Accelerated stochastic first-order method for convex optimization under heavy-tailed noise

Chuan He*   Zhaosong Lu†

October 13, 2025

Abstract

We study convex composite optimization problems, where the objective function is given by the sum of a prox-friendly function and a convex function whose subgradients are estimated under heavy-tailed noise. Existing work often employs gradient clipping or normalization techniques in stochastic first-order methods to address heavy-tailed noise. In this paper, we demonstrate that a vanilla stochastic algorithm—without additional modifications such as clipping or normalization—can achieve optimal complexity for these problems. In particular, we establish that an accelerated stochastic proximal subgradient method achieves a first-order oracle complexity that is universally optimal for smooth, weakly smooth, and nonsmooth convex optimization, as well as for stochastic convex optimization under heavy-tailed noise. Numerical experiments are further provided to validate our theoretical results.

Keywords: Convex composite optimization, heavy-tailed noise, accelerated stochastic proximal subgradient method, first-order oracle complexity

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

1 Introduction

In this paper, we consider a class of convex composite optimization problems of the form

$$F^* := \min_{x \in \mathbb{R}^n} \{F(x) := f(x) + h(x)\}, \tag{1}$$

where $f, h : \mathbb{R}^n \to (-\infty, \infty]$ are proper lower semicontinuous convex functions such that $\operatorname{dom} h \subseteq \operatorname{dom} f$. We assume that f satisfies a hybrid of smooth and nonsmooth conditions:

$$\|f'(y) - f'(x)\| \le L_f \|y - x\| + H_f \|y - x\|^\nu + M_f \qquad \forall f'(y) \in \partial f(y), f'(x) \in \partial f(x), x, y \in \operatorname{dom} f \tag{2}$$

for some constants $L_f, H_f, M_f \ge 0$ and $\nu \in (0, 1)$. In addition, we assume that the proximal operator associated with h can be computed exactly. Clearly, this class of functions f includes Lipschitz smooth, Hölder smooth, and Lipschitz continuous functions, as well as any nonnegative combination of functions from these three subclasses. As recently observed in [25, Example 1], the sum of a Lipschitz smooth

---

function and a Hölder smooth function is not necessarily a Hölder smooth function. Moreover, the sum of a Hölder smooth function and a Lipschitz continuous function is not necessarily a Lipschitz continuous function. Consequently, the class of problems under consideration is broader than the class of problems studied in [24] with f satisfying

$$\|f'(y) - f'(x)\| \le H_f \|y - x\|^\nu \quad \forall f'(y) \in \partial f(y), f'(x) \in \partial f(x), x, y \in \mathrm{dom}\, f$$

for some $H_f > 0$ and $\nu \in [0, 1]$.

With the rise of data science, instances of problem (1) are increasingly common in modern, often large-scale, applications. As a result, the subgradients of f are typically expensive to compute exactly and can usually only be approximated using stochastic estimators. Stochastic first-order methods have been extensively studied for solving (1) and its variants; see, e.g., [1, 2, 4, 6, 15, 17, 18, 21, 22, 23, 27, 28, 30]. Remarkably, an optimal method has been developed in [15] for solving a special case of (1) with $H_f = 0$ and h being the indicator function of a simple closed convex set, under the assumption that the stochastic subgradient estimator $G(\cdot; \xi)$ of $f(\cdot)$ is unbiased and has bounded variance—that is, $G(\cdot; \xi)$ satisfies the following conditions:

$$\mathbb{E}[G(x; \xi)] \in \partial f(x), \quad \mathbb{E}\big[\|G(x; \xi) - \mathbb{E}[G(x; \xi)]\|^2\big] \le \sigma^2 \quad \forall x \in \mathbb{R}^n \tag{3}$$

for some $\sigma > 0$. Under these conditions, it has been shown in [15] that a projected stochastic subgradient method with Nesterov's acceleration scheme achieves an optimal first-order oracle complexity of

$$\mathcal{O}\Big(\Big(\frac{L_f}{\epsilon}\Big)^{\frac{1}{2}} + \Big(\frac{M_f + \sigma}{\epsilon}\Big)^2\Big) \quad \text{and} \quad \mathcal{O}\Big(\Big(\frac{L_f}{\epsilon}\Big)^{\frac{1}{2}} + \Big(\frac{M_f + \sigma \log(1/\delta)}{\epsilon}\Big)^2\Big) \tag{4}$$

for finding an ϵ-optimal solution of (1) in expectation, and an ϵ-optimal solution with probability at least $1 - \delta$, respectively (see Definition 1 for precise definitions). The first complexity bound above recovers the optimal results achieved by first-order methods for smooth, nonsmooth, and stochastic convex optimization in a unified manner.

In recent years, with the development of machine learning and related fields, challenging stochastic optimization problems often extend beyond those satisfying classical assumption imposed in (3). Recent numerical evidence [12, 31, 32, 36] demonstrates that the stochastic estimator $G(\cdot; \xi)$ in these problems satisfies the following conditions, which include heavy-tailed noise scenarios:

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

for some $\sigma > 0$ and $\alpha \in (1, 2]$, generalizing the classical assumptions in (3). Indeed, when $\alpha < 2$, gradient estimators $G(\cdot; \xi)$ can exhibit unbounded variance, which may preclude the applicability of many classic algorithmic frameworks for them that are specifically developed for problems under condition (3). Notably, most existing algorithmic developments in stochastic optimization under heavy-tailed noise rely on gradient clipping [3, 8, 19, 26, 29, 36] or normalization techniques [13, 14, 20, 33], providing theoretical justification for their empirical success in deep learning. Nevertheless, a recent study [5] shows that vanilla SGD, without using gradient clipping or gradient normalization, can be applied to a special case of (1) with $H_f = 0$ and h being an indicator function, achieving a first-order oracle complexity of $\mathcal{O}(\epsilon^{-\alpha/(\alpha-1)})$. Given that no acceleration scheme is used in [5], the following natural question arises:

Is an accelerated vanilla stochastic algorithm without clipping or normalization applicable to the general problem (1) *under heavy-tailed noise?*

This paper provides an affirmative answer to this question. Specifically, we show that an accelerated stochastic proximal subgradient method (SPGM) achieves optimal complexity guarantees for solving problem (1) under heavy-tailed noise. Our main contributions are summarized below.

- We show that a vanilla SPGM and its accelerated counterpart, without any modifications such as clipping or normalization, can find an approximate optimal solution of (1) both in expectation and with high probability.

- We show that the vanilla SPGM (Algorithm 1) achieves a first-order oracle complexity of

$$\mathcal{O}\Big(\frac{L_f}{\epsilon} + \Big(\frac{H_f}{\epsilon}\Big)^{\frac{2}{1+\nu}} + \Big(\frac{M_f}{\epsilon}\Big)^2 + \Big(\frac{\sigma}{\epsilon}\Big)^{\frac{\alpha}{\alpha-1}}\Big), \tag{5a}$$

$$\text{and}\ \ \mathcal{O}\Big(\frac{L_f}{\epsilon} + \Big(\frac{H_f}{\epsilon}\Big)^{\frac{2}{1+\nu}} + \Big(\frac{M_f}{\epsilon}\Big)^2 + \Big(\frac{\sigma\ln(1/\delta)^{1/\alpha}}{\epsilon}\Big)^{\frac{\alpha}{\alpha-1}}\Big) \tag{5b}$$

for finding an ϵ-stochastic optimal solution and an (ϵ, δ)-stochastic optimal solution, respectively. In addition, we establish that the accelerated SPGM (Algorithm 2) achieves a first-order oracle complexity of

$$\mathcal{O}\Big(\Big(\frac{L_f}{\epsilon}\Big)^{\frac{1}{2}} + \Big(\frac{H_f}{\epsilon}\Big)^{\frac{2}{1+3\nu}} + \Big(\frac{M_f}{\epsilon}\Big)^2 + \Big(\frac{\sigma}{\epsilon}\Big)^{\frac{\alpha}{\alpha-1}}\Big), \tag{6a}$$

$$\text{and}\ \ \mathcal{O}\Big(\Big(\frac{L_f}{\epsilon}\Big)^{\frac{1}{2}} + \Big(\frac{H_f}{\epsilon}\Big)^{\frac{2}{1+3\nu}} + \Big(\frac{M_f}{\epsilon}\Big)^2 + \Big(\frac{\sigma\ln(1/\delta)^{1/\alpha}}{\epsilon}\Big)^{\frac{\alpha}{\alpha-1}}\Big) \tag{6b}$$

for finding an ϵ-stochastic optimal solution and an (ϵ, δ)-stochastic optimal solution, respectively.

It shall be mentioned that the accelerated SPGM achieves universally optimal complexity results for smooth, weakly smooth, and nonsmooth convex optimization, as well as for stochastic convex optimization under heavy-tailed noise. Moreover, for the aforementioned special case of problem (1) studied in [15], our complexity bounds (5b) and (6b) with $\alpha = 2$ enjoy an improved dependence on $\ln(1/\delta)$ compared to the bound in (4) obtained in [15].

The rest of this paper is organized as follows. Section 1.1 presents notation and assumptions. In Sections 2 and 3, we present SPGM and its accelerated counterpart along with their first-order oracle complexity results for finding an approximate solution of problem (1) under heavy-tailed noise. Section 4 presents preliminary numerical results illustrating the performance of the proposed methods. Finally, we provide the proof of the main results in Section 5.

1.1 Notation and assumptions

Throughout this section, we use $\mathbb{R}^n$ to stand for the n-dimensional Euclidean space, and $\|\cdot\|$ to denote the Euclidean norm for vectors. For any proper closed convex function φ, we denote its subdifferential by $\partial\varphi$ and define the proximal mapping associated with φ, with parameter $\eta > 0$, as

$$\text{prox}_{\eta\varphi}(x) := \arg\min_{z\in\mathbb{R}^n}\Big\{\varphi(z) + \frac{1}{2\eta}\|z - x\|^2\Big\}.$$

We denote the domain of φ as $\text{dom}\,\varphi$. For any $s \in \mathbb{R}$ and $\mathcal{A} \subseteq \mathbb{R}$, we define the Boolean indicator function $\mathbb{1}_{\mathcal{A}}(s)$ to be 1 if $s \in \mathcal{A}$ and 0 otherwise. In addition, we use $\mathcal{O}(\cdot)$ to denote the standard big-O notation.

We now make the following assumption throughout this paper.

Assumption 1. (a) *The function f satisfies (2) for some constants $L_f, H_f, M_f \geq 0$ and $\nu \in (0, 1)$.*

(b) *The proximal operator associated with h can be exactly evaluated, and its domain $\text{dom}\,h$ is bounded.*

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

$$\mathbb{E}[G(x;\xi)] \in \partial f(x), \quad \mathbb{E}[\|G(x;\xi) - \mathbb{E}[G(x;\xi)]\|^\alpha] \leq \sigma^\alpha \quad \forall x \in \text{dom}\,f \tag{7}$$

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

We next make some remarks on Assumption 1.

Remark 1. (i) The class of f satisfying Assumption 1(a) is broad, which includes smooth (gradient Lipschitz continuous), weakly smooth (gradient Hölder continuous), and nonsmooth (Lipschitz continuous) functions, as well as any nonnegative combination of functions from these subclasses. Problem (1) with f from these subclasses have been extensively studied in the literature (e.g., [10, 11, 15]). However, there was no study on problem (1) with f satisfying Assumption 1(a) except a very recent work [25]. In particular, [25] proposed first-order methods and established complexity guarantees for such problem under a deterministic first-order oracle, where the exact gradient or an exact subgradient of f is used.

(ii) By a standard argument for deriving the descent inequality, (2) implies

$$f(y) \le f(x) + f'(x)^T(y - x) + \frac{L_f}{2}\|y - x\|^2 + \frac{H_f}{1 + \nu}\|y - x\|^{1+\nu} + M_f\|y - x\| \tag{8}$$

holds for all $f'(x) \in \partial f(x), x, y \in \operatorname{dom} f$. It follows from [24, Lemma 2] that

$$\frac{H_f}{1 + \nu}\|y - x\|^{1+\nu} \le \frac{1}{2}L(\varepsilon)\|y - x\|^2 + \frac{\varepsilon}{8},$$

where

$$L(\varepsilon) := H_f^{\frac{2}{1+\nu}}\left(\frac{4}{\varepsilon}\right)^{\frac{1-\nu}{1+\nu}} \qquad \forall \varepsilon > 0. \tag{9}$$

This together with (8) implies that

$$f(y) \le f(x) + f'(x)^T(y - x) + \frac{1}{2}\big(L_f + L(\varepsilon)\big)\|y - x\|^2 + M_f\|y - x\| + \frac{\varepsilon}{8} \tag{10}$$

holds for any $\varepsilon > 0$ and all $f'(x) \in \partial f(x), x, y \in \operatorname{dom} f$.

(iii) Assumption 1(b) is quite common in stochastic optimization. We define the diameter of $\operatorname{dom} h$ as

$$D_h := \max_{x,y \in \operatorname{dom} h}\{\|x - y\|\}. \tag{11}$$

Moreover, Assumption 1(c) states that $G(x; \xi)$ is an unbiased estimator of a subgradient of $f(x)$, and its αth central moment is uniformly bounded. It is weaker than the commonly used variance bounded assumption corresponding to the case $\alpha = 2$. When $\alpha \in (1, 2)$, the stochastic subgradient noise exhibits heavy-tailed behavior (see, e.g., [36]), a phenomenon commonly encountered in machine learning applications. For ease of presentation, we introduce two related quantities, $\Lambda(\varepsilon)^2$ and $\widetilde{\Lambda}(\delta, \varepsilon)^2$, as follows:

$$\Lambda(\varepsilon)^2 := 8(\alpha - 1)^2\left(\frac{\sigma}{\alpha}\right)^{\frac{\alpha}{\alpha-1}}\left(\frac{8D_h}{\varepsilon}\right)^{\frac{2-\alpha}{\alpha-1}}, \quad \widetilde{\Lambda}(\delta, \varepsilon)^2 := \left(1 + \ln\left(\frac{2}{\delta}\right)\right)^{\frac{1}{\alpha-1}}\Lambda(\varepsilon)^2 \qquad \forall \varepsilon, \delta > 0, \tag{12}$$

which will be used to analyze stochastic algorithms under heavy-tailed noise.

We next introduce another assumption, which will be used to establish complexity bounds for finding approximate solutions of problem (1) with high-probability guarantees.

Assumption 2. *The stochastic subgradient estimator $G : \mathbb{R}^n \times \Xi \to \mathbb{R}^n$ satisfies*

$$\mathbb{E}[\exp\{\|G(x; \xi) - \mathbb{E}[G(x; \xi)]\|^\alpha/\sigma^\alpha\}] \le \exp\{1\} \qquad \forall x \in \operatorname{dom} f, \tag{13}$$

where $\sigma > 0$ and $\alpha \in (1, 2]$ are given in Assumption 1(c).

We now make some remarks regarding Assumption 2.

Remark 2. Assumption 2 states that the stochastic subgradient noise follows a sub-Weibull distribution (see, e.g., [35]). This assumption is weaker than the standard sub-Gaussian assumption imposed in [15, Assumption A2], which corresponds to the case with $\alpha = 2$. When $\alpha \in (1, 2)$, condition (13) implies the second condition in (7) and indicates that the stochastic subgradient noise has heavy tails.

We next give formal definitions for approximate stochastic optimal solutions of problem (1).

Definition 1. Let $\epsilon, \delta \in (0, 1)$. We say that

- $x \in \mathbb{R}^n$ is *an ϵ-stochastic optimal solution* of (1) if it satisfies $\mathbb{E}[F(x) - F^*] \leq \epsilon$; and

- $x \in \mathbb{R}^n$ is *an (ϵ, δ)-stochastic optimal solution x* of (1) if it satisfying $F(x) - F^* \leq \epsilon$ with probability at least $1 - \delta$.

2 A stochastic proximal subgradient method

In this section, we present an SPGM and establish its first-order oracle complexity for solving problem (1) under heavy-tailed noise.

The SPGM was originally proposed for solving a special case of (1) with $H_f = 0$ under the conditions (3) (see, e.g., [15, 22]). We now extend it to address the general problem (1) in the presence of heavy-tailed noise. In particular, the SPGM generates two sequences, $\{x^k\}$ and $\{z^k\}$. At each iteration $k \geq 0$, SPGM first updates x^{k+1} by performing a stochastic proximal subgradient step. It then computes z^{k+1} as a weighted average of the past iterates $\{x^t\}_{t=1}^{k+1}$. The details of this method are presented in Algorithm 1, with specific choices of step sizes provided in Theorem 1.

Algorithm 1 A stochastic proximal subgradient method

Input: starting point $x^0 \in \mathrm{dom}\, h$, step sizes $\{\eta_k\} \subset (0, \infty)$.

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

Update the next iterate:

$$x^{k+1} = \mathrm{prox}_{\eta_k h}(x^k - \eta_k G(x^k; \xi_k)). \tag{14}$$

Compute the weighted average:

$$z^{k+1} = \Big(\sum_{t=0}^{k} \eta_t \Big)^{-1} \sum_{t=0}^{k} \eta_t x^{t+1}. \tag{15}$$

end for

The theorem below establishes a complexity bound for Algorithm 1 to compute an ϵ-stochastic optimal solution and an (ϵ, δ)-stochastic optimal solution of (1), respectively. Its proof is deferred to Section 5.1.

Theorem 1. *Suppose that Assumption 1 holds. Let $\epsilon, \delta \in (0, 1)$ be arbitrarily chosen, and let K be a pre-chosen maximum iteration number for running Algorithm 1. Let $L(\cdot), D_h$, and $(\Lambda(\cdot), \widetilde{\Lambda}(\cdot, \cdot))$ be defined in (9), (11), and (12), respectively, L_f, M_f be given in Assumption 1(a), and let*

$$\eta = \min \left\{ \frac{1}{4(L_f + L(\epsilon))}, \frac{D_h}{[2K(M_f^2 + \Lambda(\epsilon)^2)]^{1/2}} \right\}, \quad \tilde{\eta} = \min \left\{ \frac{1}{4(L_f + L(\epsilon))}, \frac{D_h}{[2K(M_f^2 + \widetilde{\Lambda}(\delta, \epsilon)^2)]^{1/2}} \right\}. \tag{16}$$

Then the following statements hold.

(i) *Let $\{z^k\}$ be generated by Algorithm 1 with $\eta_k \equiv \eta$ for all $k \geq 0$. Then, $\mathbb{E}[F(z^K) - F^*] \leq \epsilon$ for all K satisfying*

$$K \geq \max\left\{\frac{8D_h^2(L_f + L(\epsilon))}{\epsilon}, \frac{8D_h^2(M_f + \Lambda(\epsilon))^2}{\epsilon^2}, 1\right\}. \tag{17}$$

(ii) *Suppose additionally that Assumption 2 holds. Let $\{z^k\}$ be generated by Algorithm 1 with $\eta_k \equiv \tilde{\eta}$ for all $k \geq 0$. Then, with probability at least $1 - \delta$, $F(z^K) - F^* \leq \epsilon$ holds for all K satisfying*

$$K \geq \max\left\{\frac{8D_h^2(L_f + L(\epsilon))}{\epsilon}, \frac{32D_h^2(M_f + \widetilde{\Lambda}(\delta,\epsilon))^2}{\epsilon^2}, \left(\left(\frac{4\alpha D_h \sigma}{\epsilon}\right)^{\frac{\alpha}{\alpha-1}} + \mathbb{1}_{(1,2)}(\alpha)\right) \cdot \frac{\ln(2/\delta)}{\alpha - 1}, 1\right\}. \tag{18}$$

Remark 3. From Theorem 1 and (12), we see that Algorithm 1 achieves a first-order oracle complexity of

$$\mathcal{O}\left(\frac{L_f + L(\epsilon)}{\epsilon} + \left(\frac{M_f + \Lambda(\epsilon)}{\epsilon}\right)^2\right) \quad \text{and} \quad \mathcal{O}\left(\frac{L_f + L(\epsilon)}{\epsilon} + \left(\frac{M_f + \widetilde{\Lambda}(\delta,\epsilon)}{\epsilon}\right)^2\right)$$

for finding an ϵ-stochastic optimal solution and an (ϵ, δ)-stochastic optimal solution of (1), respectively. Further, in view of the definitions of $L(\cdot)$ and $(\Lambda(\cdot), \widetilde{\Lambda}(\cdot, \cdot))$ in (9) and (12), these bounds reduce to (5), which achieves the optimal dependence on ϵ for nonsmooth convex problems [23] and for stochastic convex optimization under heavy-tailed noise [5, 19]. However, for smooth and weakly smooth convex problems, the above bounds are not optimal.

3 An accelerated stochastic proximal subgradient method

In this section, we present an accelerated SPGM and show that it achieves a universally optimal first-order oracle complexity for solving smooth, weakly smooth, and nonsmooth convex problems, as well as stochastic convex problems under heavy-tailed noise.

Algorithm 2 An accelerated stochastic proximal subgradient method
---
Input: starting point $x^0 = z^0 \in \text{dom}\, h$, step sizes $\{\eta_k\} \subset (0, \infty)$, weighting parameters $\{\gamma_k\} \subset (0, 1]$.
  for $k = 0, 1, 2, \ldots$ **do**
  Compute the intermediate point:

$$y^k = (1 - \gamma_k)z^k + \gamma_k x^k \tag{19}$$

  Update the next iterate:

$$x^{k+1} = \text{prox}_{\eta_k h}(x^k - \eta_k G(y^k; \xi_k)). \tag{20}$$

  Compute the weighted average:

$$z^{k+1} = (1 - \gamma_k)z^k + \gamma_k x^{k+1}. \tag{21}$$

  end for

The accelerated SPGM was originally proposed in [15] for solving a special case of problem (1) with $H_f = 0$ under conditions (3). We now extend it to handle the general problem (1) under heavy-tailed noise. The accelerated SPGM can be viewed as a stochastic variant of Nesterov's accelerated proximal gradient method [34, Algorithm 1], obtained by replacing the deterministic gradient with a stochastic

subgradient. Specifically, the method generates three sequences, $\{x^k\}$, $\{y^k\}$, and $\{z^k\}$. The sequence $\{x^k\}$ represents the main iterates, updated via a proximal operator. The sequence $\{z^k\}$ denotes the aggregated iterates, where each z^k is a weighted average of $\{x^t\}_{t=0}^k$. The sequence $\{y^k\}$ serves as an intermediate sequence, with each y^k computed as an average of x^k and z^k. The complete description of the method is given in Algorithm 2, and the specific choices of step sizes are provided in Theorem 2.

The next theorem establishes a complexity bound for Algorithm 2 to compute an ϵ-stochastic optimal solution and an (ϵ, δ)-stochastic optimal solution of (1), respectively. Its proof is deferred to Section 5.2.

Theorem 2. *Suppose that Assumption 1 holds. Let $\epsilon, \delta \in (0, 1)$ be arbitrarily chosen, and let K be a pre-chosen maximum iteration number for running Algorithm 2. Let $L(\cdot)$, D_h, and $(\Lambda(\cdot), \widetilde{\Lambda}(\cdot, \cdot))$ be defined in (9), (11), and (12), respectively, L_f, M_f be given in Assumption 1(a), and let*

$$\eta = \min\left\{\frac{1}{4(L_f + L(\epsilon/K))}, \left(\frac{6}{(M_f^2 + \Lambda(\epsilon)^2)(2K + 3)(K + 2)K}\right)^{\frac{1}{2}} D_h\right\}, \tag{22}$$

$$\tilde{\eta} = \min\left\{\frac{1}{4(L_f + L(\epsilon/K))}, \left(\frac{2}{(M_f^2 + \widetilde{\Lambda}(\delta, \epsilon)^2)(K + 2)^2 K}\right)^{\frac{1}{2}} D_h\right\}. \tag{23}$$

Then the following statements hold.

(i) *Let $\{z^k\}$ be generated by Algorithm 2 with $(\gamma_k, \eta_k) = (2/(k+2), (k+2)\eta/2)$ for all $k \geq 0$. Then, $\mathbb{E}[F(z^K) - F^*] \leq \epsilon$ for all K satisfying*

$$K \geq \max\left\{\left(\frac{48D_h^2 L_f}{\epsilon}\right)^{\frac{1}{2}}, \left(\frac{48D_h^2 L(\epsilon)}{\epsilon}\right)^{\frac{1+\nu}{1+3\nu}}, \frac{(24D_h)^2(M_f + \Lambda(\epsilon))^2}{3\epsilon^2}, 2\right\}. \tag{24}$$

(ii) *Suppose additionally that Assumption 2 holds. Let $\{z^k\}$ be generated by Algorithm 2 with $(\gamma_k, \eta_k) = (2/(k+2), (k+2)\tilde{\eta}/2)$ for all $k \geq 0$. Then, with probability at least $1 - \delta$, $F(z^K) - F^* \leq \epsilon$ holds for all K satisfying*

$$K \geq \max\left\{\left(\frac{64D_h^2 L_f}{\epsilon}\right)^{\frac{1}{2}}, \left(\frac{64D_h^2 L(\epsilon)}{\epsilon}\right)^{\frac{1+\nu}{1+3\nu}}, \frac{2(16D_h)^2(M_f + \widetilde{\Lambda}(\delta, \epsilon))^2}{\epsilon^2},\right.$$
$$\left.\left(\left(\frac{16\alpha D_h \sigma}{\epsilon}\right)^{\frac{\alpha}{\alpha-1}} + \mathbb{1}_{(1,2)}(\alpha)\right) \cdot \frac{\ln(2/\delta)}{\alpha - 1}, 2\right\}. \tag{25}$$

Remark 4. From Theorem 2 and (12), we see that Algorithm 2 achieves a first-order oracle complexity of

$$\mathcal{O}\left(\left(\frac{L_f}{\epsilon}\right)^{\frac{1}{2}} + \left(\frac{L(\epsilon)}{\epsilon}\right)^{\frac{1+\nu}{1+3\nu}} + \left(\frac{M_f + \Lambda(\epsilon)}{\epsilon}\right)^2\right) \quad \text{and} \quad \mathcal{O}\left(\left(\frac{L_f}{\epsilon}\right)^{\frac{1}{2}} + \left(\frac{L(\epsilon)}{\epsilon}\right)^{\frac{1+\nu}{1+3\nu}} + \left(\frac{M_f + \widetilde{\Lambda}(\delta, \epsilon)}{\epsilon}\right)^2\right)$$

for finding an ϵ-stochastic optimal solution and an (ϵ, δ)-stochastic optimal solution of (1), respectively. Further, by virtue of the definitions of $L(\cdot)$ and $(\Lambda(\cdot), \widetilde{\Lambda}(\cdot, \cdot))$ in (9) and (12), these bounds reduce to the ones in (6), which match the universal optimal bound for smooth, weakly smooth, and nonsmooth convex problems [7, 23, 24], as well as for stochastic convex optimization with heavy-tailed noise [5, 19]. Moreover, for a special case of problem (1) with $H_f = 0$, the complexity bounds in (6) with $\alpha = 2$ have an improved dependence on $\log(1/\delta)$ compared to the ones in (4) obtained in [15].

4 Numerical experiments

In this section, we present preliminary numerical experiments to evaluate the performance of Algorithms 1 and 2, where Algorithm 2 is referred to as SPGM-A. We also compare these methods with the clipped version of SPGM, denoted as SPGM-C. All algorithms are implemented in MATLAB, and all computations are conducted on a laptop equipped with an Intel Core i9-14900HX processor (2.20 GHz) and 32 GB of RAM.

4.1 ℓ_1-regularized ℓ_2-ℓ_p regression with box constraints

In this subsection, we consider the ℓ_1-regularized ℓ_2-ℓ_p regression with box constraints:

$$\min_{l \le x \le u} \frac{1}{2}\|Ax - b\|^2 + \frac{1}{p}\|Ax - b\|_p^p + \lambda\|x\|_1, \tag{26}$$

where $A \in \mathbb{R}^{n \times n}$, $b \in \mathbb{R}^n$, $u = -l = -100 \cdot \mathbf{1}$ with $\mathbf{1}$ being the all-ones vector, $p = 1.5$, and $\lambda = 1$. We simulate noisy gradient evaluations by setting the stochastic gradient estimator as $G(x; \xi) = \nabla f(x) + \rho\xi$, where $\rho > 0$ is a deterministic scalar, and $\xi \in \mathbb{R}^n$ has independently distributed coordinates, each following a heavy-tailed distribution with density function $p(t) = \omega/(2(1 + |t|)^{1+\omega})$. One can verify that such $G(\cdot; \xi)$ satisfies Assumption 1(c) for every $\alpha \in (1, \omega)$, and that the αth central moment of $G(\cdot; \xi)$ is unbounded for all $\alpha \ge \omega$.

For each triple (n, ρ, ω), we randomly generate 10 instances of problem (26). In particular, we first randomly generate A with all its elements sampled from the standard normal distribution. We then randomly generate $\bar{x}^*$, with all its components sampled from the standard normal distribution, and construct a sparse solution x^* by randomly setting half of its components to zero. Finally, we set $b = Ax^*$.

We apply SPGM, SPGM-A, and SPGM-C to problem (26) to find an approximate solution x^k such that its relative objective value gap $(F(x^k) - F^*)/(F(x^0) - F^*)$ is less than 10^{-4}, where F^* is estimated by CVX [9]. All methods are initialized at the zero vector. Other algorithmic parameters are selected to suit each method well in terms of computational performance.

n	ρ	ω	CPU time (seconds)			Iterations		
			SPGM	SPGM-A	SPGM-C	SPGM	SPGM-A	SPGM-C
500	1	1.8	3.15	1.50	3.16	2602.2	1284.0	2606.8
500	1	1.5	3.54	1.65	3.68	2734.8	1280.3	2727.6
500	1	1.2	3.53	1.73	3.76	2810.5	1311.6	2825.6
500	100	1.8	3.28	1.57	3.24	2528.1	1233.8	2520.5
500	100	1.5	4.38	2.22	3.90	3226.5	1711.5	2689.3
500	100	1.2	4.65	3.16	3.22	4158.1	2740.8	2872.0
1000	1	1.8	18.82	8.11	19.40	5006.9	2130.6	5040.6
1000	1	1.5	20.97	8.99	21.69	4952.5	2149.6	5059.7
1000	1	1.2	21.14	8.77	22.52	5085.2	2175.4	5322.0
1000	100	1.8	20.90	8.95	21.68	5154.1	2174.9	5355.7
1000	100	1.5	24.09	9.32	22.72	5695.5	2167.0	5048.7
1000	100	1.2	24.85	19.08	22.41	6779.0	5165.0	6106.3

Table 1: Numerical results for problem (26).

The computational results of SPGM, SPGM-A, and SPGM-C for solving (26) are presented in Table 1. Specifically, the first three columns list the values of n, ρ, and ω, respectively, while the remaining columns report the average CPU time and the average number of iterations for each triple (n, ρ, ω). We observe that, except for the case $(\rho, \omega) = (100, 1.2)$, SPGM-A substantially outperforms both SPGM and SPGM-C. When $(\rho, \omega) = (100, 1.2)$, the performance gap between SPGM-A and SPGM-C becomes much smaller, although both methods still outperform SPGM. These observations suggest that when the noise level is not too high, the accelerated SPGM can achieve significantly faster convergence than the vanilla SPGM and the clipped SPGM, which is consistent with our theoretical findings.

4.2 ℓ_2-ℓ_p-ℓ_1 regression with ℓ_2-ball constraint

In this subsection, we consider the ℓ_2-ℓ_p-ℓ_1 regression with ℓ_2-ball constraint:

$$\min_{\|x\| \leq u} \frac{1}{2}\|Ax - b\|^2 + \frac{1}{p}\|Ax - b\|_p^p + \lambda\|Ax - b\|_1, \tag{27}$$

where $A \in \mathbb{R}^{n \times n}$, $b \in \mathbb{R}^n$, $u = 100$, $p = 1.5$, and $\lambda = 0.1$. We simulate the noisy gradient evaluations by setting the stochastic gradient estimator as $G(x;\xi) = \nabla f(x) + \rho\xi$, where $\rho > 0$ is a deterministic scalar, and $\xi \in \mathbb{R}^n$ has independently distributed coordinates, each following a heavy-tailed distribution with density function $p(t) = \omega/(2(1 + |t|)^{1+\omega})$. One can verify that such $G(\cdot;\xi)$ satisfies Assumption 1(c) for every $\alpha \in (1, \omega)$, and that the αth central moment of $G(\cdot;\xi)$ is unbounded for all $\alpha \geq \omega$.

For each triple (n, ρ, ω), we randomly generate 10 instances of problem (27). In particular, we first randomly generate A with all its elements sampled from the standard normal distribution. We then randomly generate x^*, with all its components sampled from the standard normal distribution, and set $b = Ax^*$.

We apply SPGM, SPGM-A, and SPGM-C to problem (27) to find an approximate solution x^k such that its relative objective value gap $F(x^k)/F(x^0)$ is less than 10^{-4} (note that $F^* = 0$ due to the data generation setup). All methods are initialized at the zero vector. Other algorithmic parameters are selected to suit each method well in terms of computational performance.

n	ρ	ω	CPU time (seconds)			Iterations		
			SPGM	SPGM-A	SPGM-C	SPGM	SPGM-A	SPGM-C
500	1	1.8	18.00	9.65	26.80	9772.6	3640.0	9374.6
500	1	1.5	12.10	4.11	11.76	9126.5	3167.2	8659.1
500	1	1.2	12.02	3.95	11.85	8838.4	2936.4	8319.8
500	100	1.8	13.30	5.06	13.42	9671.7	3594.2	9411.8
500	100	1.5	16.32	5.19	13.11	12190.4	3828.9	9640.9
500	100	1.2	21.36	12.54	13.00	15843.0	9051.7	9730.7
1000	1	1.8	48.61	18.62	59.23	8524.6	2769.9	8512.5
1000	1	1.5	39.42	13.03	39.54	8305.2	2753.9	8300.3
1000	1	1.2	50.64	12.66	40.22	10454.3	2700.5	8386.8
1000	100	1.8	40.62	13.83	43.13	8126.3	2638.7	8119.6
1000	100	1.5	82.54	24.76	77.66	10729.2	2871.3	9076.5
1000	100	1.2	124.10	52.07	80.57	17992.7	8854.2	9496.5

Table 2: Numerical results for problem (27).

The computational results of SPGM, SPGM-A, and SPGM-C for solving (27) are presented in Table 2. The first three columns list the values of n, ρ, and ω, respectively, while the remaining columns report the average CPU time and the average number of iterations for each triple (n, ρ, ω). We observe that, except for the case $(\rho, \omega) = (100, 1.2)$, SPGM-A significantly outperforms both SPGM and SPGM-C. When $(\rho, \omega) = (100, 1.2)$, the performance gap between SPGM-A and SPGM-C narrows, although both methods still outperform SPGM. These observations suggest that when the noise level is moderate, the accelerated SPGM can achieve substantially faster convergence than both the vanilla and clipped versions of SPGM, which aligns well with our theoretical results.

5   Proof of the main results

In this section, we present the proofs of the main results stated in Sections 2 and 3, namely, Theorems 1 and 2. Throughout this section, let x^* denote an arbitrary but fixed optimal solution to (1).

Before proceeding, we establish several technical lemmas below. The following lemma provides a useful inequality for handling the heavy-tailed noise condition.

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

$$c\sigma^\alpha \hat{\eta}^{\alpha-1} \leq (\alpha-1)c^{\frac{1}{\alpha-1}}\left(\frac{8}{\varepsilon}\right)^{\frac{2-\alpha}{\alpha-1}} \sigma^{\frac{\alpha}{\alpha-1}}\hat{\eta} + \frac{\varepsilon}{8} \qquad \forall c, \sigma, \hat{\eta}, \varepsilon > 0. \tag{28}$$

Proof. When $\alpha = 2$, this inequality holds trivially. We next prove (28) for the case when $\alpha \in (1,2)$. By Young's inequality, one has that $\tau s \leq \tau^p/p + s^q/q$ holds for all $\tau, s > 0$ and $p, q \geq 1$ satisfying $1/p + 1/q = 1$. Letting $\tau = c\sigma^\alpha \hat{\eta}^{\alpha-1}/s$, $p = 1/(\alpha-1)$, and $q = 1/(2-\alpha)$, we obtain that

$$c\sigma^\alpha \hat{\eta}^{\alpha-1} \leq \frac{(\alpha-1)c^{1/(\alpha-1)}\sigma^{\alpha/(\alpha-1)}\hat{\eta}}{s^{1/(\alpha-1)}} + (2-\alpha)s^{1/(2-\alpha)}.$$

Further, let $s > 0$ be such that $\varepsilon/8 = (2-\alpha)s^{1/(2-\alpha)}$. Then, one has $s^{1/(\alpha-1)} = (\varepsilon/(8(2-\alpha)))^{(2-\alpha)/(\alpha-1)}$. Combining these and the above inequality, and using $(\alpha-2)^{\alpha-2} \leq 1$, we conclude that (28) holds. $\qquad\square$

The next lemma will be used to derive the complexity bounds.

Lemma 2. *Let $a, b, c > 0$ be given, $t^* = \min\{c, (a/b)^{1/2}\}$, and $\varphi(t) = a/t + bt$ for $t \in (0, \infty)$. Then, it holds that*

$$\min_{t \in (0,c]} \varphi(t) = \varphi(t^*) \leq a/c + 2(ab)^{1/2}. \tag{29}$$

Proof. It is easy to see that the first relation in (29) holds. We now prove the second relation in (29) by considering two separate cases. If $c \leq (a/b)^{1/2}$, one has $\varphi(t^*) = a/c + bc \leq a/c + (ab)^{1/2}$. On the other hand, if $c > (a/b)^{1/2}$, one has $\varphi(t^*) = 2(ab)^{1/2} < a/c + 2(ab)^{1/2}$. Combining these cases, we conclude that the second relation in (29) holds as desired. $\qquad\square$

The lemma below provides an inequality that will used to establish a concentration inequality subsequently.

Lemma 3. *Let $\alpha \in (1,2]$ be given. Then, $e^t \leq t + e^{|t|^\alpha}$ holds for all $t \in \mathbb{R}$.*

Proof. We prove this inequality by considering three separate cases.

Case 1) $t \in (1, \infty)$. This along with $\alpha > 1$ implies that $e^t < e^{t^\alpha} = e^{|t|^\alpha}$, and hence $e^t \leq t + e^{|t|^\alpha}$ holds.

Case 2) $t \in [-1, 1]$. By this and $\alpha \in (1,2]$, one has

$$e^t = 1 + t + \sum_{s=2}^\infty \frac{t^s}{s!} \leq 1 + t + t^2 \sum_{s=2}^\infty \frac{1}{s!} = 1 + t + t^2 \left(\sum_{s=0}^\infty \frac{1}{s!} - 2\right) = 1 + t + (e-2)t^2$$

$$< 1 + t + t^2 \leq 1 + t + |t|^\alpha \leq t + e^{|t|^\alpha},$$

where the first and third inequalities follow from $t \in [-1,1]$ and $\alpha \leq 2$, and the last inequality is due to the convexity of the exponential function.

Case 3) $t \in (-\infty, -1)$. Using this and $\alpha \in (1,2]$, we have

$$e^{|t|^\alpha} \geq e^{-t} \geq 1 - t \geq e^t - t,$$

where the first and last inequalities are due to $t < -1$ and $\alpha > 1$, and the second inequality follows from the convexity of the exponential function.

Combining the above three cases, we conclude that $e^t \leq t + e^{|t|^\alpha}$ holds for all $t \in \mathbb{R}$. $\qquad\square$

The next lemma provides a concentration inequality for a martingale difference sequence of sub-Weibull random variables, which generalizes the result established in [16, Lemma 2] for sub-Gaussian random variables. It will be used to establish high-probability complexity bounds.

Lemma 4. *Let $\{\xi_k\}$ be a sequence of i.i.d. random variables, $\varsigma > 0$ and $\alpha \in (1, 2]$ be given, and $\{\phi_k(\cdot)\}$ be a sequence of deterministic functions. Define $\phi^k = \phi_k(\xi_{[k]})$ with $\xi_{[k]} = \{\xi_t\}_{t=0}^k$ for all $k = 0, 1, \ldots,$ and let $\mathbb{E}_{\xi_0}[\cdot \,|\xi_{[-1]}] := \mathbb{E}_{\xi_0}[\cdot]$. Suppose that $\mathbb{E}_{\xi_k}[\phi^k \,|\xi_{[k-1]}] = 0$ and $\mathbb{E}_{\xi_k}[\exp\{|\phi^k/\varsigma|^\alpha\} \,|\xi_{[k-1]}] \leq \exp\{1\}$ hold for all $k = 1, 2, \ldots.$ Then, we have*

$$\mathbb{P}\left(\sum_{k=0}^{K-1} \phi^k > \Omega \varsigma K^{\frac{1}{\alpha}} \right) \leq \exp\left\{ (1-\alpha)\left(\frac{\Omega}{\alpha}\right)^{\frac{\alpha}{\alpha-1}} \right\} \qquad \forall \Omega \geq 0, K \geq \max\left\{ 1, \mathbb{1}_{(1,2)}(\alpha)\left(\frac{\Omega}{\alpha}\right)^{\frac{\alpha}{\alpha-1}} \right\}. \tag{30}$$

Proof. When $\alpha = 2$, (30) holds due to [16, Lemma 2]. It remains to show that (30) holds for any $\alpha \in (1, 2)$. To this end, we first show that

$$\mathbb{E}_{\xi_k}\left[\exp\{\tau \phi^k\} \,|\xi_{[k-1]} \right] \leq \exp\{\max\{(\tau\varsigma)^\alpha, (\tau\varsigma)^{\alpha'}\}\} \qquad \forall \tau \geq 0, k \geq 0, \tag{31}$$

where $\alpha' = \alpha/(\alpha-1)$. Indeed, for notational convenience, we denote $\bar{\phi}^k = \phi^k/\varsigma$ for all $k \geq 0$. Then, one has $\mathbb{E}_{\xi_k}[\bar{\phi}^k \,|\xi_{[k-1]}] = 0$ and $\mathbb{E}_{\xi_k}[\exp\{|\bar{\phi}^k|^\alpha\} \,|\xi_{[k-1]}] \leq \exp\{1\}$ for all $k \geq 0$. By Jensen's inequality and the concavity of $\psi(s) = s^\beta$ for any $\beta \in [0, 1]$, one has that

$$\begin{aligned}
\mathbb{E}_{\xi_k}\left[\exp\{\beta|\bar{\phi}^k|^\alpha\} \,|\xi_{[k-1]} \right] &= \mathbb{E}_{\xi_k}\left[(\exp\{|\bar{\phi}^k|^\alpha\})^\beta \,|\xi_{[k-1]} \right] \\
&\leq \left(\mathbb{E}_{\xi_k}\left[\exp\{|\bar{\phi}^k|^\alpha\} \,|\xi_{[k-1]} \right] \right)^\beta \leq \exp\{\beta\} \qquad \forall \beta \in [0,1], k \geq 0, \tag{32}
\end{aligned}$$

where the last inequality follows from $\mathbb{E}_{\xi_k}[\exp\{|\bar{\phi}^k|^\alpha\} \,|\xi_{[k-1]}] \leq \exp\{1\}$. Using this and Lemma 3 with $t = \lambda \bar{\phi}^k$, we obtain that

$$\mathbb{E}_{\xi_k}\left[\exp\{\lambda \bar{\phi}^k\} \,|\xi_{[k-1]} \right] \leq \mathbb{E}_{\xi_k}\left[\lambda \bar{\phi}^k \,|\xi_{[k-1]} \right] + \mathbb{E}_{\xi_k}\left[\exp\{|\lambda \bar{\phi}^k|^\alpha\} \,|\xi_{[k-1]} \right] \leq \exp\{\lambda^\alpha\} \qquad \forall \lambda \in [0,1], k \geq 0, \tag{33}$$

where the last inequality is due to $\mathbb{E}_{\xi_k}[\bar{\phi}^k \,|\xi_{[k-1]}] = 0$ and (32) with $\beta = \lambda^\alpha$. In addition, by $\alpha' = \alpha/(\alpha-1)$ and Young's inequality, one has $ts \leq |t|^\alpha/\alpha + |s|^{\alpha'}/\alpha'$ for all $s, t \in \mathbb{R}$. It follows from this and (32) with $\beta = 1/\alpha$ that

$$\begin{aligned}
\mathbb{E}_{\xi_k}\left[\exp\{\lambda \bar{\phi}^k\} \,|\xi_{[k-1]} \right] &\leq \mathbb{E}_{\xi_k}\left[\exp\{|\bar{\phi}^k|^\alpha/\alpha\} \,|\xi_{[k-1]} \right] \cdot \exp\{\lambda^{\alpha'}/\alpha'\} \overset{(32)}{\leq} \exp\{1/\alpha + \lambda^{\alpha'}/\alpha'\} \\
&\leq \exp\{(1/\alpha + 1/\alpha')\lambda^{\alpha'}\} = \exp\{\lambda^{\alpha'}\} \qquad \forall \lambda \geq 1, k \geq 0, \tag{34}
\end{aligned}$$

where the third inequality is due to $\alpha' > 0$ and $\lambda \geq 1$, and the last relation follows from $1/\alpha + 1/\alpha' = 1$. Using (33) and (34), we obtain that $\mathbb{E}_{\xi_k}\left[\exp\{\lambda \bar{\phi}^k\} \,|\xi_{[k-1]} \right] \leq \exp\{\max\{\lambda^\alpha, \lambda^{\alpha'}\}\}$ for all $\lambda \geq 0$ and $k \geq 0$. Substituting $\bar{\phi}^k = \phi^k/\varsigma$ into this inequality, and rearranging the terms, we obtain that

$$\mathbb{E}_{\xi_k}\left[\exp\{\tau \phi^k\} \,|\xi_{[k-1]} \right] = \mathbb{E}_{\xi_k}\left[\exp\{\tau\varsigma \bar{\phi}^k\} \,|\xi_{[k-1]} \right] \leq \exp\{\max\{(\tau\varsigma)^\alpha, (\tau\varsigma)^{\alpha'}\}\} \qquad \forall \tau \geq 0, k \geq 0,$$

and hence (31) holds as desired.

Further, using (31) and the definition of ϕ^k for all $k \geq 0$, we obtain that for all $\tau \geq 0$ and $K \geq 1$,

$$\mathbb{E}\left[\exp\left\{ \tau \sum_{k=0}^{K-1} \phi^k \right\} \right] = \mathbb{E}_{\xi_{[K-2]}}\left[\mathbb{E}_{\xi_{K-1}}\left[\exp\left\{ \tau \sum_{k=0}^{K-1} \phi^k \right\} \,\Big|\xi_{[K-2]} \right] \right]$$

$$= \mathbb{E}_{\xi_{[K-2]}} \left[\exp \left\{ \tau \sum_{k=0}^{K-2} \phi^k \right\} \cdot \mathbb{E}_{\xi_{K-1}} \left[\exp \left\{ \tau \phi^{K-1} \right\} \big| \xi_{[K-2]} \right] \right]$$

$$= \mathbb{E} \left[\exp \left\{ \tau \sum_{k=0}^{K-2} \phi^k \right\} \cdot \mathbb{E}_{\xi_{K-1}} \left[\exp \left\{ \tau \phi^{K-1} \right\} \big| \xi_{[K-2]} \right] \right]$$

$$\overset{(31)}{\leq} \mathbb{E} \left[\exp \left\{ \tau \sum_{k=0}^{K-2} \phi^k \right\} \right] \cdot \exp\{\max\{(\tau\varsigma)^\alpha, (\tau\varsigma)^{\alpha'}\}\}.$$

This recursion implies that

$$\mathbb{E} \left[\exp \left\{ \tau \sum_{k=0}^{K-1} \phi^k \right\} \right] \leq \exp \left\{ \max\{(\tau\varsigma)^\alpha, (\tau\varsigma)^{\alpha'}\} K \right\} \qquad \forall \tau \geq 0, K \geq 1.$$

Using this and Markov's inequality, we have that for all $\tau > 0$, $\Omega \geq 0$, and $K \geq 1$,

$$\mathbb{P} \left(\sum_{k=0}^{K-1} \phi^k > \Omega \varsigma K^{\frac{1}{\alpha}} \right) = \mathbb{P} \left(\exp \left\{ \tau \sum_{k=0}^{K-1} \phi^k \right\} > \exp \left\{ \tau \Omega \varsigma K^{\frac{1}{\alpha}} \right\} \right)$$

$$\leq \exp \left\{ -\tau \Omega \varsigma K^{\frac{1}{\alpha}} \right\} \mathbb{E} \left[\exp \left\{ \tau \sum_{k=0}^{K-1} \phi^k \right\} \right] \leq \exp \left\{ \max\{(\tau\varsigma)^\alpha, (\tau\varsigma)^{\alpha'}\} K - \tau \Omega \varsigma K^{\frac{1}{\alpha}} \right\}. \qquad (35)$$

Let $\tau = (\Omega/\alpha)^{1/(\alpha-1)}/(\varsigma K^{1/\alpha})$. It follows that $\tau\varsigma \leq 1$ for all $K \geq (\Omega/\alpha)^{\alpha/(\alpha-1)}$, which together with $\alpha' > \alpha$ implies that $\max\{(\tau\varsigma)^\alpha, (\tau\varsigma)^{\alpha'}\} = (\tau\varsigma)^\alpha$. By this, (35), and the expression of τ, one has that for all $\Omega \geq 0$ and $K \geq \max\{1, (\Omega/\alpha)^{\alpha/(\alpha-1)}\}$,

$$\mathbb{P} \left(\sum_{k=0}^{K-1} \phi^k > \Omega \varsigma K^{\frac{1}{\alpha}} \right) \leq \exp \left\{ (\tau\varsigma)^\alpha K - \tau \Omega \varsigma K^{\frac{1}{\alpha}} \right\} = \exp \left\{ (1-\alpha) \left(\frac{\Omega}{\alpha} \right)^{\frac{\alpha}{\alpha-1}} \right\}.$$

Hence, (30) also holds for any $\alpha \in (1,2)$. $\qquad \square$

5.1 Proof of the main results in Section 2

In this subsection, we first establish a lemma and then use it to prove Theorem 1.

Lemma 5. *Suppose that Assumption 1 holds. Let $\epsilon \in (0,1)$ be arbitrarily chosen, L_f, M_f, α, and σ be given in Assumption 1, and $L(\cdot)$ be defined in (9). Let $\{x^k\}$ be the sequence generated by Algorithm 1 with step sizes $\{\eta_k\}$ satisfying $\eta_k \in \left(0, \frac{1}{4(L_f + L(\epsilon))}\right]$ for all $k \geq 0$. Then, it holds that for all $k \geq 0$,*

$$\eta_k(F(x^{k+1}) - F^*) \leq (\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2)/2 + M_f^2 \eta_k^2 + \eta_k \Delta_k$$

$$+ \frac{(8(\alpha-1))^{\alpha-1} D_h^{2-\alpha} \eta_k^\alpha \delta_k^\alpha}{\alpha^\alpha} + \frac{\epsilon \eta_k}{8}, \qquad (36)$$

where

$$\Delta_k = (\mathbb{E}[G(x^k; \xi_k)] - G(x^k; \xi_k))^T (x^k - x^*), \quad \delta_k = \|\mathbb{E}[G(x^k; \xi_k)] - G(x^k; \xi_k)\| \qquad \forall k \geq 0. \qquad (37)$$

Proof. Fix any $k \geq 0$. By the optimality condition of (14), there exists $h'(x^{k+1}) \in \partial h(x^{k+1})$ such that

$$G(x^k; \xi_k) + \eta_k^{-1}(x^{k+1} - x^k) + h'(x^{k+1}) = 0,$$

12

which along with the convexity of h implies that

$$h(x^{k+1}) \le h(x^*) + h'(x^{k+1})^T(x^{k+1} - x^*) = h(x^*) + G(x^k; \xi_k)^T(x^* - x^{k+1}) + \eta_k^{-1}(x^{k+1} - x^k)^T(x^* - x^{k+1})$$
$$= h(x^*) + G(x^k; \xi_k)^T(x^* - x^{k+1}) + (2\eta_k)^{-1}(\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2 - \|x^{k+1} - x^k\|^2). \qquad (38)$$

Denote $f'(x^k) = \mathbb{E}_{\xi_k}[G(x^k; \xi_k)]$. It follows from (7) that $f'(x^k) \in \partial f(x^k)$. By this, (10), and the convexity of f, one has

$$f(x^{k+1}) \overset{(10)}{\le} f(x^k) + f'(x^k)^T(x^{k+1} - x^k) + \frac{L_f + L(\epsilon)}{2}\|x^{k+1} - x^k\|^2 + \frac{\epsilon}{8} + M_f\|x^{k+1} - x^k\|$$
$$\le f(x^*) + f'(x^k)^T(x^{k+1} - x^*) + \frac{L_f + L(\epsilon)}{2}\|x^{k+1} - x^k\|^2 + \frac{\epsilon}{8} + M_f\|x^{k+1} - x^k\|.$$

Using this, (37), and (38), we obtain that

$$F(x^{k+1}) \overset{(37)(38)}{\le} F(x^*) + \Delta_k + (f'(x^k) - G(x^k; \xi_k))^T(x^{k+1} - x^k) + (2\eta_k)^{-1}(\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2)$$
$$+ \left(\frac{L_f + L(\epsilon)}{2} - \frac{1}{2\eta_k}\right)\|x^{k+1} - x^k\|^2 + \frac{\epsilon}{8} + M_f\|x^{k+1} - x^k\|$$
$$\le F(x^*) + \Delta_k + (f'(x^k) - G(x^k; \xi_k))^T(x^{k+1} - x^k) + (2\eta_k)^{-1}(\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2)$$
$$+ \left(\frac{L_f + L(\epsilon)}{2} - \frac{1}{4\eta_k}\right)\|x^{k+1} - x^k\|^2 + \frac{\epsilon}{8} + M_f^2\eta_k, \qquad (39)$$

where the last inequality follows from $M_f\|x^{k+1} - x^k\| \le \|x^{k+1} - x^k\|^2/(4\eta_k) + M_f^2\eta_k$. In addition, let $\alpha' = \alpha/(\alpha - 1)$. Observe that $\alpha' \ge 2$ due to $\alpha \in (1, 2]$. This together with (11) and $x^{k+1}, x^k \in \text{dom } h$ implies that $\|x^{k+1} - x^k\|^{\alpha'} \le D_h^{\alpha'-2}\|x^{k+1} - x^k\|^2$. Using this, (37), and Young's inequality, we have

$$(f'(x^k) - G(x^k; \xi_k))^T(x^{k+1} - x^k) \le \frac{\left(\left(\frac{\alpha'}{8D_h^{\alpha'-2}\eta_k}\right)^{1/\alpha'}\|x^{k+1} - x^k\|\right)^{\alpha'}}{\alpha'} + \frac{\left(\left(\frac{8D_h^{\alpha'-2}\eta_k}{\alpha'}\right)^{1/\alpha'}\delta_k\right)^{\alpha}}{\alpha}$$
$$= \frac{\|x^{k+1} - x^k\|^{\alpha'}}{8D_h^{\alpha'-2}\eta_k} + \frac{(8(\alpha-1))^{\alpha-1}D_h^{2-\alpha}\eta_k^{\alpha-1}\delta_k^{\alpha}}{\alpha^{\alpha}} \le \frac{\|x^{k+1} - x^k\|^2}{8\eta_k} + \frac{(8(\alpha-1))^{\alpha-1}D_h^{2-\alpha}\eta_k^{\alpha-1}\delta_k^{\alpha}}{\alpha^{\alpha}}.$$

By this inequality, (39), Lemma 1, and $\eta_k \in \left(0, \frac{1}{4(L_f+L(\epsilon))}\right]$, we obtain that

$$F(x^{k+1}) - F(x^*) \le (2\eta_k)^{-1}(\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2) + M_f^2\eta_k + \Delta_k$$
$$+ \left(\frac{L_f + L(\epsilon)}{2} - \frac{1}{8\eta_k}\right)\|x^{k+1} - x^k\|^2 + \frac{(8(\alpha-1))^{\alpha-1}D_h^{2-\alpha}\eta_k^{\alpha-1}\delta_k^{\alpha}}{\alpha^{\alpha}} + \frac{\epsilon}{8}$$
$$\le (2\eta_k)^{-1}(\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2) + M_f^2\eta_k + \Delta_k + \frac{(8(\alpha-1))^{\alpha-1}D_h^{2-\alpha}\eta_k^{\alpha-1}\delta_k^{\alpha}}{\alpha^{\alpha}} + \frac{\epsilon}{8}.$$

Hence, the conclusion (36) holds. $\qquad\square$

We are now ready to provide a proof of Theorem 1.

Proof of Theorem 1. Using (15), (36), and the convexity of f, we obtain that

$$F(z^K) - F^* \overset{(15)}{\le} \frac{\sum_{k=0}^{K-1}\eta_k F(x^{k+1})}{\sum_{k=0}^{K-1}\eta_k} - F^* = \frac{1}{\sum_{k=0}^{K-1}\eta_k}\sum_{k=0}^{K-1}\eta_k(F(x^{k+1}) - F^*)$$

13

$$\overset{(36)}{\leq} \frac{\|x^0 - x^*\|^2}{2\sum_{k=0}^{K-1}\eta_k} + \frac{(8(\alpha-1))^{\alpha-1}D_h^{2-\alpha}\sum_{k=0}^{K-1}\eta_k^\alpha\delta_k^\alpha}{\alpha^\alpha\sum_{k=0}^{K-1}\eta_k} + \frac{M_f^2\sum_{k=0}^{K-1}\eta_k^2}{\sum_{k=0}^{K-1}\eta_k} + \frac{\sum_{k=0}^{K-1}\eta_k\Delta_k}{\sum_{k=0}^{K-1}\eta_k} + \frac{\epsilon}{8}. \quad (40)$$

We now prove statement (i) of Theorem 1. By (12) and Lemma 1 with $c = (8(\alpha-1))^{\alpha-1}D_h^{2-\alpha}/\alpha^\alpha$, $\hat{\eta} = \eta$, and $\varepsilon = \epsilon$, one has

$$\frac{(8(\alpha-1))^{\alpha-1}D_h^{2-\alpha}}{\alpha^\alpha} \cdot \sigma^\alpha\eta^{\alpha-1} \overset{(28)}{\leq} 8(\alpha-1)^2\Big(\frac{\sigma}{\alpha}\Big)^{\frac{\alpha}{\alpha-1}}\Big(\frac{8D_h}{\epsilon}\Big)^{\frac{2-\alpha}{\alpha-1}}\eta + \frac{\epsilon}{8} \overset{(12)}{=} \Lambda(\epsilon)^2\eta + \frac{\epsilon}{8}. \quad (41)$$

In addition, recall from (37) and Assumption 1(c) that $\mathbb{E}_{\xi_k}[\Delta_k] = 0$ and $\mathbb{E}_{\xi_k}[\delta_k^\alpha] \leq \sigma^\alpha$. Using these, (11), (16), (29), (41), $\eta_k \equiv \eta$ for all k, and taking expectation on (40) with respect to $\{\xi_k\}_{k=0}^{K-1}$, we obtain that for all $k \geq 0$,

$$\mathbb{E}[F(z^K) - F(x^*)] \overset{(40)}{\leq} \frac{\|x^0 - x^*\|^2}{2K\eta} + \frac{(8(\alpha-1))^{\alpha-1}D_h^{2-\alpha}}{\alpha^\alpha} \cdot \sigma^\alpha\eta^{\alpha-1} + M_f^2\eta + \frac{\epsilon}{8}$$

$$\overset{(11)(41)}{\leq} \frac{D_h^2}{2K\eta} + (M_f^2 + \Lambda(\epsilon)^2)\eta + \frac{\epsilon}{4} \overset{(16)}{=} \min_{\hat{\eta}}\left\{\frac{D_h^2}{2K\hat{\eta}} + (M_f^2 + \Lambda(\epsilon)^2)\hat{\eta} : \hat{\eta} \in \Big(0, \frac{1}{4(L_f + L(\epsilon))}\Big]\right\} + \frac{\epsilon}{4},$$

$$\overset{(29)}{\leq} \frac{2D_h^2(L_f + L(\epsilon))}{K} + \sqrt{2}D_h\Big(\frac{M_f^2 + \Lambda(\epsilon)^2}{K}\Big)^{1/2} + \frac{\epsilon}{4} \leq \frac{2D_h^2(L_f + L(\epsilon))}{K} + \frac{\sqrt{2}D_h(M_f + \Lambda(\epsilon))}{K^{1/2}} + \frac{\epsilon}{4},$$

where the third inequality follows from (29) with $(a, b, c) = \big(\frac{D_h^2}{2K}, M_f^2 + \Lambda(\epsilon)^2, \frac{1}{4(L_f+L(\epsilon))}\big)$. Then, by this, one can see that $\mathbb{E}[F(z^K) - F^*] \leq \epsilon/4 + \epsilon/2 + \epsilon/4 = \epsilon$ holds for all K satisfying (17). Hence, statement (i) of Theorem 1 holds.

We next prove statement (ii) of Theorem 1. Recall from the definition of Δ_k in (37) and Assumption 1(c) that $\mathbb{E}_{\xi_k}[\Delta_k] = 0$. In addition, by $x^k, x^* \in \text{dom } h$, (11), and (37), one has

$$|\Delta_k| \overset{(37)}{=} \big|(G(x^k; \xi_k) - \mathbb{E}[G(x^k; \xi_k)])^T(x^k - x^*)\big| \leq D_h\|G(x^k; \xi_k) - \mathbb{E}[G(x^k; \xi_k)]\| \qquad \forall k \geq 0,$$

which along with Assumption 2 implies that

$$\mathbb{E}_{\xi_k}\big[\exp\{|\Delta_k/(\sigma D_h)|^\alpha\}\big] \leq \mathbb{E}_{\xi_k}\big[\exp\{\|G(x^k; \xi_k) - \mathbb{E}[G(x^k; \xi_k)]\|^\alpha/\sigma^\alpha\}\big] \leq \exp\{1\} \qquad \forall k \geq 0.$$

Hence, the assumptions of Lemma 4 hold with $\phi^k = \Delta_k$ and $\varsigma = \sigma D_h$. It then follows from Lemma 4 with $\Omega = \alpha(\ln(2/\delta)/(\alpha-1))^{(\alpha-1)/\alpha}$ that for any $\delta \in (0, 1)$ and $K \geq \max\{1, \mathbb{1}_{(1,2)}(\alpha)\ln(2/\delta)/(\alpha-1)\}$,

$$\mathbb{P}\left(\frac{1}{K}\sum_{k=0}^{K-1}\Delta_k > \frac{\alpha D_h\sigma}{K^{(\alpha-1)/\alpha}} \cdot \Big(\frac{\ln(2/\delta)}{\alpha-1}\Big)^{\frac{\alpha-1}{\alpha}}\right) \leq \frac{\delta}{2}. \quad (42)$$

In addition, it follows from the convexity of the exponential function and Assumption 2 that for all $K \geq 1$,

$$\mathbb{E}\left[\exp\left\{\frac{1}{K}\sum_{k=0}^{K-1}\frac{\delta_k^\alpha}{\sigma^\alpha}\right\}\right] \leq \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\exp\left\{\frac{\delta_k^\alpha}{\sigma^\alpha}\right\}\right] \leq \exp\{1\}.$$

Using this and Markov's inequality, we obtain that for all $\delta \in (0, 1)$ and $K \geq 1$,

$$\mathbb{P}\left(\frac{1}{K}\sum_{k=0}^{K-1}\delta_k^\alpha > \Big(1 + \ln\Big(\frac{2}{\delta}\Big)\Big)\sigma^\alpha\right) = \mathbb{P}\left(\exp\left\{\frac{1}{K}\sum_{k=0}^{K-1}\frac{\delta_k^\alpha}{\sigma^\alpha}\right\} > \exp\left\{1 + \ln\Big(\frac{2}{\delta}\Big)\right\}\right)$$

$$\leq \exp\left\{-1 - \ln\Big(\frac{2}{\delta}\Big)\right\} \cdot \mathbb{E}\left[\exp\left\{\frac{1}{K}\sum_{k=0}^{K-1}\frac{\delta_k^\alpha}{\sigma^\alpha}\right\}\right] \leq \exp\left\{-1 - \ln\Big(\frac{2}{\delta}\Big)\right\} \cdot \exp\{1\} = \frac{\delta}{2}. \quad (43)$$

In view of (42) and (43), we can see that for all $K \geq \max\{1, \mathbb{1}_{(1,2)}(\alpha) \ln(2/\delta)/(\alpha - 1)\}$,

$$\mathbb{P}\left(\left\{\frac{1}{K}\sum_{k=0}^{K-1}\Delta_k > \frac{\alpha D_h \sigma}{K^{(\alpha-1)/\alpha}} \cdot \left(\frac{\ln(2/\delta)}{\alpha-1}\right)^{\frac{\alpha-1}{\alpha}}\right\} \bigcup \left\{\frac{1}{K}\sum_{k=0}^{K-1}\delta_k^\alpha > \left(1 + \ln\left(\frac{2}{\delta}\right)\right)\sigma^\alpha\right\}\right) \leq \delta,$$

which implies that

$$\mathbb{P}\left(\left\{\frac{1}{K}\sum_{k=0}^{K-1}\Delta_k \leq \frac{\alpha D_h \sigma}{K^{(\alpha-1)/\alpha}} \cdot \left(\frac{\ln(2/\delta)}{\alpha-1}\right)^{\frac{\alpha-1}{\alpha}}\right\} \bigcap \left\{\frac{1}{K}\sum_{k=0}^{K-1}\delta_k^\alpha \leq \left(1 + \ln\left(\frac{2}{\delta}\right)\right)\sigma^\alpha\right\}\right) \geq 1 - \delta. \quad (44)$$

On the other hand, by (12) and (28) with $c = (8(\alpha - 1))^{\alpha-1}D_h^{2-\alpha}(1 + \ln(2/\delta))/\alpha^\alpha$, $\hat{\eta} = \tilde{\eta}$, and $\varepsilon = \epsilon$, one has that

$$\frac{(8(\alpha-1))^{\alpha-1}D_h^{2-\alpha}(1+\ln(2/\delta))}{\alpha^\alpha} \cdot \sigma^\alpha\tilde{\eta}^{\alpha-1} \overset{(28)}{\leq} 8(\alpha-1)^2\left(\frac{\sigma}{\alpha}\right)^{\frac{\alpha}{\alpha-1}}\left(\frac{8D_h}{\epsilon}\right)^{\frac{2-\alpha}{\alpha-1}}\left(1+\ln\left(\frac{2}{\delta}\right)\right)^{\frac{1}{\alpha-1}}\tilde{\eta} + \frac{\epsilon}{8}$$

$$\overset{(12)}{=} \widetilde{\Lambda}(\delta,\epsilon)^2\tilde{\eta} + \frac{\epsilon}{8}. \quad (45)$$

In view of (11), (16), (29), (44), (45), and (40) with $\eta_k \equiv \tilde{\eta}$ for all k, we have that for all $K \geq \max\{1, \mathbb{1}_{(1,2)}(\alpha) \ln(2/\delta)/(\alpha - 1)\}$, it holds that with probability at least $1 - \delta$,

$$F(z^K) - F^* \overset{(40)}{\leq} \frac{\|x^0 - x^*\|^2}{2K\tilde{\eta}} + \frac{(8(\alpha-1))^{\alpha-1}D_h^{2-\alpha}}{K\alpha^\alpha} \cdot \tilde{\eta}^{\alpha-1}\sum_{k=0}^{K-1}\delta_k^\alpha + \frac{1}{K}\sum_{k=0}^{K-1}\Delta_k + M_f^2\tilde{\eta} + \frac{\epsilon}{8}$$

$$\overset{(11)(44)}{\leq} \frac{D_h^2}{2K\tilde{\eta}} + \frac{(8(\alpha-1))^{\alpha-1}D_h^{2-\alpha}(1+\ln(2/\delta))}{\alpha^\alpha} \cdot \sigma^\alpha\tilde{\eta}^{\alpha-1} + \frac{\alpha D_h \sigma}{K^{(\alpha-1)/\alpha}}\left(\frac{\ln(2/\delta)}{\alpha-1}\right)^{\frac{\alpha-1}{\alpha}} + M_f^2\tilde{\eta} + \frac{\epsilon}{8}$$

$$\overset{(45)}{\leq} \frac{D_h^2}{2K\tilde{\eta}} + (M_f^2 + \widetilde{\Lambda}(\delta,\epsilon)^2)\tilde{\eta} + \frac{\alpha D_h \sigma}{K^{(\alpha-1)/\alpha}}\left(\frac{\ln(2/\delta)}{\alpha-1}\right)^{\frac{\alpha-1}{\alpha}} + \frac{\epsilon}{4}$$

$$\overset{(16)}{=} \min_{\hat{\eta}}\left\{\frac{D_h^2}{2K\hat{\eta}} + (M_f^2 + \widetilde{\Lambda}(\delta,\epsilon)^2)\hat{\eta} : \hat{\eta} \in \left(0, \frac{1}{4(L_f + L(\epsilon))}\right]\right\} + \frac{\alpha D_h \sigma}{K^{(\alpha-1)/\alpha}}\left(\frac{\ln(2/\delta)}{\alpha-1}\right)^{\frac{\alpha-1}{\alpha}} + \frac{\epsilon}{4}$$

$$\overset{(29)}{\leq} \frac{2D_h^2(L_f + L(\epsilon))}{K} + \sqrt{2}D_h\left(\frac{M_f^2 + \widetilde{\Lambda}(\delta,\epsilon)^2}{K}\right)^{1/2} + \frac{\alpha D_h \sigma}{K^{(\alpha-1)/\alpha}}\left(\frac{\ln(2/\delta)}{\alpha-1}\right)^{\frac{\alpha-1}{\alpha}} + \frac{\epsilon}{4}$$

$$\leq \frac{2D_h^2(L_f + L(\epsilon))}{K} + \frac{\sqrt{2}D_h(M_f + \widetilde{\Lambda}(\delta,\epsilon))}{K^{1/2}} + \frac{\alpha D_h \sigma}{K^{(\alpha-1)/\alpha}}\left(\frac{\ln(2/\delta)}{\alpha-1}\right)^{\frac{\alpha-1}{\alpha}} + \frac{\epsilon}{4},$$

where the fourth inequality is due to (29) with $(a, b, c) = \left(\frac{D_h^2}{2K}, M_f^2 + \widetilde{\Lambda}(\delta,\epsilon)^2, \frac{1}{4(L_f+L(\epsilon))}\right)$. It then follows that $F(z^K) - F^* \leq \epsilon/4 + \epsilon/4 + \epsilon/4 + \epsilon/4 = \epsilon$ holds with probability at least $1 - \delta$ for all K satisfying (18). Hence, statement (ii) of Theorem 1 holds. $\qquad\square$

5.2 Proof of the main results in Section 3

In this subsection, we first establish a lemma and then use it to prove Theorem 2.

Lemma 6. *Suppose that Assumption 1 holds. Let $\epsilon \in (0, 1)$ be arbitrarily chosen, L_f, M_f, α, and σ be given in Assumption 1, and $L(\cdot)$ be defined in (9). Let $\{(x^k, y^k, z^k)\}$ be the sequence generated by Algorithm 2 with input parameters $\{(\eta_k, \gamma_k)\}$ satisfying $4(L_f + L(\epsilon\gamma_k))\eta_k\gamma_k \leq 1$ and $\eta_{k+1}(\gamma_{k+1}^{-1} - 1) \leq \eta_k\gamma_k^{-1}$ for all $k \geq 0$. Then, it holds that for all $k \geq 0$,*

$$\eta_{k+1}(\gamma_{k+1}^{-1} - 1)(F(z^{k+1}) - F^*) \leq \eta_k(\gamma_k^{-1} - 1)(F(z^k) - F^*) + (\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2)/2$$

15

$$+ M_f^2\eta_k^2 + \eta_k\Delta_k + \frac{(8(\alpha-1))^{\alpha-1}D_h^{2-\alpha}\delta_k^\alpha\eta_k^\alpha}{\alpha^\alpha} + \frac{\epsilon\eta_k}{8}, \tag{46}$$

where

$$\Delta_k = (\mathbb{E}[G(y^k;\xi_k)] - G(y^k;\xi_k))^T(x^k - x^*), \quad \delta_k = \|\mathbb{E}[G(y^k;\xi_k)] - G(y^k;\xi_k)\| \quad \forall k \geq 0. \tag{47}$$

Proof. Fix any $k \geq 0$. Using (20) and similar arguments as for deriving (38) with x^k replaced by y^k, we can deduce that

$$h(x^{k+1}) \leq h(x^*) + G(y^k;\xi_k)^T(x^* - x^{k+1}) + (2\eta_k)^{-1}(\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2 - \|x^k - x^{k+1}\|^2).$$

By this, (21), and the convexity of h, one has

$$\eta_k\gamma_k^{-1}h(z^{k+1}) \leq \eta_k(\gamma_k^{-1} - 1)h(z^k) + \eta_k h(x^{k+1})$$
$$\leq \eta_k(\gamma_k^{-1} - 1)h(z^k) + \eta_k h(x^*) + (\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2 - \|x^k - x^{k+1}\|^2)/2$$
$$+ \eta_k G(y^k;\xi_k)^T(x^* - x^{k+1}). \tag{48}$$

In addition, notice from (19) and (21) that $z^{k+1} - y^k = \gamma_k(x^{k+1} - x^k)$. Denote $f'(y^k) = \mathbb{E}_{\xi_k}[G(y^k;\xi_k)]$. It follows from (7) that $f'(y^k) \in \partial f(y^k)$. By these and (10) with $(y,x) = (z^{k+1}, y^k)$, one has that

$$\eta_k\gamma_k^{-1}f(z^{k+1}) \leq \eta_k\gamma_k^{-1}\left(f(y^k) + f'(y^k)^T(z^{k+1} - y^k) + \frac{L_f + L(\epsilon\gamma_k)}{2}\|z^{k+1} - y^k\|^2 + \frac{\epsilon\gamma_k}{8} + M_f\|z^{k+1} - y^k\|\right)$$
$$= \eta_k\gamma_k^{-1}(f(y^k) + f'(y^k)^T(z^{k+1} - y^k)) + \frac{(L_f + L(\epsilon\gamma_k))\eta_k\gamma_k}{2}\|x^{k+1} - x^k\|^2 + \frac{\epsilon\eta_k}{8} + M_f\eta_k\|x^{k+1} - x^k\|. \tag{49}$$

Also, it follows from (21) and the convexity of f that

$$\eta_k\gamma_k^{-1}(f(y^k) + f'(y^k)^T(z^{k+1} - y^k)) \stackrel{(21)}{=} \eta_k(\gamma_k^{-1} - 1)(f(y^k) + f'(y^k)^T(z^k - y^k))$$
$$+ \eta_k(f(y^k) + f'(y^k)^T(x^{k+1} - y^k))$$
$$\leq \eta_k(\gamma_k^{-1} - 1)f(z^k) + \eta_k(f(y^k) + f'(y^k)^T(x^{k+1} - y^k))$$
$$= \eta_k(\gamma_k^{-1} - 1)f(z^k) + \eta_k(f(y^k) + f'(y^k)^T(x^* - y^k))$$
$$+ \eta_k f'(y^k)^T(x^{k+1} - x^*)$$
$$\leq \eta_k(\gamma_k^{-1} - 1)f(z^k) + \eta_k f(x^*) + \eta_k f'(y^k)^T(x^{k+1} - x^*),$$

where the first and second inequalities are due to the convexity of f. This along with (49) implies that

$$\eta_k\gamma_k^{-1}f(z^{k+1}) \stackrel{(49)}{\leq} \eta_k(\gamma_k^{-1} - 1)f(z^k) + \eta_k f(x^*) + \eta_k f'(y^k)^T(x^{k+1} - x^*)$$
$$+ \frac{(L_f + L(\epsilon\gamma_k))\eta_k\gamma_k}{2}\|x^{k+1} - x^k\|^2 + \frac{\epsilon\eta_k}{8} + M_f\eta_k\|x^{k+1} - x^k\|.$$

By this, (21) and (48), one has that

$$\eta_k\gamma_k^{-1}F(z^{k+1}) \leq \eta_k(\gamma_k^{-1} - 1)F(z^k) + \eta_k F(x^*) + (\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2)/2 + M_f\eta_k\|x^{k+1} - x^k\|$$
$$+ \left(\frac{(L_f + L(\epsilon\gamma_k))\eta_k\gamma_k}{2} - \frac{1}{2}\right)\|x^{k+1} - x^k\|^2 + \eta_k(f'(y^k) - G(y^k;\xi_k))^T(x^{k+1} - x^*) + \frac{\epsilon\eta_k}{8}$$
$$\leq \eta_k(\gamma_k^{-1} - 1)F(z^k) + \eta_k F(x^*) + (\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2)/2 + \eta_k\Delta_k + M_f^2\eta_k^2$$

16

$$+\left(\frac{(L_f + L(\epsilon\gamma_k))\eta_k\gamma_k}{2} - \frac{1}{4}\right)\|x^{k+1} - x^k\|^2 + \eta_k(f'(y^k) - G(y^k;\xi_k))^T(x^{k+1} - x^k) + \frac{\epsilon\eta_k}{8}, \tag{50}$$

where the last inequality is due to (47) and $M_f\eta_k\|x^{k+1} - x^k\| \le \|x^{k+1} - x^k\|^2/4 + M_f^2\eta_k^2$. In addition, let $\alpha' = \alpha/(\alpha - 1)$. Observe that $\alpha' \ge 2$ due to $\alpha \in (1, 2]$. This together with (11) and $x^{k+1}, x^k \in \text{dom}\, h$ implies that $\|x^{k+1} - x^k\|^{\alpha'} \le D_h^{\alpha'-2}\|x^{k+1} - x^k\|^2$. Using this, (47), and the Young's inequality, we obtain that

$$(f'(y^k) - G(y^k;\xi_k))^T(x^{k+1} - x^k) \le \frac{\left(\left(\frac{\alpha'}{8D_h^{\alpha'-2}\eta_k}\right)^{1/\alpha'}\|x^{k+1} - x^k\|\right)^{\alpha'}}{\alpha'} + \frac{\left(\left(\frac{8D_h^{\alpha'-2}\eta_k}{\alpha'}\right)^{1/\alpha'}\delta_k\right)^\alpha}{\alpha}$$

$$= \frac{\|x^{k+1} - x^k\|^{\alpha'}}{8D_h^{\alpha'-2}\eta_k} + \frac{(8(\alpha-1))^{\alpha-1}D_h^{2-\alpha}\delta_k^\alpha\eta_k^{\alpha-1}}{\alpha^\alpha} \le \frac{\|x^{k+1} - x^k\|^2}{8\eta_k} + \frac{(8(\alpha-1))^{\alpha-1}D_h^{2-\alpha}\delta_k^\alpha\eta_k^{\alpha-1}}{\alpha^\alpha}.$$

This along with (50) and $4(L_f + L(\epsilon\gamma_k))\eta_k\gamma_k \le 1$ implies that

$$\eta_k\gamma_k^{-1}(F(z^{k+1}) - F^*) \le \eta_k(\gamma_k^{-1} - 1)(F(z^k) - F^*) + (\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2)/2 + \eta_k\Delta_k + M_f^2\eta_k^2$$

$$+ \left(\frac{(L_f + L(\epsilon\gamma_k))\eta_k\gamma_k}{2} - \frac{1}{8}\right)\|x^{k+1} - x^k\|^2 + \frac{(8(\alpha-1))^{\alpha-1}D_h^{2-\alpha}\delta_k^\alpha\eta_k^\alpha}{\alpha^\alpha} + \frac{\epsilon\eta_k}{8}$$

$$\le \eta_k(\gamma_k^{-1} - 1)(F(z^k) - F^*) + (\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2)/2 + \eta_k\Delta_k + M_f^2\eta_k^2$$

$$+ \frac{(8(\alpha-1))^{\alpha-1}D_h^{2-\alpha}\delta_k^\alpha\eta_k^\alpha}{\alpha^\alpha} + \frac{\epsilon\eta_k}{8}.$$

This inequality together with $\eta_{k+1}(\gamma_{k+1}^{-1} - 1) \le \eta_k\gamma_k^{-1}$ implies that the conclusion (46) holds. $\qquad\square$

We are now ready to provide a proof of Theorem 2.

Proof of Theorem 2. Summing up (46) over $k = 0, \ldots, K - 1$, and rearranging terms, we obtain that

$$\eta_K(\gamma_K^{-1} - 1)(F(z^K) - F^*) \le \eta_0(\gamma_0^{-1} - 1)(F(z^0) - F^*) + \frac{\|x^0 - x^*\|^2}{2} + \sum_{k=0}^{K-1}\eta_k\Delta_k + M_f^2\sum_{k=0}^{K-1}\eta_k^2$$

$$+ \frac{(8(\alpha-1))^{\alpha-1}D_h^{2-\alpha}}{\alpha^\alpha}\sum_{k=0}^{K-1}\delta_k^\alpha\eta_k^\alpha + \frac{\epsilon}{8}\sum_{k=0}^{K-1}\eta_k. \tag{51}$$

We now prove statement (i) of Theorem 2. Recall that $\gamma_k = 2/(k+2)$ and $\eta_k = (k+2)\eta/2$ for all $k \ge 0$. Using these, the expression of η in (22), and the fact that $L(\cdot)$ is nonincreasing, we obtain that $4(L_f + L(\epsilon\gamma_k))\eta_k\gamma_k = 4(L_f + L(\epsilon\gamma_k))\eta \le 4(L_f + L(\epsilon/K))\eta \le 1$ holds for all $0 \le k \le K$ and $K \ge 2$. Also, one verify that

$$\eta_{k+1}(\gamma_{k+1}^{-1} - 1) = \eta \cdot \frac{k+3}{2} \cdot \frac{k+1}{2} = \frac{\eta(k+3)(k+1)}{4} \le \frac{\eta(k+2)^2}{4} = \eta_k\gamma_k^{-1} \qquad \forall k \ge 0.$$

Hence, the assumptions of Lemma 6 hold for (η_k, γ_k) given in statement (i) of of Theorem 2. Also, by (12) and (28) with $c = (8(\alpha-1))^{\alpha-1}D_h^{2-\alpha}/\alpha^\alpha$, $\hat{\eta} = \eta_k$, and $\varepsilon = \epsilon$, one has

$$\frac{(8(\alpha-1))^{\alpha-1}D_h^{2-\alpha}}{\alpha^\alpha} \cdot \sigma^\alpha\eta_k^{\alpha-1} \stackrel{(28)}{\le} 8(\alpha-1)^2\left(\frac{\sigma}{\alpha}\right)^{\frac{\alpha}{\alpha-1}}\left(\frac{8D_h}{\epsilon}\right)^{\frac{2-\alpha}{\alpha-1}}\eta_k + \frac{\epsilon}{8} \stackrel{(12)}{=} \Lambda(\epsilon)^2\eta_k + \frac{\epsilon}{8} \qquad \forall k \ge 0. \tag{52}$$

Using this, $\gamma_0 = 1$, (11), (52), Assumption 1(c), and taking expectation on (51) with respect to $\{\xi_k\}_{k=0}^{K-1}$, we obtain that for all $K \geq 2$,

$$\eta_K(\gamma_K^{-1} - 1)\mathbb{E}[F(z^K) - F^*] \leq \eta_0(\gamma_0^{-1} - 1)(F(z^0) - F^*) + \frac{\|x^0 - x^*\|^2}{2} + M_f^2 \sum_{k=0}^{K-1} \eta_k^2$$

$$+ \frac{(8(\alpha - 1))^{\alpha-1} D_h^{2-\alpha}}{\alpha^\alpha} \cdot \sigma^\alpha \sum_{k=0}^{K-1} \eta_k^\alpha + \frac{\epsilon}{8} \sum_{k=0}^{K-1} \eta_k$$

$$\overset{(11)(52)}{\leq} \eta_0(\gamma_0^{-1} - 1)(F(z^0) - F^*) + \frac{D_h^2}{2} + (M_f^2 + \Lambda(\epsilon)^2) \sum_{k=0}^{K-1} \eta_k^2 + \frac{\epsilon}{4} \sum_{k=0}^{K-1} \eta_k$$

$$= \frac{D_h^2}{2} + (M_f^2 + \Lambda(\epsilon)^2) \sum_{k=0}^{K-1} \eta_k^2 + \frac{\epsilon}{4} \sum_{k=0}^{K-1} \eta_k,$$

where the last equality is due to $\gamma_0 = 1$. Further, using this, (22), (28), $\gamma_k = 2/(k + 2)$, $\eta_k = (k+2)\eta/2$, and rearranging the terms, we obtain that for all $K \geq 2$,

$$\mathbb{E}[F(z^K) - F^*] \leq \frac{2D_h^2}{(K+2)K\eta} + \frac{(M_f^2 + \Lambda(\epsilon)^2)\eta}{(K+2)K} \sum_{k=0}^{K-1} (k+2)^2 + \frac{\epsilon}{2(K+2)K} \sum_{k=0}^{K-1} (k+2)$$

$$= \frac{2D_h^2}{(K+2)K\eta} + \frac{(M_f^2 + \Lambda(\epsilon)^2)((K+1)(K+2)(2K+3)/6 - 1)\eta}{(K+2)K} + \frac{\epsilon((K+1)(K+2)/2 - 1)}{2(K+2)K}$$

$$\leq \frac{2D_h^2}{(K+2)K\eta} + \frac{(M_f^2 + \Lambda(\epsilon)^2)(2K+3)\eta}{3} + \frac{\epsilon}{2}$$

$$\overset{(22)}{=} \min_{\hat\eta} \left\{ \frac{2D_h^2}{(K+2)K\hat\eta} + \frac{(M_f^2 + \Lambda(\epsilon)^2)(2K+3)\hat\eta}{3} : \hat\eta \in \left(0, \frac{1}{4(L_f + L(\epsilon/K))}\right] \right\} + \frac{\epsilon}{2},$$

$$\overset{(28)}{\leq} \frac{8D_h^2(L_f + L(\epsilon/K))}{(K+2)K} + 4D_h \left(\frac{M_f^2 + \Lambda(\epsilon)^2}{3K}\right)^{1/2} + \frac{\epsilon}{2}$$

$$\leq \frac{8D_h^2 L_f}{K^2} + \frac{8D_h^2 L(\epsilon)}{K^{(1+3\nu)/(1+\nu)}} + \frac{4D_h(M_f + \Lambda(\epsilon))}{\sqrt{3}K^{1/2}} + \frac{\epsilon}{2},$$

where the third inequality follows from (28) with $(a, b, c) = \left(\frac{2D_h^2}{(K+2)K}, \frac{(M_f^2 + \Lambda(\epsilon)^2)(2K+3)}{3}, \frac{1}{4(L_f + L(\epsilon/K))}\right)$, and the last equality is due to (9) and $K \geq 2$. It then follows that $\mathbb{E}[F(z^K) - F^*] \leq \epsilon/6 + \epsilon/6 + \epsilon/6 + \epsilon/2 = \epsilon$ holds for all K satisfying (24). Hence, statement (i) of Theorem 2 holds.

We next prove statement (ii) of Theorem 2. Similar to the proof of statement (i), one can show that the assumptions of Lemma 6 hold for (η_k, γ_k) defined in statement (ii) of Theorem 2. Recall from the expression of Δ_k in (47) and Assumption 1(c) that $\mathbb{E}_{\xi_k}[\Delta_k] = 0$. In addition, by $x^k, x^* \in \text{dom } h$, (11), (47), and $\eta_k = (k+2)\tilde\eta/2$, one has that

$$|\eta_k \Delta_k| \overset{(47)}{=} |\eta_k(G(y^k; \xi_k) - \mathbb{E}[G(y^k; \xi_k)])^T(x^k - x^*)| \leq \eta_K D_h \|G(y^k; \xi_k) - \mathbb{E}[G(y^k; \xi_k)]\| \qquad \forall 0 \leq k \leq K,$$

which along with Assumption 2 implies that

$$\mathbb{E}_{\xi_k}\left[\exp\{|\eta_k \Delta_k/(\sigma\eta_K D_h)|^\alpha\}\right] \leq \mathbb{E}_{\xi_k}\left[\exp\{\|G(y^k; \xi_k) - \mathbb{E}[G(y^k; \xi_k)]\|^\alpha/\sigma^\alpha\}\right] \leq \exp\{1\} \qquad \forall 0 \leq k \leq K.$$

Hence, the assumptions of Lemma 4 hold with $\phi^k = \eta_k \Delta_k$ and $\varsigma = \sigma\eta_K D_h$. In addition, it follows from the convexity of the exponential function and Assumption 2 that for all $K \geq 1$,

$$\mathbb{E}\left[\exp\left\{\frac{1}{K} \sum_{k=0}^{K-1} \frac{\delta_k^\alpha}{\sigma^\alpha}\right\}\right] \leq \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}\left[\exp\left\{\frac{\delta_k^\alpha}{\sigma^\alpha}\right\}\right] \leq \exp\{1\}.$$

18

Using these and similar arguments as for proving (44), we obtain that for all $K \geq \max\{2, \mathbb{1}_{(1,2)}(\alpha)\ln(2/\delta)/(\alpha-1)\}$, it holds that

$$\mathbb{P}\left(\frac{1}{K}\sum_{k=0}^{K-1}\eta_k\Delta_k \leq \frac{\alpha D_h\sigma}{K^{(\alpha-1)/\alpha}}\cdot\left(\frac{\ln(2/\delta)}{\alpha-1}\right)^{\frac{\alpha-1}{\alpha}}\cdot\eta_K\right\}\bigcap\left\{\frac{1}{K}\sum_{k=0}^{K-1}\delta_k^\alpha \leq \left(1+\ln\left(\frac{2}{\delta}\right)\right)\sigma^\alpha\right\}\right) \geq 1-\delta. \quad (53)$$

On the other hand, by (12) and (28) with $c = (8(\alpha-1))^{\alpha-1}D_h^{2-\alpha}(1+\ln(2/\delta))/\alpha^\alpha$, $\hat\eta = \eta_K$, and $\varepsilon = \epsilon$, one has that

$$\frac{(8(\alpha-1))^{\alpha-1}D_h^{2-\alpha}(1+\ln(2/\delta))}{\alpha^\alpha}\cdot\sigma^\alpha\eta_K^{\alpha-1} \overset{(28)}{\leq} 8(\alpha-1)^2\left(\frac{\sigma}{\alpha}\right)^{\frac{\alpha}{\alpha-1}}\left(\frac{8D_h}{\epsilon}\right)^{\frac{2-\alpha}{\alpha-1}}\left(1+\ln\left(\frac{2}{\delta}\right)\right)^{\frac{1}{\alpha-1}}\eta_K + \frac{\epsilon}{8}$$

$$\overset{(12)}{=} \widetilde\Lambda(\delta,\epsilon)^2\eta_K + \frac{\epsilon}{8}. \quad (54)$$

Using these, (51), $\gamma_0 = 1$, $\gamma_k = 2/(k+2)$, and $\eta_k = (k+2)\tilde\eta/2$, we obtain that for all $K \geq \max\{2, \mathbb{1}_{(1,2)}(\alpha)\ln(2/\delta)/(\alpha-1)\}$, it holds that with probability at least $1-\delta$,

$$\tilde\eta(K+2)K(F(z^K)-F^*)/4 = \eta_K(\gamma_K^{-1}-1)(F(z^K)-F^*)$$

$$\overset{(51)}{\leq} \eta_0(\gamma_0^{-1}-1)(F(z^0)-F^*) + \frac{D_h^2}{2} + \frac{(8(\alpha-1))^{\alpha-1}D_h^{2-\alpha}}{\alpha^\alpha}\eta_K^\alpha\sum_{k=0}^{K-1}\delta_k^\alpha + M_f^2\sum_{k=0}^{K-1}\eta_k^2 + \sum_{k=0}^{K-1}\eta_k\Delta_k + \frac{\epsilon}{8}\sum_{k=0}^{K-1}\eta_k$$

$$\overset{(53)}{\leq} \eta_0(\gamma_0^{-1}-1)(F(z^0)-F^*) + \frac{D_h^2}{2} + \frac{(8(\alpha-1))^{\alpha-1}D_h^{2-\alpha}(1+\ln(2/\delta))}{\alpha^\alpha}\sigma^\alpha\eta_K^\alpha K + M_f^2\sum_{k=0}^{K-1}\eta_k^2$$

$$+ \frac{\alpha\sigma D_h}{K^{(\alpha-1)/\alpha}}\cdot\left(\frac{\ln(2/\delta)}{\alpha-1}\right)^{\frac{\alpha-1}{\alpha}}\cdot K\eta_K + \frac{\epsilon}{8}\sum_{k=0}^{K-1}\eta_k$$

$$\overset{(54)}{\leq} \eta_0(\gamma_0^{-1}-1)(F(z^0)-F^*) + \frac{D_h^2}{2} + (M_f^2+\widetilde\Lambda(\delta,\epsilon)^2)K\eta_K^2 + \frac{\alpha\sigma D_h}{K^{(\alpha-1)/\alpha}}\cdot\left(\frac{\ln(2/\delta)}{\alpha-1}\right)^{\frac{\alpha-1}{\alpha}}\cdot K\eta_K$$

$$+ \frac{\epsilon}{8}\left(\sum_{k=0}^{K-1}\eta_k + K\eta_K\right)$$

$$\leq \frac{D_h^2}{2} + (M_f^2+\widetilde\Lambda(\delta,\epsilon)^2)\left(\frac{K+2}{2}\right)^2 K\tilde\eta^2 + \frac{\alpha\sigma D_h}{K^{(\alpha-1)/\alpha}}\cdot\left(\frac{\ln(2/\delta)}{\alpha-1}\right)^{\frac{\alpha-1}{\alpha}}\cdot\frac{(K+2)K}{2}\tilde\eta + \frac{\epsilon(K+2)K\tilde\eta}{8},$$

where the last inequality is due to $\gamma_0 = 1$ and $\eta_k = (k+2)\tilde\eta/2$ for all $k \geq 0$. Further, by (23), (28), and rearranging the terms in the above inequality, one has that for all $K \geq \max\{2, \mathbb{1}_{(1,2)}(\alpha)\ln(2/\delta)/(\alpha-1)\}$, it holds that with probability at least $1-\delta$,

$$F(z^K)-F^* \leq \frac{2D_h^2}{(K+2)K\tilde\eta} + (M_f^2+\widetilde\Lambda(\delta,\epsilon)^2)(K+2)\tilde\eta + \frac{2\alpha\sigma D_h}{K^{(\alpha-1)/\alpha}}\left(\frac{\ln(2/\delta)}{\alpha-1}\right)^{\frac{\alpha-1}{\alpha}} + \frac{\epsilon}{2}$$

$$\overset{(23)}{=} \min_{\hat\eta}\left\{\frac{2D_h^2}{(K+2)K\hat\eta} + (M_f^2+\widetilde\Lambda(\delta,\epsilon)^2)(K+2)\hat\eta : \hat\eta \in \left(0, \frac{1}{4(L_f+L(\epsilon/K))}\right]\right\}$$

$$+ \frac{2\alpha\sigma D_h}{K^{(\alpha-1)/\alpha}}\left(\frac{\ln(2/\delta)}{\alpha-1}\right)^{\frac{\alpha-1}{\alpha}} + \frac{\epsilon}{2},$$

$$\leq \frac{8D_h^2(L_f+L(\epsilon/K))}{(K+2)K} + 2\sqrt{2}D_h\left(\frac{M_f^2+\widetilde\Lambda(\delta,\epsilon)^2}{K}\right)^{1/2} + \frac{2\alpha\sigma D_h}{K^{(\alpha-1)/\alpha}}\left(\frac{\ln(2/\delta)}{\alpha-1}\right)^{\frac{\alpha-1}{\alpha}} + \frac{\epsilon}{2}$$

$$\leq \frac{8D_h^2 L_f}{K^2} + \frac{8D_h^2 L(\epsilon)}{K^{(1+3\nu)/(1+\nu)}} + \frac{2\sqrt{2}D_h(M_f+\widetilde\Lambda(\delta,\epsilon))}{K^{1/2}} + \frac{2\alpha\sigma D_h}{K^{(\alpha-1)/\alpha}}\left(\frac{\ln(2/\delta)}{\alpha-1}\right)^{\frac{\alpha-1}{\alpha}} + \frac{\epsilon}{2},$$

19

where the second inequality is due to (28) with $(a, b, c) = \left(\frac{2D_h^2}{(K+2)K}, (M_f^2 + \widetilde{\Lambda}(\delta, \epsilon)^2)(K+2), \frac{1}{4(L_f + L(\epsilon/K))}\right)$. It then follows that $F(z^K) - F^* \leq \epsilon/8 + \epsilon/8 + \epsilon/8 + \epsilon/8 + \epsilon/2 = \epsilon$ holds with probability at least $1 - \delta$ for all K satisfying (25). Hence, statement (ii) of Theorem 2 holds. $\qquad\square$

References

[1] A. Alacaoglu, Y. Malitsky, and S. J. Wright. Towards weaker variance assumptions for stochastic optimization. *arXiv preprint arXiv:2504.09951*, 2025.

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

[3] 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.

[4] D. Davis, D. Drusvyatskiy, L. Xiao, and J. Zhang. From low probability to high confidence in stochastic convex optimization. *Journal of Machine Learning Research*, 22(49):1–38, 2021.

[5] I. Fatkhullin, F. Hübler, and G. Lan. Can SGD handle heavy-tailed noise? *arXiv preprint arXiv:2508.04860*, 2025.

[6] D. J. Foster, A. Sekhari, O. Shamir, N. Srebro, K. Sridharan, and B. Woodworth. The complexity of making the gradient small in stochastic convex optimization. In *Conference on Learning Theory*, pages 1319–1345, 2019.

[7] S. Ghadimi, G. Lan, and H. Zhang. Generalized uniformly optimal methods for nonlinear programming. *Journal of Scientific Computing*, 79(3):1854–1881, 2019.

[8] E. Gorbunov, M. Danilova, and A. Gasnikov. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. In *Advances in Neural Information Processing Systems*, volume 33, pages 15042–15053, 2020.

[9] M. Grant, S. Boyd, and Y. Ye. CVX: Matlab software for disciplined convex programming, 2008.

[10] B. Grimmer. On optimal universal first-order methods for minimizing heterogeneous sums. *Optimization Letters*, 18(2):427–445, 2024.

[11] V. Guigues, J. Liang, and R. D. Monteiro. Universal subgradient and proximal bundle methods for convex and strongly convex hybrid composite optimization. *arXiv preprint arXiv:2407.10073*, 2024.

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

[13] C. He, Z. Lu, D. Sun, and Z. Deng. Complexity of normalized stochastic first-order methods with momentum under heavy-tailed noise. *arXiv preprint arXiv:2506.11214*, 2025.

[14] 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.

[15] G. Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1):365–397, 2012.

[16] G. Lan, A. Nemirovski, and A. Shapiro. Validation analysis of mirror descent stochastic approximation method. *Mathematical Programming*, 134(2):425–458, 2012.

[17] J. Liang, V. Guigues, and R. D. Monteiro. A single cut proximal bundle method for stochastic convex composite optimization. *Mathematical Programming*, 208(1):173–208, 2024.

[18] 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.

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

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

[21] E. Moulines and F. Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

[22] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[23] A. Nemirovski and D. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience Series in Discrete Mathematics. John Wiley & Sons, 1983.

[24] Y. Nesterov. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 152(1-2):381–404, 2015.

[25] Y. Nesterov. Universal complexity bounds for universal gradient methods in nonlinear optimization. *arXiv preprint arXiv:2509.20902*, 2025.

[26] 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.

[27] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

[28] H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.

[29] 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.

[30] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Stochastic convex optimization. In *Conference on Learning Theory*, volume 2, page 5, 2009.

[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] P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. Manuscript, May 2008.

[35] M. Vladimirova, S. Girard, H. Nguyen, and J. Arbel. Sub-Weibull distributions: Generalizing sub-Gaussian and sub-exponential properties to heavier tailed distributions. *Stat*, 9(1):e318, 2020.

[36] 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.