Complexity of normalized stochastic first-order methods with momentum under heavy-tailed noise

Chuan He* Zhaosong Lu† Defeng Sun‡ Zhanwang Deng§

June 12, 2025

Abstract

In this paper, we propose practical normalized stochastic first-order methods with Polyak momentum, multi-extrapolated momentum, and recursive momentum for solving unconstrained optimization problems. These methods employ dynamically updated algorithmic parameters and do not require explicit knowledge of problem-dependent quantities such as the Lipschitz constant or noise bound. We establish first-order oracle complexity results for finding approximate stochastic stationary points under heavy-tailed noise and weakly average smoothness conditions—both of which are weaker than the commonly used bounded variance and mean-squared smoothness assumptions. Our complexity bounds either improve upon or match the best-known results in the literature. Numerical experiments are presented to demonstrate the practical effectiveness of the proposed methods.

Keywords: Stochastic first-order methods, momentum, heavy-tailed noise, first-order oracle complexity

Mathematics Subject Classification: 49M05, 49M37, 90C25, 90C30

1 Introduction

In this paper, we consider the smooth unconstrained optimization problem:

$$\min_{x\in\mathbb{R}^n} f(x), \tag{1}$$

where $f:\mathbb{R}^n\to\mathbb{R}$ is continuously differentiable. We assume that problem (1) has at least one optimal solution. In many emerging applications—particularly in machine learning and related fields—instances of (1) are often large- or even huge-scale, which poses significant challenges to classical first-order methods due to the high cost of computing the exact gradient of f. To address this issue, stochastic first-order methods (SFOMs) have been extensively studied, as they employ stochastic estimators of the gradient that are typically much cheaper to compute. The goal of this paper is to propose practical

*Department of Mathematics, Linköping University, Sweden (email: chuan.he@liu.se). The work of Chuan He was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

†Department of Industrial and Systems Engineering, University of Minnesota, USA (email: zhaosong@umn.edu).

‡Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, People's Republic of China (email: defeng.sun@polyu.edu.hk). The research of Defeng Sun was partially supported by the Research Center for Intelligent Operations Research at The Hong Kong Polytechnic University (P0051214).

§Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, People's Republic of China (email: dzw_opt2022@stu.pku.edu.cn).

normalized stochastic first-order methods with momentum for solving problem (1), and to analyze their complexity under *heavy-tailed noise*.

Recently, a variety of SFOMs (e.g., [1, 3, 5, 6, 7, 8, 9, 15, 17, 23]) have been developed for solving problem (1) where the stochastic gradient $G(\cdot;\xi)$ is an unbiased estimator of $\nabla f(\cdot)$ with bounded variance, i.e., $G(\cdot;\xi)$ satisfies the conditions:

$$\mathbb{E}[G(x;\xi)] = \nabla f(x), \quad \mathbb{E}[\|G(x;\xi) - \nabla f(x)\|^2] \le \sigma^2 \qquad \forall x \in \mathbb{R}^n \tag{2}$$

for some $\sigma > 0$. The complexity of these SFOMs has been extensively studied under various smoothness conditions. In particular, under the assumption that ∇f is Lipschitz continuous, the methods proposed in [3, 8, 9] achieve a first-order oracle complexity of $\mathcal{O}(\epsilon^{-4})$ for finding an *ϵ-stochastic stationary point* (SSP) of (1), i.e., a point x satisfying

$$\mathbb{E}[\|\nabla f(x)\|] \le \epsilon.$$

In addition, assuming that $\nabla^2 f$ is Lipschitz continuous, the method in [3] achieves a first-order oracle complexity of $\mathcal{O}(\epsilon^{-7/2})$ for finding an ϵ-SSP of (1). Furthermore, several variance-reduced methods [5, 6, 18] have been shown to achieve a first-order oracle complexity of $\mathcal{O}(\epsilon^{-3})$ for finding an ϵ-SSP of (1), under the assumption that the stochastic gradient estimator $G(\cdot;\xi)$ satisfies a mean-squared smoothness condition.

Despite extensive studies on SFOMs under the bounded variance assumption (2), recent empirical findings in machine learning (e.g., [10, 31, 32, 35]) suggest that in practice the stochastic gradient estimator $G(\cdot;\xi)$ is typically unbiased but exhibits a bounded αth central moment for some $\alpha \in (1, 2]$, rather than bounded variance. Specifically, $G(\cdot;\xi)$ satisfies

$$\mathbb{E}[G(x;\xi)] = \nabla f(x), \quad \mathbb{E}[\|G(x;\xi) - \nabla f(x)\|^\alpha] \le \sigma^\alpha \qquad \forall x \in \mathbb{R}^n$$

for some $\sigma > 0$, which is commonly referred to as the *heavy-tailed noise regime*. The convergence behavior and analysis of SFOMs under this condition differ significantly from those based on the bounded variance assumption (2), as vanilla stochastic gradient methods can diverge when $\alpha \in (1, 2)$; see [35, Remark 1]. To address this divergence, gradient clipping has been widely adopted in algorithm design (e.g., [4, 20, 21, 24, 28, 35]). This approach replaces the stochastic gradient estimator with

$$\overline{G}_\tau(x;\xi) := \min\left\{1, \frac{\tau}{\|G(x;\xi)\|}\right\} G(x;\xi)$$

for some suitable clipping threshold $\tau > 0$. Numerous recent works have analyzed the complexity of SFOMs with gradient clipping for solving (1) under heavy-tailed noise (e.g., [4, 21, 24, 35]). Specifically, assuming that ∇f is Lipschitz continuous, [35] established a first-order oracle complexity of $\mathcal{O}(\epsilon^{-(3\alpha-2)/(\alpha-1)})$ for finding an ϵ-SSP of (1), while [4, 24] established a first-order oracle complexity of $\widetilde{\mathcal{O}}(\epsilon^{-(3\alpha-2)/(\alpha-1)})$ for finding a point x satisfying $\|\nabla f(x)\| \le \epsilon$ with high probability. Furthermore, under the additional assumption that $\nabla^2 f$ is Lipschitz continuous, [4] has shown that the complexity bound can be improved to $\widetilde{\mathcal{O}}(\epsilon^{-(5\alpha-3)/(2\alpha-2)})$. In addition, assuming that $G(\cdot;\xi)$ is almost surely Lipschitz continuous for all $\xi \in \Xi$ with the same Lipschitz constant, namely,

$$\|G(y;\xi) - G(x;\xi)\| \le L\|y - x\| \qquad \forall x, y \in \mathbb{R}^n, \text{ a.s. } \forall \xi \in \Xi \tag{3}$$

for some $L > 0$, where Ξ is the sample space of ξ, [21] established a first-order oracle complexity of $\widetilde{\mathcal{O}}(\epsilon^{-(2\alpha-1)/(\alpha-1)})$ for finding a point x satisfying $\|\nabla f(x)\| \le \epsilon$ with high probability.

While the aforementioned SFOMs with gradient clipping provide provable convergence guarantees, they suffer from several practical limitations. In particular, these methods require sufficiently large clipping thresholds τ to ensure theoretical convergence, whereas relatively small thresholds are typically used in practice, especially in the training of deep neural networks (see, e.g., [34]). Furthermore, setting an appropriate value for τ often depends on explicit knowledge of problem-dependent parameters, such as Lipschitz constants and noise bounds, which are generally unavailable or difficult to estimate in real-world applications. A more detailed discussion of these drawbacks can be found in [13].

Several recent studies have sought to avoid gradient clipping and have shown that, assuming ∇f is Lipschitz continuous, normalized SFOMs without gradient clipping can achieve a first-order oracle complexity of $\mathcal{O}(\epsilon^{-(3\alpha-2)/(\alpha-1)})$ for finding an ϵ-SSP of (1) under heavy-tailed noise [13, 22, 33]. Moreover, it has been shown in [13, 22] that this complexity bound becomes $\mathcal{O}(\epsilon^{-2\alpha/(\alpha-1)})$ if the tail exponent α is unknown. In addition, under the assumption that $G(\cdot;\xi)$ is almost surely Lipschitz continuous with a common Lipschitz constant, as described in (3), [33] demonstrated that the complexity bound of normalized SFOMs can be further improved to $\mathcal{O}(\epsilon^{-(2\alpha-1)/(\alpha-1)})$. While these methods successfully avoid gradient clipping and achieve complexity bounds comparable to those of SFOMs that rely on it, they still suffer from several practical limitations. In particular, the methods in [22, 33] require explicit knowledge of problem-specific quantities, such as Lipschitz constant and noise bound, in order to properly set algorithmic parameters like step sizes and momentum coefficients. In contrast, [13] proposed a fully parameter-free SFOM and established a first-order complexity result under the assumption that ∇f is Lipschitz continuous. However, the achieved complexity is significantly worse than the best-known results when the tail exponent α is known. In addition, the almost sure Lipschitz continuity assumption (3) may be restrictive in practice, as it is stronger than the commonly used assumption of Lipschitz continuity in expectation, such as the mean-squared smoothness condition often imposed in the bounded variance setting (see [5, 6, 18]).

Despite significant recent advances, existing SFOMs for solving problem (1) under the heavy-tailed noise regime still face the practical limitations discussed above. These limitations motivate the following open questions:

- Can we develop practical SFOMs that do not require explicit knowledge of problem-specific quantities, while still achieving the best-known complexity?
- Can we design practical SFOMs that achieve improved complexity under higher-order smoothness condition on f?
- Can we develop practical SFOMs under weaker conditions than the commonly used mean-squared smoothness condition on $G(\cdot;\xi)$?

In this paper, we address these questions by proposing three practical SFOMs that use dynamically updated algorithmic parameters, without requiring explicit knowledge of the Lipschitz constant or noise bounds. We establish first-order oracle complexity results for these methods in finding an ϵ-SSP of (1) under the heavy-tailed noise regime. Moreover, we show that two of our proposed methods achieve improved complexity under higher-order smoothness and a weakly average smoothness condition, respectively. Our main contributions are summarized below.

- We propose a practical *normalized SFOM with Polyak momentum* for solving problem (1), and show that it achieves the best-known complexity for finding an ϵ-SSP under heavy-tailed noise and the Lipschitz smoothness condition on f.

- We propose a practical *normalized SFOM with multi-extrapolated momentum* for solving problem (1), and establish a new complexity for finding an ϵ-SSP under heavy-tailed noise and a higher-order smoothness condition on f. To the best of our knowledge, this is the first SFOM that leverages higher-order smoothness of f to achieve acceleration. The resulting complexity significantly improves upon the best-known results under the standard Lipschitz smoothness condition.

- We develop a practical *normalized SFOM with recursive momentum* for solving problem (1), and show that it achieves a new complexity for finding an ϵ-SSP under heavy-tailed noise and a weakly average smoothness condition on $G(\cdot;\xi)$. This complexity generalizes existing complexity results under the mean-squared smoothness condition.

The rest of this paper is organized as follows. In Section 2, we introduce the notation and assumptions used throughout the paper. In Section 3, we propose normalized SFOMs with momentum and establish complexity bounds for them. Section 4 presents preliminary numerical results. In Section 5, we provide the proofs of the main results.

2 Notation and assumptions

Throughout this paper, we use $\mathbb{R}^n$ to denote the n-dimensional Euclidean space and $\langle\cdot,\cdot\rangle$ to represent the standard inner product. We use $\|\cdot\|$ to denote the Euclidean norm for vectors and the spectral norm for matrices. For any positive integer p and a pth-order continuously differentiable function φ, we denote by $\mathcal{D}^p\varphi(x)[h_1,\ldots,h_p]$ the pth-order directional derivative of φ at x along $h_i\in\mathbb{R}^n$, $1\le i\le p$, and by $\mathcal{D}^p\varphi(x)[\cdot]$ the associated symmetric p-linear form. For any symmetric p-linear form $\mathcal{T}[\cdot]$, we define its norm as

$$\|\mathcal{T}\| := \max_{h_1,\ldots,h_p}\{\mathcal{T}[h_1,\ldots,h_p] : \|h_i\|\le 1, 1\le i\le p\}. \tag{4}$$

For any $x\in\mathbb{R}^n$ and $h_i\in\mathbb{R}^n$ with $1\le i\le p-1$, we define $\nabla^p\varphi(x)(h_1,\ldots,h_{p-1})\in\mathbb{R}^n$ by

$$\langle\nabla^p\varphi(x)(h_1,\ldots,h_{p-1}),h_p\rangle := \mathcal{D}^p\varphi(x)[h_1,\ldots,h_p] \qquad \forall h_p\in\mathbb{R}^n.$$

For any $x,h\in\mathbb{R}^n$, we denote $\mathcal{D}^p\varphi(x)[h]^p := \mathcal{D}^p\varphi(x)[h,\ldots,h]$ and $\nabla^p\varphi(x)(h)^{p-1} := \nabla^p\varphi(x)(h,\ldots,h)$. For any $s\in\mathbb{R}$, we let $\mathrm{sgn}(s)$ be 1 if $s\ge 0$ and -1 otherwise. For any positive integer p, we define the residual of the pth-order Taylor expansion of ∇f as:

$$\mathcal{R}_p(y,x) := \nabla f(y) - \sum_{r=1}^{p}\frac{1}{(r-1)!}\nabla^r f(x)(y-x)^{r-1} \qquad \forall x,y\in\mathbb{R}^n. \tag{5}$$

In addition, we use $\widetilde{\mathcal{O}}(\cdot)$ to denote $\mathcal{O}(\cdot)$ with logarithmic terms omitted.

We now make the following assumption throughout this paper.

Assumption 1. (a) *There exists a finite $f_{\rm low}$ such that $f(x)\ge f_{\rm low}$ for all $x\in\mathbb{R}^n$.*

(b) *There exists $L_1>0$ such that $\|\nabla f(x)-\nabla f(y)\|\le L_1\|x-y\|$ for all $x,y\in\mathbb{R}^n$.*

(c) *The stochastic gradient estimator $G:\mathbb{R}^n\times\Xi\to\mathbb{R}^n$ satisfies*

$$\mathbb{E}[G(x;\xi)] = \nabla f(x), \quad \mathbb{E}[\|G(x;\xi)-\nabla f(x)\|^\alpha]\le\sigma^\alpha \qquad \forall x\in\mathbb{R}^n$$

for some $\sigma>0$ and $\alpha\in(1,2]$.

We next make some remarks on Assumption 1.

Remark 1. Assumptions 1(a) and 1(b) are standard. In particular, Assumption 1(b) implies that

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L_1}{2}\|y - x\|^2 \qquad \forall x, y \in \mathbb{R}^n. \tag{6}$$

Assumption 1(c) states that $G(\cdot;\xi)$ is an unbiased estimator of $\nabla f(\cdot)$, and its αth central moment is uniformly bounded. This is weaker than the commonly used variance bounded assumption corresponding to the case $\alpha = 2$. When $\alpha \in (1, 2)$, the stochastic gradient noise is said to be heavy-tailed (see, e.g., [35]), a phenomenon frequently observed in modern machine learning applications.

3 Normalized stochastic first-order methods with momentum

In this section, we propose practical normalized SFOMs with Polyak momentum, multi-extrapolated momentum, and recursive momentum for solving problem (1). We also establish their first-order oracle complexities for finding an ϵ-SSP of (1) under heavy-tailed noise.

3.1 A normalized SFOM with Polyak momentum

In this subsection, we propose a practical normalized SFOM with Polyak momentum for solving problem (1) and establish its first-order oracle complexity for finding an ϵ-SSP of (1) under heavy-tailed noise.

Our proposed method follows the same general framework as that in [3], but differs in the choice of algorithmic parameters—particularly the momentum weights and step sizes. Unlike [3], which requires explicit knowledge of the Lipschitz constant and the noise bound, our method uses dynamically updated parameters that do not depend on such problem-specific quantities. This feature enhances its practical applicability, especially in scenarios where these constants are unavailable or hard to estimate.

Specifically, our practical normalized SFOM with Polyak momentum generates two sequences, $\{m^k\}$ and $\{x^k\}$. At each iteration $k \geq 0$, the direction m^k is computed as a weighted average of stochastic gradients of f evaluated at the iterates $x^0, \ldots, x^k$. The next iterate x^{k+1} is obtained by performing a line search update from x^k along the normalized direction $-m^k/\|m^k\|$, using a suitable step size. The detailed procedure is presented in Algorithm 1, with the specific momentum weights and step sizes defined in Theorems 1 and 2.

Algorithm 1 A normalized SFOM with Polyak momentum

Input: starting point $x^0 \in \mathbb{R}^n$, step sizes $\{\eta_k\} \subset (0, +\infty)$, weighting parameters $\{\theta_k\} \subset (0, 1]$.
Initialize: $m^{-1} = 0$ and $\theta_{-1} = 1$.
for $k = 0, 1, 2, \ldots$ **do**

Compute the search direction:

$$m^k = (1 - \theta_{k-1})m^{k-1} + \theta_{k-1}G(x^k; \xi^k).^1 \tag{7}$$

Update the next iterate:

$$x^{k+1} = x^k - \eta_k \frac{m^k}{\|m^k\|}.$$

end for

---

[1] $\{\xi^k\}$ is a sequence of independently drawn samples.

The following theorem establishes a complexity bound for Algorithm 1 to compute an ϵ-SSP of problem (1) under the assumption that the tail exponent α is known. Its proof is deferred to Section 5.1.

Theorem 1 (**complexity with known** α)**.** *Suppose that Assumption 1 holds. Let* $f_{\rm low}$*,* L_1*,* σ*, and* α *be given in Assumption 1, and define*

$$M_{1,\alpha} := 2(f(x^0) - f_{\rm low} + \sigma^\alpha + L_1 + (\alpha-1)(2/\alpha)^{\alpha/(\alpha-1)} + 3L_1^\alpha + 2\sigma^\alpha). \tag{8}$$

Let $\{x^k\}$ *be generated by Algorithm 1 with input parameters* $\{(\eta_k,\theta_k)\}$ *given by*

$$\eta_k = \frac{1}{(k+1)^{(2\alpha-1)/(3\alpha-2)}}, \quad \theta_k = \frac{1}{(k+1)^{\alpha/(3\alpha-2)}} \qquad \forall k \ge 0. \tag{9}$$

Then, for any $\epsilon \in (0,1)$*, it holds that* $\mathbb{E}[\|\nabla f(x^{\iota_K})\|] \le \epsilon$ *for all* K *satisfying*

$$K \ge \max\Big\{\Big(\frac{2(3\alpha-2)M_{1,\alpha}}{(\alpha-1)\epsilon}\ln\Big(\frac{2(3\alpha-2)M_{1,\alpha}}{(\alpha-1)\epsilon}\Big)\Big)^{(3\alpha-2)/(\alpha-1)}, 3\Big\},$$

where ι_K *is uniformly drawn from* $\{0,\ldots,K-1\}$*.*

The next theorem establishes a complexity bound for Algorithm 1 to compute an ϵ-SSP of problem (1) without requiring prior knowledge of the tail exponent α. Its proof is deferred to Section 5.1.

Theorem 2 (**complexity with unknown** α)**.** *Suppose that Assumption 1 holds. Let* $f_{\rm low}$*,* L_1*,* σ*, and* α *be given in Assumption 1, and define*

$$\widetilde{M}_{1,\alpha} := 2(f(x^0) - f_{\rm low} + \sigma^\alpha + L_1/2 + 3L_1^\alpha), \quad \widehat{M}_{1,\alpha} := 2((\alpha-1)(2/\alpha)^{\alpha/(\alpha-1)} + 2\sigma^\alpha). \tag{10}$$

Let $\{x^k\}$ *be generated by Algorithm 1 with input parameters* $\{(\eta_k,\theta_k)\}$ *given by*

$$\eta_k = \frac{1}{(k+1)^{3/4}}, \quad \theta_k = \frac{1}{(k+1)^{1/2}} \qquad \forall k \ge 0. \tag{11}$$

Then, for any $\epsilon \in (0,1)$*, it holds that* $\mathbb{E}[\|\nabla f(x^{\iota_K})\|] \le \epsilon$ *for all* K *satisfying*

$$K \ge \max\Big\{\Big(\frac{16\widetilde{M}_{1,\alpha}}{\epsilon}\ln\Big(\frac{16\widetilde{M}_{1,\alpha}}{\epsilon}\Big)\Big)^4, \Big(\frac{8\alpha\widehat{M}_{1,\alpha}}{(\alpha-1)\epsilon}\ln\Big(\frac{8\alpha\widehat{M}_{1,\alpha}}{(\alpha-1)\epsilon}\Big)\Big)^{2\alpha/(\alpha-1)}, 3\Big\},$$

where ι_K *is uniformly drawn from* $\{0,\ldots,K-1\}$*.*

Remark 2. (i) From Theorems 1 and 2, we observe that under Assumption 1, Algorithm 1 achieves a first-order oracle complexity of $\widetilde{\mathcal{O}}(\epsilon^{-(3\alpha-2)/(\alpha-1)})$ for finding an ϵ-SSP of problem (1) when the tail exponent α is known, and $\widetilde{\mathcal{O}}(\epsilon^{-2\alpha/(\alpha-1)})$ when α is unknown. These results match, up to logarithmic factors, the best-known complexities for SFOMs with gradient clipping [4, 24, 35] and normalized SFOMs without gradient clipping [13, 22, 33]. Nevertheless, our algorithm is more practical as it uses dynamically updated parameters that do not rely on knowledge of the Lipschitz constant or the noise bound—quantities that are often unavailable or hard to estimate in practice.

(ii) When the tail exponent α is unknown, Algorithm 1 with $\{(\eta_k,\theta_k)\}$ specified by (11) resembles the parameter-free SFOM with momentum proposed in [13, Appendix D]. Nevertheless, our complexity analysis is fundamentally different from that of [13] and other existing works, as it is based on descent properties of a novel potential sequence defined in (34).

3.2 A normalized SFOM with multi-extrapolated momentum

In this subsection, we propose a practical normalized SFOM with multi-extrapolated momentum for solving problem (1) and establish its first-order oracle complexity for finding an ϵ-SSP of (1) under heavy-tailed noise.

Specifically, our practical normalized SFOM with multi-extrapolated momentum generates three sequences: $\{z^{k,t}\}$, $\{m^k\}$, and $\{x^k\}$. At each iteration $k \geq 0$, the points $z^{k,1}, \ldots, z^{k,q}$ are computed by extrapolating x^{k-1} and x^k using a set of extrapolation weights. The direction m^k is then formed as a weighted average of stochastic gradients of f evaluated at the extrapolated points $\{z^{i,t}\}_{0\leq i\leq k, 1\leq t\leq q}$. Finally, x^{k+1} is computed via a line search update at x^k using a suitable step size and the normalized direction $-m^k/\|m^k\|$. The detailed procedure is described in Algorithm 2, with the extrapolation weights, momentum weights, and step sizes specified in Theorems 3 and 4.

Algorithm 2 A normalized SFOM with multi-extrapolated momentum

Input: starting point $x^0 \in \mathbb{R}^n$, step sizes $\{\eta_k\} \subset (0, +\infty)$, extrapolation count $q \geq 1$, extrapolation parameters $\{\gamma_{k,t}\} \subset (0,1)$, weighting parameters $\{\theta_{k,t}\}$ with $\sum_{t=1}^q \theta_{k,t} \in (0,1)$ for all $k \geq 0$.

Initialize: $x^{-1} = x^0$, $m^{-1} = 0$, and $(\gamma_{-1,t}, \theta_{-1,t}) = (1, 1/q)$ for all $1 \leq t \leq q$.

for $k = 0, 1, 2, \ldots$ **do**

Perform q separate extrapolations:

$$z^{k,t} = x^k + \frac{1-\gamma_{k-1,t}}{\gamma_{k-1,t}}(x^k - x^{k-1}) \qquad \forall 1 \leq t \leq q. \tag{12}$$

Compute the search direction:

$$m^k = \Big(1 - \sum_{t=1}^q \theta_{k-1,t}\Big)m^{k-1} + \sum_{t=1}^q \theta_{k-1,t} G(z^{k,t}; \xi^k).^2 \tag{13}$$

Update the next iterate:

$$x^{k+1} = x^k - \eta_k \frac{m^k}{\|m^k\|}.$$

end for

Before analyzing the complexity of Algorithm 2 for computing an approximate solution to problem (1), we introduce an additional assumption regarding the high-order smoothness of the objective function f.

Assumption 2. *The function f is pth-order continuously differentiable in $\mathbb{R}^n$ for some $p \geq 2$, and moreover, there exists some $L_p > 0$ such that $\|\mathcal{D}^p f(x) - \mathcal{D}^p f(y)\| \leq L_p\|x - y\|$ for all $x, y \in \mathbb{R}^n$.*

The following theorem establishes a complexity bound for Algorithm 2 to compute an ϵ-SSP of problem (1) under the assumption that the tail exponent α is known. Its proof is deferred to Section 5.2.

Theorem 3 **(complexity with known α).** *Suppose that Assumptions 1 and 2 hold. Let $f_{\rm low}$, L_1, σ, α, p and L_p be given in Assumptions 1 and 2, and define*

$$M_{p,\alpha} := 4\Big(f(x^0) - f_{\rm low} + 4^{1/3}\sigma^\alpha + L_1/2 + 30^{1/(\alpha-1)}(\alpha-1)(2/\alpha)^{\alpha/(\alpha-1)} + 306p^{2\alpha p}L_p^\alpha/(p!)^\alpha$$

[2] $\{\xi^k\}$ is a sequence of independently drawn samples. Alternatively, one may draw q independent samples $\xi^{k,t}$, $1 \leq t \leq q$, at every kth iteration for computing $G(z^{k,t}; \xi^{k,t})$, $1 \leq t \leq q$.

$$+ 64(p-1)^\alpha\sigma^\alpha\Big). \tag{14}$$

Let $\{x^k\}$ be generated by Algorithm 2 with input parameters $q = p-1$, and $\{\eta_k\}$ and $\{(\gamma_{k,t},\theta_{k,t})\}$ given by

$$\eta_k = \frac{1}{(k+4)^{(p\alpha+\alpha-1)/(p(2\alpha-1)+\alpha-1)}} \qquad \forall k \geq 0, \tag{15}$$

$$\gamma_{k,t} = \frac{\gamma_k}{t^2}, \quad \theta_{k,t} = \frac{\prod_{1\leq s\leq p-1, s\neq t}(1-s^2/\gamma_k)}{(t^2/\gamma_k)\prod_{1\leq s\leq p-1,s\neq t}((t^2-s^2)/\gamma_k)} \qquad \forall 1\leq t\leq p-1, k\geq 0, \tag{16}$$

where

$$\gamma_k = \frac{1}{(k+4)^{p\alpha/(p(2\alpha-1)+\alpha-1)}} \qquad \forall k\geq 0. \tag{17}$$

Then, for any $\epsilon\in(0,1)$, it holds that $\mathbb{E}[\|\nabla f(x^{\iota_K})\|]\leq\epsilon$ for all K satisfying

$$K \geq \max\Big\{\Big(\frac{2(p(2\alpha-1)+\alpha-1)M_{p,\alpha}}{p(\alpha-1)\epsilon}\ln\Big(\frac{2(p(2\alpha-1)+\alpha-1)M_{p,\alpha}}{p(\alpha-1)\epsilon}\Big)\Big)^{(p(2\alpha-1)+\alpha-1)/(p(\alpha-1))}, 5\Big\},$$

where ι_K is uniformly drawn from $\{0,\ldots,K-1\}$.

The next theorem establishes a complexity bound for Algorithm 2 to compute an ϵ-SSP of problem (1) without requiring prior knowledge of the tail exponent α. Its proof is deferred to Section 5.2.

Theorem 4 (complexity with unknown α). *Suppose that Assumptions 1 and 2 hold. Let $f_{\rm low}$, L_1, σ, α, p and L_p be given in Assumptions 1 and 2, and define*

$$\widetilde{M}_{p,\alpha} := 4(f(x^0) - f_{\rm low} + 4^{1/3}\sigma^\alpha + L_1/2 + 306p^{2p\alpha}L_p^\alpha/(p!)^\alpha), \tag{18}$$

$$\widehat{M}_{p,\alpha} := 8(30^{1/(\alpha-1)}(\alpha-1)(2/\alpha)^{\alpha/(\alpha-1)} + 64(p-1)^\alpha\sigma^\alpha). \tag{19}$$

Let $\{x^k\}$ be generated by Algorithm 2 with input parameters $q=p-1$, $\{\eta_k\}$, and $\{(\gamma_{k,t},\theta_{k,t})\}$ given by

$$\eta_k = \frac{1}{(k+4)^{(2p+1)/(3p+1)}} \qquad \forall k\geq 0, \tag{20}$$

$$\gamma_{k,t} = \frac{\gamma_k}{t^2}, \quad \theta_{k,t} = \frac{\prod_{1\leq s\leq p-1, s\neq t}(1-s^2/\gamma_k)}{(t^2/\gamma_k)\prod_{1\leq s\leq p-1,s\neq t}((t^2-s^2)/\gamma_k)} \qquad \forall 1\leq t\leq p-1, k\geq 0, \tag{21}$$

where

$$\gamma_k = \frac{1}{(k+4)^{2p/(3p+1)}} \qquad \forall k\geq 0. \tag{22}$$

Then, for any $\epsilon\in(0,1)$, it holds that $\mathbb{E}[\|\nabla f(x^{\iota_K})\|]\leq\epsilon$ for all K satisfying

$$K \geq \max\Bigg\{\Big(\frac{4(3p+1)\widetilde{M}_{p,\alpha}}{p\epsilon}\ln\Big(\frac{4(3p+1)\widetilde{M}_{p,\alpha}}{p\epsilon}\Big)\Big)^{(3p+1)/p}, \Big(\frac{2(3p\alpha+\alpha)\widehat{M}_{p,\alpha}}{p(\alpha-1)\epsilon}\ln\Big(\frac{2(3p\alpha+\alpha)\widehat{M}_{p,\alpha}}{p(\alpha-1)\epsilon}\Big)\Big)^{(3p\alpha+\alpha)/(2p(\alpha-1))}, 5\Bigg\},$$

where ι_K is uniformly drawn from $\{0,\ldots,K-1\}$.

Remark 3. (i) To achieve acceleration by leveraging the higher-order smoothness of f, the extrapolation parameters $\gamma_{k,t}$ and the momentum parameters $\theta_{k,t}$ must satisfy the following conditions:

$$\begin{bmatrix} 1/\gamma_{k,1} & 1/\gamma_{k,2} & \cdots & 1/\gamma_{k,q} \\ 1/\gamma_{k,1}^2 & 1/\gamma_{k,2}^2 & \cdots & 1/\gamma_{k,q}^2 \\ \vdots & \vdots & \ddots & \vdots \\ 1/\gamma_{k,1}^q & 1/\gamma_{k,2}^q & \cdots & 1/\gamma_{k,q}^q \end{bmatrix} \begin{bmatrix} \theta_{k,1} \\ \theta_{k,2} \\ \vdots \\ \theta_{k,q} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \qquad \forall k \geq 0, \tag{23}$$

$$\sum_{t=1}^{q} \theta_{k,t} \in (0,1) \qquad \forall k \geq 0, \tag{24}$$

where $q = p - 1$. Note that the coefficient matrix in (23) is a Vandermonde matrix (see, e.g., [12, 14]). As will be proven in Lemma 9 in Subsection 5.2, when choosing $\gamma_{k,t} = \gamma_k/t^2$, the resulting $\theta_{k,t}$ that satisfy (23) takes the form given in (16) or (21).

(ii) From Theorems 3 and 4, we observe that under Assumptions 1 and 2, Algorithm 2 achieves a first-order oracle complexity of $\widetilde{\mathcal{O}}(\epsilon^{-(p(2\alpha-1)+\alpha-1)/(p(\alpha-1))})$ for finding an ϵ-SSP of (1) when the tail exponent α is known, and $\widetilde{\mathcal{O}}(\epsilon^{-(3p\alpha+\alpha)/(2p(\alpha-1))})$ when α is unknown. For $p = 2$, these complexity results match, up to a logarithmic factor, the best-known bound established in [4]. Moreover, for $p \geq 3$, our results are entirely new and provide significantly improved complexities over the case $p = 2$.

3.3 A normalized SFOM with recursive momentum

In this subsection, we propose a practical normalized SFOM with recursive momentum for solving problem (1) and establish its first-order oracle complexity for finding an ϵ-SSP of (1) under heavy-tailed noise.

Our proposed method follows the same general framework as that in [5], but differs in both the search direction and the choice of algorithmic parameters—particularly the momentum weights and step sizes. Specifically, our method employs a normalized search direction, whereas [5] uses the unnormalized (full) direction. Moreover, unlike [5], which requires explicit knowledge of the Lipschitz constant and the noise bound, our method uses dynamically updated parameters that do not rely on such problem-specific quantities. This feature makes our method more practical, especially in scenarios where these constants are unknown or hard to estimate.

Specifically, our practical normalized SFOM with recursive momentum generates two sequences, $\{m^k\}$ and $\{x^k\}$. At each iteration $k \geq 0$, the direction m^k is computed as a weighted average of stochastic gradients of f evaluated at the iterates $x^0, \ldots, x^k$. The next iterate x^{k+1} is obtained by performing a line search update from x^k along the normalized direction $-m^k/\|m^k\|$, using a suitable step size. The detailed procedure is presented in Algorithm 3, with the specific momentum weights and step sizes defined in Theorems 5 and 6.

Before analyzing the complexity of Algorithm 3 for computing an approximate solution to problem (1), we introduce an additional assumption regarding the *weakly average smoothness* of the stochastic gradient estimator $G(\cdot;\xi)$.

Assumption 3. *There exists some $L > 0$ such that $\mathbb{E}[\|G(x;\xi) - G(y;\xi)\|^\alpha] \leq L^\alpha \|x-y\|^\alpha$ holds for all $x, y \in \mathbb{R}^n$, where $\alpha \in (1,2]$ is given in Assumption 1(c).*

Remark 4. (i) When $\alpha = 2$, Assumption 3 reduces to the standard mean-squared smoothness condition commonly used in the literature (see, e.g., [5, 6, 18]). For $\alpha \in (1,2)$, it is strictly weaker than the mean-squared smoothness assumption, thereby holding for a broader class of stochastic gradient estimators

[3] $\{\xi^k\}$ is a sequence of independently drawn samples.

Algorithm 3 A normalized SFOM with recursive momentum

Input: starting point $x^0 \in \mathbb{R}^n$, step sizes $\{\eta_k\} \subset (0,+\infty)$, weighting parameters $\{\theta_k\} \subset (0,1]$.
Initialize: $x^{-1} = x^0$, $m^{-1} = 0$, and $\theta_{-1} = 1$.
for $k = 0, 1, 2, \ldots$ **do**
Compute the search direction:

$$m^k = (1-\theta_{k-1})m^{k-1} + G(x^k;\xi^k) - (1-\theta_{k-1})G(x^{k-1};\xi^k).^3 \tag{25}$$

Update the next iterate:

$$x^{k+1} = x^k - \eta_k \frac{m^k}{\|m^k\|}.$$

end for

$G(\cdot;\xi)$. In addition, if $G(\cdot;\xi)$ is almost surely Lipschitz continuous for all $\xi \in \Xi$ with a uniform Lipschitz constant L, one can verify that it satisfies Assumption 3. Consequently, Assumption 3 is strictly weaker than the almost sure Lipschitz condition stated in (3), which is adopted in [21, 33].

(ii) It is reasonable to assume that the exponent α in Assumption 3 is the same as that in Assumption 1(c). Indeed, if Assumptions 1(c) and 3 hold with different exponents $\alpha_1, \alpha_2 \in (1,2]$, then both assumptions also hold with $\alpha = \min\{\alpha_1, \alpha_2\}$.

The following theorem establishes a complexity bound for Algorithm 3 to compute an ϵ-SSP of problem (1) under the assumption that the tail exponent α is known. Its proof is deferred to Section 5.3.

Theorem 5 (**complexity with known** α)**.** *Suppose that Assumptions 1 and 3 hold. Let f_{low}, L_1, σ, α, and L be given in Assumption 1 and 3, and define*

$$M_\alpha := 2(f(x^0) - f_{\mathrm{low}} + \sigma^\alpha + L_1/2 + 2^{1/(\alpha-1)}(\alpha-1)(2/\alpha)^{\alpha/(\alpha-1)} + 12(L_1^\alpha + L^\alpha) + 12\sigma^\alpha). \tag{26}$$

Let $\{x^k\}$ be generated by Algorithm 3 with input parameters $\{(\eta_k,\theta_k)\}$ given by

$$\eta_k = \theta_k = \frac{1}{(k+1)^{\alpha/(2\alpha-1)}} \qquad \forall k \geq 0. \tag{27}$$

Then, for any $\epsilon \in (0,1)$, it holds that $\mathbb{E}[\|\nabla f(x^{\iota_K})\|] \leq \epsilon$ for all K satisfying

$$\forall K \geq \max\Big\{\Big(\frac{2(2\alpha-1)M_\alpha}{(\alpha-1)\epsilon}\ln\Big(\frac{2(2\alpha-1)M_\alpha}{(\alpha-1)\epsilon}\Big)\Big)^{(2\alpha-1)/(\alpha-1)}, 3\Big\},$$

where ι_K is uniformly drawn from $\{0,\ldots,K-1\}$.

The following theorem establishes a complexity bound for Algorithm 3 to compute an ϵ-SSP of problem (1) without requiring prior knowledge of the tail exponent α. Its proof is deferred to Section 5.3.

Theorem 6 (**complexity with unknown** α)**.** *Suppose that Assumptions 1 and 3 hold. Let f_{low}, L_1, σ, α, and L be given in Assumption 1 and 3, and define*

$$\widetilde{M}_\alpha := 2(f(x^0) - f_{\mathrm{low}} + \sigma^\alpha + L_1/2), \tag{28}$$
$$\widehat{M}_\alpha := 2(2^{1/(\alpha-1)}(\alpha-1)(2/\alpha)^{\alpha/(\alpha-1)} + 12(L_1^\alpha + L^\alpha) + 12\sigma^\alpha). \tag{29}$$

Let $\{x^k\}$ be generated by Algorithm 3 with input parameters $\{(\eta_k, \theta_k)\}$ given by

$$\eta_k = \theta_k = \frac{1}{(k+1)^{2/3}} \qquad \forall k \geq 0. \tag{30}$$

Then, for any $\epsilon \in (0,1)$, it holds that $\mathbb{E}[\|\nabla f(x^{\iota_K})\|] \leq \epsilon$ for all K satisfying

$$K \geq \max\Big\{\Big(\frac{12\widetilde{M}_\alpha}{\epsilon}\ln\Big(\frac{12\widetilde{M}_\alpha}{\epsilon}\Big)\Big)^3, \Big(\frac{6\alpha\widehat{M}_\alpha}{(\alpha-1)\epsilon}\ln\Big(\frac{6\alpha\widehat{M}_\alpha}{(\alpha-1)\epsilon}\Big)\Big)^{3\alpha/(2(\alpha-1))}, 3\Big\},$$

where ι_K is uniformly drawn from $\{0,\ldots,K-1\}$.

Remark 5. From Theorems 5 and 6, we observe that under Assumptions 1 and 3, Algorithm 3 achieves a first-order oracle complexity of $\widetilde{\mathcal{O}}(\epsilon^{-(2\alpha-1)/(\alpha-1)})$ for finding an ϵ-SSP of problem (1) when the tail exponent α is known, and $\widetilde{\mathcal{O}}(\epsilon^{-3\alpha/(2(\alpha-1))})$ when α is unknown. The complexity bound for the known-α case matches, up to logarithmic factors, the best-known results in [21, 33], while the bound for the unknown-α case is, to the best of our knowledge, new to the literature. Moreover, these are the first complexity results established under the weakly average smoothness condition stated in Assumption 3. This condition generalizes and relaxes both the almost sure Lipschitz condition (see (3)) used in prior works on SFOMs under heavy-tailed noise [21, 33], and the commonly adopted mean-squared smoothness assumption [5, 6, 18].

4 Numerical experiments

In this section, we present preliminary numerical experiments to evaluate the performance of Algorithm 1, Algorithm 2 with $q=1$, and Algorithm 3, abbreviated as NSFOM-PM, NSFOM-EM, and NSFOM-RM, respectively. We compare these methods against their counterparts without normalization—SFOM-PM, SFOM-EM, and SFOM-RM—in the presence of heavy-tailed noise. The experiments are conducted on three problem classes: a data fitting problem (Section 4.1), a robust regression problem (Section 4.2), and a multimodal contrastive learning problem (Section 4.3). The first two are run on a standard PC equipped with a 3.20 GHz AMD R7 5800H processor and 16 GB of memory, while the last is executed on a server with an NVIDIA A100 GPU (80 GB). The code for reproducing our numerical results is publicly available at: `https://github.com/ChuanH6/SFOM-HT`.

4.1 Data fitting problem

In this subsection, we consider the data fitting problem:

$$\min_{x\in\mathbb{R}^n}\Big\{f(x) = \sum_{i=1}^m \big(s(a_i^Tx) - b_i\big)^2\Big\}, \tag{31}$$

where $s(t) = e^t/(1+e^t)$ is the sigmoid function, and $\{(a_i,b_i)\}_{1\leq i\leq n} \subset \mathbb{R}^n \times \mathbb{R}$ denotes the given dataset. We simulate the noisy gradient evaluations by setting the stochastic gradient estimator as $G(x;\xi) = \nabla f(x) + \xi e$, where $e \in \mathbb{R}^n$ is the all-ones vector and $\xi \in \mathbb{R}$ is drawn from a heavy-tailed distribution with density function $p(t) = 3/(4(1+|t|)^{5/2})$. One can verify that such $G(\cdot;\xi)$ satisfies Assumption 1(c) for every $\alpha \in (1, 1.5)$, and that the αth central moment of $G(\cdot;\xi)$ is unbounded for all $\alpha \geq 1.5$.

For each pair (n,m), we randomly generate a_i, $1 \leq i \leq m$, with all entries independently drawn from the standard normal distribution. We also generate a ground truth solution x^* in the same manner and set $b_i = s(a_i^T x^*) + 10^{-4} e_i$ for each $1 \le i \le m$, where e_i's are independently drawn from the standard normal distribution.

We apply NSFOM-PM, NSFOM-EM, and NSFOM-RM, along with their unnormalized variants, to solve problem (31). All methods are initialized at the zero vector. We compare them based on two metrics: the relative objective value gap $(f(x^k) - f^*)/(f(x^0) - f^*)$ and the relative gradient norm $\|\nabla f(x^k)\|/\|\nabla f(x^0)\|$, computed over the first 500 stochastic gradient evaluations. Here, f^* denotes the minimum objective value found during the first 600 stochastic gradient evaluations across all methods. The algorithmic parameters are selected to suit each method well in terms of computational performance.

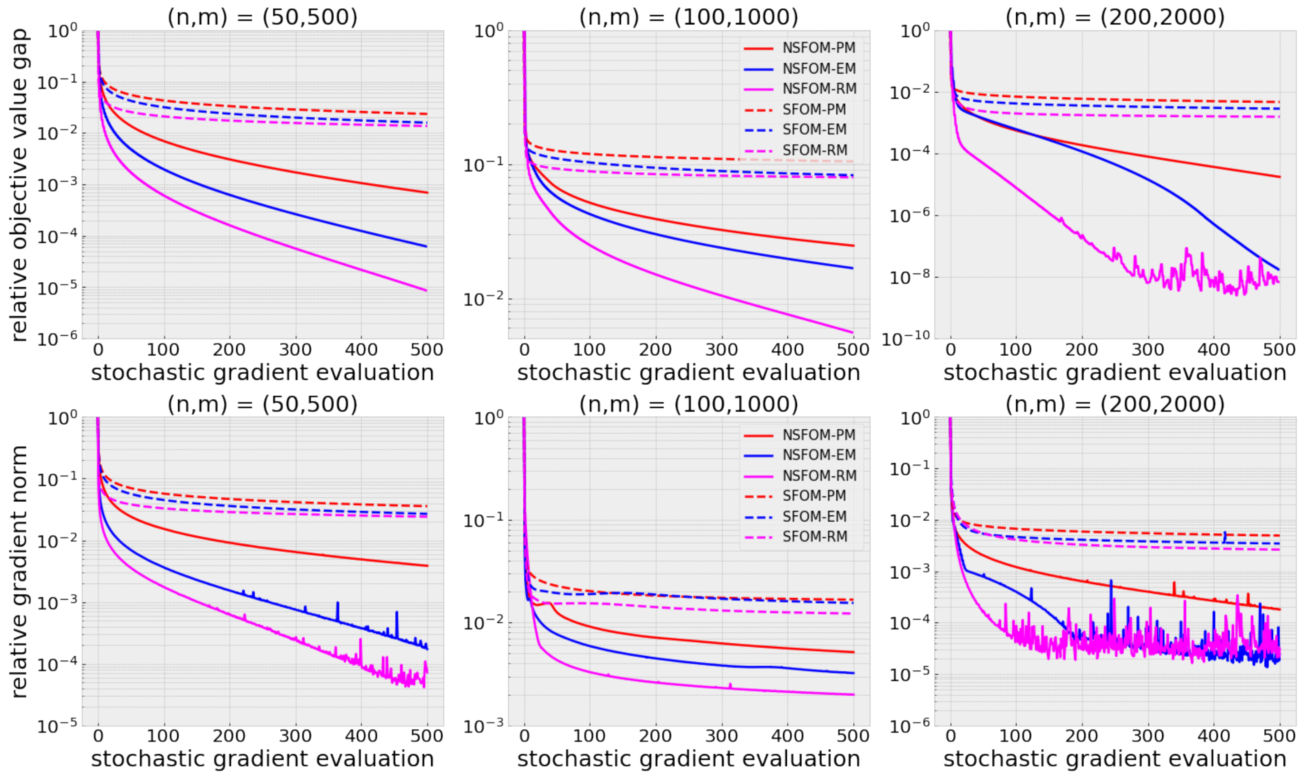

Figure 1: Convergence behavior of the relative objective value gap (first row) and relative gradient norm (second row) for all methods in solving problem (31).

For each pair (n, m), we plot the relative objective value gap and the relative gradient norm in Figure 1 to illustrate the convergence behavior of all the SFOMs. As shown in Figure 1, the SFOMs with normalization consistently outperform their unnormalized counterparts. Furthermore, among the normalized variants, those using extrapolated momentum and recursive momentum converge faster than the one using Polyak momentum. Notably, the normalized SFOM with recursive momentum achieves the best performance, even outperforming the extrapolated momentum variant. These observations are consistent with our theoretical results.

4.2 Robust regression problem

In this subsection, we consider the robust regression problem:

$$\min_{x \in \mathbb{R}^n} \Big\{ f(x) = \sum_{k=1}^{m} \sum_{i=1}^{N} \phi(a_{ik}^T x - b_{ik}) \Big\}, \tag{32}$$

where $\phi(t) = t^2/(1+t^2)$ is a robust loss function [2], and $\{(a_{ik}, b_{ik})\}_{1\le i\le N} \subset \mathbb{R}^n \times \mathbb{R}$, $1 \le k \le m$, is the kth batch of the training set. We consider this problem on three real datasets, 'red wine quality', 'white wine quality', and 'covtype' from the UCI repository.[4] For each dataset, we rescale both the features and predictions to lie in $[0, 1]$, and set the batch size $N = 100$.

We apply NSFOM-PM, NSFOM-EM, and NSFOM-RM, along with their unnormalized variants, to solve (32). All methods are initialized at the zero vector. Similar to Section 4.1, we compare these methods in terms of the relative objective value gap and the relative gradient norm, defined respectively as $(f(x^k) - f^*)/(f(x^0) - f^*)$ and $\|\nabla f(x^k)\|/\|\nabla f(x^0)\|$, over the first 500 stochastic gradient evaluations, where f^* is the minimum objective value found during the first 600 stochastic gradient evaluations across all the SFOMs. The algorithmic parameters are selected to suit each method well in terms of computational performance.

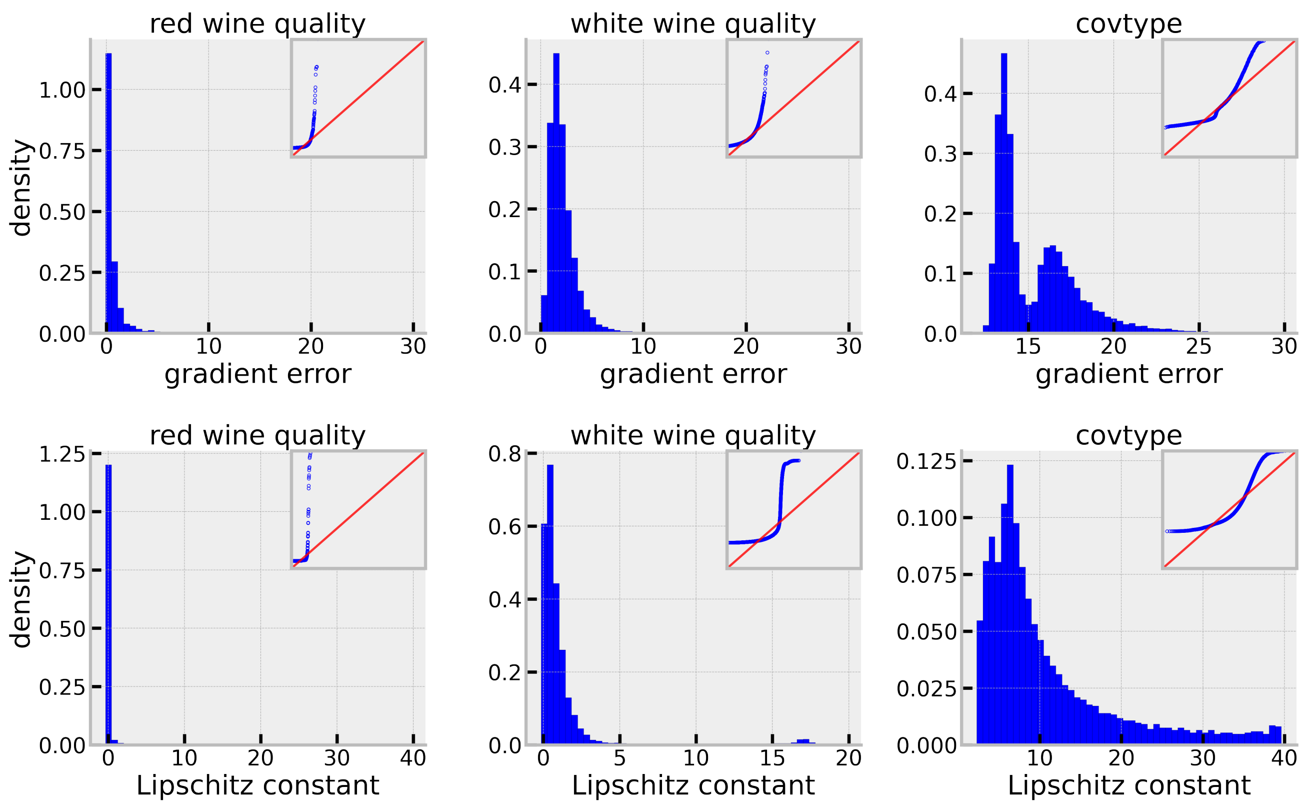

Figure 2: Distributions of gradient errors $\|G(x;\xi) - \nabla f(x)\|$ (first row) and Lipschitz constant estimates $\|G(y;\xi) - G(x;\xi)\|/\|y - x\|$ (second row) compared against a normal distribution (QQ-plot), when solving (32). Here, the gradient errors are calculated for the first epoch of optimization, and the Lipschitz constant estimates are taken over every two consecutive iterates within the first epoch of optimization for all methods.

For each dataset, we visualize the distributions of gradient errors and Lipschitz constant estimates, compared against a normal distribution (QQ-plot) in Figure 2, to illustrate their heavy-tailed behavior. This visualizations partly justify the heavy-tailed noise condition in Assumption 1(c) and the weakly average smoothness condition in Assumption 3 when solving the regression problem (32).

In addition, for each dataset, we plot the relative objective value gap and the relative gradient norm in Figure 3 to illustrate the convergence behavior of all the SFOMs. From Figure 3, we observe that

[4] see `archive.ics.uci.edu/datasets`

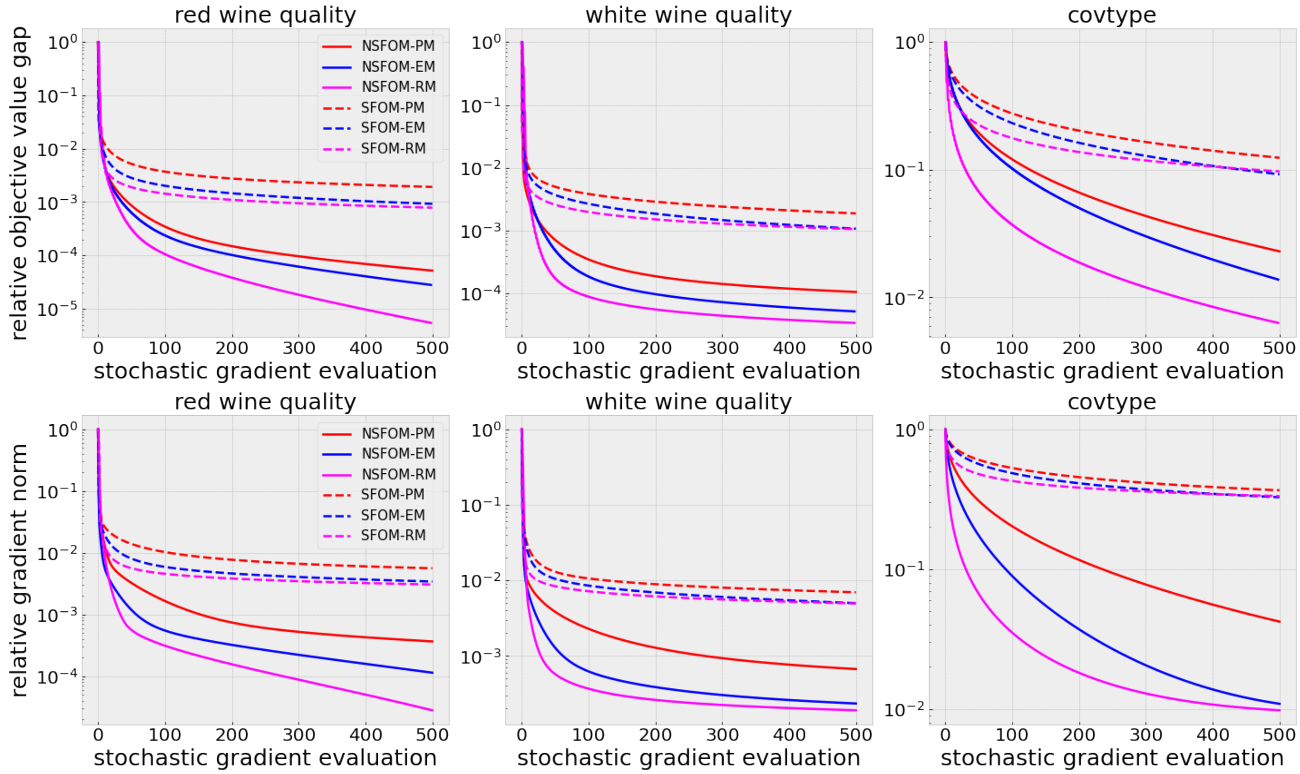

Figure 3: Convergence behavior of the relative objective value gap (first row) and relative gradient norm (second row) for all method in solving problem (32).

SFOMs with normalization consistently outperform those without normalization. We also observe that the normalized SFOMs with extrapolated and recursive momentum are faster than the SFOM with Polyak momentum, while the SFOM with recursive momentum outperforms the SFOM with extrapolated momentum. These observations align with our theoretical results.

4.3 Multimodal contrastive learning problem

In this subsection, we consider the multimodal contrastive learning problem (see [26]):

$$\min_{x_I\in\mathbb{R}^{n_I},x_T\in\mathbb{R}^{n_T}} -\sum_{k=1}^{m}\sum_{i=1}^{N}\left(\ln\left(\frac{\exp(f_{x_I}(a_{ik})^T f_{x_T}(b_{ik})/\tau)}{\sum_{j=1}^{N}\exp(f_{x_I}(a_{ik})^T f_{x_T}(b_{jk})/\tau)}\right)+\ln\left(\frac{\exp(f_{x_I}(a_{ik})^T f_{x_T}(b_{ik})/\tau)}{\sum_{j=1}^{N}\exp(f_{x_I}(a_{jk})^T f_{x_T}(b_{ik})/\tau)}\right)\right), \tag{33}$$

where $\{(a_{ik},b_{ik})\}_{1\le i\le N}$, $1\le k\le m$, denotes the image-caption pairs for the kth batch of training dataset, f_{x_I} and f_{x_T} are the image and text encoders, respectively, and $\tau>0$ is a temperature parameter. Here, we consider problem (33) on three real text-image datasets, Flickr [25], MSCOCO [19], and CC3M [30], and choose the network structure for image encoder f_{x_I} and text encoder f_{x_T} as ResNet50 [11] and DistilBERT [29], respectively.

We apply NSFOM-PM, NSFOM-EM, and NSFOM-RM, along with their variants without normalization, to solve problem (33). We use the same initial weights for all methods as those of the pretrained models ResNet50 and DistilBERT. Similar to Section 4.1, we compare these methods in terms of the relative objective value gap and the relative gradient norm, defined respectively as $(f(x^k)-f^*)/(f(x^0)-f^*)$

and $\|\nabla f(x^k)\|/\|\nabla f(x^0)\|$, over the first $50,000$ stochastic gradient evaluations, where f denotes the objective function of (33) and f^* is the minimum objective value found during the first $60,000$ stochastic gradient evaluations across all the SFOMs. The other algorithmic parameters are selected to suit each method well in terms of computational performance.

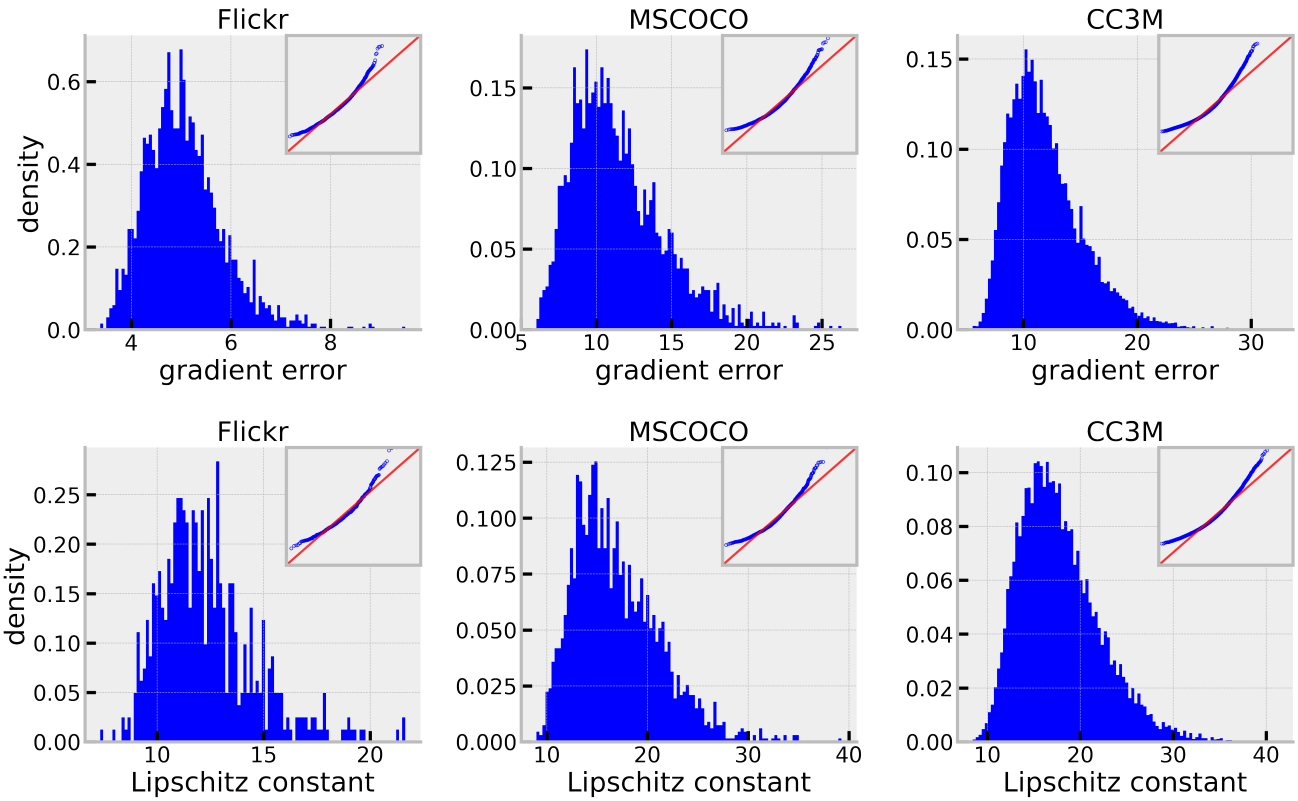

Figure 4: Distributions of gradient errors $\|G(x;\xi) - \nabla f(x)\|$ (first row) and Lipschitz constant estimates $\|G(x;\xi) - G(y;\xi)\|/\|x - y\|$ (second row) compared against a normal distribution (QQ-plot), when solving (33). Here, the gradient errors are calculated for the first epoch of training, and the Lipschitz constant estimates are taken over every two consecutive iterates within the first epoch of training for all methods.

For each dataset, we visualize the distributions of gradient errors and Lipschitz constant estimates, compared against a normal distribution (QQ-plot) in Figure 4, to illustrate their heavy-tailed behavior. These visualizations provide partial justification for the heavy-tailed noise condition in Assumption 1(c) and the weakly average smoothness condition in Assumption 3 when solving the multimodal contrastive learning problem (33).

In addition, for each dataset, we plot the relative objective value gap and the relative gradient norm in Figure 5 to illustrate the convergence behavior of the SFOMs. From Figure 5, we can observe that SFOMs with normalization tend to outperform their unnormalized counterparts. We also observe that the normalized SFOMs with extrapolated and recursive momentum converge faster than the SFOM with Polyak momentum, while the SFOM with recursive momentum slightly outperforms the one with extrapolated momentum. These phenomena are generally consistent with our theoretical results.

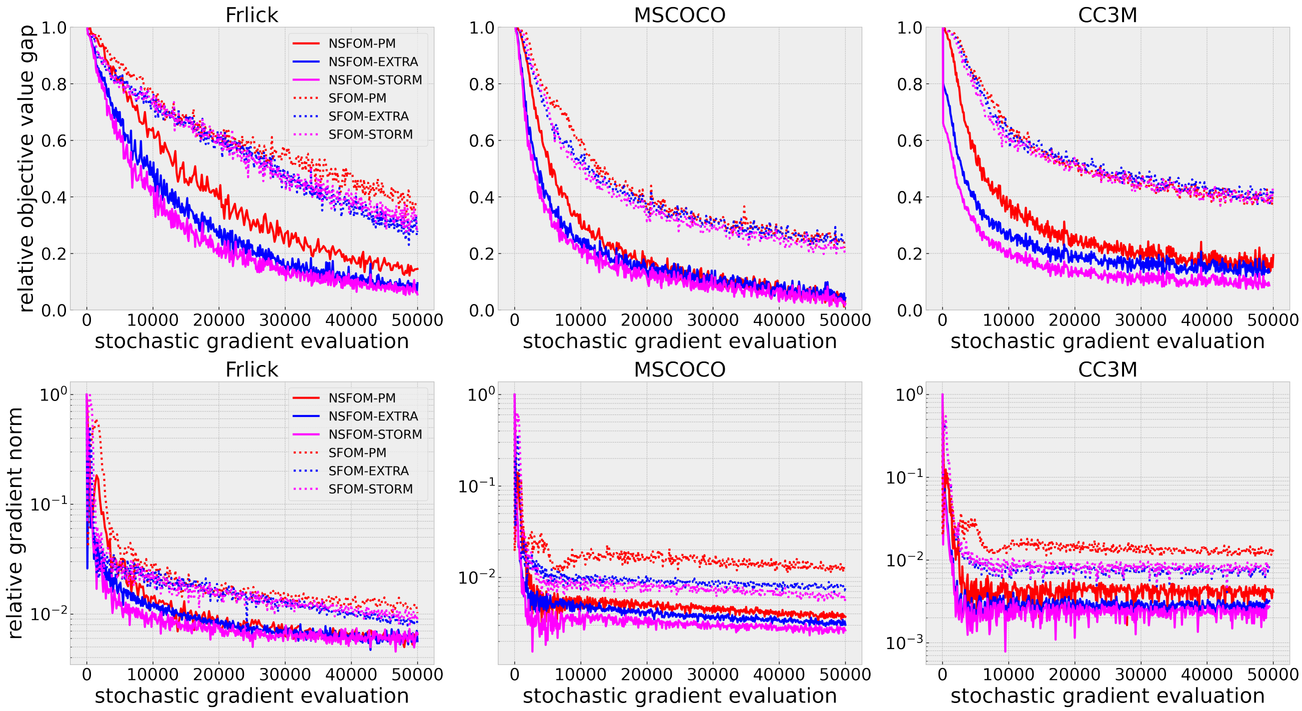

Figure 5: Convergence behavior of the relative objective value gap (first row) and relative gradient norm (second row) for all method in solving problem (33).

5 Proof of the main results

In this section, we provide the proofs of our main results presented in Section 3, specifically, Theorems 1 to 6.

For notational convenience, we define a sequence of potentials for Algorithms 1, 2, and 3 as

$$\mathcal{P}_k := f(x^k) + p_k \|m^k - \nabla f(x^k)\|^\alpha \qquad \forall k \geq 0, \tag{34}$$

where the sequence $\{(x^k, m^k)\}$ is generated by each respective algorithm, and $\{p_k\}$ is a sequence of positive scalars that will be specified separately for each case.

The following lemma provides an expansion for the α-power of the Euclidean norm, generalizing the well-known identity $\|u+v\|^2 = \|u\|^2 + 2u^Tv + \|v\|^2$ and inequality $\|u+v\|^2 \leq (1+c)\|u\|^2 + (1+1/c)\|v\|^2$ for all $u, v \in \mathbb{R}^n$ and $c > 0$.

Lemma 1. *For any $\alpha \in (1,2]$, it holds that*

$$\|u+v\|^\alpha \leq \|u\|^\alpha + \alpha\|u\|^{\alpha-2}u^Tv + 2\|v\|^\alpha \qquad \forall u, v \in \mathbb{R}^n, \tag{35}$$

$$\|u+v\|^\alpha \leq (1+c)\|u\|^\alpha + (2 + (\alpha-1)^{\alpha-1}c^{1-\alpha})\|v\|^\alpha \qquad \forall u, v \in \mathbb{R}^n, c > 0. \tag{36}$$

Proof. Let $\alpha \in (1,2]$ be arbitrarily chosen, and $\psi(w) = \|w\|^\alpha$ for all $w \in \mathbb{R}^n$. It follows from [27, Theorem 6.3] that

$$\|\nabla\psi(w^1) - \nabla\psi(w^2)\| \leq 2^{2-\alpha}\alpha\|w^1 - w^2\|^{\alpha-1} \quad \forall w^1, w^2 \in \mathbb{R}^n.$$

By this, one has that for any $u, v \in \mathbb{R}^n$,

$$|\psi(u+v) - \psi(u) - \nabla\psi(u)^Tv| = \Big|\int_0^1 (\nabla\psi(u+tv) - \nabla\psi(u))^T v \mathrm{d}t\Big|$$

$$\leq \int_0^1 \|\nabla\psi(u+tv)-\nabla\psi(u)\|\mathrm{d}t\cdot\|v\| \leq 2^{2-\alpha}\alpha\int_0^1 \|tv\|^{\alpha-1}\mathrm{d}t\cdot\|v\| = 2^{2-\alpha}\|v\|^\alpha,$$

This along with $\alpha\in(1,2]$ and the fact that $\psi(w)=\|w\|^\alpha$ and $\nabla\psi(w)=\alpha\|w\|^{\alpha-2}w$ implies that (35) holds. We next prove (36). Let $\alpha'=\alpha/(\alpha-1)$. By the Young's inequality, one has that for all $c>0$,

$$\alpha\|u\|^{\alpha-2}u^Tv \leq \frac{\big((c\alpha')^{1/\alpha'}\big\|\|u\|^{\alpha-2}u\big\|\big)^{\alpha'}}{\alpha'} + \frac{\big(\alpha\|v\|/(c\alpha')^{1/\alpha'}\big)^{\alpha}}{\alpha} = c\|u\|^\alpha + \frac{(\alpha-1)^{\alpha-1}\|v\|^\alpha}{c^{\alpha-1}},$$

which together with (35) implies that (36) holds. □

The following lemma provides an estimation of the partial sums of series.

Lemma 2. *Let $\zeta(\cdot)$ be a convex univariate function. Then it holds that $\sum_{r=a}^b\zeta(r)\leq\int_{a-1/2}^{b+1/2}\zeta(\tau)\mathrm{d}\tau$ for any integers a,b satisfying $[a-1/2,b+1/2]\subset\mathrm{dom}\,\zeta$. Consequently, one has*

$$\sum_{r=a}^b\frac{1}{r^\beta}\leq\begin{cases}\ln\big(b+\frac{1}{2}\big)-\ln\big(a-\frac{1}{2}\big) & \text{if } \beta=1,\\ \frac{1}{1-\beta}\Big(\big(b+\frac{1}{2}\big)^{1-\beta}-\big(a-\frac{1}{2}\big)^{1-\beta}\Big) & \text{if } \beta\in(0,1)\cup(1,+\infty).\end{cases} \tag{37}$$

Proof. Let a,b be integers satisfying $[a-1/2,b+1/2]\subset\mathrm{dom}\,\zeta$. Since ζ is convex, one has $\zeta(\tau)\geq\zeta(r)+s^T(\tau-r)$ for all $s\in\partial\zeta(r)$ and $r\in[a,b]$. It then follows that

$$\int_{r-1/2}^{r+1/2}\zeta(\tau)\mathrm{d}\tau\geq\int_{r-1/2}^{r+1/2}\big(\zeta(r)+s^T(\tau-r)\big)\mathrm{d}\tau=\zeta(r),$$

which implies that $\sum_{r=a}^b\zeta(r)\leq\sum_{r=a}^b\int_{r-1/2}^{r+1/2}\zeta(\tau)\mathrm{d}\tau=\int_{a-1/2}^{b+1/2}\zeta(\tau)\mathrm{d}\tau$. By this and $\zeta(\tau)=1/\tau^\beta$, one can see that (37) holds. □

We next provide a lemma that will be used to derive complexity bounds subsequently.

Lemma 3. *Let $\beta\in(0,1)$ and $u\in(0,1/e)$ be given. Then $v^{-\beta}\ln v\leq 2u/\beta$ holds for all $v\geq(u^{-1}\ln(1/u))^{1/\beta}$.*

Proof. Fix any v satisfying $v\geq(u^{-1}\ln(1/u))^{1/\beta}$. It then follows from $u\in(0,1/e)$ that

$$v\geq(u^{-1}\ln(1/u))^{1/\beta}>e^{1/\beta}. \tag{38}$$

Let $\phi(\tau)=\tau^{-\beta}\ln\tau$. It can be verified that ϕ is decreasing on $(e^{1/\beta},\infty)$. By this and (38), one has that

$$v^{-\beta}\ln v=\phi(v)\leq\phi((u^{-1}\ln(1/u))^{1/\beta})=\frac{u}{\beta}\Big(1+\frac{\ln\ln(1/u)}{\ln(1/u)}\Big)\leq\frac{2u}{\beta},$$

where the last inequality follows from $\ln\ln(1/u)\leq\ln(1/u)$ due to $u\in(0,1/e)$. Hence, the conclusion of this lemma holds. □

We next establish a descent property for f along a normalized direction.

Lemma 4. *Suppose that Assumption 1 holds. Let $x,m\in\mathbb{R}^n$ and $\eta>0$ be given, and let $x^+=x-\eta m/\|m\|$. Then we have*

$$f(x^+)\leq f(x)-\eta\|\nabla f(x)\|+2\eta\|\nabla f(x)-m\|+\frac{L_1}{2}\eta^2,$$

where L_1 is given in Assumption 1(b).

Proof. Using (6) with $y = x^+$, we obtain that

$$
\begin{aligned}
f(x^+) &\overset{(6)}{\leq} f(x) + \nabla f(x)^T(x^+ - x) + \frac{L_1}{2}\|x^+ - x\|^2 \\
&= f(x) + m^T(x^+ - x) + (\nabla f(x) - m)^T(x^+ - x) + \frac{L_1}{2}\|x^+ - x\|^2 \\
&= f(x) - \eta\|m\| - \frac{\eta}{\|m\|}(\nabla f(x) - m)^T m + \frac{L_1}{2}\eta^2 \leq f(x) - \eta\|m\| + \eta\|\nabla f(x) - m\| + \frac{L_1}{2}\eta^2 \\
&\leq f(x) - \eta\|\nabla f(x)\| + 2\eta\|\nabla f(x) - m\| + \frac{L_1}{2}\eta^2,
\end{aligned}
$$

where the second equality follows from $x^+ = x - \eta m/\|m\|$, the second inequality is due to the Cauchy-Schwarz inequality, and the last inequality follows from the triangular inequality $\|m\| \geq \|\nabla f(x)\| - \|\nabla f(x) - m\|$. Hence, this lemma holds as desired. $\square$

The following lemma provides an upper bound on the residual of the pth-order Taylor expansion of ∇f.

Lemma 5. *Suppose that Assumption 2 holds. Let $\mathcal{R}_p(\cdot,\cdot)$ and L_p be given in (5) and Assumption 2, respectively. Then it holds that $\|\mathcal{R}_p(y,x)\| \leq L_p\|y - x\|^p/p!$ for all $x, y \in \mathbb{R}^n$.*

Proof. Fix any $u \in \mathbb{R}^n$. Let $\phi(x) = \langle \nabla f(x), u\rangle$. By this and the definition of $\nabla^{r+1} f(x)(h)^r$, one has

$$
\mathcal{D}^r\phi(x)[v]^r = \langle \nabla^{r+1} f(x)(v)^r, u\rangle \qquad \forall 1 \leq r \leq p-1, v \in \mathbb{R}^n. \tag{39}
$$

Using this and (4), we have

$$
\|\mathcal{D}^{p-1}\phi(y) - \mathcal{D}^{p-1}\phi(x)\| \leq \|u\|\|\mathcal{D}^p f(y) - \mathcal{D}^p f(x)\| \qquad \forall x, y \in \mathbb{R}^n. \tag{40}
$$

Fix any $x, y \in \mathbb{R}^n$. By Taylor's expansion, one has

$$
\begin{aligned}
\phi(y) &= \phi(x) + \sum_{r=1}^{p-2}\frac{1}{r!}\mathcal{D}^r\phi(x)[y-x]^r + \frac{1}{(p-2)!}\int_0^1 (1-t)^{p-2}\mathcal{D}^{p-1}\phi(x + t(y-x))[y-x]^{p-1}\mathrm{d}t \\
&= \phi(x) + \sum_{r=1}^{p-1}\frac{1}{r!}D^r\phi(x)[y-x]^r + \frac{1}{(p-2)!}\int_0^1 (1-t)^{p-2}(\mathcal{D}^{p-1}\phi(x + t(y-x)) - \mathcal{D}^{p-1}\phi(x))[y-x]^{p-1}\mathrm{d}t.
\end{aligned}
$$

Using this, (4), (39), and (40), we obtain that

$$
\begin{aligned}
&\Big|\Big\langle \nabla f(y) - \nabla f(x) - \sum_{r=1}^{p-1}\frac{1}{r!}\nabla^{r+1} f(x)(y-x)^r, u\Big\rangle\Big| \overset{(39)}{=} \Big|\phi(y) - \phi(x) - \sum_{r=1}^{p-1}\frac{1}{r!}\mathcal{D}^r\phi(x)[y-x]^r\Big| \\
&= \Big|\frac{1}{(p-2)!}\int_0^1 (1-t)^{p-2}(\mathcal{D}^{p-1}\phi(x + t(y-x)) - \mathcal{D}^{p-1}\phi(x))[y-x]^{p-1}\mathrm{d}t\Big| \\
&\overset{(4)}{\leq} \frac{1}{(p-2)!}\|y-x\|^{p-1}\int_0^1 (1-t)^{p-2}\|\mathcal{D}^{p-1}\phi(x + t(y-x)) - \mathcal{D}^{p-1}\phi(x)\|\mathrm{d}t \\
&\overset{(40)}{\leq} \frac{1}{(p-2)!}\|y-x\|^{p-1}\|u\|\int_0^1 (1-t)^{p-2}\|\mathcal{D}^p f(x + t(y-x)) - \mathcal{D}^p f(x)\|\mathrm{d}t \\
&\leq \frac{1}{(p-2)!}L_p\|y-x\|^p\|u\|\int_0^1 (1-t)^{p-2}t\mathrm{d}t = \frac{1}{p!}L_p\|y-x\|^p\|u\|,
\end{aligned}
$$

where the last inequality follows from Assumption 2, and the last equality is due to $\int_0^1 (1-t)^{p-2}t\mathrm{d}t = 1/(p(p-1))$. Taking the maximum of this inequality over all u with $\|u\| \leq 1$, we conclude that this lemma holds. $\square$

5.1 Proof of the main results in Section 3.1

In this subsection, we first establish several technical lemmas and then use them to prove Theorems 1 and 2. The following lemma presents a recurrence relation for the estimation error of the gradient estimators $\{m^k\}$ generated by Algorithm 1.

Lemma 6. *Suppose that Assumption 1 holds. Let $\{(x^k, m^k)\}$ be the sequence generated by Algorithm 1 with input parameters $\{(\eta_k, \theta_k)\}$. Then we have*

$$\mathbb{E}_{\xi^{k+1}}[\|m^{k+1} - \nabla f(x^{k+1})\|^\alpha] \le (1-\theta_k)\|m^k - \nabla f(x^k)\|^\alpha + 3L_1^\alpha \eta_k^\alpha \theta_k^{1-\alpha} + 2\sigma^\alpha \theta_k^\alpha \qquad \forall k \ge 0, \tag{41}$$

where L_1, σ, and α are given in Assumption 1.

Proof. Fix any $k \ge 0$. It follows from (7) that

$$\begin{aligned} m^{k+1} - \nabla f(x^{k+1}) &\overset{(7)}{=} (1-\theta_k)m^k + \theta_k G(x^{k+1};\xi^{k+1}) - \nabla f(x^{k+1}) \\ &= (1-\theta_k)(m^k - \nabla f(x^k)) + (1-\theta_k)(\nabla f(x^k) - \nabla f(x^{k+1})) + \theta_k(G(x^{k+1};\xi^{k+1}) - \nabla f(x^{k+1})). \end{aligned} \tag{42}$$

Observe from Algorithm 1 and Assumption 1 that $\|x^{k+1} - x^k\| = \eta_k$, $\mathbb{E}_{\xi^{k+1}}[G(x^{k+1};\xi^{k+1}) - \nabla f(x^{k+1})] = 0$, $\mathbb{E}_{\xi^{k+1}}[\|G(x^{k+1};\xi^{k+1}) - \nabla f(x^{k+1})\|^\alpha] \le \sigma^\alpha$, and $\|\nabla f(x^k) - \nabla f(x^{k+1})\| \le L_1\eta_k$. Using these, (35), (36), and (42), we obtain that for all $c > 0$,

$$\begin{aligned} &\mathbb{E}_{\xi^{k+1}}[\|m^{k+1} - \nabla f(x^{k+1})\|^\alpha] \\ &\overset{(42)}{=} \mathbb{E}_{\xi^{k+1}}[\|(1-\theta_k)(m^k - \nabla f(x^k)) + (1-\theta_k)(\nabla f(x^k) - \nabla f(x^{k+1})) + \theta_k(G(x^{k+1};\xi^{k+1}) - \nabla f(x^{k+1}))\|^\alpha] \\ &\overset{(35)}{\le} \|(1-\theta_k)(m^k - \nabla f(x^k)) + (1-\theta_k)(\nabla f(x^k) - \nabla f(x^{k+1}))\|^\alpha + 2\mathbb{E}_{\xi^{k+1}}[\|\theta_k(G(x^{k+1};\xi^{k+1}) - \nabla f(x^{k+1}))\|^\alpha] \\ &\overset{(36)}{\le} (1+c)(1-\theta_k)^\alpha\|m^k - \nabla f(x^k)\|^\alpha + (2 + (\alpha-1)^{\alpha-1}c^{1-\alpha})(1-\theta_k)^\alpha\|\nabla f(x^k) - \nabla f(x^{k+1})\|^\alpha + 2\sigma^\alpha\theta_k^\alpha \\ &\le (1+c)(1-\theta_k)^\alpha\|m^k - \nabla f(x^k)\|^\alpha + L_1^\alpha(2 + (\alpha-1)^{\alpha-1}c^{1-\alpha})(1-\theta_k)^\alpha\eta_k^\alpha + 2\sigma^\alpha\theta_k^\alpha, \end{aligned} \tag{43}$$

where the first inequality is due to (35) and $\mathbb{E}_{\xi^{k+1}}[G(x^{k+1};\xi^{k+1}) - \nabla f(x^{k+1})] = 0$, the second inequality is due to (36) and $\mathbb{E}_{\xi^{k+1}}[\|G(x^{k+1};\xi^{k+1}) - \nabla f(x^{k+1})\|^\alpha] \le \sigma^\alpha$, and the last inequality follows from $\|x^{k+1} - x^k\| = \eta_k$ and $\|\nabla f(x^k) - \nabla f(x^{k+1})\| \le L_1\eta_k$.

When $\theta_k = 1$, (41) clearly holds. For $\theta_k \in (0,1)$, letting $c = (1-\theta_k)^{1-\alpha} - 1$ in (43), and using the fact that $\alpha \in (1,2]$, we have

$$\begin{aligned} c^{1-\alpha} = ((1-\theta_k)^{1-\alpha} - 1)^{1-\alpha} &= \Big(\frac{1}{(1-\theta_k)^{\alpha-1}} - 1\Big)^{1-\alpha} \le \Big(\frac{1}{1-(\alpha-1)\theta_k} - 1\Big)^{1-\alpha} \\ &= \Big(\frac{1-(\alpha-1)\theta_k}{(\alpha-1)\theta_k}\Big)^{\alpha-1} \le ((\alpha-1)\theta_k)^{1-\alpha}, \end{aligned}$$

where the first inequality follows from $(1-\tau)^\beta \le 1 - \beta\tau$ for all $\tau \in (-\infty, 1)$ and $\beta \in [0,1]$. Combining this inequality with (43), one has

$$\mathbb{E}_{\xi^{k+1}}[\|m^{k+1} - \nabla f(x^{k+1})\|^\alpha] \le (1-\theta_k)\|m^k - \nabla f(x^k)\|^\alpha + L_1^\alpha(2 + \theta_k^{1-\alpha})(1-\theta_k)^\alpha\eta_k^\alpha + 2\sigma^\alpha\theta_k^\alpha,$$

which together with $\theta_k \in (0,1]$ and $\alpha \in (1,2]$ implies that (41) holds. □

The following lemma establishes a descent property for the potential sequence $\{\mathcal{P}_k\}$ defined below.

Lemma 7. *Suppose that Assumption 1 holds. Let $\{(x^k,m^k)\}$ be the sequence generated by Algorithm 1 with input parameters $\{(\eta_k,\theta_k)\}$, and L_1, σ, and α be given in Assumption 1, and let $\{\mathcal{P}_k\}$ be defined in* (34) *for $\{(x^k,m^k)\}$ and any nonincreasing positive sequence $\{p_k\}$. Then it holds that for all $k\geq 0$,*

$$\mathbb{E}_{\xi^{k+1}}[\mathcal{P}_{k+1}]\leq \mathcal{P}_k-\eta_k\|\nabla f(x^k)\|+\frac{L_1}{2}\eta_k^2+\frac{(\alpha-1)(2\eta_k/\alpha)^{\alpha/(\alpha-1)}}{(\theta_k p_k)^{1/(\alpha-1)}}+3L_1^\alpha\theta_k^{1-\alpha}\eta_k^\alpha p_k+2\sigma^\alpha\theta_k^\alpha p_k. \tag{44}$$

Proof. Fix any $k\geq 0$. By Lemma 4 with $(x^+,x,m,\eta)=(x^{k+1},x^k,m^k,\eta_k)$, one has

$$f(x^{k+1})\leq f(x^k)-\eta_k\|\nabla f(x^k)\|+2\eta_k\|\nabla f(x^k)-m^k\|+\frac{L_1}{2}\eta_k^2. \tag{45}$$

Combining this with (34) and (41), we obtain that

$$\begin{aligned}\mathbb{E}_{\xi^{k+1}}[\mathcal{P}_{k+1}] &\overset{(34)}{=} \mathbb{E}_{\xi^{k+1}}[f(x^{k+1})+p_{k+1}\|m^{k+1}-\nabla f(x^{k+1})\|^\alpha]\\ &\overset{(41)(45)}{\leq} f(x^k)-\eta_k\|\nabla f(x^k)\|+2\eta_k\|\nabla f(x^k)-m^k\|+\frac{L_1}{2}\eta_k^2\\ &\quad +(1-\theta_k)p_{k+1}\|m^k-\nabla f(x^k)\|^\alpha+3L_1^\alpha\theta_k^{1-\alpha}\eta_k^\alpha p_{k+1}+2\sigma^\alpha\theta_k^\alpha p_{k+1}\\ &\leq f(x^k)-\eta_k\|\nabla f(x^k)\|+2\eta_k\|\nabla f(x^k)-m^k\|+\frac{L_1}{2}\eta_k^2\\ &\quad +(1-\theta_k)p_k\|m^k-\nabla f(x^k)\|^\alpha+3L_1^\alpha\theta_k^{1-\alpha}\eta_k^\alpha p_k+2\sigma^\alpha\theta_k^\alpha p_k,\end{aligned} \tag{46}$$

where the last inequality follows from the fact that $\{p_k\}$ is nonincreasing. In addition, letting $\alpha'=\alpha/(\alpha-1)$ and using the Young's inequality, we have

$$\begin{aligned}2\eta_k\|\nabla f(x^k)-m^k\| &\leq \frac{\big((\alpha\theta_k p_k)^{1/\alpha}\|\nabla f(x^k)-m^k\|\big)^\alpha}{\alpha}+\frac{\big(2\eta_k/(\alpha\theta_k p_k)^{1/\alpha}\big)^{\alpha'}}{\alpha'}\\ &=\theta_k p_k\|\nabla f(x^k)-m^k\|^\alpha+\frac{(\alpha-1)(2\eta_k)^{\alpha/(\alpha-1)}}{\alpha^{\alpha/(\alpha-1)}(\theta_k p_k)^{1/(\alpha-1)}}.\end{aligned}$$

This together with (46) implies that

$$\begin{aligned}\mathbb{E}_{\xi^{k+1}}[\mathcal{P}_{k+1}] &\leq f(x^k)+p_k\|m^k-\nabla f(x^k)\|^\alpha-\eta_k\|\nabla f(x^k)\|+\frac{L_1}{2}\eta_k^2\\ &\quad +\frac{(\alpha-1)(2\eta_k)^{\alpha/(\alpha-1)}}{\alpha^{\alpha/(\alpha-1)}(\theta_k p_k)^{1/(\alpha-1)}}+3L_1^\alpha\theta_k^{1-\alpha}\eta_k^\alpha p_k+2\sigma^\alpha\theta_k^\alpha p_k.\end{aligned}$$

The conclusion (44) then follows from this and (34). □

We are now ready to prove Theorem 1.

Proof of Theorem 1. Let $\{(x^k,m^k)\}$ be generated by Algorithm 1 with $\{(\eta_k,\theta_k)\}$ given in (9), and $\{\mathcal{P}_k\}$ be defined in (34) with such $\{(x^k,m^k)\}$ and the following $\{p_k\}$:

$$p_k=(k+1)^{(\alpha^2-3\alpha+2)/(3\alpha-2)}\qquad \forall k\geq 0. \tag{47}$$

Since $\alpha\in(1,2]$, one can see that $\{p_k\}$ is nonincreasing. Also, observe from (9) that $\{\eta_k\}\subset(0,+\infty)$ and $\{\theta_k\}\subset(0,1]$. Hence, $\{(\eta_k,\theta_k,p_k)\}$ satisfies the assumptions in Lemma 7 and Algorithm 1. In addition, by (34) and (47), one has that

$$\mathbb{E}[\mathcal{P}_0]=f(x^0)+p_0\mathbb{E}[\|m^0-\nabla f(x^0)\|^\alpha]=f(x^0)+\mathbb{E}[\|G(x^0;\xi^0)-\nabla f(x^0)\|^\alpha]\leq f(x^0)+\sigma^\alpha, \tag{48}$$

$$\mathbb{E}[\mathcal{P}_K] = \mathbb{E}[f(x^K) + p_K\|m^K - \nabla f(x^K)\|^\alpha] \geq \mathbb{E}[f(x^K)] \geq f_{\text{low}}. \tag{49}$$

Taking expectation on both sides of (44) with respect to $\{\xi^i\}_{i=0}^{k+1}$, we have

$$\mathbb{E}[\mathcal{P}_{k+1}] \leq \mathbb{E}[\mathcal{P}_k] - \eta_k\mathbb{E}[\|\nabla f(x^k)\|] + \frac{L_1}{2}\eta_k^2 + \frac{(\alpha-1)(2\eta_k/\alpha)^{\alpha/(\alpha-1)}}{(\theta_k p_k)^{1/(\alpha-1)}} + 3L_1^\alpha\theta_k^{1-\alpha}\eta_k^\alpha p_k + 2\sigma^\alpha\theta_k^\alpha p_k \quad \forall k \geq 0.$$

Summing up this inequality over $k = 0, \ldots, K-1$, and using (48) and (49), we obtain that for all $K \geq 1$,

$$\begin{aligned}
f_{\text{low}} &\overset{(49)}{\leq} \mathbb{E}[\mathcal{P}_K] \\
&\leq \mathbb{E}[\mathcal{P}_0] - \sum_{k=0}^{K-1}\eta_k\mathbb{E}[\|\nabla f(x^k)\|] + \sum_{k=0}^{K-1}\Big(\frac{L_1}{2}\eta_k^2 + \frac{(\alpha-1)(2\eta_k)^{\alpha/(\alpha-1)}}{\alpha^{\alpha/(\alpha-1)}(\theta_k p_k)^{1/(\alpha-1)}} + 3L_1^\alpha\theta_k^{1-\alpha}\eta_k^\alpha p_k + 2\sigma^\alpha\theta_k^\alpha p_k\Big) \\
&\overset{(48)}{\leq} f(x^0) + \sigma^\alpha - \eta_{K-1}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(x^k)\|] + \sum_{k=0}^{K-1}\Big(\frac{L_1}{2}\eta_k^2 + \frac{(\alpha-1)(2\eta_k)^{\alpha/(\alpha-1)}}{\alpha^{\alpha/(\alpha-1)}(\theta_k p_k)^{1/(\alpha-1)}} + 3L_1^\alpha\theta_k^{1-\alpha}\eta_k^\alpha p_k + 2\sigma^\alpha\theta_k^\alpha p_k\Big),
\end{aligned} \tag{50}$$

where the last inequality follows from (48) and the fact that $\{\eta_k\}$ is nonincreasing. Rearranging the terms in (50), and using (8), (9), and (47), we obtain that for all $K \geq 3$,

$$\begin{aligned}
&\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(x^k)\|] \\
&\overset{(50)}{\leq} \frac{f(x^0) - f_{\text{low}} + \sigma^\alpha}{K\eta_{K-1}} + \frac{1}{K\eta_{K-1}}\sum_{k=0}^{K-1}\Big(\frac{L_1}{2}\eta_k^2 + \frac{(\alpha-1)(2\eta_k)^{\alpha/(\alpha-1)}}{\alpha^{\alpha/(\alpha-1)}(\theta_k p_k)^{1/(\alpha-1)}} + 3L_1^\alpha\theta_k^{1-\alpha}\eta_k^\alpha p_k + 2\sigma^\alpha\theta_k^\alpha p_k\Big) \\
&\overset{(9)(47)}{=} \frac{f(x^0) - f_{\text{low}} + \sigma^\alpha}{K^{(\alpha-1)/(3\alpha-2)}} + \frac{1}{K^{(\alpha-1)/(3\alpha-2)}}\sum_{k=0}^{K-1}\Big(\frac{L_1}{2(k+1)^{2(2\alpha-1)/(3\alpha-2)}} + \frac{(\alpha-1)(2/\alpha)^{\alpha/(\alpha-1)} + 3L_1^\alpha + 2\sigma^\alpha}{k+1}\Big) \\
&\leq \frac{2(f(x^0) - f_{\text{low}} + \sigma^\alpha + L_1 + (\alpha-1)(2/\alpha)^{\alpha/(\alpha-1)} + 3L_1^\alpha + 2\sigma^\alpha)\ln K}{K^{(\alpha-1)/(3\alpha-2)}} \overset{(8)}{=} \frac{M_{1,\alpha}\ln K}{K^{(\alpha-1)/(3\alpha-2)}},
\end{aligned}$$

where the second inequality follows from $\sum_{k=0}^{K-1} 1/(k+1) \leq 2\ln K$ due to (37) and $K \geq 3$, and $\sum_{k=0}^{K-1} 1/(k+1)^{2(2\alpha-1)/(3\alpha-2)} \leq (3\alpha-2)2^{\alpha/(3\alpha-2)}/\alpha < 4$ due to (37) and $(3\alpha-2)/\alpha \in (1,2]$. Recall that ι_K is uniformly selected from $\{0,\ldots,K-1\}$. It then follows from this and the above inequality that

$$\mathbb{E}[\|\nabla f(x^{\iota_K})\|] = \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(x^k)\|] \leq \frac{M_{1,\alpha}\ln K}{K^{(\alpha-1)/(3\alpha-2)}} \qquad \forall K \geq 3. \tag{51}$$

In addition, by Lemma 3 with $(\beta, u, v) = ((\alpha-1)/(3\alpha-2), (\alpha-1)\epsilon/(2(3\alpha-2)M_{1,\alpha}), K)$, one can see that

$$K^{-(\alpha-1)/(3\alpha-2)}\ln K \leq \frac{\epsilon}{M_{1,\alpha}} \qquad \forall K \geq \Big(\frac{2(3\alpha-2)M_{1,\alpha}}{(\alpha-1)\epsilon}\ln\Big(\frac{2(3\alpha-2)M_{1,\alpha}}{(\alpha-1)\epsilon}\Big)\Big)^{(3\alpha-2)/(\alpha-1)},$$

which together with (51) implies that Theorem 1 holds. □

We next prove Theorem 2.

Proof of Theorem 2. Let $\{(x^k, m^k)\}$ be generated by Algorithm 1 with $\{(\eta_k, \theta_k)\}$ given in (11), and $\{\mathcal{P}_k\}$ be defined in (34) with such $\{(x^k, m^k)\}$ and the following $\{p_k\}$:

$$p_k = (k+1)^{(2\alpha^2-5\alpha+2)/(4\alpha)} \qquad \forall k \geq 0. \tag{52}$$

Since $\alpha \in (1,2]$, one can see that $\{p_k\}$ is nonincreasing. Also, observe from (11) that $\{\eta_k\} \subset (0,+\infty)$ and $\{\theta_k\} \subset (0,1]$. Hence, $\{(\eta_k, \theta_k, p_k)\}$ defined in (11) and (52) satisfies the assumptions in Lemma 7 and Algorithm 1. By (11), (52), and similar arguments as those for deriving (50), one has that for all $K \geq 1$,

$$f_{\text{low}} \leq f(x^0) + \sigma^\alpha - \eta_{K-1}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(x^k)\|] + \sum_{k=0}^{K-1}\Big(\frac{L_1}{2}\eta_k^2 + \frac{(\alpha-1)(2\eta_k)^{\alpha/(\alpha-1)}}{\alpha^{\alpha/(\alpha-1)}(\theta_k p_k)^{1/(\alpha-1)}} + 3L_1^\alpha\theta_k^{1-\alpha}\eta_k^\alpha p_k + 2\sigma^\alpha\theta_k^\alpha p_k\Big).$$

Rearranging the terms of this inequality, and using (10), (11), and (52), we obtain that for all $K \geq 3$,

$$\begin{aligned}
&\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(x^k)\|]\\
&\leq \frac{f(x^0) - f_{\text{low}} + \sigma^\alpha}{K\eta_{K-1}} + \frac{1}{K\eta_{K-1}}\sum_{k=0}^{K-1}\Big(\frac{L_1}{2}\eta_k^2 + \frac{(\alpha-1)(2\eta_k)^{\alpha/(\alpha-1)}}{\alpha^{\alpha/(\alpha-1)}(\theta_k p_k)^{1/(\alpha-1)}} + 3L_1^\alpha\theta_k^{1-\alpha}\eta_k^\alpha p_k + 2\sigma^\alpha\theta_k^\alpha p_k\Big)\\
&\overset{(11)(52)}{=} \frac{f(x^0) - f_{\text{low}} + \sigma^\alpha}{K^{1/4}} + \frac{1}{K^{1/4}}\sum_{k=0}^{K-1}\Big(\frac{L_1}{2(k+1)^{3/2}} + \frac{(\alpha-1)(2/\alpha)^{\alpha/(\alpha-1)} + 2\sigma^\alpha}{(k+1)^{(5\alpha-2)/(4\alpha)}} + \frac{3L_1^\alpha}{(k+1)^{(7\alpha-\alpha^2-2)/(4\alpha)}}\Big)\\
&\leq \frac{f(x^0) - f_{\text{low}} + \sigma^\alpha}{K^{1/4}} + \frac{1}{K^{1/4}}\sum_{k=0}^{K-1}\Big(\frac{L_1/2 + 3L_1^\alpha + ((\alpha-1)(2/\alpha)^{\alpha/(\alpha-1)} + 2\sigma^\alpha)K^{(2-\alpha)/(4\alpha)}}{k+1}\Big)\\
&\leq \frac{2(f(x^0) - f_{\text{low}} + \sigma^\alpha + L_1/2 + 3L_1^\alpha)\ln K}{K^{1/4}} + \frac{2((\alpha-1)(2/\alpha)^{\alpha/(\alpha-1)} + 2\sigma^\alpha)\ln K}{K^{(\alpha-1)/(2\alpha)}}\\
&\overset{(10)}{=} \frac{\widetilde{M}_{1,\alpha}\ln K}{K^{1/4}} + \frac{\widehat{M}_{1,\alpha}\ln K}{K^{(\alpha-1)/(2\alpha)}},
\end{aligned}$$

where the second inequality follows from $(5\alpha-2)/(4\alpha) \leq 1$ and $(7\alpha-\alpha^2-2)/(4\alpha) \geq 1$ due to $\alpha \in (1,2]$, and the last inequality follows from $\sum_{k=0}^{K-1} 1/(k+1) \leq 2\ln K$ due to (37) and $K \geq 3$. Recall that ι_K is uniformly selected from $\{0,\ldots,K-1\}$. It follows from this and the above inequality that

$$\mathbb{E}[\|\nabla f(x^{\iota_K})\|] = \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(x^k)\|] \leq \frac{\widetilde{M}_{1,\alpha}\ln K}{K^{1/4}} + \frac{\widehat{M}_{1,\alpha}\ln K}{K^{(\alpha-1)/(2\alpha)}} \qquad \forall K \geq 3. \tag{53}$$

In addition, by Lemma 3 with $(\beta, u, v) = (1/4, \epsilon/(16\widetilde{M}_{1,\alpha}), K)$ and $(\beta, u, v) = ((\alpha-1)/(2\alpha), (\alpha-1)\epsilon/(8\alpha\widehat{M}_{1,\alpha}), K)$, one can see that

$$\begin{aligned}
&K^{-1/4}\ln K \leq \frac{\epsilon}{2\widetilde{M}_{1,\alpha}} \qquad \forall K \geq \Big(\frac{16\widetilde{M}_{1,\alpha}}{\epsilon}\ln\Big(\frac{16\widetilde{M}_{1,\alpha}}{\epsilon}\Big)\Big)^4,\\
&K^{-(\alpha-1)/(2\alpha)}\ln K \leq \frac{\epsilon}{2\widehat{M}_{1,\alpha}} \qquad \forall K \geq \Big(\frac{8\alpha\widehat{M}_{1,\alpha}}{(\alpha-1)\epsilon}\ln\Big(\frac{8\alpha\widehat{M}_{1,\alpha}}{(\alpha-1)\epsilon}\Big)\Big)^{2\alpha/(\alpha-1)},
\end{aligned}$$

which together with (53) imply that Theorem 2 holds. $\square$

5.2 Proof of the main results in Section 3.2

In this subsection, we first establish several technical lemmas and then use them to prove Theorems 3 and 4.

Before proceeding, we present the well-known Weierstrass product inequalities (see, e.g., [16]). For any given $\{a_t\}_{t=1}^m \subset (0,1)$, it holds that

$$1 - \sum_{t=1}^m a_t \le \prod_{t=1}^m (1-a_t) \le \frac{1}{1+\sum_{t=1}^m a_t}. \tag{54}$$

We next present an auxiliary lemma that will be used subsequently.

Lemma 8. *$\prod_{s\ge 1, s\ne t}(1-t^2/s^2) = (-1)^{t-1}/2$ holds for any positive integer t.*

Proof. Fix any positive integer t. Let $\phi(a) = \prod_{s\ge 1, s\ne t}(1-a^2/s^2)$ for any $a \ge 0$. Observe that the sequence $\{u_r\}$ is decreasing and $u_r \ge 0$ when $r \ge a+1$, where $u_r = \prod_{a+1\le s\le r, s\ne t}(1-a^2/s^2)$. This implies that $\phi(a)$ is well-defined for all $a \ge 0$. In addition, it is well-known that the normalized sinc function $\sin(\pi y)/(\pi y)$ can be represented as the infinite product $\prod_{s=1}^\infty(1-y^2/s^2)$ for all $y \in \mathbb{R}$. By this, one can see that

$$\phi(a) = \frac{\prod_{s=1}^\infty (1-a^2/s^2)}{1-a^2/t^2} = \frac{\sin(\pi a)}{\pi a(1-a^2/t^2)} \qquad \forall a \ne t.$$

It then follows that

$$\lim_{a\to t}\phi(a) = \lim_{a\to t}\frac{\sin(\pi a)}{\pi a(1-a^2/t^2)} = \lim_{a\to t}\frac{\cos(\pi a)}{1-3a^2/t^2} = \frac{(-1)^{t-1}}{2}, \tag{55}$$

where the second equality is due to L'Hôpital's rule. Thus, to prove this lemma, it suffices to show that $\phi(a)$ is continuous at $a = t$. To this end, we define a sequence of functions $\{\phi_r\}$ as follows:

$$\phi_r(a) = \prod_{1\le s\le r, s\ne t}(1-a^2/s^2) \qquad \forall a \ge 0$$

for each $r \ge 1$. We next show that ϕ_r converges uniformly to ϕ on $[0,2t]$. Indeed, let $R \in [4t,\infty)$ be arbitrarily chosen, and fix any r_1, r_2 satisfying $r_2 > r_1 > R$. Observe that

$$|\phi_{r_1}(a) - \phi_{r_2}(a)| = \Big(1 - \prod_{r_1+1\le s\le r_2}(1-a^2/s^2)\Big)\Big|\prod_{1\le s\le r_1, s\ne t}(1-a^2/s^2)\Big| \qquad \forall a \in [0,2t]. \tag{56}$$

In addition, one has

$$\prod_{r_1+1\le s\le r_2}(1-a^2/s^2) \overset{(54)}{\ge} 1 - \sum_{s=r_1+1}^\infty (a^2/s^2) \ge 1 - \sum_{s=r_1+1}^\infty \frac{a^2}{s(s-1)} = 1 - \frac{a^2}{r_1} \ge 1 - \frac{a^2}{R} \qquad \forall a \in [0,2t],$$

which, together with (56) and $r_2 > r_1 > R$, implies that

$$\begin{aligned} &|\phi_{r_1}(a) - \phi_{r_2}(a)| \le \frac{a^2|\prod_{1\le s\le r_1, s\ne t}(1-a^2/s^2)|}{R} \\ &\le \frac{a^2\prod_{1\le s\le \lceil a\rceil -1, s\ne t}(a^2/s^2-1)}{R} \le \frac{4t^2\prod_{1\le s\le 2t-1}(4t^2/s^2-1)}{R} \qquad \forall a \in [0,2t],^5 \end{aligned}$$

[5] $\prod_{1\le s\le \lceil a\rceil -1, s\ne t}(a^2/s^2-1)$ is set to 1 if $a \in [0,1]$.

where the second inequality follows from $(1 - a^2/s^2) \in [0, 1]$ for all $s \geq \lceil a \rceil$, and the last inequality is due to $a \leq 2t$. By this and the choice of r_1, r_2, and R, one can conclude that $\{\phi_r\}$ converges uniformly to ϕ on $[0, 2t]$, which together with the continuity of ϕ_r for each $r \geq 1$ implies that ϕ is continuous on $[0, 2t]$. Hence, one has $\phi(t) = \lim_{a \to t} \phi(a)$, which along with (55) and the definition of ϕ implies that this lemma holds. $\square$

The following lemma provides a set of choices for $(\gamma_{k,t}, \theta_{k,t})$ that satisfy (23) and (24).

Lemma 9. *Let $\{\gamma_k\} \subset (0, 1/2]$ and a positive integer q be given, and*

$$\gamma_{k,t} = \gamma_k/t^2, \quad \theta_{k,t} = \frac{\prod_{1 \leq s \leq q, s \neq t}(1 - s^2/\gamma_k)}{(t^2/\gamma_k)\prod_{1 \leq s \leq q, s \neq t}((t^2 - s^2)/\gamma_k)} \qquad \forall 1 \leq t \leq q, k \geq 0. \tag{57}$$

Then $\{(\gamma_{k,t}, \theta_{k,t})\}$ satisfies (23). Moreover, it holds that

$$\sum_{t=1}^{q} \theta_{k,t} \in \Big(\frac{\gamma_k}{1 + \pi^2/6}, 2\gamma_k\Big) \subset (0, 1), \quad |\theta_{k,t}| \leq \frac{4\gamma_k}{t^2} \qquad \forall 1 \leq t \leq q, k \geq 0. \tag{58}$$

Proof. Fix any $k \geq 0$. We first prove that $\{(\gamma_{k,t}, \theta_{k,t})\}$ satisfies (23). For convenience, we denote the coefficient matrix in (23) as

$$\Gamma = \begin{bmatrix} 1/\gamma_{k,1} & 1/\gamma_{k,2} & \cdots & 1/\gamma_{k,q} \\ 1/\gamma_{k,1}^2 & 1/\gamma_{k,2}^2 & \cdots & 1/\gamma_{k,q}^2 \\ \vdots & \vdots & \vdots & \vdots \\ 1/\gamma_{k,1}^q & 1/\gamma_{k,2}^q & \cdots & 1/\gamma_{k,q}^q \end{bmatrix} \in \mathbb{R}^{q \times q}.$$

In addition, we define a matrix $V \in \mathbb{R}^{q \times q}$, whose t-th row $[v_{t1} \ \cdots \ v_{tq}]$ consists of the coefficients of the polynomial

$$h_t(\alpha) = \frac{\alpha \prod_{1 \leq s \leq q, s \neq t}(\alpha - 1/\gamma_{k,s})}{(1/\gamma_{k,t})\prod_{1 \leq s \leq q, s \neq t}(1/\gamma_{k,t} - 1/\gamma_{k,s})} = v_{t1}\alpha + v_{t2}\alpha^2 + \cdots + v_{tq}\alpha^q \qquad \forall 1 \leq t \leq q,$$

which satisfies $h_t(1/\gamma_{k,t}) = 1$ and $h_t(1/\gamma_{k,s}) = 0$ for all $s \neq t$. By the definitions of h_t, Γ, and V, one has

$$\begin{aligned} V\Gamma &= \begin{bmatrix} \sum_{s=1}^q v_{1s}/\gamma_{k,1}^s & \sum_{s=1}^q v_{1s}/\gamma_{k,2}^s & \cdots & \sum_{s=1}^q v_{1s}/\gamma_{k,q}^s \\ \sum_{s=1}^q v_{2s}/\gamma_{k,1}^s & \sum_{s=1}^q v_{2s}/\gamma_{k,2}^s & \cdots & \sum_{s=1}^q v_{2s}/\gamma_{k,q}^s \\ \vdots & \vdots & \vdots & \vdots \\ \sum_{s=1}^q v_{qs}/\gamma_{k,1}^s & \sum_{s=1}^q v_{qs}/\gamma_{k,2}^s & \cdots & \sum_{s=1}^q v_{qs}/\gamma_{k,q}^s \end{bmatrix} \\ &= \begin{bmatrix} h_1(1/\gamma_{k,1}) & h_1(1/\gamma_{k,2}) & \cdots & h_1(1/\gamma_{k,q}) \\ h_2(1/\gamma_{k,1}) & h_2(1/\gamma_{k,2}) & \cdots & h_2(1/\gamma_{k,q}) \\ \vdots & \vdots & \vdots & \vdots \\ h_q(1/\gamma_{k,1}) & h_q(1/\gamma_{k,2}) & \cdots & h_q(1/\gamma_{k,q}) \end{bmatrix} = I, \end{aligned}$$

where I is the $q \times q$ identity matrix. Hence, we have $V = \Gamma^{-1}$. In view of this and the definition of V, one can see that the solution to (23) is unique and can be written as

$$\begin{bmatrix} \theta_{k,1} \\ \theta_{k,2} \\ \vdots \\ \theta_{k,q} \end{bmatrix} = V \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = \begin{bmatrix} h_1(1) \\ h_2(1) \\ \vdots \\ h_q(1) \end{bmatrix}, \tag{59}$$

which together with the definition of h_t implies that $\{(\gamma_{k,t},\theta_{k,t})\}$ satisfies (23).

To prove the first relation in (58), we first establish the following equality:

$$\sum_{t=1}^{q}\theta_{k,t} = 1-\frac{\prod_{t=1}^{q}(1/\gamma_{k,t}-1)}{\prod_{t=1}^{q}1/\gamma_{k,t}}. \tag{60}$$

Notice from (59) that $\theta_{k,t}=h_t(1)$ for each $1\le t\le q$. Let $h(\alpha)=\sum_{t=1}^{q}h_t(\alpha)$. It then follows that

$$\sum_{t=1}^{q}\theta_{k,t}=\sum_{t=1}^{q}h_t(1)=h(1). \tag{61}$$

Also, observe that $h_t(0)=0$ for $1\le t\le q$. By this, $h_t(1/\gamma_{k,t})=1$ and $h_t(1/\gamma_{k,s})=0$ for all $1\le s\le q$ and $s\neq t$, one can see that h satisfies $h(0)=0$ and $h(1/\gamma_{k,t})=1$ for all $1\le t\le q$. Using these and the fact that $1/\gamma_{k,t}$, $1\le t\le q$, are distinct, we conclude that h is uniquely given by

$$h(\alpha)=1-\frac{\prod_{t=1}^{q}(1/\gamma_{k,t}-\alpha)}{\prod_{t=1}^{q}1/\gamma_{k,t}},$$

which along with (61) implies that (60) holds as desired.

We are now ready to prove the first relation in (58). Substituting the definition of $\{\gamma_{k,t}\}$ given in (57) into (60), we obtain

$$\sum_{t=1}^{q}\theta_{k,t}\overset{(57)}{=}1-\frac{\prod_{1\le t\le q}(t^2/\gamma_k-1)}{\prod_{1\le t\le q}(t^2/\gamma_k)}=1-\prod_{1\le t\le q}\Big(1-\frac{\gamma_k}{t^2}\Big). \tag{62}$$

It follows from (54) that

$$1-\gamma_k\sum_{t=1}^{q}\frac{1}{t^2}\le\prod_{1\le t\le q}\Big(1-\frac{\gamma_k}{t^2}\Big)\le\frac{1}{1+\gamma_k\sum_{t=1}^{q}(1/t^2)}.$$

Using these, (62), and the identity $\sum_{t=1}^{\infty}(1/t^2)=\pi^2/6$, we obtain that

$$\frac{\gamma_k}{1+\pi^2/6}<\frac{\gamma_k\sum_{t=1}^{q}(1/t^2)}{1+\gamma_k\sum_{t=1}^{q}(1/t^2)}\le\sum_{t=1}^{q}\theta_{k,t}\le\gamma_k\sum_{t=1}^{q}\frac{1}{t^2}\le\frac{\pi^2\gamma_k}{6}<2\gamma_k,$$

where the first inequality is due to $\gamma_k\sum_{t=1}^{q}(1/t^2)<\pi^2/6$ and $\sum_{t=1}^{q}(1/t^2)\ge1$. The above inequalities along with $\gamma_k\in(0,1/2]$ implies that the first relation in (58) holds. In addition, by (57), $\gamma_k\in(0,1/2]$, and Lemma 8, one can see that for all $1\le t\le q$,

$$\begin{aligned}
|\theta_{k,t}| &= \frac{\prod_{1\le s\le q}(s^2/\gamma_k-1)}{\prod_{1\le s\le q}(s^2/\gamma_k)}\cdot\frac{\prod_{1\le s\le q,s\neq t}(s^2/\gamma_k)}{|\prod_{1\le s\le q,s\neq t}((t^2-s^2)/\gamma_k)|}\cdot\frac{1}{t^2/\gamma_k-1}\\
&= \frac{\prod_{1\le s\le q}(1-\gamma_k/s^2)}{|\prod_{1\le s\le q,s\neq t}(t^2/s^2-1)|}\cdot\frac{1}{t^2/\gamma_k-1}\le\frac{\prod_{1\le s\le q}(1-\gamma_k/s^2)}{|\prod_{s\ge1,s\neq t}(t^2/s^2-1)|}\cdot\frac{1}{t^2/\gamma_k-1}\\
&= \frac{2\prod_{1\le s\le q}(1-\gamma_k/s^2)}{t^2/\gamma_k-1}\le\frac{2}{t^2/\gamma_k-1}\le\frac{4\gamma_k}{t^2},
\end{aligned}$$

where the first inequality is due to $|\prod_{s\ge q+1}(t^2/s^2-1)|\le1$, the third equality follows from Lemma 8, the second inequality is due to $\prod_{1\le s\le q}(1-\gamma_k/s^2)\in(0,1)$, and the last inequality is due to $\gamma_k\in(0,1/2]$ and $t\ge1$. Hence, the second relation in (58) also holds, which completes the proof of this lemma. □

The next lemma shows that $\{(\gamma_{k,t},\theta_{k,t})\}$ satisfying (23) can lead to the following important identity that will be used for the subsequent analysis.

Lemma 10. *Suppose that Assumptions 1 and 2 hold. Let $\mathcal{R}_p(\cdot,\cdot)$ be defined in (5), and $\{x^k\}$ and $\{z^{k,t}\}$ be generated by Algorithm 2 with input parameters $q=p-1$ and $\{(\gamma_{k,t},\theta_{k,t})\}$ satisfying (23), where p is given in Assumption 2. Then it holds that for all $k\ge 0$,*

$$\nabla f(x^{k+1}) = \Big(1-\sum_{t=1}^{p-1}\theta_{k,t}\Big)\nabla f(x^k) + \sum_{t=1}^{p-1}\theta_{k,t}\nabla f(z^{k+1,t}) + \mathcal{R}_p(x^{k+1},x^k) - \sum_{t=1}^{p-1}\theta_{k,t}\mathcal{R}_p(z^{k+1,t},x^k). \quad (63)$$

Proof. Fix any $k\ge 0$. It follows from (12) with $q=p-1$ that

$$z^{k+1,t}-x^k = \frac{1}{\gamma_{k,t}}(x^{k+1}-x^k) \qquad \forall 1\le t\le p-1. \quad (64)$$

By this and (23), one has that

$$\begin{aligned}
\nabla f(x^{k+1}) &\overset{(23)}{=} \nabla f(x^{k+1}) - \sum_{r=2}^{p}\Big(\Big(1-\sum_{t=1}^{p-1}\frac{\theta_{k,t}}{\gamma_{k,t}^{r-1}}\Big)\frac{1}{(r-1)!}\nabla^r f(x^k)(x^{k+1}-x^k)^{r-1}\Big)\\
&= \nabla f(x^{k+1}) - \sum_{r=2}^{p}\frac{1}{(r-1)!}\nabla^r f(x^k)(x^{k+1}-x^k)^{r-1} + \sum_{t=1}^{p-1}\sum_{r=2}^{p}\frac{\theta_{k,t}}{(r-1)!\gamma_{k,t}^{r-1}}\nabla^r f(x^k)(x^{k+1}-x^k)^{r-1}\\
&\overset{(64)}{=} \nabla f(x^{k+1}) - \sum_{r=2}^{p}\frac{1}{(r-1)!}\nabla^r f(x^k)(x^{k+1}-x^k)^{r-1} + \sum_{t=1}^{p-1}\sum_{r=2}^{p}\frac{\theta_{k,t}}{(r-1)!}\nabla^r f(x^k)(z^{k+1,t}-x^k)^{r-1}\\
&= \Big(1-\sum_{t=1}^{p-1}\theta_{k,t}\Big)\nabla f(x^k) + \sum_{t=1}^{p-1}\theta_{k,t}\nabla f(z^{k+1,t}) + \Big(\nabla f(x^{k+1}) - \sum_{r=1}^{p}\frac{1}{(r-1)!}\nabla^r f(x^k)(x^{k+1}-x^k)^{r-1}\Big)\\
&\quad - \sum_{t=1}^{p-1}\theta_{k,t}\Big(\nabla f(z^{k+1,t}) - \sum_{r=1}^{p}\frac{1}{(r-1)!}\nabla^r f(x^k)(z^{k+1,t}-x^k)^{r-1}\Big),
\end{aligned}$$

which along with the definition of $\mathcal{R}_p$ in (5) implies that (63) holds. □

The following lemma presents a recurrence relation for the estimation error of the gradient estimators $\{m^k\}$ generated by Algorithm 2.

Lemma 11. *Suppose that Assumptions 1 and 2 hold. Let $\{(x^k,m^k)\}$ be the sequence generated by Algorithm 2 with input parameters $q=p-1$, $\{\eta_k\}$, and $\{(\gamma_{k,t},\theta_{k,t})\}$ satisfying (23) and (24). Then we have*

$$\begin{aligned}
\mathbb{E}_{\xi^{k+1}}[\|m^{k+1}-\nabla f(x^{k+1})\|^\alpha] &\le \Big(1-\sum_{t=1}^{p-1}\theta_{k,t}\Big)\|m^k-\nabla f(x^k)\|^\alpha\\
&\quad + \frac{3p^{\alpha-1}L_p^\alpha}{(p!)^\alpha}\Big(\sum_{t=1}^{p-1}\theta_{k,t}\Big)^{1-\alpha}\eta_k^{\alpha p}\Big(1+\sum_{t=1}^{p-1}\frac{|\theta_{k,t}|^\alpha}{\gamma_{k,t}^{\alpha p}}\Big) + 2(p-1)^{\alpha-1}\sigma^\alpha\sum_{t=1}^{p-1}|\theta_{k,t}|^\alpha \qquad \forall k\ge 0, \quad (65)
\end{aligned}$$

where L_1, σ, α, p, and L_p are given in Assumptions 1 and 2, respectively.

Proof. Fix any $k\ge 0$. It follows from (13), (63), and $q=p-1$ that

$$\nabla f(x^{k+1}) - m^{k+1} = \Big(1-\sum_{t=1}^{p-1}\theta_{k,t}\Big)(\nabla f(x^k)-m^k) + \sum_{t=1}^{p-1}\theta_{k,t}(\nabla f(z^{k+1,t}) - G(z^{k+1,t};\xi^{k+1}))$$

$$
+\mathcal{R}_p(x^{k+1},x^k)-\sum_{t=1}^{p-1}\theta_{k,t}\mathcal{R}_p(z^{k+1,t},x^k). \tag{66}
$$

Observe from Algorithm 2 and Assumption 1 that $\|x^{k+1}-x^k\|=\eta_k$, $\|z^{k+1,t}-x^k\|=\eta_k/\gamma_{k,t}$, $\mathbb{E}_{\xi^{k+1}}[G(z^{k+1,t};\xi^{k+1})-\nabla f(z^{k+1,t})]=0$, and $\mathbb{E}_{\xi^{k+1}}[\|\nabla f(z^{k+1,t})-G(z^{k+1,t};\xi^{k+1})\|^\alpha]\le\sigma^\alpha$ for all $1\le t\le p-1$. Using these, (35), (36), (42), (66), and Lemma 5, we obtain that for all $c>0$,

$$
\begin{aligned}
&\mathbb{E}_{\xi^{k+1}}[\|\nabla f(x^{k+1})-m^{k+1}\|^\alpha]\\
&\overset{(66)}{=}\mathbb{E}_{\xi^{k+1}}\Big[\Big\|\Big(1-\sum_{t=1}^{p-1}\theta_{k,t}\Big)(\nabla f(x^k)-m^k)+\sum_{t=1}^{p-1}\theta_{k,t}(\nabla f(z^{k+1,t})-G(z^{k+1,t};\xi^{k+1}))\\
&\qquad+\mathcal{R}_p(x^{k+1},x^k)-\sum_{t=1}^{p-1}\theta_{k,t}\mathcal{R}_p(z^{k+1,t},x^k)\Big\|^\alpha\Big]\\
&\overset{(35)}{\le}\Big\|\Big(1-\sum_{t=1}^{p-1}\theta_{k,t}\Big)(\nabla f(x^k)-m^k)+\mathcal{R}_p(x^{k+1},x^k)-\sum_{t=1}^{p-1}\theta_{k,t}\mathcal{R}_p(z^{k+1,t},x^k)\Big\|^\alpha\\
&\qquad+2\mathbb{E}_{\xi^{k+1}}\Big[\Big\|\sum_{t=1}^{p-1}\theta_{k,t}(\nabla f(z^{k+1,t})-G(z^{k+1,t};\xi^{k+1}))\Big\|^\alpha\Big]\\
&\overset{(36)}{\le}(1+c)\Big(1-\sum_{t=1}^{p-1}\theta_{k,t}\Big)^\alpha\|\nabla f(x^k)-m^k\|^\alpha+(2+(\alpha-1)^{\alpha-1}c^{1-\alpha})\Big\|\mathcal{R}_p(x^{k+1},x^k)-\sum_{t=1}^{p-1}\theta_{k,t}\mathcal{R}_p(z^{k+1,t},x^k)\Big\|^\alpha\\
&\qquad+2\mathbb{E}_{\xi^{k+1}}\Big[\Big\|\sum_{t=1}^{p-1}\theta_{k,t}(\nabla f(z^{k+1,t})-G(z^{k+1,t};\xi^{k+1}))\Big\|^\alpha\Big]\\
&\le(1+c)\Big(1-\sum_{t=1}^{p-1}\theta_{k,t}\Big)^\alpha\|\nabla f(x^k)-m^k\|^\alpha\\
&\qquad+p^{\alpha-1}(2+(\alpha-1)^{\alpha-1}c^{1-\alpha})\Big(\|\mathcal{R}_p(x^{k+1},x^k)\|^\alpha+\sum_{t=1}^{p-1}|\theta_{k,t}|^\alpha\|\mathcal{R}_p(z^{k+1,t},x^k)\|^\alpha\Big)\\
&\qquad+2(p-1)^{\alpha-1}\sum_{t=1}^{p-1}|\theta_{k,t}|^\alpha\mathbb{E}_{\xi^{k+1}}[\|\nabla f(z^{k+1,t})-G(z^{k+1,t};\xi^{k+1})\|^\alpha]\\
&\le(1+c)\Big(1-\sum_{t=1}^{p-1}\theta_{k,t}\Big)^\alpha\|\nabla f(x^k)-m^k\|^\alpha\\
&\qquad+\frac{p^{\alpha-1}L_p^\alpha}{(p!)^\alpha}(2+(\alpha-1)^{\alpha-1}c^{1-\alpha})\eta_k^{\alpha p}\Big(1+\sum_{t=1}^{p-1}\frac{|\theta_{k,t}|^\alpha}{\gamma_{k,t}^{\alpha p}}\Big)+2(p-1)^{\alpha-1}\sigma^\alpha\sum_{t=1}^{p-1}|\theta_{k,t}|^\alpha,
\end{aligned}\tag{67}
$$

where the first inequality is due to (35), and the fact that $\mathbb{E}_{\xi^{k+1}}[G(z^{k+1,t};\xi^{k+1})-\nabla f(z^{k+1,t})]=0$ for all $1\le t\le p-1$, the third inequality follows from $\|\sum_{t=1}^m w_t\|^\alpha\le m^{\alpha-1}\sum_{t=1}^m\|w_t\|^\alpha$ because of the convexity of $\|\cdot\|^\alpha$, and the last inequality follows from Lemma 5, $\|x^{k+1}-x^k\|=\eta_k$, $\|z^{k+1,t}-x^k\|=\eta_k/\gamma_{k,t}$, and $\mathbb{E}_{\xi^{k+1}}[\|\nabla f(z^{k+1,t})-G(z^{k+1,t};\xi^{k+1})\|^\alpha]\le\sigma^\alpha$ for all $1\le t\le p-1$.

Letting $c=(1-\sum_{t=1}^{p-1}\theta_{k,t})^{1-\alpha}-1$ in (67), and using $\sum_{t=1}^{p-1}\theta_{k,t}\in(0,1)$ (see (24)) and $\alpha\in(1,2]$, we obtain that

$$
c^{1-\alpha}=\Big(\frac{1}{(1-\sum_{t=1}^{p-1}\theta_{k,t})^{\alpha-1}}-1\Big)^{1-\alpha}\le\Big(\frac{1}{1-(\alpha-1)\sum_{t=1}^{p-1}\theta_{k,t}}-1\Big)^{1-\alpha}
$$

$$= \Big(\frac{1-(\alpha-1)\sum_{t=1}^{p-1}\theta_{k,t}}{(\alpha-1)\sum_{t=1}^{p-1}\theta_{k,t}}\Big)^{\alpha-1} \le \Big((\alpha-1)\sum_{t=1}^{p-1}\theta_{k,t}\Big)^{1-\alpha},$$

where the first inequality follows from $(1-\tau)^\beta \le 1-\beta\tau$ for all $\tau\in(-\infty,1)$ and $\beta\in[0,1]$. Combining the above inequality with (67), one can obtain that

$$\mathbb{E}_{\xi^{k+1}}[\|\nabla f(x^{k+1}) - m^{k+1}\|^\alpha] \le \Big(1-\sum_{t=1}^{p-1}\theta_{k,t}\Big)\|\nabla f(x^k)-m^k\|^\alpha$$
$$+\frac{p^{\alpha-1}L_p^\alpha}{(p!)^\alpha}\Big(2+\Big(\sum_{t=1}^{p-1}\theta_{k,t}\Big)^{1-\alpha}\Big)\eta_k^{\alpha p}\Big(1+\sum_{t=1}^{p-1}\frac{|\theta_{k,t}|^\alpha}{\gamma_{k,t}^{\alpha p}}\Big)+2(p-1)^{\alpha-1}\sigma^\alpha\sum_{t=1}^{p-1}|\theta_{k,t}|^\alpha,$$

which together with $\sum_{t=1}^{p-1}\theta_{k,t}\in(0,1)$ and $\alpha\in(1,2]$ implies that (65) holds. □

The following lemma establishes a descent property for the potential sequence $\{\mathcal{P}_k\}$ defined below.

Lemma 12. *Suppose that Assumptions 1 and 2 hold. Let $\{(x^k,m^k)\}$ be generated by Algorithm 2 with input parameters $q=p-1$, $\{\eta_k\}$, and $\{(\gamma_{k,t},\theta_{k,t})\}$ satisfying (23) and (24). Let L_1, σ, and α be given in Assumption 1, p and L_p be given in Assumption 2, and $\{\mathcal{P}_k\}$ be defined in (34) for $\{(x^k,m^k)\}$ and any positive sequence $\{p_k\}$ that satisfies $(1-\sum_{t=1}^{p-1}\theta_{k,t})p_{k+1}\le(1-\sum_{t=1}^{p-1}\theta_{k,t}/10)p_k$ for all $k\ge 0$. Then it holds that for all $k\ge 0$,*

$$\mathbb{E}_{\xi^{k+1}}[\mathcal{P}_{k+1}] \le \mathcal{P}_k - \eta_k\|\nabla f(x^k)\| + \frac{L_1}{2}\eta_k^2 + \frac{(\alpha-1)(2\eta_k)^{\alpha/(\alpha-1)}}{\alpha^{\alpha/(\alpha-1)}(p_k\sum_{t=1}^{p-1}\theta_{k,t}/10)^{1/(\alpha-1)}}$$
$$+\frac{3p^{\alpha-1}L_p^\alpha}{(p!)^\alpha}p_{k+1}\Big(\sum_{t=1}^{p-1}\theta_{k,t}\Big)^{1-\alpha}\eta_k^{\alpha p}\Big(1+\sum_{t=1}^{p-1}\frac{|\theta_{k,t}|^\alpha}{\gamma_{k,t}^{\alpha p}}\Big)+2(p-1)^{\alpha-1}\sigma^\alpha p_{k+1}\sum_{t=1}^{p-1}|\theta_{k,t}|^\alpha. \quad (68)$$

Proof. Fix any $k\ge0$. By Lemma 4 with $(x^+,x,m,\eta)=(x^{k+1},x^k,m^k,\eta_k)$, one has

$$f(x^{k+1}) \le f(x^k) - \eta_k\|\nabla f(x^k)\| + 2\eta_k\|\nabla f(x^k)-m^k\| + \frac{L_1}{2}\eta_k^2. \quad (69)$$

Combining this with (34) and (65), we obtain that

$$\mathbb{E}_{\xi^{k+1}}[\mathcal{P}_{k+1}] \overset{(34)}{=} \mathbb{E}_{\xi^{k+1}}[f(x^{k+1}) + p_{k+1}\|\nabla f(x^{k+1}) - m^{k+1}\|^\alpha]$$
$$\overset{(65)(69)}{\le} f(x^k) - \eta_k\|\nabla f(x^k)\| + 2\eta_k\|\nabla f(x^k)-m^k\| + \frac{L_1}{2}\eta_k^2 + \Big(1-\sum_{t=1}^{p-1}\theta_{k,t}\Big)p_{k+1}\|\nabla f(x^k)-m^k\|^\alpha$$
$$+\frac{3p^{\alpha-1}L_p^\alpha}{(p!)^\alpha}p_{k+1}\Big(\sum_{t=1}^{p-1}\theta_{k,t}\Big)^{1-\alpha}\eta_k^{\alpha p}\Big(1+\sum_{t=1}^{p-1}\frac{|\theta_{k,t}|^\alpha}{\gamma_{k,t}^{\alpha p}}\Big)+2(p-1)^{\alpha-1}\sigma^\alpha p_{k+1}\sum_{t=1}^{p-1}|\theta_{k,t}|^\alpha$$
$$\le f(x^k) - \eta_k\|\nabla f(x^k)\| + 2\eta_k\|\nabla f(x^k)-m^k\| + \frac{L_1}{2}\eta_k^2 + \Big(1-\sum_{t=1}^{p-1}\theta_{k,t}/10\Big)p_k\|\nabla f(x^k)-m^k\|^\alpha$$
$$+\frac{3p^{\alpha-1}L_p^\alpha}{(p!)^\alpha}p_{k+1}\Big(\sum_{t=1}^{p-1}\theta_{k,t}\Big)^{1-\alpha}\eta_k^{\alpha p}\Big(1+\sum_{t=1}^{p-1}\frac{|\theta_{k,t}|^\alpha}{\gamma_{k,t}^{\alpha p}}\Big)+2(p-1)^{\alpha-1}\sigma^\alpha p_{k+1}\sum_{t=1}^{p-1}|\theta_{k,t}|^\alpha, \quad (70)$$

where the last inequality is due to $(1-\sum_{t=1}^{p-1}\theta_{k,t})p_{k+1}\le(1-\sum_{t=1}^{p-1}\theta_{k,t}/10)p_k$. In addition, letting $\alpha'=\alpha/(\alpha-1)$ and using the Young's inequality, we have that

$$2\eta_k\|\nabla f(x^k)-m^k\| \le \frac{((\alpha p_k\sum_{t=1}^{p-1}\theta_{k,t}/10)^{1/\alpha}\|\nabla f(x^k)-m^k\|)^\alpha}{\alpha} + \frac{(2\eta_k/(\alpha p_k\sum_{t=1}^{p-1}\theta_{k,t}/10)^{1/\alpha})^{\alpha'}}{\alpha'}$$

$$= \frac{p_k \sum_{t=1}^{p-1} \theta_{k,t}}{10} \|\nabla f(x^k) - m^k\|^\alpha + \frac{(\alpha-1)(2\eta_k)^{\alpha/(\alpha-1)}}{\alpha^{\alpha/(\alpha-1)}(p_k \sum_{t=1}^{p-1} \theta_{k,t}/10)^{1/(\alpha-1)}}.$$

This together with (70) implies that

$$\mathbb{E}_{\xi^{k+1}}[\mathcal{P}_{k+1}] \le f(x^k) + p_k\|m^k - \nabla f(x^k)\|^\alpha - \eta_k\|\nabla f(x^k)\| + \frac{L_1}{2}\eta_k^2 + \frac{(\alpha-1)(2\eta_k)^{\alpha/(\alpha-1)}}{\alpha^{\alpha/(\alpha-1)}(p_k \sum_{t=1}^{p-1} \theta_{k,t}/10)^{1/(\alpha-1)}}$$
$$+ \frac{3p^{\alpha-1}L_p^\alpha}{(p!)^\alpha} p_{k+1} \Big(\sum_{t=1}^{p-1} \theta_{k,t}\Big)^{1-\alpha} \eta_k^{\alpha p}\Big(1 + \sum_{t=1}^{p-1} \frac{|\theta_{k,t}|^\alpha}{\gamma_{k,t}^{\alpha p}}\Big) + 2(p-1)^{\alpha-1}\sigma^\alpha p_{k+1} \sum_{t=1}^{p-1} |\theta_{k,t}|^\alpha.$$

The conclusion (68) then follows from this relation and (34). □

We next establish some properties for a specific choice of $\{(\gamma_{k,t}, \theta_{k,t})\}$ and $\{p_k\}$, which will be used to prove Theorem 3 subsequently.

Lemma 13. *Let $\{(\gamma_{k,t}, \theta_{k,t})\}$ be defined in (16) and (17), and let $\{p_k\}$ be defined as*

$$p_k = (k+4)^{(p(\alpha-1)^2-\alpha+1)/(p(2\alpha-1)+\alpha-1)} \qquad \forall k \ge 0. \tag{71}$$

Then (23) and (24) hold for such $\{(\gamma_{k,t}, \theta_{k,t})\}$, and moreover, $(1-\sum_{t=1}^{p-1} \theta_{k,t})p_{k+1} \le (1-\sum_{t=1}^{p-1} \theta_{k,t}/10)p_k$ holds for all $k \ge 0$.

Proof. Fix any $k \ge 0$. Notice that $p\alpha/(p(2\alpha-1)+\alpha-1) \in (1/2, 1)$ for all $p \ge 2$ and $\alpha \in (1,2]$. It then follows from (17) that $\gamma_k = 1/(k+4)^{p\alpha/(p(2\alpha-1)+\alpha-1)} \in (0, 1/2)$. By this, (16), and Lemma 9, one can see that (23) and (24) hold for $\{(\gamma_{k,t}, \theta_{k,t})\}$ that is defined in (16) and (17). In addition, observe that

$$\frac{1-\sum_{t=1}^{p-1}\theta_{k,t}/10}{1-\sum_{t=1}^{p-1}\theta_{k,t}} = 1 + \frac{9\sum_{t=1}^{p-1}\theta_{k,t}}{10(1-\sum_{t=1}^{p-1}\theta_{k,t})} \ge 1 + \frac{9\sum_{t=1}^{p-1}\theta_{k,t}}{10} \overset{(58)}{\ge} 1 + \frac{9\gamma_k}{10(1+\pi^2/6)}$$
$$\overset{(17)}{=} 1 + \frac{9}{10(1+\pi^2/6)(k+4)^{p\alpha/(p(2\alpha-1)+\alpha-1)}} > 1 + \frac{9}{10(1+\pi^2/6)(k+4)} > 1 + \frac{1}{3(k+4)}, \tag{72}$$

where the first inequality follows from $\sum_{t=1}^{p-1}\theta_{k,t} \in (0,1)$, and the third inequality is due to $p\alpha/(p(2\alpha-1)+\alpha-1) < 1$ for all $p \ge 1$ and $\alpha \in (1,2]$. Also, note that

$$\frac{p_{k+1}}{p_k} = \Big(1 + \frac{1}{k+4}\Big)^{(p(\alpha-1)^2-\alpha+1)/(p(2\alpha-1)+\alpha-1)} \le \Big(1 + \frac{1}{k+4}\Big)^{1/3} \le 1 + \frac{1}{3(k+4)},$$

where the first inequality follows from $(p(\alpha-1)^2-\alpha+1)/(p(2\alpha-1)+\alpha-1) \le 1/3$ for all $p \ge 1$ and $\alpha \in (1,2]$, and the last inequality is due to $(1+\tau)^\beta \le 1+\tau\beta$ for all $\tau > -1$ and $\beta \in [0,1]$. The above relation together with (72) implies that $(1-\sum_{t=1}^{p-1}\theta_{k,t})p_{k+1} \le (1-\sum_{t=1}^{p-1}\theta_{k,t}/10)p_k$ holds. □

We are now ready to prove Theorem 3.

***Proof of Theorem 3*.** Let $\{(x^k, m^k)\}$ be generated by Algorithm 2 with $\{(\eta_k, \gamma_{k,t}, \theta_{k,t})\}$ defined in (15) and (16), and let $\{\mathcal{P}_k\}$ be defined in (34) with such $\{(x^k, m^k)\}$ and $\{p_k\}$ given in (71). By Lemma 13, one can see that such $\{(\eta_k, \gamma_{k,t}, \theta_{k,t}, p_k)\}$ satisfies the assumptions in Lemma 12 and Algorithm 2. In addition, by (34) and (71), one has that

$$\mathbb{E}[\mathcal{P}_0] = f(x^0) + p_0\mathbb{E}[\|m^0 - \nabla f(x^0)\|^\alpha]$$
$$\le f(x^0) + 4^{(p(\alpha-1)^2-\alpha+1)/(p(2\alpha-1)+\alpha-1)}\mathbb{E}[\|G(x^0;\xi^0) - \nabla f(x^0)\|^\alpha] \le f(x^0) + 4^{1/3}\sigma^\alpha, \tag{73}$$

$$\mathbb{E}[\mathcal{P}_K] = \mathbb{E}[f(x^K) + p_K \|m^K - \nabla f(x^K)\|^\alpha] \ge \mathbb{E}[f(x^K)] \ge f_{\mathrm{low}}, \tag{74}$$

where the inequality in (73) is due to $(p(\alpha-1)^2 - \alpha + 1)/(p(2\alpha-1)+\alpha-1) \le 1/3$ for all $p \ge 1$ and $\alpha \in (1,2]$, and $\mathbb{E}[\|G(x^0;\xi^0) - \nabla f(x^0)\|^\alpha] \le \sigma^\alpha$ for all $1 \le t \le p-1$. Taking expectation on both sides of (68) with respect to $\{\xi^i\}_{0\le i\le k+1}$, we have

$$\begin{aligned}
\mathbb{E}[\mathcal{P}_{k+1}] &\le \mathbb{E}[\mathcal{P}_k] - \eta_k \mathbb{E}[\|\nabla f(x^k)\|] + \frac{L_1}{2}\eta_k^2 + \frac{(\alpha-1)(2\eta_k)^{\alpha/(\alpha-1)}}{\alpha^{\alpha/(\alpha-1)}(p_k\sum_{t=1}^{p-1}\theta_{k,t}/10)^{1/(\alpha-1)}} \\
&\quad + \frac{3p^{\alpha-1}L_p^\alpha}{(p!)^\alpha} p_{k+1}\Big(\sum_{t=1}^{p-1}\theta_{k,t}\Big)^{1-\alpha}\eta_k^{\alpha p}\Big(1 + \sum_{t=1}^{p-1}\frac{|\theta_{k,t}|^\alpha}{\gamma_{k,t}^{\alpha p}}\Big) + 2(p-1)^{\alpha-1}\sigma^\alpha p_{k+1}\sum_{t=1}^{p-1}|\theta_{k,t}|^\alpha \quad \forall k \ge 0.
\end{aligned}$$

Summing up this inequality over $k = 0,\ldots,K-1$, and using (73) and (74), we obtain that for all $K \ge 1$,

$$\begin{aligned}
f_{\mathrm{low}} &\overset{(74)}{\le} \mathbb{E}[\mathcal{P}_K] \le \mathbb{E}[\mathcal{P}_0] - \sum_{k=0}^{K-1}\eta_k\mathbb{E}[\|\nabla f(x^k)\|] + \sum_{k=0}^{K-1}\Big(\frac{L_1}{2}\eta_k^2 + \frac{(\alpha-1)(2\eta_k)^{\alpha/(\alpha-1)}}{\alpha^{\alpha/(\alpha-1)}(p_k\sum_{t=1}^{p-1}\theta_{k,t}/10)^{1/(\alpha-1)}} \\
&\quad + \frac{3p^{\alpha-1}L_p^\alpha}{(p!)^\alpha}p_{k+1}\Big(\sum_{t=1}^{p-1}\theta_{k,t}\Big)^{1-\alpha}\eta_k^{\alpha p}\Big(1+\sum_{t=1}^{p-1}\frac{|\theta_{k,t}|^\alpha}{\gamma_{k,t}^{\alpha p}}\Big) + 2(p-1)^{\alpha-1}\sigma^\alpha p_{k+1}\sum_{t=1}^{p-1}|\theta_{k,t}|^\alpha\Big) \\
&\overset{(73)}{\le} f(x^0) + 4^{1/3}\sigma^\alpha - \eta_{K-1}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(x^k)\|] + \sum_{k=0}^{K-1}\Big(\frac{L_1}{2}\eta_k^2 + \frac{(\alpha-1)(2\eta_k)^{\alpha/(\alpha-1)}}{\alpha^{\alpha/(\alpha-1)}(p_k\sum_{t=1}^{p-1}\theta_{k,t}/10)^{1/(\alpha-1)}} \\
&\quad + \frac{6p^{\alpha-1}L_p^\alpha}{(p!)^\alpha}p_k\Big(\sum_{t=1}^{p-1}\theta_{k,t}\Big)^{1-\alpha}\eta_k^{\alpha p}\Big(1+\sum_{t=1}^{p-1}\frac{|\theta_{k,t}|^\alpha}{\gamma_{k,t}^{\alpha p}}\Big) + 4(p-1)^{\alpha-1}\sigma^\alpha p_k\sum_{t=1}^{p-1}|\theta_{k,t}|^\alpha\Big) \\
&\le f(x^0) + 4^{1/3}\sigma^\alpha - \eta_{K-1}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(x^k)\|] + \sum_{k=0}^{K-1}\Big(\frac{L_1}{2}\eta_k^2 + \frac{30^{1/(\alpha-1)}(\alpha-1)(2/\alpha)^{\alpha/(\alpha-1)}\eta_k^{\alpha/(\alpha-1)}}{(p_k\gamma_k)^{1/(\alpha-1)}} \\
&\quad + \frac{18p^{\alpha-1}L_p^\alpha}{(p!)^\alpha}p_k\gamma_k^{1-\alpha}\eta_k^{\alpha p}\Big(1+\sum_{t=1}^{p-1}\frac{|\theta_{k,t}|^\alpha}{\gamma_{k,t}^{\alpha p}}\Big) + 64(p-1)^\alpha\sigma^\alpha p_k\gamma_k^\alpha\Big) \\
&\le f(x^0) + 4^{1/3}\sigma^\alpha - \eta_{K-1}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(x^k)\|] + \sum_{k=0}^{K-1}\Big(\frac{L_1}{2}\eta_k^2 + \frac{30^{1/(\alpha-1)}(\alpha-1)(2/\alpha)^{\alpha/(\alpha-1)}\eta_k^{\alpha/(\alpha-1)}}{(p_k\gamma_k)^{1/(\alpha-1)}} \\
&\quad + \frac{18p^{\alpha-1}L_p^\alpha}{(p!)^\alpha}p_k\gamma_k^{1-\alpha}\eta_k^{\alpha p}\Big(1+16\gamma_k^{\alpha-\alpha p}\sum_{t=1}^{p-1}t^{2\alpha(p-1)}\Big) + 64(p-1)^\alpha\sigma^\alpha p_k\gamma_k^\alpha\Big) \\
&\le f(x^0) + 4^{1/3}\sigma^\alpha - \eta_{K-1}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(x^k)\|] + \sum_{k=0}^{K-1}\Big(\frac{L_1}{2}\eta_k^2 + \frac{30^{1/(\alpha-1)}(\alpha-1)(2/\alpha)^{\alpha/(\alpha-1)}\eta_k^{\alpha/(\alpha-1)}}{(p_k\gamma_k)^{1/(\alpha-1)}} \\
&\quad + \frac{306p^{2\alpha p}L_p^\alpha}{(p!)^\alpha}p_k\gamma_k^{1-\alpha p}\eta_k^{\alpha p} + 64(p-1)^\alpha\sigma^\alpha p_k\gamma_k^\alpha\Big),
\end{aligned} \tag{75}$$

where the third inequality follows from (73), $p_{k+1} \le 2p_k$ for all $k \ge 0$ due to (71), and the fact that $\{\eta_k\}$ is nonincreasing, the fourth inequality is due to $\sum_{t=1}^{p-1}\theta_{k,t} \in (\gamma_k/3, 2\gamma_k)$ and $|\theta_{k,t}| \le 4\gamma_k/t^2 \le 4\gamma_k$ for all $1 \le t \le p-1$ because of (58), the fifth inequality follows from $|\theta_{k,t}|^\alpha/\gamma_{k,t}^{\alpha p} \le 4^\alpha\gamma_k^{\alpha-\alpha p}t^{2\alpha(p-1)} \le 16\gamma_k^{\alpha-\alpha p}t^{2\alpha(p-1)}$ because of (58) and (16), and the last inequality is due to $1 + 16\gamma_k^{\alpha-\alpha p}\sum_{t=1}^{p-1}t^{2\alpha(p-1)} \le 17\gamma_k^{\alpha-\alpha p}\sum_{t=1}^{p-1}t^{2\alpha(p-1)} \le 17p^{2\alpha p-2\alpha+1}\gamma_k^{\alpha-\alpha p}$. Rearranging the terms in (75), and using (14), (15), (17), and (71), we obtain that for all $K \ge 5$,

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(x^k)\|]$$

$$
\begin{aligned}
&\overset{(75)}{\leq} \frac{f(x^0)-f_{\text{low}}+4^{1/3}\sigma^\alpha}{K\eta_{K-1}} + \frac{1}{K\eta_{K-1}}\sum_{k=0}^{K-1}\Big(\frac{L_1}{2}\eta_k^2 + \frac{30^{1/(\alpha-1)}(\alpha-1)(2/\alpha)^{\alpha/(\alpha-1)}\eta_k^{\alpha/(\alpha-1)}}{(p_k\gamma_k)^{1/(\alpha-1)}} \\
&\quad + \frac{306p^{2\alpha p}L_p^\alpha}{(p!)^\alpha}p_k\gamma_k^{1-\alpha p}\eta_k^{\alpha p} + 64(p-1)^\alpha\sigma^\alpha p_k\gamma_k^\alpha\Big) \\
&\overset{(15)(17)(71)}{=} \frac{(f(x^0)-f_{\text{low}}+4^{1/3}\sigma^\alpha)(K+3)^{(p\alpha+\alpha-1)/(p(2\alpha-1)+\alpha-1)}}{K} \\
&\quad + \frac{(K+3)^{(p\alpha+\alpha-1)/(p(2\alpha-1)+\alpha-1)}}{K}\sum_{k=0}^{K-1}\Big(\frac{L_1}{2(k+4)^{2(p\alpha+\alpha-1)/(p(2\alpha-1)+\alpha-1)}} \\
&\quad + \frac{30^{1/(\alpha-1)}(\alpha-1)(2/\alpha)^{\alpha/(\alpha-1)} + 306p^{2\alpha p}L_p^\alpha/(p!)^\alpha + 64(p-1)^\alpha\sigma^\alpha}{k+4}\Big) \\
&\leq \frac{2(f(x^0)-f_{\text{low}}+4^{1/3}\sigma^\alpha)}{K^{p(\alpha-1)/(p(2\alpha-1)+\alpha-1)}} + \frac{2}{K^{p(\alpha-1)/(p(2\alpha-1)+\alpha-1)}}\sum_{k=0}^{K-1}\Big(\frac{L_1}{2(k+4)^{2(p\alpha+\alpha-1)/(p(2\alpha-1)+\alpha-1)}} \\
&\quad + \frac{30^{1/(\alpha-1)}(\alpha-1)(2/\alpha)^{\alpha/(\alpha-1)} + 306p^{2\alpha p}L_p^\alpha/(p!)^\alpha + 64(p-1)^\alpha\sigma^\alpha}{k+4}\Big) \\
&\leq \frac{2(f(x^0)-f_{\text{low}}+4^{1/3}\sigma^\alpha)}{K^{p(\alpha-1)/(p(2\alpha-1)+\alpha-1)}} + \frac{2}{K^{p(\alpha-1)/(p(2\alpha-1)+\alpha-1)}} \\
&\quad \times \sum_{k=0}^{K-1}\Big(\frac{L_1/2 + 30^{1/(\alpha-1)}(\alpha-1)(2/\alpha)^{\alpha/(\alpha-1)} + 306p^{2\alpha p}L_p^\alpha/(p!)^\alpha + 64(p-1)^\alpha\sigma^\alpha}{k+4}\Big) \\
&\leq \frac{4(f(x^0)-f_{\text{low}}+4^{1/3}\sigma^\alpha + L_1/2 + 30^{1/(\alpha-1)}(\alpha-1)(2/\alpha)^{\alpha/(\alpha-1)} + 306p^{2\alpha p}L_p^\alpha/(p!)^\alpha + 64(p-1)^\alpha\sigma^\alpha)\ln K}{K^{p(\alpha-1)/(p(2\alpha-1)+\alpha-1)}} \\
&\overset{(14)}{=} \frac{M_{p,\alpha}\ln K}{K^{p(\alpha-1)/(p(2\alpha-1)+\alpha-1)}},
\end{aligned}
$$

where the second inequality follows from $(K+3)^{(p\alpha+\alpha-1)/(p(2\alpha-1)+\alpha-1)} \leq 2K^{(p\alpha+\alpha-1)/(p(2\alpha-1)+\alpha-1)}$ for all $K \geq 5$, the third inequality is due to $2(p\alpha+\alpha-1)/(p(2\alpha-1)+\alpha-1) \geq 1$ for all $p \geq 1$ and $\alpha \in (1,2]$, and the last inequality follows from $\sum_{k=0}^{K-1} 1/(k+4) \leq \ln(2K/5+1) \leq 2\ln K$ for all $K \geq 5$ due to (37). Recall that ι_K is uniformly selected from $\{0,\ldots,K-1\}$. It then follows from this and the above inequality that

$$
\mathbb{E}[\|\nabla f(x^{\iota_K})\|] = \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(x^k)\|] \leq \frac{M_{p,\alpha}\ln K}{K^{p(\alpha-1)/(p(2\alpha-1)+\alpha-1)}} \qquad \forall K \geq 5. \tag{76}
$$

By Lemma 3 with $(\beta,u,v) = (p(\alpha-1)/(p(2\alpha-1)+\alpha-1), p(\alpha-1)\epsilon/(2(p(2\alpha-1)+\alpha-1)M_{p,\alpha}), K)$, one can see that $K^{-p(\alpha-1)/(p(2\alpha-1)+\alpha-1)}\ln K \leq \epsilon/M_{p,\alpha}$ for all K satisfying

$$
K \geq \Big(\frac{2(p(2\alpha-1)+\alpha-1)M_{p,\alpha}}{p(\alpha-1)\epsilon}\ln\Big(\frac{2(p(2\alpha-1)+\alpha-1)M_{p,\alpha}}{p(\alpha-1)\epsilon}\Big)\Big)^{(p(2\alpha-1)+\alpha-1)/(p(\alpha-1))},
$$

which together with (76) implies that Theorem 3. $\square$

The following lemma establishes some properties for a specific choice of $\{(\gamma_{k,t},\theta_{k,t})\}$ and $\{p_k\}$, which will be used to prove Theorem 4 subsequently.

Lemma 14. *Let $\{(\gamma_{k,t},\theta_{k,t})\}$ be defined in (21) and (22), and let $\{p_k\}$ be defined as*

$$
p_k = (k+4)^{(2p(\alpha-1)^2-\alpha)/(3p\alpha+\alpha)} \qquad \forall k \geq 0. \tag{77}
$$

Then (23) and (24) hold for such $\{(\gamma_{k,t},\theta_{k,t})\}$, and moreover, $(1-\sum_{t=1}^{p-1}\theta_{k,t})p_{k+1}\le(1-\sum_{t=1}^{p-1}\theta_{k,t}/10)p_k$ holds for all $k\ge 0$.

Proof. Fix any $k\ge 0$. Notice that $2p/(3p+1)>1/2$ for all $p\ge 2$. It then follows from (22) that $\gamma_k=1/(k+4)^{2p/(3p+1)}\in(0,1/2)$. By this, (21) and Lemma 9, one can see that (23) and (24) hold for $\{(\gamma_{k,t},\theta_{k,t})\}$ that is defined in (21) and (22). In addition, observe that

$$\begin{aligned}\frac{1-\sum_{t=1}^{p-1}\theta_{k,t}/10}{1-\sum_{t=1}^{p-1}\theta_{k,t}}&=1+\frac{9\sum_{t=1}^{p-1}\theta_{k,t}}{10(1-\sum_{t=1}^{p-1}\theta_{k,t})}\ge 1+\frac{9\sum_{t=1}^{p-1}\theta_{k,t}}{10}\overset{(58)}{\ge}1+\frac{9\gamma_k}{10(1+\pi^2/6)}\\&\overset{(22)}{=}1+\frac{9}{10(1+\pi^2/6)(k+4)^{2p/(3p+1)}}>1+\frac{9}{10(1+\pi^2/6)(k+4)}>1+\frac{1}{3(k+4)},\end{aligned}\tag{78}$$

where the first inequality follows from $\sum_{t=1}^{p-1}\theta_{k,t}\in(0,1)$, the third inequality follows from $2p/(3p+1)<1$ for all $p\ge 2$. Also, note that

$$\frac{p_{k+1}}{p_k}=\Big(1+\frac{1}{k+4}\Big)^{(2p(\alpha-1)^2-\alpha)/(3p\alpha+\alpha)}\le\Big(1+\frac{1}{k+4}\Big)^{1/3}\le 1+\frac{1}{3(k+4)},$$

where the first inequality follows from $(2p(\alpha-1)^2-\alpha)/(3p\alpha+\alpha)\le 1/3$ for all $p\ge 2$ and $\alpha\in(1,2]$, and the last inequality is due to $(1+\tau)^\beta\le 1+\tau\beta$ for all $\tau>-1$ and $\beta\in[0,1]$. The above relation along with (78) implies that $(1-\sum_{t=1}^{p-1}\theta_{k,t})p_{k+1}\le(1-\sum_{t=1}^{p-1}\theta_{k,t}/10)p_k$ holds. $\square$

We are now ready to prove Theorem 4.

Proof of Theorem 4. Let $\{(x^k,m^k)\}$ be generated by Algorithm 2 with $\{(\eta_k,\gamma_{k,t},\theta_{k,t})\}$ defined in (20), (21), and (22), and let $\{\mathcal{P}_k\}$ be defined in (34) with such $\{(x^k,m^k)\}$ and $\{p_k\}$ given in (77). By Lemma 14, one can see that such $\{(\eta_k,\gamma_{k,t},\theta_{k,t},p_k)\}$ satisfies the assumptions in Lemma 12 and Algorithm 2. Using this and similar arguments as those for deriving (75), we obtain that for all $K\ge 1$,

$$\begin{aligned}f_{\mathrm{low}}\le f(x^0)+4^{1/3}\sigma^\alpha-\eta_{K-1}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(x^k)\|]+\sum_{k=0}^{K-1}\Big(\frac{L_1}{2}\eta_k^2+\frac{30^{1/(\alpha-1)}(\alpha-1)(2/\alpha)^{\alpha/(\alpha-1)}\eta_k^{\alpha/(\alpha-1)}}{(p_k\gamma_k)^{1/(\alpha-1)}}\\+\frac{306p^{2\alpha p}L_p^\alpha}{(p!)^\alpha}p_k\gamma_k^{1-\alpha p}\eta_k^{\alpha p}+64(p-1)^\alpha\sigma^\alpha p_k\gamma_k^\alpha\Big).\end{aligned}$$

Rearranging the terms in this inequality, and using (18), (19), (20), (22), and (77), we obtain that for all $K\ge 5$,

$$\begin{aligned}\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(x^k)\|]&\le\frac{f(x^0)-f_{\mathrm{low}}+4^{1/3}\sigma^\alpha}{K\eta_{K-1}}+\frac{1}{K\eta_{K-1}}\sum_{k=0}^{K-1}\Big(\frac{L_1}{2}\eta_k^2\\&\quad+\frac{30^{1/(\alpha-1)}(\alpha-1)(2/\alpha)^{\alpha/(\alpha-1)}\eta_k^{\alpha/(\alpha-1)}}{(p_k\gamma_k)^{1/(\alpha-1)}}+\frac{306p^{2\alpha p}L_p^\alpha}{(p!)^\alpha}p_k\gamma_k^{1-\alpha p}\eta_k^{\alpha p}+64(p-1)^\alpha\sigma^\alpha p_k\gamma_k^\alpha\Big)\\&\overset{(20)(22)(77)}{=}\frac{(f(x^0)-f_{\mathrm{low}}+4^{1/3}\sigma^\alpha)(K+3)^{(2p+1)/(3p+1)}}{K}+\frac{(K+3)^{(2p+1)/(3p+1)}}{K}\sum_{k=0}^{K-1}\Big(\frac{L_1}{2(k+4)^{2(2p+1)/(3p+1)}}\\&\quad+\frac{30^{1/(\alpha-1)}(\alpha-1)(2/\alpha)^{\alpha/(\alpha-1)}+64(p-1)^\alpha\sigma^\alpha}{(k+4)^{(2p(2\alpha-1)+\alpha)/(3p\alpha+\alpha)}}+\frac{306p^{2p\alpha}L_p^\alpha/(p!)^\alpha}{(k+4)^{(p(6\alpha-\alpha^2-2)+\alpha)/(3p\alpha+\alpha)}}\Big)\\&\le\frac{2(f(x^0)-f_{\mathrm{low}}+4^{1/3}\sigma^\alpha)}{K^{p/(3p+1)}}+\frac{2}{K^{p/(3p+1)}}\sum_{k=0}^{K-1}\Big(\frac{L_1}{2(k+4)^{2(2p+1)/(3p+1)}}\end{aligned}$$

$$
\begin{aligned}
&\quad + \frac{30^{1/(\alpha-1)}(\alpha-1)(2/\alpha)^{\alpha/(\alpha-1)} + 64(p-1)^\alpha\sigma^\alpha}{(k+4)^{(2p(2\alpha-1)+\alpha)/(3p\alpha+\alpha)}} + \frac{306p^{2p\alpha}L_p^\alpha/(p!)^\alpha}{(k+4)^{(p(6\alpha-\alpha^2-2)+\alpha)/(3p\alpha+\alpha)}}\Big) \\
&\le \frac{2(f(x^0)-f_{\text{low}}+4^{1/3}\sigma^\alpha)}{K^{p/(3p+1)}} + \frac{2}{K^{p/(3p+1)}}\sum_{k=0}^{K-1}\frac{L_1/2+306p^{2p\alpha}L_p^\alpha/(p!)^\alpha}{k+4} \\
&\quad + \frac{2}{K^{p/(3p+1)}}\sum_{k=0}^{K-1}\frac{(30^{1/(\alpha-1)}(\alpha-1)(2/\alpha)^{\alpha/(\alpha-1)} + 64(p-1)^\alpha\sigma^\alpha)(K+3)^{p(2-\alpha)/(3p\alpha+\alpha)}}{k+4} \\
&\le \frac{2(f(x^0)-f_{\text{low}}+4^{1/3}\sigma^\alpha)}{K^{p/(3p+1)}} + \frac{2}{K^{p/(3p+1)}}\sum_{k=0}^{K-1}\frac{L_1/2+306p^{2p\alpha}L_p^\alpha/(p!)^\alpha}{k+4} \\
&\quad + \frac{4}{K^{p/(3p+1)}}\sum_{k=0}^{K-1}\frac{(30^{1/(\alpha-1)}(\alpha-1)(2/\alpha)^{\alpha/(\alpha-1)} + 64(p-1)^\alpha\sigma^\alpha)K^{p(2-\alpha)/(3p\alpha+\alpha)}}{k+4} \\
&\le \frac{4(f(x^0)-f_{\text{low}}+4^{1/3}\sigma^\alpha + L_1/2 + 306p^{2p\alpha}L_p^\alpha/(p!)^\alpha)\ln K}{K^{p/(3p+1)}} \\
&\quad + \frac{8(30^{1/(\alpha-1)}(\alpha-1)(2/\alpha)^{\alpha/(\alpha-1)} + 64(p-1)^\alpha\sigma^\alpha)\ln K}{K^{2p(\alpha-1)/(3p\alpha+\alpha)}} \overset{(18)(19)}{=} \frac{\widetilde{M}_{p,\alpha}\ln K}{K^{p/(3p+1)}} + \frac{\widehat{M}_{p,\alpha}\ln K}{K^{2p(\alpha-1)/(3p\alpha+\alpha)}},
\end{aligned}
$$

where the second inequality is due to $(K+3)^{(2p+1)/(3p+1)} \le 2K^{(2p+1)/(3p+1)}$ for all $K \ge 5$, the third inequality follows from $2(2p+1)/(3p+1) \ge 1$, $(p(6\alpha-\alpha^2-2)+\alpha)/(3p\alpha+\alpha) \ge 1$, and $(2p(2\alpha-1)+\alpha)/(3p\alpha+\alpha) \le 1$ for all $p \ge 1$ and $\alpha \in (1,2]$, the fourth inequality is due to $(K+3)^{p(2-\alpha)/(3p\alpha+\alpha)} \le 2K^{p(2-\alpha)/(3p\alpha+\alpha)}$ for all $K \ge 5$, and the last inequality follows from $\sum_{k=0}^{K-1} 1/(k+4) \le \ln(2K/5+1) \le 2\ln K$ for all $K \ge 5$ due to (37). Recall that ι_K is uniformly selected from $\{0,\ldots,K-1\}$. It then follows from this and the above relation that

$$
\mathbb{E}[\|\nabla f(x^{\iota_K})\|] = \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(x^k)\|] \le \frac{\widetilde{M}_{p,\alpha}\ln K}{K^{p/(3p+1)}} + \frac{\widehat{M}_{p,\alpha}\ln K}{K^{2p(\alpha-1)/(3p\alpha+\alpha)}} \qquad \forall K \ge 5. \tag{79}
$$

By Lemma 3 with $(\beta,u,v) = (p/(3p+1), p\epsilon/(4(3p+1)\widetilde{M}_{p,\alpha}), K)$ and $(\beta,u,v) = (2p(\alpha-1)/(3p\alpha+\alpha), p(\alpha-1)\epsilon/(2(3p\alpha+\alpha)\widehat{M}_{p,\alpha}), K)$, one can see that

$$
K^{-p/(3p+1)}\ln K \le \frac{\epsilon}{2\widetilde{M}_{p,\alpha}} \qquad \forall K \ge \Big(\frac{4(3p+1)\widetilde{M}_{p,\alpha}}{p\epsilon}\ln\Big(\frac{4(3p+1)\widetilde{M}_{p,\alpha}}{p\epsilon}\Big)\Big)^{(3p+1)/p},
$$

$$
K^{-2p(\alpha-1)/(3p\alpha+\alpha)}\ln K \le \frac{\epsilon}{2\widehat{M}_{p,\alpha}} \qquad \forall K \ge \Big(\frac{2(3p\alpha+\alpha)\widehat{M}_{p,\alpha}}{p(\alpha-1)\epsilon}\ln\Big(\frac{2(3p\alpha+\alpha)\widehat{M}_{p,\alpha}}{p(\alpha-1)\epsilon}\Big)\Big)^{(3p\alpha+\alpha)/(2p(\alpha-1))},
$$

which together with (79) implies that Theorem 4 holds. □

5.3 Proof of the main results in Section 3.3

In this subsection, we first establish several technical lemmas and then use them to prove Theorems 5 and 6.

The next lemma presents a recurrence relation for the estimation error of the gradient estimators $\{m^k\}$ generated by Algorithm 3.

Lemma 15. *Suppose that Assumptions 1 and 3 hold. Let $\{(x^k, m^k)\}$ be the sequence generated by Algorithm 3 with input parameters $\{(\eta_k,\theta_k)\}$. Then we have*

$$
\mathbb{E}_{\xi^{k+1}}[\|m^{k+1}-\nabla f(x^{k+1})\|^\alpha] \le (1-\theta_k)\|m^k-\nabla f(x^k)\|^\alpha + 6(L_1^\alpha + L^\alpha)\eta_k^\alpha + 6\sigma^\alpha\theta_k^\alpha \qquad \forall k \ge 0, \tag{80}
$$

where L_1, σ, α, and L are given in Assumptions 1 and 3, respectively.

Proof. Fix any $k \geq 0$. It follows from (25) that

$$m^{k+1} - \nabla f(x^{k+1}) = (1-\theta_k)(m^k - \nabla f(x^k)) + G(x^{k+1};\xi^{k+1}) - \nabla f(x^{k+1}) \\ + (1-\theta_k)(\nabla f(x^k) - G(x^k;\xi^{k+1})). \tag{81}$$

Observe from Algorithm 3 and Assumptions 1 and 3 that $\|x^{k+1} - x^k\| = \eta_k$, $\mathbb{E}_{\xi^{k+1}}[G(x^{k+1};\xi^{k+1}) - \nabla f(x^{k+1})] = 0$, $\mathbb{E}_{\xi^{k+1}}[G(x^k;\xi^{k+1}) - \nabla f(x^k)] = 0$, $\mathbb{E}_{\xi^{k+1}}[\|\nabla f(x^k) - G(x^k;\xi^{k+1}))\|^\alpha] \leq \sigma^\alpha$, $\|\nabla f(x^k) - \nabla f(x^{k+1})\| \leq L_1\eta_k$, and $\mathbb{E}_{\xi^{k+1}}[\|G(x^{k+1};\xi^{k+1}) - G(x^k;\xi^{k+1})\|^\alpha] \leq L^\alpha\eta_k^\alpha$. Using these, (35), and (81), we obtain that

$$\begin{aligned}
&\mathbb{E}_{\xi^{k+1}}[\|m^{k+1} - \nabla f(x^{k+1})\|^\alpha] \\
&\overset{(81)}{=} \mathbb{E}_{\xi^{k+1}}[\|(1-\theta_k)(m^k - \nabla f(x^k)) + G(x^{k+1};\xi^{k+1}) - \nabla f(x^{k+1}) + (1-\theta_k)(\nabla f(x^k) - G(x^k;\xi^{k+1}))\|^\alpha] \\
&\leq (1-\theta_k)^\alpha\|m^k - \nabla f(x^k)\|^\alpha + 2\mathbb{E}_{\xi^{k+1}}[\|G(x^{k+1};\xi^{k+1}) - \nabla f(x^{k+1}) + (1-\theta_k)(\nabla f(x^k) - G(x^k;\xi^{k+1}))\|^\alpha] \\
&= (1-\theta_k)^\alpha\|m^k - \nabla f(x^k)\|^\alpha + 2\mathbb{E}_{\xi^{k+1}}\big[\|G(x^{k+1};\xi^{k+1}) - G(x^k;\xi^{k+1}) + \nabla f(x^k) - \nabla f(x^{k+1}) \\
&\quad - \theta_k(\nabla f(x^k) - G(x^k;\xi^{k+1}))\|^\alpha\big] \\
&\leq (1-\theta_k)^\alpha\|m^k - \nabla f(x^k)\|^\alpha + 6\mathbb{E}_{\xi^{k+1}}\big[\|G(x^{k+1};\xi^{k+1}) - G(x^k;\xi^{k+1})\|^\alpha\big] + 6\|\nabla f(x^{k+1}) - \nabla f(x^k)\|^\alpha \\
&\quad + 6\theta_k^\alpha\mathbb{E}_{\xi^{k+1}}[\|\nabla f(x^k) - G(x^k;\xi^{k+1}))\|^\alpha] \\
&\leq (1-\theta_k)^\alpha\|m^k - \nabla f(x^k)\|^\alpha + 6(L_1^\alpha + L^\alpha)\eta_k^\alpha + 6\sigma^\alpha\theta_k^\alpha,
\end{aligned}$$

where the first inequality is due to (35), $\mathbb{E}_{\xi^{k+1}}[G(x^{k+1};\xi^{k+1}) - \nabla f(x^{k+1})] = 0$, and $\mathbb{E}_{\xi^{k+1}}[G(x^k;\xi^{k+1}) - \nabla f(x^k)] = 0$, the second inequality follows from $\|a+b+c\|^\alpha \leq 3(\|a\|^\alpha + \|b\|^\alpha + \|c\|^\alpha)$ for all $a, b, c \in \mathbb{R}^n$ due to $\alpha \in (1,2]$ and the convexity of $\|\cdot\|^\alpha$, and the last inequality is due to $\mathbb{E}_{\xi^{k+1}}[\|\nabla f(x^k) - G(x^k;\xi^{k+1}))\|^\alpha] \leq \sigma^\alpha$, $\|\nabla f(x^k) - \nabla f(x^{k+1})\| \leq L_1\eta_k$, and $\mathbb{E}_{\xi^{k+1}}[\|G(x^{k+1};\xi^{k+1}) - G(x^k;\xi^{k+1})\|^\alpha] \leq L^\alpha\eta_k^\alpha$. The above relation together with $\theta_k \in (0,1]$ and $\alpha \in (1,2]$ implies that this lemma holds. □

The following lemma establishes a descent property for the potential sequence $\{\mathcal{P}_k\}$ defined below.

Lemma 16. *Suppose that Assumptions 1 and 3 hold. Let $\{(x^k, m^k)\}$ be the sequence generated by Algorithm 3 with input parameters $\{(\eta_k, \theta_k)\}$. Let L_1, α, and σ be given in Assumption 1, L be given in Assumption 3, and $\{\mathcal{P}_k\}$ be defined in (34) for $\{(x^k, m^k)\}$ and any positive sequence $\{p_k\}$ that satisfies $(1-\theta_k)p_{k+1} \leq (1-\theta_k/2)p_k$ for all $k \geq 0$. Then it holds that for all $k \geq 0$,*

$$\mathbb{E}_{\xi^{k+1}}[\mathcal{P}_{k+1}] \leq \mathcal{P}_k - \eta_k\|\nabla f(x^k)\| + \frac{L_1}{2}\eta_k^2 + \frac{(\alpha-1)(2\eta_k)^{\alpha/(\alpha-1)}}{\alpha^{\alpha/(\alpha-1)}(\theta_k p_k/2)^{1/(\alpha-1)}} + 6(L_1^\alpha + L^\alpha)\eta_k^\alpha p_{k+1} + 6\sigma^\alpha\theta_k^\alpha p_{k+1}. \tag{82}$$

Proof. Fix any $k \geq 0$. By Lemma 4 with $(x^+, x, m, \eta) = (x^{k+1}, x^k, m^k, \eta_k)$, one has

$$f(x^{k+1}) \leq f(x^k) - \eta_k\|\nabla f(x^k)\| + 2\eta_k\|\nabla f(x^k) - m^k\| + \frac{L_1}{2}\eta_k^2. \tag{83}$$

Combining this with (34) and (80), we obtain that

$$\begin{aligned}
\mathbb{E}_{\xi^{k+1}}[\mathcal{P}_{k+1}] &\overset{(34)}{=} \mathbb{E}_{\xi^{k+1}}[f(x^{k+1}) + p_{k+1}\|m^{k+1} - \nabla f(x^{k+1})\|^\alpha] \\
&\overset{(80)(83)}{\leq} f(x^k) - \eta_k\|\nabla f(x^k)\| + 2\eta_k\|\nabla f(x^k) - m^k\| + \frac{L_1}{2}\eta_k^2 \\
&\quad + (1-\theta_k)p_{k+1}\|m^k - \nabla f(x^k)\|^\alpha + 6(L_1^\alpha + L^\alpha)\eta_k^\alpha p_{k+1} + 6\sigma^\alpha\theta_k^\alpha p_{k+1}
\end{aligned}$$

$$
\begin{aligned}
&\leq f(x^k) - \eta_k\|\nabla f(x^k)\| + 2\eta_k\|\nabla f(x^k) - m^k\| + \frac{L_1}{2}\eta_k^2 \\
&\quad + (1-\theta_k/2)p_k\|m^k - \nabla f(x^k)\|^\alpha + 6(L_1^\alpha + L^\alpha)\eta_k^\alpha p_{k+1} + 6\sigma^\alpha\theta_k^\alpha p_{k+1},
\end{aligned} \tag{84}
$$

where the last inequality follows from $(1-\theta_k)p_{k+1} \leq (1-\theta_k/2)p_k$. In addition, letting $\alpha' = \alpha/(\alpha-1)$, and using the Young's inequality, one has that

$$
\begin{aligned}
2\eta_k\|\nabla f(x^k) - m^k\| &\leq \frac{((\alpha\theta_k p_k/2)^{1/\alpha}\|\nabla f(x^k) - m^k\|)^\alpha}{\alpha} + \frac{(2\eta_k/(\alpha\theta_k p_k/2)^{1/\alpha})^{\alpha'}}{\alpha'} \\
&= \frac{\theta_k p_k}{2}\|\nabla f(x^k) - m^k\|^\alpha + \frac{(\alpha-1)(2\eta_k)^{\alpha/(\alpha-1)}}{\alpha^{\alpha/(\alpha-1)}(\theta_k p_k/2)^{1/(\alpha-1)}}.
\end{aligned}
$$

This together with (84) implies that

$$
\begin{aligned}
\mathbb{E}_{\xi^{k+1}}[\mathcal{P}_{k+1}] &\leq f(x^k) + p_k\|\nabla f(x^k) - m^k\|^\alpha - \eta_k\|\nabla f(x^k)\| + \frac{L_1}{2}\eta_k^2 + \frac{(\alpha-1)(2\eta_k)^{\alpha/(\alpha-1)}}{\alpha^{\alpha/(\alpha-1)}(\theta_k p_k/2)^{1/(\alpha-1)}} \\
&\quad + 6(L_1^\alpha + L^\alpha)\eta_k^\alpha p_{k+1} + 6\sigma^\alpha\theta_k^\alpha p_{k+1} \\
&\leq \mathcal{P}_k - \eta_k\|\nabla f(x^k)\| + \frac{L_1}{2}\eta_k^2 + \frac{(\alpha-1)(2\eta_k)^{\alpha/(\alpha-1)}}{\alpha^{\alpha/(\alpha-1)}(\theta_k p_k/2)^{1/(\alpha-1)}} + 6(L_1^\alpha + L^\alpha)\eta_k^\alpha p_{k+1} + 6\sigma^\alpha\theta_k^\alpha p_{k+1}.
\end{aligned}
$$

By this and (34), one can see that (82) holds. □

The following lemma establishes some property for a specific choice of $\{(\theta_k, p_k)\}$, which will be used to prove Theorem 5 subsequently.

Lemma 17. *Let $\{\theta_k\}$ be given in (27), and $\{p_k\}$ be defined as*

$$
p_k = (k+1)^{(\alpha-1)^2/(2\alpha-1)} \qquad \forall k \geq 0. \tag{85}
$$

Then $(1-\theta_k)p_{k+1} \leq (1-\theta_k/2)p_k$ holds for all $k \geq 0$.

Proof. Fix any $k \geq 0$. It follows from (27) that

$$
\frac{1-\theta_k/2}{1-\theta_k} = 1 + \frac{\theta_k}{2(1-\theta_k)} \geq 1 + \frac{\theta_k}{2} \overset{(27)}{=} 1 + \frac{1}{2(k+1)^{\alpha/(2\alpha-1)}} \geq 1 + \frac{1}{2(k+1)}, \tag{86}
$$

where the last inequality is due to $\alpha/(2\alpha-1) < 1$ for all $\alpha \in (1,2]$. In addition, notice from (85) that

$$
\frac{p_{k+1}}{p_k} = \Big(1 + \frac{1}{k+1}\Big)^{(\alpha-1)^2/(2\alpha-1)} \leq \Big(1 + \frac{1}{k+1}\Big)^{1/3} \leq 1 + \frac{1}{3(k+1)},
$$

where the first inequality is due to $(\alpha-1)^2/(2\alpha-1) \leq 1/3$ for all $\alpha \in (1,2]$, and the last inequality is due to $(1+\tau)^\beta \leq 1 + \tau\beta$ for all $\tau > -1$ and $\beta \in [0,1]$. The above relation along with (86) implies that $(1-\theta_k)p_{k+1} \leq (1-\theta_k/2)p_k$ holds. □

We are now ready to prove Theorem 5.

***Proof of Theorem 5*.** Let $\{(x^k, m^k)\}$ be generated by Algorithm 3 with $\{(\eta_k, \theta_k)\}$ defined in (27), and let $\{\mathcal{P}_k\}$ be defined in (34) with such $\{(x^k, m^k)\}$ and $\{p_k\}$ given in (85). By Lemma 17, one can see that such $\{(\eta_k, \theta_k, p_k)\}$ satisfies the assumptions in Lemma 16 and Algorithm 3. In addition, by (34) and (85), one has that

$$
\mathbb{E}[\mathcal{P}_0] = f(x^0) + p_0\mathbb{E}[\|m^0 - \nabla f(x^0)\|^\alpha] = f(x^0) + \mathbb{E}[\|G(x^0;\xi^0) - \nabla f(x^0)\|^\alpha] \leq f(x^0) + \sigma^\alpha, \tag{87}
$$

$$\mathbb{E}[\mathcal{P}_K] = \mathbb{E}[f(x^K) + p_K\|m^K - \nabla f(x^K)\|^\alpha] \geq \mathbb{E}[f(x^K)] \geq f_{\text{low}}. \tag{88}$$

Taking expectation on both sides of (82) with respect to $\{\xi^i\}_{i=0}^{k+1}$, we have that for all $k \geq 0$,

$$\mathbb{E}[\mathcal{P}_{k+1}] \leq \mathbb{E}[\mathcal{P}_k] - \eta_k\mathbb{E}[\|\nabla f(x^k)\|] + \frac{L_1}{2}\eta_k^2 + \frac{(\alpha-1)(2\eta_k)^{\alpha/(\alpha-1)}}{\alpha^{\alpha/(\alpha-1)}(\theta_k p_k/2)^{1/(\alpha-1)}} + 6(L_1^\alpha + L^\alpha)\eta_k^\alpha p_{k+1} + 6\sigma^\alpha\theta_k^\alpha p_{k+1}.$$

Summing up this inequality over $k = 0, \ldots, K-1$, and using (87) and (88), we obtain that for all $K \geq 1$,

$$\begin{aligned}
f_{\text{low}} &\overset{(88)}{\leq} \mathbb{E}[\mathcal{P}_K] \leq \mathbb{E}[\mathcal{P}_0] - \sum_{k=0}^{K-1}\eta_k\mathbb{E}[\|\nabla f(x^k)\|] + \sum_{k=0}^{K-1}\Big(\frac{L_1}{2}\eta_k^2 + \frac{(\alpha-1)(2\eta_k)^{\alpha/(\alpha-1)}}{\alpha^{\alpha/(\alpha-1)}(\theta_k p_k/2)^{1/(\alpha-1)}} \\
&\qquad + 6(L_1^\alpha + L^\alpha)\eta_k^\alpha p_{k+1} + 6\sigma^\alpha\theta_k^\alpha p_{k+1}\Big) \\
&\overset{(87)}{\leq} f(x^0) + \sigma^\alpha - \eta_{K-1}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(x^k)\|] + \sum_{k=0}^{K-1}\Big(\frac{L_1}{2}\eta_k^2 + \frac{(\alpha-1)(2\eta_k)^{\alpha/(\alpha-1)}}{\alpha^{\alpha/(\alpha-1)}(\theta_k p_k/2)^{1/(\alpha-1)}} \\
&\qquad + 12(L_1^\alpha + L^\alpha)\eta_k^\alpha p_k + 12\sigma^\alpha\theta_k^\alpha p_k\Big),
\end{aligned} \tag{89}$$

where the last inequality follows from (87) and the fact that $\{\eta_k\}$ is nonincreasing and $p_{k+1} \leq 2p_k$ for all $k \geq 0$. Rearranging the terms in (89), and using (26), (27), and (85), we obtain that for all $K \geq 3$,

$$\begin{aligned}
\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(x^k)\|] &\overset{(89)}{\leq} \frac{f(x^0) - f_{\text{low}} + \sigma^\alpha}{K\eta_{K-1}} \\
&\quad + \frac{1}{K\eta_{K-1}}\sum_{k=0}^{K-1}\Big(\frac{L_1}{2}\eta_k^2 + \frac{(\alpha-1)(2\eta_k)^{\alpha/(\alpha-1)}}{\alpha^{\alpha/(\alpha-1)}(\theta_k p_k/2)^{1/(\alpha-1)}} + 12(L_1^\alpha + L^\alpha)\eta_k^\alpha p_k + 12\sigma^\alpha\theta_k^\alpha p_k\Big) \\
&\overset{(27)(85)}{=} \frac{f(x^0) - f_{\text{low}} + \sigma^\alpha}{K^{(\alpha-1)/(2\alpha-1)}} \\
&\quad + \frac{1}{K^{(\alpha-1)/(2\alpha-1)}}\sum_{k=0}^{K-1}\Big(\frac{L_1}{2(k+1)^{2\alpha/(2\alpha-1)}} + \frac{2^{1/(\alpha-1)}(\alpha-1)(2/\alpha)^{\alpha/(\alpha-1)} + 12(L_1^\alpha + L^\alpha) + 12\sigma^\alpha}{k+1}\Big) \\
&\leq \frac{f(x^0) - f_{\text{low}} + \sigma^\alpha}{K^{(\alpha-1)/(2\alpha-1)}} + \frac{L_1/2 + 2^{1/(\alpha-1)}(\alpha-1)(2/\alpha)^{\alpha/(\alpha-1)} + 12(L_1^\alpha + L^\alpha) + 12\sigma^\alpha}{K^{(\alpha-1)/(2\alpha-1)}}\sum_{k=0}^{K-1}\frac{1}{k+1} \\
&\leq \frac{2(f(x^0) - f_{\text{low}} + \sigma^\alpha + L_1/2 + 2^{1/(\alpha-1)}(\alpha-1)(2/\alpha)^{\alpha/(\alpha-1)} + 12(L_1^\alpha + L^\alpha) + 12\sigma^\alpha)\ln K}{K^{(\alpha-1)/(2\alpha-1)}} \\
&\overset{(26)}{=} \frac{M_\alpha \ln K}{K^{(\alpha-1)/(2\alpha-1)}},
\end{aligned}$$

where the second inequality follows from $2\alpha/(2\alpha-1) > 1$, the third inequality follows from $\sum_{k=0}^{K-1} 1/(k+1) \leq \ln(2K+1) \leq 2\ln K$ due to (37) and $K \geq 3$. Recall that ι_K is uniformly selected from $\{0, \ldots, K-1\}$. It follows from this and the above relation that

$$\mathbb{E}[\|\nabla f(x^{\iota_K})\|] = \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(x^k)\|] \leq \frac{M_\alpha \ln K}{K^{(\alpha-1)/(2\alpha-1)}} \qquad \forall K \geq 3. \tag{90}$$

By Lemma 3 with $(\beta, u, v) = ((\alpha-1)/(2\alpha-1), (\alpha-1)\epsilon/(2(2\alpha-1)M_\alpha), K)$, one can see that

$$K^{-(\alpha-1)/(2\alpha-1)}\ln K \leq \frac{\epsilon}{M_\alpha} \qquad \forall K \geq \Big(\frac{2(2\alpha-1)M_\alpha}{(\alpha-1)\epsilon}\ln\Big(\frac{2(2\alpha-1)M_\alpha}{(\alpha-1)\epsilon}\Big)\Big)^{(2\alpha-1)/(\alpha-1)},$$

which together with (90) implies that Theorem 5 holds. $\square$

The next lemma establishes some property for a specific choice of $\{(\theta_k, p_k)\}$, which will be used to prove Theorem 6 subsequently.

Lemma 18. *Let $\{\theta_k\}$ be given in (30), and $\{p_k\}$ be defined as*

$$p_k = (k+1)^{2(\alpha-1)^2/(3\alpha)} \qquad \forall k \geq 0. \tag{91}$$

Then $(1-\theta_k)p_{k+1} \leq (1-\theta_k/2)p_k$ holds for all $k \geq 0$.

Proof. Fix any $k \geq 0$. Observe that

$$\frac{1-\theta_k/2}{1-\theta_k} = 1 + \frac{\theta_k}{2(1-\theta_k)} \geq 1 + \frac{\theta_k}{2} \overset{(30)}{=} 1 + \frac{1}{2(k+1)^{2/3}} \geq 1 + \frac{1}{2(k+1)}, \tag{92}$$

In addition, notice from (91) that

$$\frac{p_{k+1}}{p_k} = \Big(1 + \frac{1}{k+1}\Big)^{2(\alpha-1)^2/(3\alpha)} \leq \Big(1 + \frac{1}{k+1}\Big)^{1/3} \leq 1 + \frac{1}{3(k+1)},$$

where the first inequality is due to $2(\alpha-1)^2/(3\alpha) \leq 1/3$ for all $\alpha \in (1,2]$, and the last inequality is due to $(1+\tau)^\beta \leq 1 + \tau\beta$ for all $\tau > -1$ and $\beta \in [0,1]$. The above relation together with (92) implies that $(1-\theta_k)p_{k+1} \leq (1-\theta_k/2)p_k$ holds. $\square$

We are now ready to prove Theorem 6.

Proof of Theorem 6. Let $\{(x^k, m^k)\}$ be generated by Algorithm 3 with $\{(\eta_k, \theta_k)\}$ defined in (30), and let $\{\mathcal{P}_k\}$ be defined in (34) with such $\{(x^k, m^k)\}$ and $\{p_k\}$ given in (91). By Lemma 18, one can see that such $\{(\eta_k, \theta_k, p_k)\}$ satisfies the assumptions in Lemma 16 and Algorithm 3. Using this and similar arguments as those for deriving (89), we have that for all $K \geq 1$,

$$\begin{aligned} f_{\text{low}} \leq\ & f(x^0) + \sigma^\alpha - \eta_{K-1} \sum_{k=0}^{K-1} \mathbb{E}[\|\nabla f(x^k)\|] \\ & + \sum_{k=0}^{K-1} \Big(\frac{L_1}{2}\eta_k^2 + \frac{(\alpha-1)(2\eta_k)^{\alpha/(\alpha-1)}}{\alpha^{\alpha/(\alpha-1)}(\theta_k p_k/2)^{1/(\alpha-1)}} + 12(L_1^\alpha + L^\alpha)\eta_k^\alpha p_k + 12\sigma^\alpha \theta_k^\alpha p_k \Big). \end{aligned} \tag{93}$$

Rearranging the terms in (93), and using (28), (29), (30), and (91), we obtain that for all $K \geq 3$,

$$\begin{aligned} \frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}[\|\nabla f(x^k)\|] &\overset{(93)}{\leq} \frac{f(x^0) - f_{\text{low}} + \sigma^\alpha}{K\eta_{K-1}} \\ &\quad + \frac{1}{K\eta_{K-1}} \sum_{k=0}^{K-1} \Big(\frac{L_1}{2}\eta_k^2 + \frac{(\alpha-1)(2\eta_k)^{\alpha/(\alpha-1)}}{\alpha^{\alpha/(\alpha-1)}(\theta_k p_k/2)^{1/(\alpha-1)}} + 12(L_1^\alpha + L^\alpha)\eta_k^\alpha p_k + 12\sigma^\alpha\theta_k^\alpha p_k \Big) \\ &\overset{(30)(91)}{=} \frac{f(x^0) - f_{\text{low}} + \sigma^\alpha}{K^{1/3}} + \frac{1}{K^{1/3}} \sum_{k=0}^{K-1} \Big(\frac{L_1}{2(k+1)^{4/3}} + \frac{2^{1/(\alpha-1)}(\alpha-1)(2/\alpha)^{\alpha/(\alpha-1)} + 12(L_1^\alpha + L^\alpha) + 12\sigma^\alpha}{(k+1)^{2(2\alpha-1)/(3\alpha)}} \Big) \\ &\leq \frac{f(x^0) - f_{\text{low}} + \sigma^\alpha}{K^{1/3}} + \frac{1}{K^{1/3}} \sum_{k=0}^{K-1} \Big(\frac{L_1}{2(k+1)} + \frac{(2^{1/(\alpha-1)}(\alpha-1)(2/\alpha)^{\alpha/(\alpha-1)} + 12(L_1^\alpha + L^\alpha) + 12\sigma^\alpha)K^{(2-\alpha)/(3\alpha)}}{k+1} \Big) \\ &\leq \frac{2(f(x^0) - f_{\text{low}} + \sigma^\alpha + L_1/2)\ln K}{K^{1/3}} + \frac{2(2^{1/(\alpha-1)}(\alpha-1)(2/\alpha)^{\alpha/(\alpha-1)} + 12(L_1^\alpha + L^\alpha) + 12\sigma^\alpha)\ln K}{K^{2(\alpha-1)/(3\alpha)}} \end{aligned}$$

$$\overset{(28)(29)}{=} \frac{\widetilde{M}_\alpha \ln K}{K^{1/3}} + \frac{\widehat{M}_\alpha \ln K}{K^{2(\alpha-1)/(3\alpha)}},$$

where the second inequality follows from $2(2\alpha-1)/(3\alpha) \leq 1$ for all $\alpha \in (1,2]$, and the last inequality follows from $\sum_{k=0}^{K-1} 1/(k+1) \leq 2\ln K$ due to (37) and $K \geq 3$. Recall that ι_K is uniformly selected from $\{0,\ldots,K-1\}$. It follows from this and the above relation that

$$\mathbb{E}[\|\nabla f(x^{\iota_K})\|] = \frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}[\|\nabla f(x^k)\|] \leq \frac{\widetilde{M}_\alpha \ln K}{K^{1/3}} + \frac{\widehat{M}_\alpha \ln K}{K^{2(\alpha-1)/(3\alpha)}} \qquad \forall K \geq 3. \tag{94}$$

By Lemma 3 with $(\beta,u,v) = (1/3, \epsilon/(12\widetilde{M}_\alpha), K)$ and $(\beta,u,v) = (2(\alpha-1)/(3\alpha), (\alpha-1)\epsilon/(6\alpha\widehat{M}_\alpha), K)$, one can see that

$$K^{-1/3}\ln K \leq \frac{\epsilon}{2\widetilde{M}_\alpha} \qquad \forall K \geq \Big(\frac{12\widetilde{M}_\alpha}{\epsilon}\ln\Big(\frac{12\widetilde{M}_\alpha}{\epsilon}\Big)\Big)^3,$$
$$K^{-2(\alpha-1)/(3\alpha)}\ln K \leq \frac{\epsilon}{2\widehat{M}_\alpha} \qquad \forall K \geq \Big(\frac{6\alpha\widehat{M}_\alpha}{(\alpha-1)\epsilon}\ln\Big(\frac{6\alpha\widehat{M}_\alpha}{(\alpha-1)\epsilon}\Big)\Big)^{3\alpha/(2(\alpha-1))},$$

which together with (94) implies that Theorem 6 holds. $\square$

References

[1] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

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

[3] A. Cutkosky and H. Mehta. Momentum improves normalized SGD. In *International Conference on Machine Learning*, pages 2260–2268, 2020.

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

[5] A. Cutkosky and F. Orabona. Momentum-based variance reduction in non-convex SGD. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[6] C. Fang, C. J. Li, Z. Lin, and T. Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[7] Y. Gao, A. Rodomanov, and S. U. Stich. Non-convex stochastic composite optimization with Polyak momentum. *arXiv preprint arXiv:2403.02967*, 2024.

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

[9] S. Ghadimi and G. Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1):59–99, 2016.

[10] M. Gurbuzbalaban, U. Simsekli, and L. Zhu. The heavy-tail phenomenon in SGD. In *International Conference on Machine Learning*, pages 3964–3975, 2021.

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[12] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge university press, 2012.

[13] F. Hübler, I. Fatkhullin, and N. He. From gradient clipping to normalization for heavy tailed SGD. In *International Conference on Artificial Intelligence and Statistics*, 2025.

[14] J. Humpherys and T. J. Jarvis. *Foundations of Applied Mathematics Volume 2: Algorithms, Approximation, Optimization*. SIAM, 2020.

[15] A. Khaled and P. Richtárik. Better theory for SGD in the nonconvex world. *Transactions on Machine Learning Research*, 2023. Survey Certification.

[16] M. Klamkin and D. J. Newman. Extensions of the Weierstrass product inequalities. *Mathematics Magazine*, 43(3):137–141, 1970.

[17] L. Lei, C. Ju, J. Chen, and M. I. Jordan. Non-convex finite-sum optimization via SCSG methods. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[18] Z. Li, H. Bao, X. Zhang, and P. Richtárik. PAGE: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In *International Conference on Machine Learning*, pages 6286–6295, 2021.

[19] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pages 740–755, 2014.

[20] L. Liu, Y. Wang, and L. Zhang. High-probability bound for non-smooth non-convex stochastic optimization with heavy tails. In *International Conference on Machine Learning*, 2024.

[21] Z. Liu, J. Zhang, and Z. Zhou. Breaking the lower bound with (little) structure: Acceleration in non-convex stochastic optimization with heavy-tailed noise. In *Conference on Learning Theory*, pages 2266–2290, 2023.

[22] Z. Liu and Z. Zhou. Nonconvex stochastic optimization under heavy-tailed noises: Optimal convergence without gradient clipping. In *International Conference on Learning Representations*, 2025.

[23] L. M. Nguyen, J. Liu, K. Scheinberg, and M. Takáč. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *International Conference on Machine Learning*, pages 2613–2621, 2017.

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

[25] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *International Conference on Computer Vision*, pages 2641–2649, 2015.

[26] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021.

[27] A. Rodomanov and Y. Nesterov. Smoothness parameter of power of Euclidean norm. *Journal of Optimization Theory and Applications*, 185:303–326, 2020.

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

[29] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[30] P. Sharma, N. Ding, S. Goodman, and R. Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Annual Meeting of the Association for Computational Linguistics*, pages 2556–2565, 2018.

[31] U. Simsekli, M. Gürbüzbalaban, T. H. Nguyen, G. Richard, and L. Sagun. On the heavy-tailed theory of stochastic gradient descent for deep neural networks. *arXiv preprint arXiv:1912.00018*, 222, 2019.

[32] U. Simsekli, L. Sagun, and M. Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, pages 5827–5837, 2019.

[33] T. Sun, X. Liu, and K. Yuan. Gradient normalization provably benefits nonconvex SGD under heavy-tailed noise. *arXiv preprint arXiv:2410.16561*, 2024.

[34] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

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