# Stochastic first-order methods with multi-extrapolated momentum for highly smooth unconstrained optimization

Chuan He*

December 19, 2024

**Abstract**

In this paper we consider an unconstrained stochastic optimization problem where the objective function exhibits a high order of smoothness. In particular, we propose a stochastic first-order method (SFOM) with multi-extrapolated momentum, in which multiple extrapolations are performed in each iteration, followed by a momentum step based on these extrapolations. We show that our proposed SFOM with multi-extrapolated momentum can accelerate optimization by exploiting the high-order smoothness of the objective function $f$. Specifically, assuming that the gradient and the $p$th-order derivative of $f$ are Lipschitz continuous for some $p \geq 2$, and under some additional mild assumptions, we establish that our method achieves a sample complexity of $\widetilde{\mathcal{O}}(\epsilon^{-(3p+1)/p})$[1] for finding a point $x$ satisfying $\mathbb{E}[\|\nabla f(x)\|] \leq \epsilon$. To the best of our knowledge, our method is the first SFOM to leverage arbitrary order smoothness of the objective function for acceleration, resulting in a sample complexity that strictly improves upon the best-known results without assuming the average smoothness condition. Finally, preliminary numerical experiments validate the practical performance of our method and corroborate our theoretical findings.

**Keywords:** Unconstrained optimization, high-order smoothness, stochastic first-order method, extrapolation, momentum, sample 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 and has a Lipschitz continuous $p$th-order derivative for some $p \geq 2$ (see Assumption 2 for details). We assume that problem (1) has at least one optimal solution. Our goal is to develop first-order methods for solving (1) in the stochastic regime where the derivatives of $f$ are not directly accessible. Instead, our algorithm relies solely on stochastic estimators $G(\cdot;\xi)$ for the gradient $\nabla f(\cdot)$, where $\xi$ is a random variable with sample space $\Xi$ (see Assumption 1(c) for our assumptions on $G$).

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

[1] $\widetilde{\mathcal{O}}(\cdot)$ represents $\mathcal{O}(\cdot)$ with hidden logarithmic factors.

In recent years, there has been significant developments on stochastic first-order methods (SFOMs) with sample complexity guarantees for solving problem (1). Notably, when assuming the gradient $\nabla f$ is Lipschitz continuous (see Assumption 1(b)), SFOMs [3, 7, 8, 9] have been proposed with a sample complexity [2] of $\mathcal{O}(\epsilon^{-4})$ for finding a point $x$ satisfying

$$\mathbb{E}[\|\nabla f(x)\|] \leq \epsilon, \tag{2}$$

where $\epsilon \in (0,1)$ is a given tolerance parameter, and the expectation is taken over the randomness in the algorithm. This sample complexity has been proved to be optimal in [2]. Among these works, [3, 7] proposed SFOMs that incorporate Polyak momentum steps:

$$m^k = (1-\gamma_{k-1})m^{k-1} + \gamma_{k-1}G(x^k;\xi^k) \qquad \forall k \geq 0, \tag{3}$$

where $\{x^k\}$ are the algorithm iterates, $\{\gamma_k\}$ are the momentum parameters, and $\{m^k\}$ are the stochastic estimators of $\{\nabla f(x^k)\}$. It has been shown in [3, 7] that Polyak momentum promotes a variance reduction effect in gradient estimation, and it was further shown in [3] that Polyak momentum facilitates the convergence of SFOMs with normalized updates. Moreover, other benign theoretical properties of SFOMs with Polyak momentum have been studied in [11, 13, 16, 18].

Recently, many SFOMs [4, 5, 12, 14] have been proposed under the average smooth assumption on the gradient estimators, namely,

$$\mathbb{E}_\xi[\|G(y;\xi) - G(x;\xi)\|^2] \leq L^2\|y-x\|^2 \qquad \forall x, y \in \mathbb{R}^n \tag{4}$$

for some $L > 0$. The methods in [4, 5] achieve a sample complexity of $\mathcal{O}(\epsilon^{-3})$ for finding $x$ that satisfies (2), which has been proven to be optimal in [2]. In particular, [4] proposed an SFOM with the following recursive momentum steps:

$$m^k = (1-\gamma_{k-1})m^{k-1} + \gamma_{k-1}G(x^k;\xi^k) + (1-\gamma_{k-1})(G(x^k;\xi^k) - G(x^{k-1};\xi^k)) \qquad \forall k \geq 0, \tag{5}$$

which can be viewed as a modified variant of the Polyak momentum steps in (3), with an additional term $(1-\gamma_{k-1})(G(x^k;\xi^k) - G(x^{k-1};\xi^k))$. In addition, SFOMs [15, 17, 19, 20] were proposed for stochastic composite optimization problems, achieving a sample complexity of $\mathcal{O}(\epsilon^{-3})$ under the average smoothness assumption in (4). However, it shall be mentioned that the average smoothness condition in (4) implies the gradient Lipschitz condition (see Assumption 1(b)), but the reverse implication does not generally hold. The strong assumption of average smoothness in (4) appears to be crucial for achieving the sample complexity of $\mathcal{O}(\epsilon^{-3})$ for finding $x$ that satisfies (2).

Aside from assuming (4), several other attempts have been made to improve the sample complexity of SFOMs by leveraging the second-order smoothness of $f$. Assuming that the Hessian $\nabla^2 f$ is Lipschitz continuous, i.e., Assumption 2 with $p = 2$, SFOMs [1, 3, 6] have been proposed with a sample complexity of $\mathcal{O}(\epsilon^{-7/2})$ for finding $x$ satisfying (2). In particular, [3] proposed an SFOM with implicit gradient transport that performs extrapolation steps combined with Polyak momentum steps:

$$z^k = x^k + \frac{1-\gamma_{k-1}}{\gamma_{k-1}}(x^k - x^{k-1}), \quad m^k = (1-\gamma_{k-1})m^{k-1} + \gamma_{k-1}G(z^k;\xi^k) \qquad \forall k \geq 0. \tag{6}$$

It was shown that constructing $\{z^k\}$ through extrapolation and combining it with Polyak momentum achieves faster variance reduction for gradient estimators $\{m^k\}$, leading to an improved overall sample complexity. Furthermore, there appear to be no SFOMs that leverage higher-order smoothness beyond the Hessian Lipschitz condition.

[2] Sample complexity means the number of samples of $\xi$ drawn for constructing the stochastic estimator $G(\cdot,\xi)$.

In this paper we show that SFOMs can achieve acceleration by *exploiting the arbitrarily high-order smoothness of $f$* through the introduction of an SFOM with multi-extrapolated momentum (Algorithm 1). Our proposed SFOM can be viewed as a significant generalization of the SFOM proposed in [3], which uses extrapolated momentum steps described in (6), as the acceleration of our method also relies on extrapolation and momentum. Specifically, we demonstrate that for any $p \geq 2$, performing $p-1$ separate extrapolation steps in each iteration and combining them with a momentum step can accelerate variance reduction by exploiting the smoothness of the $p$th-order derivative of $f$, thereby leading to a sample complexity of $\mathcal{O}(\epsilon^{-(3p+1)/p})$, which strictly improves upon the best-known results of SFOMs without assuming average smoothness. In contrast to the straightforward parameter choices in previous SFOMs, the parameters of our proposed SFOM are determined through an innovative use of Lagrange interpolation (see Section 3). For ease of comparison, we summarize the sample complexity of several existing SFOMs, along with their associated smoothness assumptions, and those of our method in Table 1.

Table 1: Comparison of the sample complexity of several SFOMs and their associated smoothness assumptions in the literature with those of our method for finding a point $x$ that satisfies (2). Here, SG and PM stand for stochastic gradient and Polyak momentum, respectively.

| Method | Sample complexity | Smoothness assumption |
|---|---|---|
| SG [8] | $\mathcal{O}(\epsilon^{-4})$ | gradient Lipschitz |
| SG-PM [3, 7] | $\mathcal{O}(\epsilon^{-4})$ | gradient Lipschitz |
| Restarted SG [6] | $\mathcal{O}(\epsilon^{-7/2})$ | gradient & Hessian Lipschitz |
| NIGT [3] | $\mathcal{O}(\epsilon^{-7/2})$ | gradient & Hessian Lipschitz |
| **Algorithm 1 (ours)** | $\widetilde{\mathcal{O}}(\epsilon^{-(3p+1)/p})$ | gradient & $p$th-order derivative Lipschitz |
| STORM [4] | $\mathcal{O}(\epsilon^{-3})$ | average smoothness |
| SPIDER [5] | $\mathcal{O}(\epsilon^{-3})$ | average smoothness |

The main contributions of this paper are highlighted below.

- We propose an SFOM with multi-extrapolated momentum (Algorithm 1), which is the first SFOM to leverage the arbitrary order of smoothness of the objective function for acceleration. Our method is efficient to implement in practice, and its update schemes and parameter selection can be performed cheaply and neatly (see Section 3), offering insights for future algorithmic design.

- We show that, assuming the $p$th-order derivative of $f$ is Lipschitz continuous and under other mild assumptions, our proposed SFOM achieves a sample complexity of $\widetilde{\mathcal{O}}(\epsilon^{-(3p+1)/p})$. This sample complexity strictly improves upon the best-known results for SFOMs without assuming average smoothness and provides an affirmative answer to the open question raised at the end of [2] regarding SFOMs that leverage high-order smoothness for acceleration.

The rest of this paper is organized as follows. In Section 2, we introduce some notation, assumptions, and preliminaries that will be used in the paper. In Section 3, we propose an SFOM with multi-extrapolated momentum and study its sample complexity. Section 4 presents preliminary numerical results. In Section 5, we present the proofs of the main results.

## 2 Notation, assumptions, and preliminaries

Throughout this paper, let $\mathbb{R}^n$ denote the $n$-dimensional Euclidean space and $\langle \cdot, \cdot \rangle$ denote the standard inner product. We use $\|\cdot\|$ to denote the Euclidean norm of a vector or the spectral norm of a matrix.

For any $p \geq 1$ and a $p$th-order continuously differentiable function $\varphi$, we denote by $D^p\varphi(x)[h_1, \ldots, h_p]$ the $p$th-order directional derivative of $\varphi$ at $x$ along $h_i \in \mathbb{R}^n$, $1 \leq i \leq p$, and use $D^p\varphi(x)[\cdot]$ to denote the associated symmetric $p$-linear form. For any symmetric $p$-linear form $\mathcal{T}[\cdot]$, we denote its norm as

$$\|\mathcal{T}\|_{(p)} \doteq \max_{h_1,\ldots,h_p} \{\mathcal{T}[h_1, \ldots, h_p] : \|h_i\| \leq 1, 1 \leq i \leq p\}. \tag{7}$$

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

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

For any $x, h \in \mathbb{R}^n$, we denote $D^p\varphi(x)[h]^p \doteq D^p\varphi(x)[h, \ldots, h]$ and $\nabla^p\varphi(x)(h)^{p-1} \doteq \nabla^p\varphi(x)(h, \ldots, h)$. For any $s \in \mathbb{R}$, we let $\mathrm{sgn}(s)$ be 1 if $s \geq 0$ and let it be $-1$ otherwise. In addition, $\widetilde{\mathcal{O}}(\cdot)$ represents $\mathcal{O}(\cdot)$ with logarithmic terms omitted.

We now make the following assumptions throughout this paper.

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

(b) *There exists $L_1 > 0$ such that*

$$\|\nabla f(y) - \nabla f(x)\| \leq L_1\|y - x\| \qquad \forall x, y \in \mathbb{R}^n. \tag{8}$$

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

$$\mathbb{E}_\xi[G(x;\xi)] = \nabla f(x), \quad \mathbb{E}_\xi[\|G(x;\xi) - \nabla f(x)\|^2] \leq \sigma^2 \qquad \forall x \in \mathbb{R}^n \tag{9}$$

*for some $\sigma > 0$.*

We now make some remarks on Assumption 1.

**Remark 1.** (i) Assumptions 1(a) and (b) are standard. It follows from Assumption 1(b) that

$$f(y) \leq f(x) + \nabla f(x)^T(y - x) + \frac{L_1}{2}\|y - x\|^2 \qquad \forall x, y \in \mathbb{R}^n. \tag{10}$$

(ii) Assumption 1(c) is commonly used in stochastic optimization. It implies that $G(\cdot;\xi)$ is an unbiased estimator for $\nabla f(\cdot)$ with bounded variance.

We also make the following assumption regarding the Lipschitz continuity of $D^p f$.

**Assumption 2.** *The function $f$ is $p$th-order continuously differentiable, and there exists some $p \geq 2$ and $L_p > 0$ such that*

$$\|D^p f(y) - D^p f(x)\|_{(p)} \leq L_p\|y - x\| \qquad \forall x, y \in \mathbb{R}^n. \tag{11}$$

The following lemma provides a useful inequality under Assumption 2, and its proof is deferred to Section 5.1.

**Lemma 1.** *Under Assumption 2, the following inequality holds:*

$$\left\|\nabla f(y) - \sum_{t=1}^{p} \frac{1}{(t-1)!}\nabla^t f(x)(y - x)^{t-1}\right\| \leq \frac{L_p}{p!}\|y - x\|^p \qquad \forall x, y \in \mathbb{R}^n. \tag{12}$$

## 3 Stochastic first-order methods with multi-extrapolated momentum

In this section we propose an SFOM with multi-extrapolated momentum in Algorithm 1, and study its sample complexity.

Specifically, at the $k$th iteration, our method performs $q$ separate extrapolations for some $q \geq 1$ as in (15) to obtain $q$ points $\{z^{k,t}\}_{1\leq t\leq q}$, where the extrapolation parameters $\{\gamma_{k-1,t}\}_{1\leq t\leq q}$ in (15) are chosen to have distinct positive values. Then, a gradient estimator $m^k$ is constructed using the previous gradient estimator $m^{k-1}$ and the stochastic estimator $G(\cdot;\xi^k)$ evaluated at $\{z^{k,t}\}_{1\leq t\leq q}$, as described in (16). Here, the weighting parameters $\{\theta_{k-1,t}\}_{1\leq t\leq q}$ must be obtained by solving the linear system in (28) with the coefficient matrix constructed using $\{\gamma_{k-1,t}\}_{1\leq t\leq q}$ to exploit high-order smoothness. The resulting values of $\{\theta_{k-1,t}\}_{1\leq t\leq q}$ follow a pattern of alternating signs (see Lemma 5). After obtaining $m^k$, the next iterate $x^{k+1}$ is generated via a normalized update[3], as described in (17). For ease of understanding our extrapolation and momentum steps, we consider Algorithm 1 with $q=3$ and visualize the updates for $\{z^{k,t}\}_{1\leq t\leq 3}$ and $m^k$ on a two-dimensional contour plot shown in Figure 1.

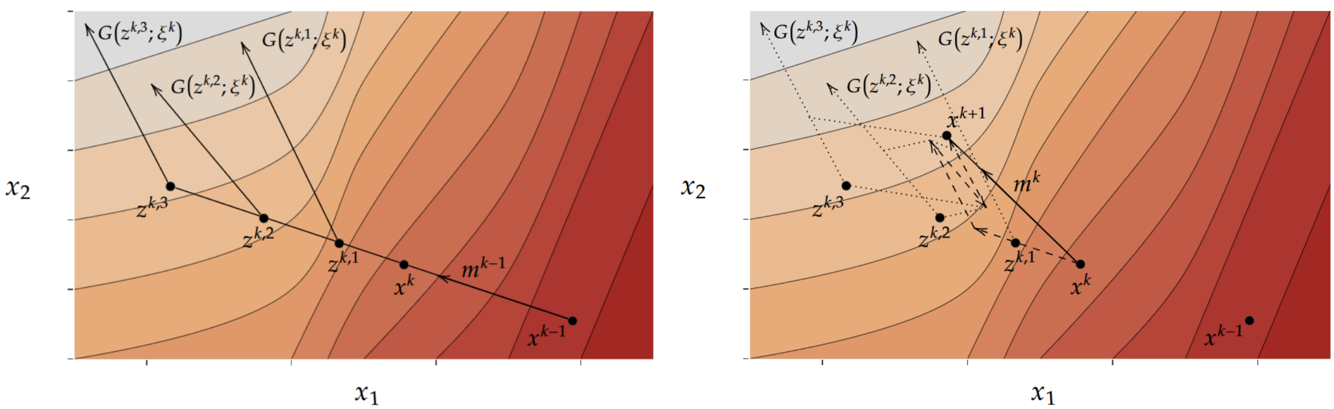


Figure 1: Visualization of the updates for $\{z^{k,t}\}_{1\leq t\leq 3}$ (left) and $m^k$ (right) on a contour plot.

Before proceeding, we present the following lemma regarding the descent of $f$ for iterates generated by Algorithm 1. Its proof is deferred to Section 5.1.

**Lemma 2.** *Suppose that Assumption 1 holds. Let $\{x^k\}_{k\geq 0}$ be generated by Algorithm 1. Then,*

$$f(x^{k+1}) \leq f(x^k) - \eta_k\|\nabla f(x^k)\| + 2\eta_k\|\nabla f(x^k) - m^k\| + \frac{L_1}{2}\eta_k^2 \qquad \forall k \geq 0, \tag{13}$$

*where $L_1$ is given in Assumption 1(b).*

### 3.1 An SFOM with double-extrapolated momentum

In this subsection we study a simple variant of Algorithm 1 with $q=2$, which is capable of exploiting the smoothness of $D^3 f$. We refer to this method as the SFOM with double-extrapolated momentum as two separate extrapolations are performed in each iteration. In the following, we establish its sample complexity under Assumption 1 and Assumption 2 with $p=3$.

Throughout this subsection, we impose the following equations on the parameters $\{(\gamma_{k,1},\gamma_{k,2},\theta_{k,1},\theta_{k,2})\}_{k\geq 0}$ of Algorithm 1:

$$\theta_{k,1}/\gamma_{k,1} + \theta_{k,2}/\gamma_{k,2} = 1, \quad \theta_{k,1}/\gamma_{k,1}^2 + \theta_{k,2}/\gamma_{k,2}^2 = 1 \qquad \forall k \geq 0. \tag{14}$$

**Algorithm 1** An SFOM with multi-extrapolated momentum

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

Initialize $m^{-1} = 0$ 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:

$$z^{k,t} = x^k + \frac{1-\gamma_{k-1,t}}{\gamma_{k-1,t}}(x^k - x^{k-1}) \qquad \forall 1 \leq t \leq q. \tag{15}$$

Compute the search direction:

$$m^k = \Big(1 - \sum_{t=1}^q \theta_{k-1,t}\Big) m^{k-1} + \sum_{t=1}^q \theta_{k-1,t} G(z^{k,t}; \xi^k). \tag{16}$$

Update the next iterate:

$$x^{k+1} = x^k - \eta_k \frac{m^k}{\|m^k\|}. \tag{17}$$

**end for**

It is noteworthy that for any two distinct positive values of $\gamma_{k,1}$ and $\gamma_{k,2}$, the values of $\theta_{k,1}$ and $\theta_{k,2}$ can be uniquely determined by solving the above equations. In addition, we require that

$$\theta_{k,1} + \theta_{k,2} \in (0,1) \qquad \forall k \geq 0. \tag{18}$$

The following lemma establishes the recurrence relation for the estimation error of the gradient estimators $\{m^k\}_{k\geq 0}$ in Algorithm 1 with $q = 2$. Its proof is deferred to Section 5.2.

**Lemma 3.** *Suppose that Assumption 1 holds, and Assumption 2 holds with $p = 3$. Let $\{(x^k, m^k)\}_{k\geq 0}$ be generated by Algorithm 1 with $q = 2$, and let $\{(\eta_k, \gamma_{k,1}, \gamma_{k,2}, \theta_{k,1}, \theta_{k,2})\}_{k\geq 0}$ be inputs of Algorithm 1. Assume that $\{(\gamma_{k,1}, \gamma_{k,2}, \theta_{k,1}, \theta_{k,2})\}_{k\geq 0}$ satisfies (14) and (18). Then,*

$$\begin{aligned}\mathbb{E}_{\xi^{k+1}}[\|m^{k+1} - \nabla f(x^{k+1})\|^2] &\leq (1 - \theta_{k,1} - \theta_{k,2})\|m^k - \nabla f(x^k)\|^2 \\ &+ \frac{L_3^2 \eta_k^6 \theta_{k,1}^2}{12\gamma_{k,1}^6(\theta_{k,1} + \theta_{k,2})} + \frac{L_3^2 \eta_k^6 \theta_{k,2}^2}{12\gamma_{k,2}^6(\theta_{k,1} + \theta_{k,2})} + \frac{L_3^2 \eta_k^6}{12(\theta_{k,1} + \theta_{k,2})} + 2(\theta_{k,1}^2 + \theta_{k,2}^2)\sigma^2 \qquad \forall k \geq 0,\end{aligned} \tag{19}$$

*where $\sigma$ and $L_3$ are given in Assumptions 1(b) and 2, respectively.*

We next derive an upper bound for the average expected error of the stationary condition across all iterates generated by Algorithm 1 with $q = 2$. Its proof is relegated to Section 5.2.

**Theorem 1.** *Suppose that Assumption 1 holds, and Assumption 2 holds with $p = 3$. Let $\{x^k\}_{k\geq 0}$ be generated by Algorithm 1 with $q = 2$, and let $\{(\eta_k, \gamma_{k,1}, \gamma_{k,2}, \theta_{k,1}, \theta_{k,2})\}_{k\geq 0}$ be inputs of Algorithm 1. Assume that $\{(\gamma_{k,1}, \gamma_{k,2}, \theta_{k,1}, \theta_{k,2})\}_{k\geq 0}$ satisfies (14) and (18), and that the sequence $\{p_k\}_{k\geq 0}$ satisfies*

$$(1 - \theta_{k,1} - \theta_{k,2})p_{k+1} \leq (1 - (\theta_{k,1} + \theta_{k,2})/2)p_k \qquad \forall k \geq 0. \tag{20}$$

[3]Normalized updates are recognized as an important technique in training deep neural networks (see, e.g., [21]).

*Then, for any $K \geq 1$,*

$$\sum_{k=0}^{K-1} \mathbb{E}[\|\nabla f(x^k)\|] \leq \frac{f(x^0) - f_{\text{low}} + p_0\sigma^2}{\eta_{K-1}} + \frac{1}{\eta_{K-1}} \sum_{k=0}^{K-1} \bigg( \frac{L_1}{2}\eta_k^2 + \frac{2\eta_k^2}{(\theta_{k,1}+\theta_{k,2})p_k} + \frac{L_3^2\eta_k^6\theta_{k,1}^2 p_{k+1}}{12\gamma_{k,1}^6(\theta_{k,1}+\theta_{k,2})} + \frac{L_3^2\eta_k^6\theta_{k,2}^2 p_{k+1}}{12\gamma_{k,2}^6(\theta_{k,1}+\theta_{k,2})} + \frac{L_3^2\eta_k^6 p_{k+1}}{12(\theta_{k,1}+\theta_{k,2})} + 2(\theta_{k,1}^2+\theta_{k,2}^2)p_{k+1}\sigma^2 \bigg), \tag{21}$$

*where $f_{\text{low}}$, $L_1$, and $\sigma$ are given in Assumption 1, $L_3$ is given in Assumption 2, and the expectation is taken over the randomness in the algorithm.*

### 3.1.1 Input parameters and convergence rate

To analyze the sample complexity of Algorithm 1 with $q = 2$, we specify $\{(\eta_k, \gamma_{k,1}, \gamma_{k,2})\}_{k\geq 0}$ as

$$\eta_k = \frac{1}{(k+3)^{7/10}}, \quad \gamma_{k,1} = \frac{1}{(k+3)^{3/5}}, \quad \gamma_{k,2} = \frac{1}{2(k+3)^{3/5}} \qquad \forall k \geq 0, \tag{22}$$

and determine $\{(\theta_{k,1}, \theta_{k,2})\}_{k\geq 0}$, by solving (14), as

$$\theta_{k,1} = \frac{2(k+3)^{3/5} - 1}{(k+3)^{6/5}}, \quad \theta_{k,2} = \frac{1-(k+3)^{3/5}}{2(k+3)^{6/5}} \qquad \forall k \geq 0. \tag{23}$$

In addition, we define the sequence $\{p_k\}_{k\geq 0}$ used in Theorem 1 as follows:

$$p_k = (k+3)^{1/5} \qquad \forall k \geq 0. \tag{24}$$

The following lemma provides some useful properties of $\{(\theta_{k,1}, \theta_{k,2})\}_{k\geq 0}$ and $\{p_k\}_{k\geq 0}$ defined in (23) and (24), respectively. Its proof is relegated to Section 5.2.

**Lemma 4.** *Let $\{(\theta_{k,1}, \theta_{k,2})\}_{k\geq 0}$ and $\{p_k\}_{k\geq 0}$ be defined in (23) and (24), respectively. Then,*

$$\theta_{k,1} + \theta_{k,2} \in \left( \frac{1}{(k+3)^{3/5}}, \frac{3}{2(k+3)^{3/5}} \right) \subset (0,1), \quad \theta_{k,1}^2 \leq \frac{4}{(k+3)^{6/5}}, \quad \theta_{k,2}^2 \leq \frac{1}{4(k+3)^{6/5}} \qquad \forall k \geq 0. \tag{25}$$

*Moreover, $\{(\theta_{k,1}, \theta_{k,2}, p_k)\}_{k\geq 0}$ satisfies (20).*

The next theorem establishes the sample complexity of Algorithm 1 with $q = 2$ and other inputs specified in (22) and (23). Its proof is deferred to Section 5.2.

**Theorem 2.** *Suppose that Assumption 1 holds, and Assumption 2 holds with $p = 3$. Let $\{x^k\}_{k\geq 0}$ be generated by Algorithm 1 with $q = 2$ and inputs $\{(\eta_k, \gamma_{k,1}, \gamma_{k,2}, \theta_{k,1}, \theta_{k,2})\}_{k\geq 0}$ specified as in (22) and (23). Define*

$$M_3 \doteq 4\big(f(x^0) - f_{\text{low}} + 19\sigma^2 + L_1 + 4L_3^2 + 2\big). \tag{26}$$

*Let $\kappa(k)$ be uniformly drawn from $\{0, \ldots, k-1\}$. Then,*

$$\mathbb{E}[\|\nabla f(x^{\kappa(k)})\|] \leq \epsilon \qquad \forall k \geq \max\left\{ \left( \frac{20M_3}{3\epsilon} \ln\left( \frac{20M_3}{3\epsilon} \right) \right)^{10/3}, 2 \right\}, \tag{27}$$

*where $\epsilon \in (0,1)$, and the expectation is taken over the randomness in the algorithm.*

### 3.2 An SFOM with multi-extrapolated momentum

In this subsection, we study Algorithm 1 with $q = p - 1$, which is capable of exploiting the smoothness of $D^p f$ for some $p \geq 2$. In the following, we establish its sample complexity under Assumption 1 and Assumption 2 for $p \geq 2$.

Throughout this subsection, we impose the following system of linear equations on the parameters $\{(\gamma_{k,t}, \theta_{k,t})\}_{1\leq t\leq q, k\geq 0}$ of Algorithm 1:

$$\begin{bmatrix} 1/\gamma_{k,1} & 1/\gamma_{k,2} & \cdots & 1/\gamma_{k,q} \\ 1/\gamma_{k,1}^2 & 1/\gamma_{k,2}^2 & \cdots & 1/\gamma_{k,q}^2 \\ \vdots & \vdots & \ddots & \vdots \\ 1/\gamma_{k,1}^q & 1/\gamma_{k,2}^q & \cdots & 1/\gamma_{k,q}^q \end{bmatrix} \begin{bmatrix} \theta_{k,1} \\ \theta_{k,2} \\ \vdots \\ \theta_{k,q} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \qquad \forall k \geq 0, \tag{28}$$

and in addition, we require that

$$\sum_{t=1}^{q} \theta_{k,t} \in (0,1) \qquad \forall k \geq 0. \tag{29}$$

The coefficient matrix in (28) is known as the Vandermonde matrix (e.g., see [10]). The following lemma demonstrates that if the values of $\{\gamma_{k,t}\}_{1\leq t\leq q}$ are positive and distinct, the values of $\{\theta_{k,t}\}_{1\leq t\leq q}$ can be uniquely determined by solving (28). In addition, the next lemma provides the explicit solution to the linear system (28), along with the elegant property of alternating signs for $\{\theta_{k,t}\}_{1\leq t\leq q}$. Its proof is deferred to Section 5.3.

**Lemma 5.** *Assume that $\{\gamma_{k,t}\}_{1\leq t\leq q} \subset (0,1)$ with $\gamma_{k,1} > \cdots > \gamma_{k,q}$ are given for some $k \geq 0$. Then, the solution $\{\theta_{k,t}\}_{1\leq t\leq q}$ to the linear system in (28) is unique and can be explicitly written as*

$$\theta_{k,t} = \frac{\prod_{1\leq s\leq q, s\neq t}(1 - 1/\gamma_{k,s})}{1/\gamma_{k,t}\prod_{1\leq s\leq q, s\neq t}(1/\gamma_{k,t} - 1/\gamma_{k,s})} \qquad \forall 1 \leq t \leq q, \tag{30}$$

*which satisfies $\theta_{k,t} > 0$ for all odd $t$ and $\theta_{k,t} < 0$ for all even $t$. Moreover, it holds that*

$$\sum_{t=1}^{q} \theta_{k,t} = 1 - \frac{\prod_{t=1}^{q}(1/\gamma_{k,t} - 1)}{\prod_{t=1}^{q} 1/\gamma_{k,t}}. \tag{31}$$

The following lemma establishes the recurrence relation for the estimation error of the gradient estimators $\{m^k\}_{k\geq 0}$ of Algorithm 1 with $q = p - 1$. Its proof is deferred to Section 5.3.

**Lemma 6.** *Suppose that Assumption 1 holds, and Assumption 2 holds for $p \geq 2$. Let $\{(x^k, m^k)\}_{k\geq 0}$ be generated by Algorithm 1 with $q = p - 1$, and let $\{(\eta_k, \gamma_{k,t}, \theta_{k,t})\}_{1\leq t\leq p-1, k\geq 0}$ be inputs of Algorithm 1. Assume that $\{(\gamma_{k,t}, \theta_{k,t})\}_{1\leq t\leq p-1, k\geq 0}$ satisfies (28) and (29). Then,*

$$\begin{aligned} &\mathbb{E}_{\xi^{k+1}}[\|m^{k+1} - \nabla f(x^{k+1})\|^2] \\ &\leq \left(1 - \sum_{t=1}^{p-1}\theta_{k,t}\right)\|m^k - \nabla f(x^k)\|^2 + \frac{pL_p^2\eta_k^{2p}}{(p!)^2\sum_{t=1}^{p-1}\theta_{k,t}}\left(1 + \sum_{t=1}^{p-1}\frac{\theta_{k,t}^2}{\gamma_{k,t}^{2p}}\right) + (p-1)\sigma^2\sum_{t=1}^{p-1}\theta_{k,t}^2, \end{aligned} \tag{32}$$

*where $\sigma$ and $L_p$ are given in Assumptions 1(b) and 2, respectively.*

We next derive an upper bound for the average expected error of the stationary condition among all iterates generated by Algorithm 1 with $q = p - 1$. Its proof is relegated to Section 5.3.

**Theorem 3.** *Suppose that Assumption 1 holds, and Assumption 2 holds for $p \geq 2$. Let $\{x^k\}_{k\geq 0}$ be generated by Algorithm 1 with $q = p-1$, and let $\{(\eta_k, \gamma_{k,t}, \theta_{k,t})\}_{1\leq t\leq p-1, k\geq 0}$ be inputs of Algorithm 1. Assume that $\{(\gamma_{k,t}, \theta_{k,t})\}_{1\leq t\leq p-1, k\geq 0}$ satisfies (28) and (29), and that the sequence $\{p_k\}_{k\geq 0}$ satisfies*

$$\Big(1 - \sum_{t=1}^{p-1} \theta_{k,t}\Big) p_{k+1} \leq \Big(1 - \sum_{t=1}^{p-1} \theta_{k,t}/(6p+2)\Big) p_k \qquad \forall k \geq 0. \tag{33}$$

*Then, for any $K \geq 1$,*

$$\sum_{k=0}^{K-1} \mathbb{E}[\|\nabla f(x^k)\|] \leq \frac{f(x^0) - f_{\mathrm{low}} + p_0\sigma^2}{\eta_{K-1}}$$
$$+ \frac{1}{\eta_{K-1}} \sum_{k=0}^{K-1} \Big( \frac{L_1}{2}\eta_k^2 + \frac{(6p+2)\eta_k^2}{p_k \sum_{t=1}^{p-1}\theta_{k,t}} + \frac{L_p^2 \eta_k^{2p} p_{k+1}}{p! \sum_{t=1}^{p-1}\theta_{k,t}} \Big(1 + \sum_{t=1}^{p-1} \frac{\theta_{k,t}^2}{\gamma_{k,t}^{2p}}\Big) + (p-1)\sigma^2 p_{k+1} \sum_{t=1}^{p-1} \theta_{k,t}^2 \Big), \tag{34}$$

*where $f_{\mathrm{low}}$, $L_1$ and $\sigma$ are given in Assumption 1, $L_p$ is given in Assumption 2, and the expectation is taken over the randomness in the algorithm.*

### 3.2.1 Input parameters and convergence rate

We now specify the input parameters of Algorithm 1 with $q = p-1$ and analyze its sample complexity. We first define a quantity that will be used to set the input parameters:

$$k_p \doteq p^{(3p+1)/(2p)}. \tag{35}$$

To analyze the sample complexity of Algorithm 1 with $q = p-1$, we specify $\{(\eta_k, \gamma_{k,t})\}_{1\leq t\leq p-1, k\geq 0}$ as

$$\eta_k = \frac{1}{(k+k_p)^{(2p+1)/(3p+1)}}, \quad \gamma_{k,t} = \frac{1}{t(k+k_p)^{2p/(3p+1)}} \qquad \forall 1 \leq t \leq p-1, k \geq 0. \tag{36}$$

Since $\{\theta_{k,t}\}_{0\leq t\leq p-1, k\geq 0}$ is solution to (28), we compute $\{\theta_{k,t}\}_{0\leq t\leq p-1, k\geq 0}$ using Lemma 5 as

$$\theta_{k,t} = \frac{\prod_{1\leq s\leq p-1, s\neq t}(1/(k+k_p)^{2p/(3p+1)} - s)}{t(k+k_p)^{2p/(3p+1)} \prod_{1\leq s\leq p-1, s\neq t}(t-s)} \qquad \forall 1 \leq t \leq p-1, k \geq 0. \tag{37}$$

In addition, we define the sequence $\{p_k\}_{k\geq 0}$ used in Theorem 3 as

$$p_k = (k+k_p)^{(p-1)/(3p+1)}. \tag{38}$$

The following lemma provides some useful properties of $\{\theta_{k,t}\}_{1\leq t\leq q, k\geq 0}$ and $\{p_k\}_{k\geq 0}$ defined in (36) and (38), respectively. Its proof is deferred to Section 5.3.

**Lemma 7.** *Let $\{\theta_{k,t}\}_{1\leq t\leq q, k\geq 0}$ and $\{p_k\}_{k\geq 0}$ be defined in (36) and (38), respectively. Then,*

$$\sum_{t=1}^{p-1} \theta_{k,t} \in \Big[\frac{1}{2(k+k_p)^{2p/(3p+1)}}, \frac{p-1}{(k+k_p)^{2p/(3p+1)}}\Big] \subset (0,1) \qquad \forall k \geq 0, \tag{39}$$

$$\theta_{k,t}^2 \leq \frac{4((p-1)!)^2}{(k+k_p)^{4p/(3p+1)}} \qquad \forall 1 \leq t \leq p-1, k \geq 0, \tag{40}$$

*where $k_p$ is defined in (35). Moreover, $\{(\theta_{k,t}, p_k)\}_{1\leq t\leq p-1, k\geq 0}$ satisfies (33).*

The next theorem establishes the sample complexity of Algorithm 1 with $q = p-1$ and other inputs specified in (36) and (37). Its proof is deferred to Section 5.3.

**Theorem 4.** *Suppose that Assumption 1 holds, and Assumption 2 holds for $p \geq 2$. Let $\{x^k\}_{k\geq 0}$ be generated by Algorithm 1 with $q = p-1$ and inputs $\{(\eta_k, \gamma_{k,t}, \theta_{k,t})\}_{1\leq t\leq q, k\geq 0}$ specified as in (36) and (37). Define*

$$M_p \doteq 4\bigg(f(x^0) - f_{\mathrm{low}} + p^{1/2}\sigma^2 + \frac{3L_1}{p^{(p+1)/(2p)}} + \frac{7L_p^2}{((p-1)!)^2} + 4(3p+1+2p^{2p}L_p^2 + 2(p!)^2\sigma^2)\bigg). \tag{41}$$

*Let $\kappa(k)$ be uniformly drawn from $\{0,\ldots,k-1\}$. Then,*

$$\mathbb{E}[\|\nabla f(x^{\kappa(k)})\|] \leq \epsilon \qquad \forall k \geq \max\Big\{\Big(\frac{(6p+2)M_p}{p\epsilon}\ln\Big(\frac{(6p+2)M_p}{p\epsilon}\Big)\Big)^{(3p+1)/p}, 2k_p\Big\}, \tag{42}$$

*where $\epsilon \in (0,1)$, and the expectation is taken over the randomness in the algorithm.*

## 4 Numerical experiments

In this section we conduct some preliminary numerical experiments to test practical performance of our SFOMs with multi-extrapolated momentum (Algorithm 1). We compare our SFOMs against the normalized stochastic gradient method with Polyak momentum (SG-PM) [3] and STORM [4] on a robust regression problem. All the algorithms are coded in Matlab, and all the computations are performed on a laptop with a 2.20 GHz Intel Core i9-14900HX processor and 32 GB of RAM.

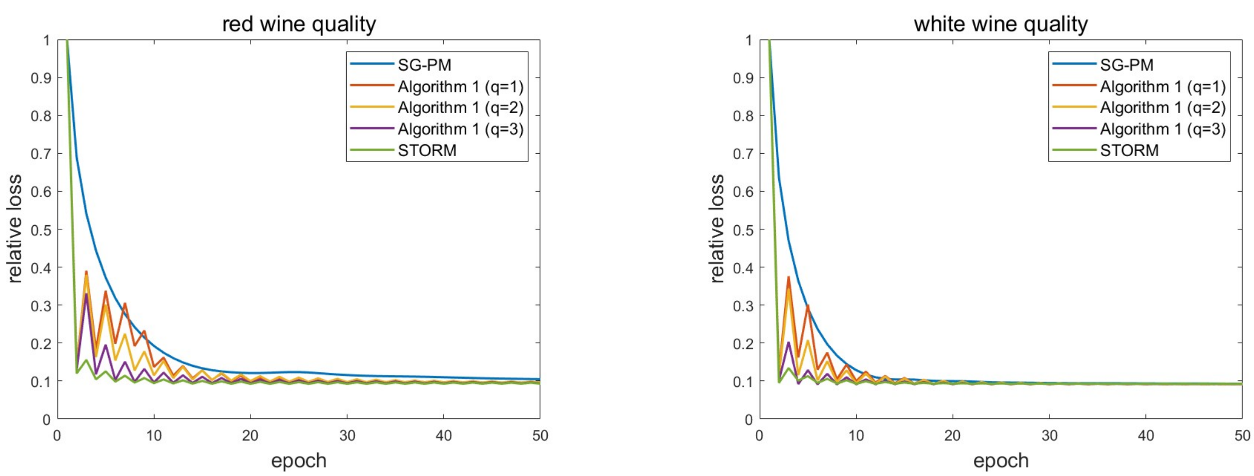


Figure 2: Convergence behavior of the relative loss per epoch for all SFOMs.

Specifically, we consider the robust regression problem:

$$\min_{x\in\mathbb{R}^n} \sum_{i=1}^m \phi(a_i^T x - b_i), \tag{43}$$

where $\phi(t) = t^2/(1+t^2)$, and $\{(a_i, b_i)\}_{1\leq i\leq m} \subset \mathbb{R}^n \times \mathbb{R}$ is the training set. It can be verified that $\phi$ is infinitely differentiable. We consider two datasets, 'red wine quality' and 'white wine quality' from the UCI repository.[4] We apply our Algorithm 1 with $q = 1, 2, 3$, as well as SG-PM and STORM to solve (43).

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

We compare these methods in terms of relative loss, which is defined as $f(x^k)/f(x^0)$. For all methods, we set the maximum number of epochs as 100, and set the initial iterate as the all-zero vector.

From Figure 2, we observe that Algorithm 1 with $q = 1, 2, 3$ slightly outperforms SG-PM and comes very close to the performance of STORM, which corroborates our theoretical results and shows that more extrapolations can achieve faster convergence.

# 5 Proof of the main results

In this section we provide proofs of the main results in Sections 2 and 3, which are particularly Lemmas 1 to 7 and Theorems 1 to 4.

To proceed, we first establish several technical lemmas. The following lemma concerns the estimation of the partial sums of series.

**Lemma 8.** *Let $\zeta(\cdot)$ be a convex univariate function. Then, for any integers $a, b$ satisfying $[a-1/2, b+1/2] \subset \mathrm{dom}\zeta$, it holds that $\sum_{s=a}^{b} \zeta(s) \le \int_{a-1/2}^{b+1/2} \zeta(\tau)\mathrm{d}\tau$.*

*Proof.* Since $f$ is convex, one has $\zeta(s) \le \int_{s-1/2}^{s+1/2} \zeta(\tau)\mathrm{d}\tau$ for all $s \in [a, b]$. It then follows that $\sum_{s=a}^{b} \zeta(s) \le \int_{a-1/2}^{b+1/2} \zeta(\tau)\mathrm{d}\tau$ holds as desired. □

As a consequence of Lemma 8, we consider $\zeta(\tau) = 1/\tau^\alpha$ for some $\alpha \in (0, \infty]$, where $\tau \in (0, \infty)$. Then, for any positive integers $a, b$, one has

$$\sum_{p=a}^{b} 1/p^\alpha \le \begin{cases} \ln(b + 1/2) - \ln(a - 1/2) & \text{if } \alpha = 1, \\ \frac{1}{1-\alpha}((b + 1/2)^{1-\alpha} - (a - 1/2)^{1-\alpha}) & \text{if } \alpha \in (0, 1) \cup (1, +\infty). \end{cases} \tag{44}$$

We next provide an auxiliary lemma that will be used to estimate the maximum number of iterations for achieving targeted approximate stationarity in expectation.

**Lemma 9.** *Let $\alpha \in (0, 1)$ and $u \in (0, 1/e)$ be given. Then, $1/v^\alpha \ln v \le 2u/\alpha$ holds for all $v$ satisfying $v \ge (1/u \ln(1/u))^{1/\alpha}$.*

*Proof.* Let $v$ be such that $v \ge (1/u \ln(1/u))^{1/\alpha}$. By this and $u \in (0, 1/e)$, one has $v \ge (1/u \ln(1/u))^{1/\alpha} > e^{1/\alpha}$. Denote $\phi(v) \doteq 1/v^\alpha \ln v$. Since $\phi$ is decreasing over $(e^{1/\alpha}, \infty)$, it follows that

$$1/v^\alpha \ln v = \phi(v) \le \phi((1/u \ln(1/u))^{1/\alpha}) = \frac{u}{\alpha}\left(1 + \frac{\ln\ln(1/u)}{\ln(1/u)}\right) \le \frac{2u}{\alpha},$$

where the last inequality is due to $\ln\ln(1/u) \le \ln(1/u)$ for all $u \in (0, 1/e)$. Hence, the conclusion of this lemma holds as desired. □

## 5.1 Proof of Lemmas 1 and 2

*Proof of Lemma 1.* For convenience, we denote $\phi(x) = \langle \nabla f(x), u \rangle$. In view of this and the definition of $\nabla^r f$, one can see that

$$D^r \phi(x)[y - x]^r = \langle \nabla^{r+1} f(x)(y - x)^r, u \rangle \qquad \forall 1 \le r \le p - 1, x, y \in \mathbb{R}^n. \tag{45}$$

Since $f$ is $p$th-order continuously differentiable, we have that $\phi$ is $(p-1)$th-order continuously differentiable, and also that

$$\|D^{p-1}\phi(y) - D^{p-1}\phi(x)\|_{(p-1)} = \|u\| \|D^p f(y) - D^p f(x)\|_{(p)} \qquad \forall x, y \in \mathbb{R}^n. \tag{46}$$

Fix any $x, y \in \mathbb{R}^n$. Using Taylor theorem, we have

$$\phi(y) = \phi(x) + \sum_{r=1}^{p-2} \frac{1}{r!} D^r \phi(x)[y-x]^r + \frac{1}{(p-2)!} \int_0^1 (1-t)^{p-2} D^{p-1} \phi(x + t(y-x))[y-x]^{p-1} \mathrm{d}t. \tag{47}$$

It then follows that

$$\begin{aligned}
&\left| \left\langle \nabla f(y) - \nabla f(x) - \sum_{r=1}^{p-1} \frac{1}{r!} \nabla^{r+1} f(x)(y-x)^r, u \right\rangle \right| \overset{(45)}{=} \left| \phi(y) - \phi(x) - \sum_{r=1}^{p-1} \frac{1}{r!} D^r \phi(x)[y-x]^r \right| \\
&\overset{(47)}{\leq} \left| \frac{1}{(p-2)!} \int_0^1 (1-t)^{p-2} (D^{p-1}\phi(x + t(y-x)) - D^{p-1}\phi(x))[y-x]^{p-1} \mathrm{d}t \right| \\
&\overset{(7)}{\leq} \frac{1}{(p-2)!} \|y-x\|^{p-1} \int_0^1 (1-t)^{p-2} \|D^{p-1}\phi(x+t(y-x)) - D^{p-1}\phi(x)\|_{(p-1)} \mathrm{d}t \\
&\overset{(46)}{=} \frac{1}{(p-2)!} \|y-x\|^{p-1} \|u\| \int_0^1 (1-t)^{p-2} \|D^p f(x + t(y-x)) - D^p f(x)\|_{(p)} \mathrm{d}t \\
&\overset{(11)}{\leq} \frac{1}{(p-2)!} L_p \|y-x\|^p \|u\| \int_0^1 (1-t)^{p-2} t \mathrm{d}t = \frac{1}{p!} L_p \|y-x\|^p \|u\|.
\end{aligned}$$

Taking the maximum of this inequality over all $u$ satisfying $\|u\| \leq 1$, we obtain that (12) holds. ☐

*Proof of Lemma 2.* Fix any $k \geq 0$. Using (10) with $(x, y) = (x^k, x^{k+1})$, we obtain that

$$\begin{aligned}
f(x^{k+1}) &\overset{(10)}{\leq} f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L_1}{2} \|x^{k+1} - x^k\|^2 \\
&= f(x^k) + \langle m^k, x^{k+1} - x^k \rangle + \langle \nabla f(x^k) - m^k, x^{k+1} - x^k \rangle + \frac{L_1}{2} \|x^{k+1} - x^k\|^2 \\
&\overset{(17)}{=} f(x^k) - \eta_k \|m^k\| + \langle \nabla f(x^k) - m^k, x^{k+1} - x^k \rangle + \frac{L_1}{2} \eta_k^2 \\
&\leq f(x^k) - \eta_k \|m^k\| + \eta_k \|\nabla f(x^k) - m^k\| + \frac{L_1}{2} \eta_k^2 \\
&\leq f(x^k) - \eta_k \|\nabla f(x^k)\| + 2\eta_k \|\nabla f(x^k) - m^k\| + \frac{L_1}{2} \eta_k^2,
\end{aligned}$$

where the second inequality follows from the Cauchy-Schwarz inequality and $\|x^{k+1} - x^k\| = \eta_k$ due to (17), and the last inequality is due to the triangular inequality. Hence, Lemma 2 holds as desired. ☐

### 5.2 Proof of the main results in Section 3.1

In this subsection we prove Lemmas 3 and 4 and Theorems 1 and 2.

When Assumption 2 holds with $p = 3$, it directly follows from Lemma 1 that

$$\left\| \nabla f(y) - \nabla f(x) - \nabla^2 f(x)(y-x) - \frac{1}{2} \nabla^3 f(x)(y-x)^2 \right\| \leq \frac{L_3}{6} \|y-x\|^3 \qquad \forall x, y \in \mathbb{R}^n. \tag{48}$$

*Proof of Lemma 3.* Fix any $k \geq 0$. Notice from (15) with $q = 2$ that

$$z^{k+1,1} - x^k = \frac{1}{\gamma_{k,1}} (x^{k+1} - x^k), \quad z^{k+1,2} - x^k = \frac{1}{\gamma_{k,2}} (x^{k+1} - x^k). \tag{49}$$

By this, (14), and (16), one has that

$$m^{k+1} - \nabla f(x^{k+1}) \overset{(16)}{=} (1 - \theta_{k,1} - \theta_{k,2}) m^k + \theta_{k,1} G(z^{k+1,1}; \xi^{k+1}) + \theta_{k,2} G(z^{k+1,2}; \xi^{k+1}) - \nabla f(x^{k+1})$$

$$
\begin{aligned}
&\overset{(14)}{=} (1-\theta_{k,1}-\theta_{k,2})(m^k-\nabla f(x^k)) \\
&\quad + \theta_{k,1}(G(z^{k+1,1};\xi^{k+1})-\nabla f(z^{k+1,1})) + \theta_{k,2}(G(z^{k+1,2};\xi^{k+1})-\nabla f(z^{k+1,2})) \\
&\quad - (\nabla f(x^{k+1})-\nabla f(x^k)-\nabla^2 f(x^k)(x^{k+1}-x^k)-\frac{1}{2}\nabla^3 f(x^k)(x^{k+1}-x^k)^2) \\
&\quad + \theta_{k,1}(\nabla f(z^{k+1,1})-\nabla f(x^k)-\frac{1}{\gamma_{k,1}}\nabla^2 f(x^k)(x^{k+1}-x^k)-\frac{1}{2\gamma_{k,1}^2}\nabla^3 f(x^k)(x^{k+1}-x^k)^2) \\
&\quad + \theta_{k,2}(\nabla f(z^{k+1,2})-\nabla f(x^k)-\frac{1}{\gamma_{k,2}}\nabla^2 f(x^k)(x^{k+1}-x^k)-\frac{1}{2\gamma_{k,2}^2}\nabla^3 f(x^k)(x^{k+1}-x^k)^2) \\
&\overset{(49)}{=} (1-\theta_{k,1}-\theta_{k,2})(m^k-\nabla f(x^k)) \\
&\quad + \theta_{k,1}(G(z^{k+1,1};\xi^{k+1})-\nabla f(z^{k+1,1})) + \theta_{k,2}(G(z^{k+1,2};\xi^{k+1})-\nabla f(z^{k+1,2})) \\
&\quad - (\nabla f(x^{k+1})-\nabla f(x^k)-\nabla^2 f(x^k)(x^{k+1}-x^k)-\frac{1}{2}\nabla^3 f(x^k)(x^{k+1}-x^k)^2) \\
&\quad + \theta_{k,1}(\nabla f(z^{k+1,1})-\nabla f(x^k)-\nabla^2 f(x^k)(z^{k+1,1}-x^k)-\frac{1}{2}\nabla^3 f(x^k)(z^{k+1,1}-x^k)^2) \\
&\quad + \theta_{k,2}(\nabla f(z^{k+1,2})-\nabla f(x^k)-\nabla^2 f(x^k)(z^{k+1,2}-x^k)-\frac{1}{2}\nabla^3 f(x^k)(z^{k+1,2}-x^k)^2).
\end{aligned}
$$

Taking the squared norm and the expectation with respect to $\xi^{k+1}$ for both sides of this inequality, we obtain that

$$
\begin{aligned}
&\mathbb{E}_{\xi^{k+1}}[\|m^{k+1}-\nabla f(x^{k+1})\|^2] \\
&= \mathbb{E}_{\xi^{k+1}}[\|\theta_{k,1}(G(z^{k+1,1};\xi^{k+1})-\nabla f(z^{k+1,1})) + \theta_{k,2}(G(z^{k+1,2};\xi^{k+1})-\nabla f(z^{k+1,2}))\|^2] \\
&\quad + \|(1-\theta_{k,1}-\theta_{k,2})(m^k-\nabla f(x^k)) \\
&\quad - (\nabla f(x^{k+1})-\nabla f(x^k)-\nabla^2 f(x^k)(x^{k+1}-x^k)-\frac{1}{2}\nabla^3 f(x^k)(x^{k+1}-x^k)^2) \\
&\quad + \theta_{k,1}(\nabla f(z^{k+1,1})-\nabla f(x^k)-\nabla^2 f(x^k)(z^{k+1,1}-x^k)-\frac{1}{2}\nabla^3 f(x^k)(z^{k+1,1}-x^k)^2) \\
&\quad + \theta_{k,2}(\nabla f(z^{k+1,2})-\nabla f(x^k)-\nabla^2 f(x^k)(z^{k+1,2}-x^k)-\frac{1}{2}\nabla^3 f(x^k)(z^{k+1,2}-x^k)^2)\|^2 \\
&\le 2\theta_{k,1}^2\mathbb{E}_{\xi^{k+1}}[\|G(z^{k+1,1};\xi^{k+1})-\nabla f(z^{k+1,1})\|^2] + 2\theta_{k,2}^2\mathbb{E}_{\xi^{k+1}}[\|G(z^{k+1,2};\xi^{k+1})-\nabla f(z^{k+1,2})\|^2] \\
&\quad + (1-\theta_{k,1}-\theta_{k,2})^2(1+a)\|m^k-\nabla f(x^k)\|^2 \\
&\quad + 3(1+1/a)\|\nabla f(x^{k+1})-\nabla f(x^k)-\nabla^2 f(x^k)(x^{k+1}-x^k)-\frac{1}{2}\nabla^3 f(x^k)(x^{k+1}-x^k)^2\|^2 \\
&\quad + 3(1+1/a)\theta_{k,1}^2\|\nabla f(z^{k+1,1})-\nabla f(x^k)-\nabla^2 f(x^k)(z^{k+1,1}-x^k)-\frac{1}{2}\nabla^3 f(x^k)(z^{k+1,1}-x^k)^2\|^2 \\
&\quad + 3(1+1/a)\theta_{k,2}^2\|\nabla f(z^{k+1,2})-\nabla f(x^k)-\nabla^2 f(x^k)(z^{k+1,2}-x^k)-\frac{1}{2}\nabla^3 f(x^k)(z^{k+1,2}-x^k)^2\|^2 \\
&\le 2(\theta_{k,1}^2+\theta_{k,2}^2)\sigma^2 + (1-\theta_{k,1}-\theta_{k,2})^2(1+a)\|m^k-\nabla f(x^k)\|^2 \\
&\quad + 3(1+1/a)\|\nabla f(x^{k+1})-\nabla f(x^k)-\nabla^2 f(x^k)(x^{k+1}-x^k)-\frac{1}{2}\nabla^3 f(x^k)(x^{k+1}-x^k)^2\|^2 \\
&\quad + 3(1+1/a)\theta_{k,1}^2\|\nabla f(z^{k+1,1})-\nabla f(x^k)-\nabla^2 f(x^k)(z^{k+1,1}-x^k)-\frac{1}{2}\nabla^3 f(x^k)(z^{k+1,1}-x^k)^2\|^2 \\
&\quad + 3(1+1/a)\theta_{k,2}^2\|\nabla f(z^{k+1,2})-\nabla f(x^k)-\nabla^2 f(x^k)(z^{k+1,2}-x^k)-\frac{1}{2}\nabla^3 f(x^k)(z^{k+1,2}-x^k)^2\|^2, \quad (50)
\end{aligned}
$$

where the equality is due to the first relation in (9), the second inequality is due to Young's inequality, and the last inequality is due to the second relation in (9). In addition, recall from (17) and (49) that

$$
\|x^{k+1}-x^k\| = \eta_k, \quad \|z^{k+1,1}-x^k\| = \eta_k/\gamma_{k,1}, \quad \|z^{k+1,2}-x^k\| = \eta_k/\gamma_{k,2}.
$$

It then follows from (48) with $(x,y)=(x^k,x^{k+1}),(x^k,z^{k+1,1}),(x^k,z^{k+1,2})$ that

$$\|\nabla f(x^{k+1})-\nabla f(x^k)-\nabla^2 f(x^k)(x^{k+1}-x^k)-\frac{1}{2}\nabla^3 f(x^k)(x^{k+1}-x^k)^2\|\le\frac{L_3}{6}\|x^{k+1}-x^k\|^3=\frac{L_3\eta_k^3}{6},$$
$$\|\nabla f(z^{k+1,1})-\nabla f(x^k)-\nabla^2 f(x^k)(z^{k+1,1}-x^k)-\frac{1}{2}\nabla^3 f(x^k)(z^{k+1,1}-x^k)^2\|\le\frac{L_3}{6}\|z^{k+1,1}-x^k\|^3=\frac{L_3\eta_k^3}{6\gamma_{k,1}^3},$$
$$\|\nabla f(z^{k+1,2})-\nabla f(x^k)-\nabla^2 f(x^k)(z^{k+1,2}-x^k)-\frac{1}{2}\nabla^3 f(x^k)(z^{k+1,2}-x^k)^2\|\le\frac{L_3}{6}\|z^{k+1,2}-x^k\|^3=\frac{L_3\eta_k^3}{6\gamma_{k,2}^3}.$$

Substituting these inequalities into (50) and letting $a=(\theta_{k,1}+\theta_{k,2})/(1-\theta_{k,1}-\theta_{k,2})$, we obtain that

$$\begin{aligned}\mathbb{E}_{\xi^{k+1}}[\|m^{k+1}-\nabla f(x^{k+1})\|^2]&\le(1-\theta_{k,1}-\theta_{k,2})\|m^k-\nabla f(x^k)\|^2\\&\quad+\frac{L_3^2\eta_k^6\theta_{k,1}^2}{12\gamma_{k,1}^6(\theta_{k,1}+\theta_{k,2})}+\frac{L_3^2\eta_k^6\theta_{k,2}^2}{12\gamma_{k,2}^6(\theta_{k,1}+\theta_{k,2})}+\frac{L_3^2\eta_k^6}{12(\theta_{k,1}+\theta_{k,2})}+2(\theta_{k,1}^2+\theta_{k,2}^2)\sigma^2.\end{aligned}$$

Hence, the conclusion of this lemma holds as desired. □

We are now ready to prove Theorem 1.

*Proof of Theorem 1.* For convenience, we define the following potentials:

$$P_k=f(x^k)+p_k\|m^k-\nabla f(x^k)\|^2\qquad\forall k\ge 0.\tag{51}$$

It then follows from (13), (19), and (20) that for all $k\ge 0$,

$$\begin{aligned}\mathbb{E}_{\xi^{k+1}}[P_{k+1}]&\overset{(51)}{=}\mathbb{E}_{\xi^{k+1}}[f(x^{k+1})+p_{k+1}\|m^{k+1}-\nabla f(x^{k+1})\|^2]\\&\overset{(13)(19)}{\le}f(x^k)-\eta_k\|\nabla f(x^k)\|+2\eta_k\|\nabla f(x^k)-m^k\|+\frac{L_1}{2}\eta_k^2+(1-\theta_{k,1}-\theta_{k,2})p_{k+1}\|m^k-\nabla f(x^k)\|^2\\&\quad+\frac{L_3^2\eta_k^6\theta_{k,1}^2p_{k+1}}{12\gamma_{k,1}^6(\theta_{k,1}+\theta_{k,2})}+\frac{L_3^2\eta_k^6\theta_{k,2}^2p_{k+1}}{12\gamma_{k,2}^6(\theta_{k,1}+\theta_{k,2})}+\frac{L_3^2\eta_k^6p_{k+1}}{12(\theta_{k,1}+\theta_{k,2})}+2(\theta_{k,1}^2+\theta_{k,2}^2)p_{k+1}\sigma^2\\&\overset{(20)}{\le}f(x^k)-\eta_k\|\nabla f(x^k)\|+2\eta_k\|\nabla f(x^k)-m^k\|+\frac{L_1}{2}\eta_k^2+(1-(\theta_{k,1}+\theta_{k,2})/2)p_k\|m^k-\nabla f(x^k)\|^2\\&\quad+\frac{L_3^2\eta_k^6\theta_{k,1}^2p_{k+1}}{12\gamma_{k,1}^6(\theta_{k,1}+\theta_{k,2})}+\frac{L_3^2\eta_k^6\theta_{k,2}^2p_{k+1}}{12\gamma_{k,2}^6(\theta_{k,1}+\theta_{k,2})}+\frac{L_3^2\eta_k^6p_{k+1}}{12(\theta_{k,1}+\theta_{k,2})}+2(\theta_{k,1}^2+\theta_{k,2}^2)p_{k+1}\sigma^2.\end{aligned}\tag{52}$$

We also notice that

$$2\eta_k\|\nabla f(x^k)-m^k\|\le\frac{(\theta_{k,1}+\theta_{k,2})p_k}{2}\|\nabla f(x^k)-m^k\|^2+\frac{2\eta_k^2}{(\theta_{k,1}+\theta_{k,2})p_k}\qquad\forall k\ge 0,$$

which together with (52) implies that for all $k\ge 0$,

$$\begin{aligned}\mathbb{E}_{\xi^{k+1}}[P_{k+1}]&\le f(x^k)+p_k\|m^k-\nabla f(x^k)\|^2-\eta_k\|\nabla f(x^k)\|+\frac{L_1}{2}\eta_k^2+\frac{2\eta_k^2}{(\theta_{k,1}+\theta_{k,2})p_k}\\&\quad+\frac{L_3^2\eta_k^6\theta_{k,1}^2p_{k+1}}{12\gamma_{k,1}^6(\theta_{k,1}+\theta_{k,2})}+\frac{L_3^2\eta_k^6\theta_{k,2}^2p_{k+1}}{12\gamma_{k,2}^6(\theta_{k,1}+\theta_{k,2})}+\frac{L_3^2\eta_k^6p_{k+1}}{12(\theta_{k,1}+\theta_{k,2})}+2(\theta_{k,1}^2+\theta_{k,2}^2)p_{k+1}\sigma^2\\&\overset{(51)}{=}P_k-\eta_k\|\nabla f(x^k)\|+\frac{L_1}{2}\eta_k^2+\frac{2\eta_k^2}{(\theta_{k,1}+\theta_{k,2})p_k}\\&\quad+\frac{L_3^2\eta_k^6\theta_{k,1}^2p_{k+1}}{12\gamma_{k,1}^6(\theta_{k,1}+\theta_{k,2})}+\frac{L_3^2\eta_k^6\theta_{k,2}^2p_{k+1}}{12\gamma_{k,2}^6(\theta_{k,1}+\theta_{k,2})}+\frac{L_3^2\eta_k^6p_{k+1}}{12(\theta_{k,1}+\theta_{k,2})}+2(\theta_{k,1}^2+\theta_{k,2}^2)p_{k+1}\sigma^2.\end{aligned}$$

By summing this inequality over $k = 0, \dots, K-1$ and using the fact that $\{\eta_k\}_{k\geq 0}$ is nonincreasing, we obtain that

$$\sum_{k=0}^{K-1} \mathbb{E}[\|\nabla f(x^k)\|] \leq \frac{f(x^0) - f_{\text{low}} + p_0\sigma^2}{\eta_{K-1}} + \frac{1}{\eta_{K-1}} \sum_{k=0}^{K-1} \Big( \frac{L_1}{2}\eta_k^2 + \frac{2\eta_k^2}{(\theta_{k,1}+\theta_{k,2})p_k} + \frac{L_3^2\eta_k^6\theta_{k,1}^2 p_{k+1}}{12\gamma_{k,1}^6(\theta_{k,1}+\theta_{k,2})} + \frac{L_3^2\eta_k^6\theta_{k,2}^2 p_{k+1}}{12\gamma_{k,2}^6(\theta_{k,1}+\theta_{k,2})} + \frac{L_3^2\eta_k^6 p_{k+1}}{12(\theta_{k,1}+\theta_{k,2})} + 2(\theta_{k,1}^2+\theta_{k,2}^2)p_{k+1}\sigma^2 \Big).$$

Hence, the conclusion of this theorem holds as desired. □

We now prove Lemma 4.

*Proof of Lemma 4.* Fix any $k \geq 0$. We first prove (25). It follows from (23) and $k \geq 0$ that

$$\theta_{k,1} + \theta_{k,2} = \frac{3(k+3)^{3/5} - 1}{2(k+3)^{6/5}} \in \left( \frac{1}{(k+3)^{3/5}}, \frac{3}{2(k+3)^{3/5}} \right) \subset (0,1), \tag{53}$$

and also that

$$|\theta_{k,1}| \leq \frac{2}{(k+3)^{3/5}}, \quad |\theta_{k,2}| \leq \frac{1}{2(k+3)^{3/5}}.$$

Therefore, (25) holds as desired. We next prove prove that (20) holds for $\{(\theta_{k,1}, \theta_{k,2}, p_k)\}_{k\geq 0}$ defined in (23) and (24). By (53), one has

$$\frac{1 - (\theta_{k,1}+\theta_{k,2})/2}{1 - (\theta_{k,1}+\theta_{k,2})} > \frac{1 - 3/(4(k+3)^{3/5})}{1 - 1/(k+3)^{3/5}} = 1 + \frac{1}{4((k+3)^{3/5}-1)} > 1 + \frac{1}{4(k+3)^{3/5}}. \tag{54}$$

In addition, using (24), we have

$$\frac{p_{k+1}}{p_k} \leq \left(1 + \frac{1}{k+3}\right)^{1/5} \leq 1 + \frac{1}{5(k+3)},$$

which together with (54) implies that (20) holds as desired. □

We next prove Theorem 2.

*Proof of Theorem 2.* Notice from (24) that $p_{r+1} \leq 2p_r$ for all $r \geq 0$. Substituting this, (22), (24), and (25) into (21), we obtain that for all $k \geq 2$,

$$\begin{aligned}
&\sum_{r=0}^{k-1} \mathbb{E}[\|\nabla f(x^r)\|] \\
&\leq (k+2)^{7/10} \Big( f(x^0) - f_{\text{low}} + 2\sigma^2 + \sum_{r=0}^{k-1} \Big( \frac{2 + 10L_3^2/3 + 17\sigma^2}{r+3} + \frac{L_1}{2(r+3)^{7/5}} + \frac{L_3^2}{6(r+3)^{17/5}} \Big) \Big) \\
&< (k+2)^{7/10} \Big( f(x^0) - f_{\text{low}} + 2\sigma^2 + \Big(2 + \frac{10L_3^2}{3} + 17\sigma^2\Big) \ln\Big(\frac{2k}{5} + 1\Big) + L_1 + \frac{L_3^2}{120} \Big) \\
&< 4\big(f(x^0) - f_{\text{low}} + 19\sigma^2 + L_1 + 4L_3^2 + 2\big) k^{7/10} \ln k \overset{(26)}{=} M_3 k^{7/10} \ln k,
\end{aligned} \tag{55}$$

where the second inequality is due to (44) with $(a,b) = (3, k+2)$ and $\alpha = 1, 7/5, 17/5$, and third inequality follows from $(k+2)^{7/10} < 2k^{7/10}$ and $1 < \ln(2k/5+1) < 2\ln k$ for all $k \geq 2$. Since $\kappa(k)$ is uniformly drawn from $\{0, \dots, k-1\}$, it follows that

$$\mathbb{E}[\|\nabla f(x^{\kappa(k)})\|] = \frac{1}{k} \sum_{r=0}^{k-1} \mathbb{E}[\|\nabla f(x^r)\|] \overset{(55)}{\leq} M_3 k^{-3/10} \ln k \qquad \forall k \geq 2. \tag{56}$$

In view of Lemma 9 with $(\alpha, u, v) = (3/10, 3\epsilon/(20M_3), k)$, one can see that

$$k^{-3/10} \ln k \le \frac{\epsilon}{M_3} \qquad \forall k \ge \Big(\frac{20M_3}{3\epsilon} \ln\Big(\frac{20M_3}{3\epsilon}\Big)\Big)^{10/3},$$

which together with (56) proves (27) as desired. Hence, this theorem holds as desired. □

### 5.3 Proof of the main results in Section 3.2

In this subsection we prove Lemmas 5 to 7 and Theorems 3 and 4.

*Proof of Lemma 5.* We first prove that the solution to (28) is unique. For convenience, we define

$$\Gamma \doteq \begin{bmatrix} 1/\gamma_{k,1} & 1/\gamma_{k,1}^2 & \cdots & 1/\gamma_{k,1}^q \\ 1/\gamma_{k,2} & 1/\gamma_{k,2}^2 & \cdots & 1/\gamma_{k,2}^q \\ \vdots & \vdots & \ddots & \vdots \\ 1/\gamma_{k,q} & 1/\gamma_{k,q}^2 & \cdots & 1/\gamma_{k,q}^q \end{bmatrix},$$

which is the transpose of the coefficient matrix in (28). To show that the solution to the linear system in (28) is unique, it suffices to prove that $\Gamma$ is invertible. To this end, we let $e_t$, $1 \le t \le q$, be the standard basis vector in $\mathbb{R}^q$, whose $t$th coordinate is 1, and other coordinates are 0. Then, we observe that for any $1 \le t \le q$, solving the linear system

$$\Gamma \begin{bmatrix} c_{1t} \\ c_{2t} \\ \vdots \\ c_{qt} \end{bmatrix} = \begin{bmatrix} 1/\gamma_{k,1} & 1/\gamma_{k,1}^2 & \cdots & 1/\gamma_{k,1}^q \\ 1/\gamma_{k,2} & 1/\gamma_{k,2}^2 & \cdots & 1/\gamma_{k,2}^q \\ \vdots & \vdots & \ddots & \vdots \\ 1/\gamma_{k,q} & 1/\gamma_{k,q}^2 & \cdots & 1/\gamma_{k,q}^q \end{bmatrix} \begin{bmatrix} c_{1t} \\ c_{2t} \\ \vdots \\ c_{qt} \end{bmatrix} = e_t \tag{57}$$

is equivalent to finding the coefficients for a polynomial $h_t(\alpha) \doteq c_{1t}\alpha + c_{2t}\alpha^2 + \cdots + c_{qt}\alpha^q$ such that $h_t(1/\gamma_{k,t}) = 1$ and $h_t(1/\gamma_{k,s}) = 0$ for all $s$ with $1 \le s \le q$ and $s \ne t$. Using Lagrange interpolation and the fact that $1/\gamma_{k,t}$, $1 \le t \le q$, take distinct values, we obtain that such polynomial $h_t(\alpha)$ can be uniquely expressed as

$$h_t(\alpha) = \frac{\alpha \prod_{1\le s\le q, s\ne t}(\alpha - 1/\gamma_{k,s})}{1/\gamma_{k,t} \prod_{1\le s\le q, s\ne t}(1/\gamma_{k,t} - 1/\gamma_{k,s})} \qquad \forall 1 \le t \le q. \tag{58}$$

Therefore, the solution to (57) is unique for each $1 \le t \le q$, and thus $\Gamma$ is invertible. Hence, the solution $\{\theta_{k,t}\}_{1\le t\le q}$ to (28) is unique.

We now prove that the unique solution to (28) can be explicitly written as in (30). For convenience, we denote $V \doteq \Gamma^{-1}$. Since $\Gamma V = I_q$, where $I_q$ is the $q \times q$ identity matrix, it follows that the $t$th column of $V$ is the solution to (57). In addition, recall from (28) that

$$\begin{bmatrix} \theta_{k,1} \\ \theta_{k,2} \\ \vdots \\ \theta_{k,q} \end{bmatrix} = \begin{bmatrix} 1/\gamma_{k,1} & 1/\gamma_{k,2} & \cdots & 1/\gamma_{k,q} \\ 1/\gamma_{k,1}^2 & 1/\gamma_{k,2}^2 & \cdots & 1/\gamma_{k,q}^2 \\ \vdots & \vdots & \ddots & \vdots \\ 1/\gamma_{k,1}^q & 1/\gamma_{k,2}^q & \cdots & 1/\gamma_{k,q}^q \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = V^T \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix},$$

which along with the fact that the $t$th column of $V$ corresponds to the coefficients of the polynomial in (58) implies that

$$\theta_{k,t} = h_t(1) = \frac{\prod_{1\le s\le q, s\ne t}(1 - 1/\gamma_{k,s})}{1/\gamma_{k,t} \prod_{1\le s\le q, s\ne t}(1/\gamma_{k,t} - 1/\gamma_{k,s})} \qquad \forall 1 \le t \le q. \tag{59}$$

Hence, (30) holds as desired.

We then prove that $\theta_{k,t} > 0$ for all odd $t$ and $\theta_{k,t} < 0$ for all even $t$. Since $0 < \gamma_{k,q} < \cdots < \gamma_{k,1} < 1$, it follows that

$$\mathrm{sgn}\Big(\prod_{1\le s\le q, s\ne t}(1-1/\gamma_{k,s})\Big) = (-1)^{q-1}, \quad \mathrm{sgn}\Big(\prod_{1\le s\le q, s\ne t}(1/\gamma_{k,t}-1/\gamma_{k,s})\Big) = (-1)^{q-t},$$

which implies that $\mathrm{sgn}(\theta_{k,t}) = (-1)^{t-1}$, and thus, $\theta_{k,t} > 0$ for all odd $t$ and $\theta_{k,t} < 0$ for all even $t$.

We next prove (31). Recall from (59) that $\theta_{k,t} = h_t(1)$. Let $h(\alpha) \doteq \sum_{t=1}^{q} h_t(\alpha)$. It then follows that

$$\sum_{t=1}^{q}\theta_{k,t} = \sum_{t=1}^{q} h_t(1) = h(1). \tag{60}$$

Also, in view of the fact that the polynomial $h_t(\alpha)$ satisfies $h_t(1/\gamma_{k,t}) = 1$ and $h_t(1/\gamma_{k,s}) = 0$ for all $1 \le s \le q$ and $s \ne t$, one can see that $h(\alpha)$ satisfies $h(0) = 0$ and $h(1/\gamma_{k,t}) = 1$ for all $1 \le t \le q$. Using Lagrange interpolation and the fact that $1/\gamma_{k,t}$, $1 \le t \le q$, take distinct values, we obtain that $h(\alpha)$ can be uniquely expressed as

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

which along with (60) proves (31) as desired. Hence, this lemma holds as desired. □

*Proof of Lemma 6.* Fix any $k \ge 0$. Notice from (15) with $q = p-1$ that

$$z^{k+1,t} - x^k = \frac{1}{\gamma_{k,t}}(x^{k+1}-x^k) \qquad \forall 1 \le t \le p-1. \tag{61}$$

By this, $q = p-1$, (16), and (28), one has that

$$\begin{aligned}
m^{k+1} - \nabla f(x^{k+1}) &\overset{(16)}{=} \Big(1-\sum_{t=1}^{p-1}\theta_{k,t}\Big)m^k + \sum_{t=1}^{p-1}\theta_{k,t}G(z^{k+1,t};\xi^{k+1}) - \nabla f(x^{k+1}) \\
&\overset{(28)}{=} \Big(1-\sum_{t=1}^{p-1}\theta_{k,t}\Big)(m^k - \nabla f(x^k)) + \sum_{t=1}^{p-1}\theta_{k,t}(G(z^{k+1,t};\xi^{k+1}) - \nabla f(z^{k+1,t})) \\
&\quad - \Big(\nabla f(x^{k+1}) - \sum_{r=1}^{p}\frac{1}{(r-1)!}\nabla^r f(x^k)(x^{k+1}-x^k)^{r-1}\Big) \\
&\quad + \sum_{t=1}^{p-1}\theta_{k,t}\Big(\nabla f(z^{k+1,t}) - \sum_{r=1}^{p}\frac{1}{(r-1)!\gamma_{k,t}^{r-1}}\nabla^r f(x^k)(x^{k+1}-x^k)^{r-1}\Big) \\
&\overset{(61)}{=} \Big(1-\sum_{t=1}^{p-1}\theta_{k,t}\Big)(m^k - \nabla f(x^k)) + \sum_{t=1}^{p-1}\theta_{k,t}(G(z^{k+1,t};\xi^{k+1}) - \nabla f(z^{k+1,t})) \\
&\quad - \Big(\nabla f(x^{k+1}) - \sum_{r=1}^{p}\frac{1}{(r-1)!}\nabla^r f(x^k)(x^{k+1}-x^k)^{r-1}\Big) \\
&\quad + \sum_{t=1}^{p-1}\theta_{k,t}\Big(\nabla f(z^{k+1,t}) - \sum_{r=1}^{p}\frac{1}{(r-1)!}\nabla^r f(x^k)(z^{k+1,t}-x^k)^{r-1}\Big).
\end{aligned}$$

Taking the squared norm and the expectation with respect to $\xi^{k+1}$ for both sides of this inequality, we obtain that

$$\mathbb{E}_{\xi^{k+1}}[\|m^{k+1} - \nabla f(x^{k+1})\|^2]$$

$$
\begin{aligned}
&\le \mathbb{E}_{\xi^{k+1}}\left[\left\|\sum_{t=1}^{p-1}\theta_{k,t}(G(z^{k+1,t};\xi^{k+1})-\nabla f(z^{k+1,t}))\right\|^2\right]\\
&\quad+\left\|\left(1-\sum_{t=1}^{p-1}\theta_{k,t}\right)(m^k-\nabla f(x^k))\right.\\
&\quad-\left(\nabla f(x^{k+1})-\sum_{r=1}^{p}\frac{1}{(r-1)!}\nabla^r f(x^k)(x^{k+1}-x^k)^{r-1}\right)\\
&\quad\left.+\sum_{t=1}^{p-1}\theta_{k,t}\left(\nabla f(z^{k+1,t})-\sum_{r=1}^{p}\frac{1}{(r-1)!}\nabla^r f(x^k)(z^{k+1,t}-x^k)^{r-1}\right)\right\|^2\\
&\le (p-1)\sum_{t=1}^{p-1}\theta_{k,t}^2\mathbb{E}_{\xi^{k+1}}[\|G(z^{k+1,t};\xi^{k+1})-\nabla f(z^{k+1,t})\|^2]+\left(1-\sum_{t=1}^{p-1}\theta_{k,t}\right)^2(1+a)\|m^k-\nabla f(x^k)\|^2\\
&\quad+p(1+1/a)\left\|\nabla f(x^{k+1})-\sum_{r=1}^{p}\frac{1}{(r-1)!}\nabla^r f(x^k)(x^{k+1}-x^k)^{r-1}\right\|^2\\
&\quad+p(1+1/a)\sum_{t=1}^{p-1}\theta_{k,t}^2\left\|\nabla f(z^{k+1,t})-\sum_{r=1}^{p}\frac{1}{(r-1)!}\nabla^r f(x^k)(z^{k+1,t}-x^k)^{r-1}\right\|^2\\
&\le (p-1)\sigma^2\sum_{t=1}^{p-1}\theta_{k,t}^2+\left(1-\sum_{t=1}^{p-1}\theta_{k,t}\right)^2(1+a)\|m^k-\nabla f(x^k)\|^2\\
&\quad+p(1+1/a)\left\|\nabla f(x^{k+1})-\sum_{r=1}^{p}\frac{1}{(r-1)!}\nabla^r f(x^k)(x^{k+1}-x^k)^{r-1}\right\|^2\\
&\quad+p(1+1/a)\sum_{t=1}^{p-1}\theta_{k,t}^2\left\|\nabla f(z^{k+1,t})-\sum_{r=1}^{p}\frac{1}{(r-1)!}\nabla^r f(x^k)(z^{k+1,t}-x^k)^{r-1}\right\|^2, \qquad (62)
\end{aligned}
$$

where the equality is due to the first relation in (9), the second inequality is due to Young's inequality, and the last inequality is due to the second relation in (9). In addition, recall from (17) and (61) that

$$
\|x^{k+1}-x^k\|=\eta_k,\quad \|z^{k+1,t}-x^k\|=\eta_k/\gamma_{k,t}\qquad \forall 1\le t\le p-1.
$$

It then follows from (12) with $(x,y)=(x^k,x^{k+1})$ and $(x,y)=(x^k,z^{k+1,t})$ for all $1\le t\le p-1$ that

$$
\begin{aligned}
&\left\|\nabla f(x^{k+1})-\sum_{r=1}^{p}\frac{1}{(r-1)!}\nabla^r f(x^k)(x^{k+1}-x^k)^{r-1}\right\|\le\frac{L_p}{p!}\|x^{k+1}-x^k\|^p=\frac{L_p}{p!}\eta_k^p,\\
&\left\|\nabla f(z^{k+1,t})-\sum_{r=1}^{p}\frac{1}{(r-1)!}\nabla^r f(x^k)(z^{k+1,t}-x^k)^{r-1}\right\|\le\frac{L_p}{p!}\|z^{k+1,t}-x^k\|^p=\frac{L_p\eta_k^p}{p!\gamma_{k,t}^p}\quad\forall 1\le t\le p-1.
\end{aligned}
$$

Substituting these inequalities into (62) and letting $a=\sum_{t=1}^{p-1}\theta_{k,t}/(1-\sum_{t=1}^{p-1}\theta_{k,t})$, we obtain that

$$
\begin{aligned}
&\mathbb{E}_{\xi^{k+1}}[\|m^{k+1}-\nabla f(x^{k+1})\|^2]\\
&\le\left(1-\sum_{t=1}^{p-1}\theta_{k,t}\right)\|m^k-\nabla f(x^k)\|^2+\frac{pL_p^2\eta_k^{2p}}{(p!)^2\sum_{t=1}^{p-1}\theta_{k,t}}+\frac{pL_p^2\eta_k^{2p}}{(p!)^2\sum_{t=1}^{p-1}\theta_{k,t}}\sum_{t=1}^{p-1}\frac{\theta_{k,t}^2}{\gamma_{k,t}^{2p}}+(p-1)\sigma^2\sum_{t=1}^{p-1}\theta_{k,t}^2.
\end{aligned}
$$

Hence, the conclusion of this lemma holds as desired. □

We are now ready to prove Theorem 3.

*Proof of Theorem 3.* For convenience, we define the following potentials:

$$P_k = f(x^k) + p_k\|m^k - \nabla f(x^k)\|^2 \qquad \forall k \geq 0. \tag{63}$$

It then follows from (32), (33), and (13) that

$$\begin{aligned}
\mathbb{E}_{\xi^{k+1}}[P_{k+1}] &\overset{(63)}{=} \mathbb{E}_{\xi^{k+1}}[f(x^{k+1}) + p_{k+1}\|m^{k+1} - \nabla f(x^{k+1})\|^2] \\
&\overset{(32)(13)}{\leq} f(x^k) - \eta_k\|\nabla f(x^k)\| + 2\eta_k\|\nabla f(x^k) - m^k\| + \frac{L_1}{2}\eta_k^2 + \Big(1 - \sum_{t=1}^{p-1}\theta_{k,t}\Big)p_{k+1}\|m^k - \nabla f(x^k)\|^2 \\
&\quad + \frac{pL_p^2\eta_k^{2p}p_{k+1}}{(p!)^2\sum_{t=1}^{p-1}\theta_{k,t}}\Big(1 + \sum_{t=1}^{p-1}\frac{\theta_{k,t}^2}{\gamma_{k,t}^{2p}}\Big) + (p-1)\sigma^2 p_{k+1}\sum_{t=1}^{p-1}\theta_{k,t}^2 \\
&\overset{(33)}{\leq} f(x^k) - \eta_k\|\nabla f(x^k)\| + 2\eta_k\|\nabla f(x^k) - m^k\| + \frac{L_1}{2}\eta_k^2 + \Big(1 - \sum_{t=1}^{p-1}\theta_{k,t}/(6k+2)\Big)p_k\|m^k - \nabla f(x^k)\|^2 \\
&\quad + \frac{pL_p^2\eta_k^{2p}p_{k+1}}{(p!)^2\sum_{t=1}^{p-1}\theta_{k,t}}\Big(1 + \sum_{t=1}^{p-1}\frac{\theta_{k,t}^2}{\gamma_{k,t}^{2p}}\Big) + (p-1)\sigma^2 p_{k+1}\sum_{t=1}^{p-1}\theta_{k,t}^2.
\end{aligned} \tag{64}$$

We also notice that

$$2\eta_k\|\nabla f(x^k) - m^k\| \leq \frac{p_k\sum_{t=1}^{p-1}\theta_{k,t}}{6p+2}\|\nabla f(x^k) - m^k\|^2 + \frac{(6p+2)\eta_k^2}{p_k\sum_{t=1}^{p-1}\theta_{k,t}},$$

which together with (64) implies that

$$\begin{aligned}
\mathbb{E}_{\xi^{k+1}}[P_{k+1}] &\leq f(x^k) + p_k\|m^k - \nabla f(x^k)\|^2 - \eta_k\|\nabla f(x^k)\| + \frac{L_1}{2}\eta_k^2 \\
&\quad + \frac{(6p+2)\eta_k^2}{p_k\sum_{t=1}^{p-1}\theta_{k,t}} + \frac{pL_p^2\eta_k^{2p}p_{k+1}}{(p!)^2\sum_{t=1}^{p-1}\theta_{k,t}}\Big(1 + \sum_{t=1}^{p-1}\frac{\theta_{k,t}^2}{\gamma_{k,t}^{2p}}\Big) + (p-1)\sigma^2 p_{k+1}\sum_{t=1}^{p-1}\theta_{k,t}^2 \\
&\overset{(63)}{=} P_k - \eta_k\|\nabla f(x^k)\| + \frac{L_1}{2}\eta_k^2 + \frac{(6p+2)\eta_k^2}{p_k\sum_{t=1}^{p-1}\theta_{k,t}} + \frac{pL_p^2\eta_k^{2p}p_{k+1}}{(p!)^2\sum_{t=1}^{p-1}\theta_{k,t}}\Big(1 + \sum_{t=1}^{p-1}\frac{\theta_{k,t}^2}{\gamma_{k,t}^{2p}}\Big) + (p-1)\sigma^2 p_{k+1}\sum_{t=1}^{p-1}\theta_{k,t}^2.
\end{aligned}$$

By summing this inequality over $k = 0, \ldots, K-1$ and using the fact that $\{\eta_k\}_{k\geq 0}$ is nonincreasing, we obtain that

$$\begin{aligned}
\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(x^k)\|] &\leq \frac{f(x^0) - f_{\text{low}} + p_0\sigma^2}{\eta_{K-1}} \\
&\quad + \frac{1}{\eta_{K-1}}\sum_{k=0}^{K-1}\Big(\frac{L_1}{2}\eta_k^2 + \frac{(6p+2)\eta_k^2}{p_k\sum_{t=1}^{p-1}\theta_{k,t}} + \frac{pL_p^2\eta_k^{2p}p_{k+1}}{(p!)^2\sum_{t=1}^{p-1}\theta_{k,t}}\Big(1 + \sum_{t=1}^{p-1}\frac{\theta_{k,t}^2}{\gamma_{k,t}^{2p}}\Big) + (p-1)\sigma^2 p_{k+1}\sum_{t=1}^{p-1}\theta_{k,t}^2\Big).
\end{aligned}$$

Hence, the conclusion of this theorem holds as desired. ☐

*Proof of Lemma 7.* Fix any $k \geq 0$. We first prove (39). Recall that (37) is the solution to (28) with $q = p-1$ and $\{\gamma_{k,t}\}_{1\leq t\leq p-1}$ is specified in (36). By substituting $\{\gamma_{k,t}\}_{1\leq t\leq p-1}$ defined in (36) into (31), one has that

$$\sum_{t=1}^{p-1}\theta_{k,t} = 1 - \frac{\prod_{t=1}^{p-1}(t(k+k_p)^{2p/(3p+1)} - 1)}{\prod_{t=1}^{p-1}(t(k+k_p)^{2p/(3p+1)})} = 1 - \prod_{t=1}^{p-1}\Big(1 - \frac{1}{t(k+k_p)^{2p/(3p+1)}}\Big). \tag{65}$$

We notice that

$$\prod_{t=1}^{p-1}\left(1-\frac{1}{t(k+k_p)^{2p/(3p+1)}}\right) \geq \left(1-\frac{1}{(k+k_p)^{2p/(3p+1)}}\right)^{p-1} \geq 1-\frac{p-1}{(k+k_p)^{2p/(3p+1)}}$$
$$\geq 1-\frac{p-1}{k_p^{2p/(3p+1)}} \overset{(35)}{=} \frac{1}{p} > 0, \tag{66}$$

where the first inequality is due to $t \geq 1$, the second inequality follows from $(1+a)^r \geq 1+ra$ for all $a \geq -1$ and $r \geq 1$, and the third inequality is due to $k \geq 0$. We also notice that

$$\prod_{t=1}^{p-1}\left(1-\frac{1}{t(k+2)^{2p/(3p+1)}}\right) \leq \left(1-\frac{1}{(p-1)(k+k_p)^{2p/(3p+1)}}\right)^{p-1} \leq 1-\frac{1}{(k+k_p)^{2p/(3p+1)}+1}$$
$$\leq 1-\frac{1}{2(k+k_p)^{2p/(3p+1)}} < 1, \tag{67}$$

where the first inequality is due to $1 \leq t \leq p-1$, the second inequality is because $(1+a)^r \leq 1/(1-ra)$ for all $a \in [-1,0]$ and $r \geq 0$, and the third inequality follows from $(k+k_p)^{2p/(3p+1)} \geq k_p^{2p/(3p+1)} = p > 1$ for all $k \geq 0$. In view of (66), (67), and (65), we see that (39) holds as desired.

We now prove (40). It follows from (37) and the fact that $p \geq 2$ that

$$|\theta_{k,t}| = \frac{\prod_{s=1}^{p-1}(s-1/(k+k_p)^{2p/(3p+1)})}{t(k+k_p)^{2p/(3p+1)}(t-1/(k+k_p)^{2p/(3p+1)})|\prod_{1\leq s\leq p-1, s\neq t}(t-s)|}$$
$$\leq \frac{(p-1)!}{(1-1/p)(k+k_p)^{2p/(3p+1)}} \leq \frac{2(p-1)!}{(k+k_p)^{2p/(3p+1)}} \qquad \forall 1 \leq t \leq p-1.$$

Hence, (40) holds as desired.

We next prove (33). Using (39), we obtain that

$$\frac{1-\sum_{t=1}^{p-1}\theta_{k,t}/(6p+2)}{1-\sum_{t=1}^{p-1}\theta_{k,t}} \overset{(39)}{\geq} \frac{1-(p-1)/((6p+2)(k+k_p)^{2p/(3p+1)})}{1-1/(2(k+k_p)^{2p/(3p+1)})}$$
$$= 1+\frac{p+1}{(3p+1)((k+k_p)^{2p/(3p+1)}-1/2)} > 1+\frac{p+1}{(3p+1)(k+k_p)^{2p/(3p+1)}}. \tag{68}$$

In addition, observe that

$$\frac{p_{k+1}}{p_k} \leq \left(1+\frac{1}{k+k_p}\right)^{(p-1)/(3p+1)} \leq 1+\frac{p-1}{(3p+1)(k+k_p)}, \tag{69}$$

where the second inequality is due to $(1+a)^r \geq 1+ra$ for all $a \geq -1$ and $r \in [0,1]$. By combining (68) and (69) with the fact that $k+k_p \geq (k+k_p)^{2p/(3p+1)}$, we obtain that (33) holds as desired.

□

We next prove Theorem 4.

*Proof of Theorem 4.* Notice from (38) and $p \geq 2$ that $p_{r+1} \leq 2p_r$ for all $r \geq 0$. Substituting this, (35), (36), (38), (39), and (40) into (34), we obtain that for all $k \geq 0$,

$$\sum_{r=1}^{k-1}\mathbb{E}[\|\nabla f(x^r)\|] \leq (k+k_p-1)^{(2p+1)/(3p+1)}\Bigg(f(x^0)-f_{\text{low}}+p^{1/2}\sigma^2+\sum_{r=0}^{k-1}\Bigg(\frac{L_1}{2(r+k_p)^{(4p+2)/(3p+1)}}$$

$$
\begin{aligned}
&+ \frac{4pL_p^2}{(p!)^2(r+k_p)^{(4p^2-p+1)/(3p+1)}} + \frac{4(3p+1)+8p^{2p}L_p^2+8(p-1)^2((p-1)!)^2\sigma^2}{r+k_p}\Bigg)\Bigg) \\
&\le (k+k_p-1)^{(2p+1)/(3p+1)}\Bigg(f(x^0)-f_{\text{low}}+p^{1/2}\sigma^2+\frac{(3p+1)L_1}{2(p+1)(k_p-1/2)^{(p+1)/(3p+1)}} \\
&+ \frac{(3p+1)L_p^2}{(p!)^2(p-1)(k_p-1/2)^{(4p^2-4p)/(3p+1)}} \\
&+ (4(3p+1)+8p^{2p}L_p^2+8(p-1)^2((p-1)!)^2\sigma^2)\ln\left(\frac{2k}{2k_p-1}+1\right)\Bigg),
\end{aligned}
\tag{70}
$$

where the second inequality is due to (44) with $(a,b)=(k_p,k+k_p-1)$ and $\alpha=1,(4p+2)/(3p+1),(4p^2-p+1)/(3p+1)$. In addition, by (35) and $p\ge 2$, one has that for all $k\ge 2k_p$,

$$
\begin{aligned}
&(k+k_p-1)^{(2p+1)/(3p+1)} \le (2k)^{(2p+1)/(3p+1)} \le 2k^{(2p+1)/(3p+1)}, \\
&(k_p-1/2)^{(p+1)/(3p+1)} = \left(p^{(3p+1)/(2p)}-\frac{1}{2}\right)^{(p+1)/(3p+1)} \ge \left(\frac{p^{(3p+1)/(2p)}}{2}\right)^{(p+1)/(3p+1)} \ge \frac{p^{(p+1)/(2p)}}{2}, \\
&(k_p-1/2)^{(4p^2-4p)/(3p+1)} = \left(p^{(3p+1)/(2p)}-\frac{1}{2}\right)^{(4p^2-4p)/(3p+1)} \ge \left(\frac{p^{(3p+1)/(2p)}}{2}\right)^{(4p^2-4p)/(3p+1)} \\
&\qquad \ge \frac{p^{2p-2}}{2^{(4p^2-4p)/(3p+1)}} \ge \frac{p^{2p-2}}{2^{2p}} \ge \frac{1}{p^2}, \\
&1 < \ln\left(\frac{2}{2k_p-1}+3\right) \le \ln\left(\frac{2k}{2k_p-1}+1\right) \le \ln(2k+1) \le 2\ln k.
\end{aligned}
$$

Substituting these inequalities into (70) and using $p\ge 2$, we obtain that for all $k\ge k_p$,

$$
\begin{aligned}
\sum_{r=1}^{k-1}\mathbb{E}[\|\nabla f(x^r)\|] &\le 4\Bigg(f(x^0)-f_{\text{low}}+p^{1/2}\sigma^2+\frac{3L_1}{p^{(p+1)/(2p)}}+\frac{7L_p^2}{((p-1)!)^2} \\
&\qquad + 4(3p+1+2p^{2p}L_p^2+2(p!)^2\sigma^2)\Bigg)k^{(2p+1)/(3p+1)}\ln k \\
&\overset{(41)}{=} M_p k^{(2p+1)/(3p+1)}\ln k.
\end{aligned}
$$

Since $\kappa(k)$ is uniformly drawn from $\{0,\ldots,k-1\}$, it follows that

$$
\mathbb{E}[\|\nabla f(x^{\kappa(k)})\|] = \frac{1}{k}\sum_{r=0}^{k-1}\mathbb{E}[\|\nabla f(x^r)\|] \overset{(55)}{\le} M_p k^{-p/(3p+1)}\ln k \qquad \forall k\ge 2k_p. \tag{71}
$$

In view of Lemma 9 with $(\alpha,u,v)=(p/(3p+1),p\epsilon/((6p+2)M_p),k)$, one can see that

$$
k^{-p/(3p+1)}\ln k \le \frac{\epsilon}{M_p} \qquad \forall k\ge \left(\frac{(6p+2)M_p}{p\epsilon}\ln\left(\frac{(6p+2)M_p}{p\epsilon}\right)\right)^{(3p+1)/p},
$$

which together with (71) proves (42) as desired. Hence, this theorem holds as desired. □

## References


[1] Y. Arjevani, Y. Carmon, J. C. Duchi, D. J. Foster, A. Sekhari, and K. Sridharan. Second-order information in non-convex stochastic optimization: Power and limitations. In *Conference on Learning Theory*, pages 242–299, 2020.

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

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

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

[5] 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.

[6] C. Fang, Z. Lin, and T. Zhang. Sharp analysis for nonconvex SGD escaping from saddle points. In *Conference on Learning Theory*, pages 1192–1234, 2019.

[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 J. Optim.*, 23(4):2341–2368, 2013.

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

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

[11] S. Jelassi and Y. Li. Towards understanding how momentum improves generalization in deep learning. In *International Conference on Machine Learning*, pages 9965–10040, 2022.

[12] 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.

[13] X. Li, M. Liu, and F. Orabona. On the last iterate convergence of momentum methods. In *International Conference on Algorithmic Learning Theory*, pages 699–717, 2022.

[14] 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.

[15] N. H. Pham, L. M. Nguyen, D. T. Phan, and Q. Tran-Dinh. ProxSARAH: An efficient algorithmic framework for stochastic composite nonconvex optimization. *J. Mach. Learn. Res.*, 21(110):1–48, 2020.

[16] O. Sebbouh, R. M. Gower, and A. Defazio. Almost sure convergence rates for stochastic gradient descent and stochastic heavy ball. In *Conference on Learning Theory*, pages 3935–3971, 2021.

[17] Q. Tran-Dinh, N. H. Pham, D. T. Phan, and L. M. Nguyen. A hybrid stochastic optimization framework for composite nonconvex optimization. *Math. Program.*, 191(2):1005–1071, 2022.

[18] J.-K. Wang and J. Abernethy. Quickly finding a benign region via heavy ball momentum in non-convex optimization. *arXiv preprint arXiv:2010.01449*, 2020.

[19] Z. Wang, K. Ji, Y. Zhou, Y. Liang, and V. Tarokh. Spiderboost and momentum: Faster variance reduction algorithms. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[20] Y. Xu and Y. Xu. Momentum-based variance-reduced proximal stochastic gradient method for composite nonconvex stochastic optimization. *J. Optim. Theory Appl.*, 196(1):266–297, 2023.

[21] Y. You, J. Li, S. Reddi, J. Hseu, S. Kumar, S. Bhojanapalli, X. Song, J. Demmel, K. Keutzer, and C.-J. Hsieh. Large batch optimization for deep learning: Training BERT in 76 minutes. In *International Conference on Learning Representations*, 2020.