^{\frac{2}{4p-1}}}+\frac{20\sigma^{\frac{4}{7}}}{T^{\frac{2p-2}{4p-1}}} \\
& +\frac{2\|\boldsymbol{\epsilon}^0\|\sigma^{\frac{6}{7}}}{T^{\frac{2p-1}{4p-1}}}+\frac{\sigma^{\frac{4}{7}}}{2T^{\frac{2p+1}{4p-1}}}+\frac{16\|\nabla f(\boldsymbol{w}^0)\|^{\frac{2-p}{2}}\sigma^{\frac{4}{7}}}{T^{\frac{p}{4p-1}}}.
\end{aligned}
\tag{42}
$$

We then proved the result by combining (39), (40) and (42).

Appendix G. Accelerated Normalized Stochastic Gradient Descent under Second-Order Smoothness

---

Algorithm 7 Accelerated Normalized Stochastic Gradient Descent under Second-Order Smoothness (A-NSGD)

Require: parameters $\gamma > 0$, $0 \leq \theta < 1$, $\zeta \geq 0$
  Initialization: $\boldsymbol{w}^0 = \boldsymbol{w}^1$, $\boldsymbol{m}^0 = \boldsymbol{0}$
  for $t = 1, 2, \ldots$
  step 1 $\boldsymbol{v}^t = \boldsymbol{w}^t + \zeta(\boldsymbol{w}^t - \boldsymbol{w}^{t-1})$
  step 2: Sample the data $\xi^t \sim \mathcal{D}$ and $\boldsymbol{m}^t = \theta\boldsymbol{m}^{t-1} + (1-\theta)\nabla f(\boldsymbol{v}^t; \xi^t)$
  step 3: $\boldsymbol{w}^{t+1} = \boldsymbol{w}^t - \gamma\frac{\boldsymbol{m}^t}{\|\boldsymbol{m}^t\|}$
  end for

The convergence result of nonconvex accelerated stochastic algorithm under heavy-tailed noise is presented as follows.

Proposition 10 *Let $(\boldsymbol{w}^t)_{t\geq 0}$ be generated by Algorithm 7. Suppose Assumption 1–3 hold. If $1-\theta = \min\left\{\frac{1}{\sigma^{\frac{1}{2}}T^{\frac{2p}{4p-1}}}, 1\right\}$, $\gamma = \frac{(1-\theta)^{\frac{3}{4p}}}{L\sqrt{T}}$, $\zeta = \frac{\theta}{1-\theta}$, then it holds that*

$$
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{w}^t)\| \leq \frac{L\sigma^{\frac{3}{8p}}(f(\boldsymbol{w}^1)-\min f)+\frac{H}{L^2}\sigma^{\frac{4p-3}{4p}}+4\left(\frac{B}{L}+1\right)^{\frac{2-p}{2}}(\sigma^{\frac{1}{5}}+\sigma^{\frac{1}{2}})+\sigma^{\frac{3}{8p}}}{T^{\frac{2p-2}{4p-1}}}
$$

$$
+\frac{\|\nabla f(\boldsymbol{w}^0)\|\sigma^{\frac{1}{2}}}{T^{\frac{2p-1}{4p-1}}}+\frac{L(f(\boldsymbol{w}^0)-\min f)+1+4\left(\frac{B}{L}+1\right)^{\frac{2-p}{2}}}{\sqrt{T}}+\frac{2\|\nabla f(\boldsymbol{w}^0)\|}{T}.
$$

Proof of Proposition 10: This proof employs the following notation

$$
\boldsymbol{g}^t := \nabla f(\boldsymbol{v}^t; \xi^t), \boldsymbol{\delta}^t := \nabla f(\boldsymbol{v}^t; \xi^t) - \nabla f(\boldsymbol{v}^t),
$$

$$
\boldsymbol{\epsilon}^t := \boldsymbol{m}^t - \nabla f(\boldsymbol{w}^t), \ \mathbf{D}(\boldsymbol{x}, \boldsymbol{y}) := \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}) - [\nabla^2 f(\boldsymbol{x})](\boldsymbol{x} - \boldsymbol{y}).
$$

Similar to the proof of Theorem 5, we can derive

$$
\mathbb{E}\|\boldsymbol{\epsilon}^t\| \leq \frac{H}{2}\frac{\theta}{(1-\theta)^2}\gamma^2 + \mathbb{E}\|(1-\theta)\sum_{i=1}^{t-1}\theta^{t-i}\boldsymbol{\delta}^i\| + \theta^t\|\boldsymbol{\epsilon}^0\|.
$$

35

The bound of $\mathbb{E}\|(1-\theta)\sum_{i=1}^{t-1}\theta^{t-i}\boldsymbol{\delta}^i\|$ follows the same of that in Theorem 1. Thus, we are led to the following bound

$$\mathbb{E}\|\boldsymbol{\epsilon}^t\| \le \frac{H}{2}\frac{\theta}{(1-\theta)^2}\gamma^2 + 2\sqrt{1-\theta}(B+L\gamma T)^{\frac{2-p}{2}}\sigma^{p/2} + \theta^t\|\boldsymbol{\epsilon}^0\|,$$

which yields

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\boldsymbol{\epsilon}^t\| \le \frac{H}{2}\frac{\theta}{(1-\theta)^2}\gamma^2 + 2\sqrt{1-\theta}(B+L\gamma T)^{\frac{2-p}{2}}\sigma^{p/2} + \frac{\|\boldsymbol{\epsilon}^0\|}{(1-\theta)T}.$$

Noticing (36) still holds, aggregating the inequality over the range from $t=1$ to T,

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{w}^t)\|$$
$$\le \frac{f(\boldsymbol{w}^1)-\min f}{\gamma T} + \frac{H\theta}{(1-\theta)^2}\gamma^2 + 4\sqrt{1-\theta}(B+L\gamma T)^{\frac{2-p}{2}}\sigma^{p/2} + \frac{L\gamma}{2} + \frac{2\|\boldsymbol{\epsilon}^0\|}{(1-\theta)T}$$
$$\le \frac{f(\boldsymbol{w}^1)-\min f}{\gamma T} + \frac{H\theta}{(1-\theta)^2}\gamma^2 + 4(B+L)^{\frac{2-p}{2}}\sqrt{1-\theta}(\gamma T)^{\frac{2-p}{2}}\sigma^{p/2} + \frac{L\gamma}{2} + \frac{2\|\boldsymbol{\epsilon}^0\|}{(1-\theta)T}.$$

1) If $\sigma^{1/2}T^{\frac{2p}{4p-1}} \ge 1$, we have

$$1-\theta = \frac{1}{\sigma^{\frac{1}{2}}T^{\frac{2p}{4p-1}}}, \quad \gamma = \frac{1}{L\sigma^{\frac{3}{8p}}T^{\frac{2p+1}{4p-1}}}.$$

The bound of rate then reduces to

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{w}^t)\|$$
$$\le \frac{L\sigma^{\frac{3}{8p}}(f(\boldsymbol{w}^1)-\min f)}{T^{\frac{2p-2}{4p-1}}} + \frac{H\sigma^{\frac{4p-3}{4p}}}{L^2T^{\frac{2}{4p-1}}} + 4\left(\frac{B}{L}+1\right)^{\frac{2-p}{2}}\frac{\sigma^{\frac{8p^2-11p+6}{16p}}}{T^{\frac{p^2-2p+2}{4p-1}}} + \frac{1}{2\sigma^{\frac{3}{8p}}T^{\frac{2p+1}{4p-1}}} + \frac{\|\boldsymbol{\epsilon}^0\|\sigma^{\frac{1}{2}}}{T^{\frac{2p-1}{4p-1}}}$$
$$\le \frac{L\sigma^{\frac{3}{8p}}(f(\boldsymbol{w}^1)-\min f)}{T^{\frac{2p-2}{4p-1}}} + \frac{H\sigma^{\frac{4p-3}{4p}}}{L^2T^{\frac{2p-2}{4p-1}}} + 4\left(\frac{B}{L}+1\right)^{\frac{2-p}{2}}\frac{(\sigma^{1/5}+\sigma^{1/2})}{T^{\frac{2p-2}{4p-1}}} + \frac{\sigma^{\frac{3}{8p}}}{2T^{\frac{2p-2}{4p-1}}} + \frac{\|\boldsymbol{\epsilon}^0\|\sigma^{\frac{1}{2}}}{T^{\frac{2p-1}{4p-1}}}, \tag{43}$$

where we used $\sigma^{\frac{8p^2-11p+6}{16p}} \le \sigma^{1/5}+\sigma^{1/2}$ as $p \in (1,2]$.

2) If $\sigma^{1/2}T^{\frac{2p}{4p-1}} \le 1$, we have

$$1-\theta = 1, \quad \gamma = \frac{1}{L\sqrt{T}}.$$

Further, we are led to

$$
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{w}^t)\| \leq \frac{L(f(\boldsymbol{w}^1)-\min f)+1}{\sqrt{T}} + 4\left(\frac{B}{L}+1\right)^{\frac{2-p}{2}}T^{\frac{2-p}{4}}\sigma^{p/2} + \frac{2\|\boldsymbol{\epsilon}^0\|}{T}
$$

$$
\leq \frac{L(f(\boldsymbol{w}^1)-\min f)+1}{\sqrt{T}} + 4\left(\frac{B}{L}+1\right)^{\frac{2-p}{2}}T^{\frac{2-p}{4}-\frac{p(8-2p)}{11-2p}} + \frac{2\|\boldsymbol{\epsilon}^0\|}{T} \qquad (44)
$$

$$
\leq \frac{L(f(\boldsymbol{w}^0)-\min f)+1}{\sqrt{T}} + \frac{4\left(\frac{B}{L}+1\right)^{\frac{2-p}{2}}}{\sqrt{T}} + \frac{2\|\boldsymbol{\epsilon}^0\|}{T},
$$

where we used that

$$
\frac{p(8-2p)}{11-2p} - \frac{2-p}{4} \geq \frac{1}{2}
$$

as $p \in (1,2]$, and $\boldsymbol{w}^0 = \boldsymbol{w}^1$.
Combing the bounds (43) and (44), we have

$$
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{w}^t)\| \leq \frac{L\sigma^{\frac{3}{8p}}(f(\boldsymbol{w}^1)-\min f)+\frac{H}{L^2}\sigma^{\frac{4p-3}{4p}}+4\left(\frac{B}{L}+1\right)^{\frac{2-p}{2}}(\sigma^{\frac{1}{5}}+\sigma^{\frac{1}{2}})+\sigma^{\frac{3}{8p}}}{T^{\frac{2p-2}{4p-1}}}
$$

$$
+ \frac{\|\boldsymbol{\epsilon}^0\|\sigma^{\frac{1}{2}}}{T^{\frac{2p-1}{4p-1}}} + \frac{L(f(\boldsymbol{w}^0)-\min f)+1}{\sqrt{T}} + \frac{4\left(\frac{B}{L}+1\right)^{\frac{2-p}{2}}}{\sqrt{T}} + \frac{2\|\boldsymbol{\epsilon}^0\|}{T}.
$$

Appendix H. Proofs of Technical Lemmas

H.1 Proof of Lemma 7

From the scheme of the algorithms, we have

$$
\|\boldsymbol{w}^{t+1}-\boldsymbol{w}^0\| = \left\|\boldsymbol{w}^t - \gamma\frac{\boldsymbol{m}^t}{\|\boldsymbol{m}^t\|} - \boldsymbol{w}^0\right\| \leq \gamma + \|\boldsymbol{w}^t - \boldsymbol{w}^0\|.
$$

For any $t \in \mathbb{Z}^+$, as $1 \leq t \leq T$, it holds

$$
\|\boldsymbol{w}^t - \boldsymbol{w}^0\| \leq \gamma T. \qquad (45)
$$

The Lipschitz continuity of the gradient indicates

$$
\|\nabla f(\boldsymbol{w}^t)\| \leq \|\nabla f(\boldsymbol{w}^0)\| + \|\nabla f(\boldsymbol{w}^t)-\nabla f(\boldsymbol{w}^0)\|
$$
$$
\leq \|\nabla f(\boldsymbol{w}^0)\| + L\|\boldsymbol{w}^t - \boldsymbol{w}^0\| \leq \|\nabla f(\boldsymbol{w}^0)\| + L\gamma T
$$

as $1 \leq t \leq T$. If $h \geq 2(\|\nabla f(\boldsymbol{w}^0)\| + L\gamma T)$, we then have $\|\nabla f(\boldsymbol{w}^t)\| \leq \frac{h}{2}$. According to [Lemma 5, (Liu et al., 2023)], we further get the following result

$$
\mathbb{E}\|\mathrm{Clip}_h(\boldsymbol{g}^t)-\mathbb{E}\mathrm{Clip}_h(\boldsymbol{g}^t)\|^2 \leq \mathbb{E}\|\mathrm{Clip}_h(\boldsymbol{g}^t)\|^2 \leq 10\sigma^p h^{2-p},
$$
$$
\|\mathbb{E}\mathrm{Clip}_h(\boldsymbol{g}^t)-\nabla f(\boldsymbol{w}^t)\|^2 \leq 4\sigma^{2p}h^{-2(p-1)}.
$$

H.2 Proof of Lemma 8

Noticing that property (45) still holds, the individual Lipschitz continuity of the gradient then yields

$$\|\nabla f(\boldsymbol{w}^t; \xi^t)\| \le \|\nabla f(\boldsymbol{w}^0; \xi^t)\| + \|\nabla f(\boldsymbol{w}^t; \xi^t) - \nabla f(\boldsymbol{w}^0; \xi^t)\|$$
$$\le \|\nabla f(\boldsymbol{w}^0; \xi^t)\| + L\|\boldsymbol{w}^t - \boldsymbol{w}^0\| \le B + L\gamma T.$$

Noticing $\|\nabla f(\boldsymbol{w}^t)\| = \|\mathbb{E}_{\xi^t \sim \mathcal{D}} \nabla f(\boldsymbol{w}^t; \xi^t)\| \le \mathbb{E}_{\xi^t \sim \mathcal{D}} \|\nabla f(\boldsymbol{w}^t; \xi^t)\|$, we then derive

$$\|\nabla f(\boldsymbol{w}^t)\| \le B + L\gamma T$$

as $1 \le t \le T$. Direct computations give us

$$\|\nabla f(\boldsymbol{w}^t; \xi^t) - \nabla f(\boldsymbol{w}^t)\|^2 = \|\nabla f(\boldsymbol{w}^t; \xi^t) - \nabla f(\boldsymbol{w}^t)\|^{2-p} \cdot \|\nabla f(\boldsymbol{w}^t; \xi^t) - \nabla f(\boldsymbol{w}^t)\|^p$$
$$\le (\|\nabla f(\boldsymbol{w}^t; \xi^t)\| + \|\nabla f(\boldsymbol{w}^t)\|)^{2-p} \cdot \|\nabla f(\boldsymbol{w}^t; \xi^t) - \nabla f(\boldsymbol{w}^t)\|^p$$
$$\le (2(B + L\gamma T))^{2-p} \cdot \|\nabla f(\boldsymbol{w}^t; \xi^t) - \nabla f(\boldsymbol{w}^t)\|^p$$
$$\le 4(B + L\gamma T)^{2-p} \|\nabla f(\boldsymbol{w}^t; \xi^t) - \nabla f(\boldsymbol{w}^t)\|^p,$$

where $0 \le 2 - p \le 2$. Taking expectations of both sides, we can get

$$\mathbb{E}_{\xi^t \sim \mathcal{D}} \|\nabla f(\boldsymbol{w}^t; \xi^t) - \nabla f(\boldsymbol{w}^t)\|^2 \le 4(B + L\gamma T)^{2-p} \mathbb{E}_{\xi^t \sim \mathcal{D}} \|\nabla f(\boldsymbol{w}^t; \xi^t) - \nabla f(\boldsymbol{w}^t)\|^p$$
$$\le 4(B + L\gamma T)^{2-p} \sigma^p.$$

H.3 Proof of Lemma 9

The proof of this lemma can be found in (Cutkosky and Mehta, 2020b; Zhao et al., 2021) and is presented here for completeness. The gradient Lipschitz property can derive the following result

$$f(\boldsymbol{w}^\ddagger) - f(\boldsymbol{w}^\dagger) \le \langle \nabla f(\boldsymbol{w}^\dagger), \boldsymbol{w}^\ddagger - \boldsymbol{w}^\dagger \rangle + \frac{L}{2}\|\boldsymbol{w}^\ddagger - \boldsymbol{w}^\dagger\|^2$$
$$\le -\gamma \langle \nabla f(\boldsymbol{w}^\dagger), \frac{\boldsymbol{m}}{\|\boldsymbol{m}\|} \rangle + \frac{L\gamma^2}{2}$$
$$\overset{a)}{=} -\gamma\|\boldsymbol{m}\| + \gamma \langle \boldsymbol{m} - \nabla f(\boldsymbol{w}^\dagger), \frac{\boldsymbol{m}}{\|\boldsymbol{m}\|} \rangle + \frac{L\gamma^2}{2}$$
$$\overset{b)}{\le} -\gamma\|\nabla f(\boldsymbol{w}^\dagger)\| + 2\gamma\|\boldsymbol{m} - \nabla f(\boldsymbol{w}^\dagger)\| + \frac{L\gamma^2}{2},$$

where $a)$ comes from the following fact

$$-\langle \nabla f(\boldsymbol{w}^\dagger), \frac{\boldsymbol{m}}{\|\boldsymbol{m}\|} \rangle = -\langle \boldsymbol{m}, \frac{\boldsymbol{m}}{\|\boldsymbol{m}\|} \rangle - \langle \nabla f(\boldsymbol{w}^\dagger) - \boldsymbol{m}, \frac{\boldsymbol{m}}{\|\boldsymbol{m}\|} \rangle$$
$$= -\|\boldsymbol{m}\| + \langle \boldsymbol{m} - \nabla f(\boldsymbol{w}^\dagger), \frac{\boldsymbol{m}}{\|\boldsymbol{m}\|} \rangle$$
$$\le -\|\boldsymbol{m}\| + \|\boldsymbol{m} - \nabla f(\boldsymbol{w}^\dagger)\|,$$

and $b)$ is because

$$-\|\boldsymbol{m}\| \le -\|\nabla f(\boldsymbol{w}^\dagger)\| + \|\nabla f(\boldsymbol{w}^\dagger) - \boldsymbol{m}\|.$$

References

Naman Agarwal, Zeyuan Allen-Zhu, Brian Bullins, Elad Hazan, and Tengyu Ma. Finding approximate local minima faster than gradient descent. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1195–1199, 2017.

Zeyuan Allen-Zhu. Natasha 2: Faster non-convex optimization than sgd. *Advances in neural information processing systems*, 31, 2018.

Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1-2):165–214, 2023.

Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Accelerated methods for nonconvex optimization. *SIAM Journal on Optimization*, 28(2):1751–1772, 2018.

Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points i. *Mathematical Programming*, 184(1):71–120, 2020.

A. Cutkosky and H. Mehta. Momentum improves normalized SGD. In *International Conference on Machine Learning*. PMLR, 2020a.

Ashok Cutkosky and Harsh Mehta. Momentum improves normalized SGD. In *International conference on machine learning*, pages 2260–2268. PMLR, 2020b.

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

Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in nonconvex sgd. *Advances in neural information processing systems*, 32, 2019.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in neural information processing systems*, 27, 2014.

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal nonconvex optimization via stochastic path-integrated differential estimator. *Advances in neural information processing systems*, 31, 2018.

Cong Fang, Zhouchen Lin, and Tong Zhang. Sharp analysis for nonconvex sgd escaping from saddle points. In *Conference on Learning Theory*, pages 1192–1234. PMLR, 2019.

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

Eduard Gorbunov, Marina Danilova, and Alexander Gasnikov. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. *Advances in Neural Information Processing Systems*, 33:15042–15053, 2020.

Florian Hübler, Ilyas Fatkhullin, and Niao He. From gradient clipping to normalization for heavy tailed sgd. *arXiv preprint arXiv:2410.13849*, 2024.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26, 2013.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017.

Frederik Kunstner, Robin Yadav, Alan Milligan, Mark Schmidt, and Alberto Bietti. Heavy-tailed class imbalance and why adam outperforms gradient descent on language models. *arXiv preprint arXiv:2402.19449*, 2024.

Zhize Li, Hongyan Bao, Xiangliang Zhang, and Peter Richtárik. Page: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In *International conference on machine learning*, pages 6286–6295. PMLR, 2021.

Zijian Liu and Zhengyuan Zhou. Nonconvex stochastic optimization under heavy-tailed noises: Optimal convergence without gradient clipping. *arXiv preprint arXiv:2412.19529*, 2024.

Zijian Liu, Jiawei Zhang, and Zhengyuan Zhou. Breaking the lower bound with (little) structure: Acceleration in non-convex stochastic optimization with heavy-tailed noise. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2266–2290. PMLR, 2023.

Ta Duy Nguyen, Thien H Nguyen, Alina Ene, and Huy Nguyen. Improved convergence in high probability of clipped gradient methods with heavy tailed noise. *Advances in Neural Information Processing Systems*, 36:24191–24222, 2023.

Thanh Huy Nguyen, Umut Simsekli, Mert Gurbuzbalaban, and Gaël Richard. First exit time analysis of stochastic gradient descent under heavy-tailed gradient noise. *Advances in neural information processing systems*, 32, 2019.

Nikita Puchkin, Eduard Gorbunov, Nickolay Kutuzov, and Alexander Gasnikov. Breaking the heavy-tailed noise barrier in stochastic optimization problems. In *International Conference on Artificial Intelligence and Statistics*, pages 856–864. PMLR, 2024.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

Abdurakhmon Sadiev, Marina Danilova, Eduard Gorbunov, Samuel Horváth, Gauthier Gidel, Pavel Dvurechensky, Alexander Gasnikov, and Peter Richtárik. High-probability bounds for stochastic optimization and variational inequalities: the case of unbounded variance. In *International Conference on Machine Learning*, pages 29563–29648. PMLR, 2023.

Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14(1), 2013.

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*, pages 5827–5837. PMLR, 2019.

Tao Sun, Congliang Chen, Peng Qiao, Li Shen, Xinwang Liu, and Dongsheng Li. Rethinking sign training: Provable nonconvex acceleration without first-and second-order gradient lipschitz. *arXiv preprint arXiv:2310.14616*, 2023a.

Tao Sun, Qingsong Wang, Dongsheng Li, and Bao Wang. Momentum ensures convergence of signsgd under weaker assumptions. *ICML*, 2023b.

Tao Sun, Li Shen, and Xinwang Liu. On nonconvex sgd under unbounded noise with weak gradient lipschitz and delayed stochastic gradient. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Quoc Tran-Dinh, Nhan H Pham, Dzung T Phan, and Lam M Nguyen. Hybrid stochastic gradient descent algorithms for stochastic nonconvex optimization. *arXiv preprint arXiv:1905.05920*, 2019.

Nilesh Tripuraneni, Mitchell Stern, Chi Jin, Jeffrey Regier, and Michael I Jordan. Stochastic cubic regularization for fast nonconvex optimization. *Advances in neural information processing systems*, 31, 2018.

Nuri Mert Vural, Lu Yu, Krishna Balasubramanian, Stanislav Volgushev, and Murat A Erdogdu. Mirror descent strikes again: Optimal stochastic convex optimization under infinite noise variance. In *Conference on Learning Theory*, pages 65–102. PMLR, 2022.

Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? *Advances in Neural Information Processing Systems*, 33:15383–15393, 2020.

Shen-Yi Zhao, Yin-Peng Xie, and Wu-Jun Li. On the convergence and improvement of stochastic normalized gradient descent. *Science China Information Sciences*, 64:1–13, 2021.

Dongruo Zhou and Quanquan Gu. Lower bounds for smooth nonconvex finite-sum optimization. In *International Conference on Machine Learning*, pages 7574–7583. PMLR, 2019.

Dongruo Zhou and Quanquan Gu. Stochastic recursive variance-reduced cubic regularization methods. In *International conference on artificial intelligence and statistics*, pages 3980–3990. PMLR, 2020.

Dongruo Zhou, Pan Xu, and Quanquan Gu. Finding local minima via stochastic nested variance reduction. *arXiv preprint arXiv:1806.08782*, 2018a.

Dongruo Zhou, Pan Xu, and Quanquan Gu. Stochastic variance-reduced cubic regularized newton methods. In *International Conference on Machine Learning*, pages 5990–5999. PMLR, 2018b.

Dongruo Zhou, Pan Xu, and Quanquan Gu. Stochastic nested variance reduction for non-convex optimization. *Journal of machine learning research*, 21(103):1–63, 2020.