Vanilla SGD with Momentum Survives Heavy-Tailed Noise: Convergence Analysis without Gradient Clipping or Normalization

Ryusei Yamada[1 *] **Naoki Sato**[1 *] **Hideaki Iiduka**[1]

[1]Meiji University, Japan
[*]Equal contribution

Abstract

Stochastic gradient descent (SGD) is a cornerstone of modern optimization. While its performance under heavy-tailed noise is often addressed through specialized modifications such as gradient clipping or normalization, we investigate a more fundamental question: how does vanilla SGD, particularly with momentum, perform in the presence of heavy-tailed noise? In this paper, we refine existing convergence results for vanilla SGD and, more importantly, provide the first comprehensive convergence analysis of vanilla SGD with momentum for strongly convex, convex, and nonconvex objectives, without employing any gradient control mechanisms. Our results demonstrate that the obtained convergence rates are inferior to the optimal rates achieved by clipped or normalized variants of SGD, thereby revealing inherent limitations of vanilla methods under heavy-tailed noise. The theoretical findings are supported by experiments on synthetic functions.

1 INTRODUCTION

This paper considers the following optimization problem:

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) := \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{x}),$$

where each f_i ($i \in [n]$) is differentiable. We address three fundamental cases where f is strongly convex, convex, or nonconvex. Empirical risk minimization is a paramount objective that lies at the core of machine learning. As primary solvers for this problem, stochastic gradient descent (SGD) [Robbins and Monro, 1951] and its momentum-based variant [Polyak, 1964, Rumelhart et al., 1986] remain the gold standard in contemporary optimization. A standard assumption in the convergence analysis of such stochastic algorithms is bounded variance of stochastic noise, specifically, that the second moment of the error between the stochastic and full gradients is bounded. While numerous seminal results [Nemirovski et al., 2009, Rakhlin et al., 2012, Liu et al., 2020] rely on this assumption, recent experimental evidence suggests that stochastic noise in modern deep learning [Simsekli et al., 2019, Battash et al., 2024, Ahn et al., 2024] and reinforcement learning [Garg et al., 2021] often exhibits heavy tails. Consequently, research has shifted focus toward the convergence behavior of algorithms when the bounded variance assumption is violated, specifically, when stochastic noise possesses only bounded $\mathfrak{p}$-th moments for $\mathfrak{p} \in (1, 2]$.

Under this bounded $\mathfrak{p}$-th moment assumption, the convergence of vanilla SGD is known to break down [Zhang et al., 2020]. To mitigate this issue, clipped SGD, which scales the gradient magnitude to stay within a predefined threshold, has emerged as a robust alternative. Clipped SGD and its variants have been proven to converge in both expectation [Zhang et al., 2020] and with high probability [Cutkosky and Mehta, 2021, Liu et al., 2023, Sadiev et al., 2023, Nguyen et al., 2023a,b, Liu et al., 2024]. Notably, these studies often incorporate gradient normalization (see Table 1), and it has been widely considered essential to clip or normalize gradients to ensure convergence when noise variance is unbounded.

Recently, several studies have begun to challenge this notion by exploring the convergence of SGD (with momentum) without clipping under bounded $\mathfrak{p}$-th moment assumptions. Hübler et al. [2025] demonstrated that normalized SGD converges with high probability even without clipping. Liu and Zhou [2025] were the first to establish that normalized SGD with momentum converges in expectation, achieving rates of $\mathcal{O}\left(T^{-\frac{\mathfrak{p}-1}{3\mathfrak{p}-2}}\right)$ when the tail index $\mathfrak{p}$ is known, and $\mathcal{O}\left(T^{-\frac{\mathfrak{p}-1}{2\mathfrak{p}}}\right)$ when it is unknown. Most recently, Fatkhullin et al. [2025] and He and Lu [2025] proved that vanilla SGD without any normalization or clipping can indeed converge

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:7670–7700.

Table 1: Comparison of assumptions, algorithmic features, and convergence rates between the related and our work under heavy-tailed noise. The abbreviations "w.o. Clip.", "w.o. Norm.", and "w. Mom." denote "without gradient clipping," "without gradient normalization," and "with momentum," respectively.

Reference	Function	Smoothness	w.o. Clip.	w.o. Norm.	w. Mom.	Rates
Zhang et al. [2020]	S.C.	L-smooth	$\times$	$\times$	$\times$	$\mathcal{O}(T^{-\frac{2(\mathfrak{p}-1)}{\mathfrak{p}}})$
Sadiev et al. [2023]	S.C.	L-smooth	$\times$	✓	$\times$	$\mathcal{O}(T^{-\frac{2(\mathfrak{p}-1)}{\mathfrak{p}}})$
Fatkhullin et al. [2025]	S.C.	Hölder-smooth	✓	✓	$\times$	$\mathcal{O}(T^{-(\mathfrak{p}-1)})$
Ours (Theorem 3.1)	S.C.	Hölder-smooth	✓	✓	$\times$	$\mathcal{O}\left(T^{-(\mathfrak{p}-1)}\right)$
Ours (Theorem 3.2)	S.C.	Hölder-smooth	✓	✓	✓	$\mathcal{O}\Big((1-\eta)^T+\eta^{\mathfrak{p}}\Big)$
Sadiev et al. [2023]	C.	L-smooth	$\times$	✓	$\times$	$\mathcal{O}(T^{-\frac{\mathfrak{p}-1}{\mathfrak{p}}})$
Fatkhullin et al. [2025]	C.	Hölder-smooth	✓	✓	$\times$	$\mathcal{O}(T^{-\frac{\mathfrak{p}-1}{\mathfrak{p}}})$
Ours (Theorem 3.3)	C.	Hölder-smooth	✓	✓	$\times$	$\mathcal{O}\Big(\frac{\log T}{T^{(\mathfrak{p}-1)/\mathfrak{p}}}\Big)$
Ours (Theorem 3.4)	C.	Hölder-smooth	✓	✓	✓	$\mathcal{O}\left(\frac{1}{\eta T}+\eta^{\mathfrak{p}-1}\right)$
Zhang et al. [2020]	N.C.	L-smooth	$\times$	$\times$	$\times$	$\mathcal{O}(T^{-\frac{\mathfrak{p}-1}{3\mathfrak{p}-2}})$
Cutkosky and Mehta [2021]	N.C.	L-smooth	$\times$	$\times$	✓	$\mathcal{O}(T^{\frac{\mathfrak{p}-1}{3\mathfrak{p}-2}})$
Liu et al. [2023]	N.C.	L-smooth	$\times$	$\times$	✓	$\mathcal{O}(T^{-\frac{\mathfrak{p}-1}{3\mathfrak{p}-2}})$
Sadiev et al. [2023]	N.C.	L-smooth	$\times$	✓	$\times$	$\mathcal{O}(T^{-\frac{\mathfrak{p}-1}{\mathfrak{p}}})$
Nguyen et al. [2023b]	N.C.	L-smooth	$\times$	$\times$	$\times$	$\mathcal{O}(T^{-\frac{\mathfrak{p}-1}{3\mathfrak{p}-2}})$
Hübler et al. [2025]	N.C.	L-smooth	✓	$\times$	$\times$	$\mathcal{O}(T^{-\frac{\mathfrak{p}-1}{3\mathfrak{p}-2}})$
Liu and Zhou [2025]	N.C.	(L_0, L_1)-smooth	✓	$\times$	✓	$\mathcal{O}(T^{-\frac{\mathfrak{p}-1}{3\mathfrak{p}-2}})$
Fatkhullin et al. [2025]	N.C.	Hölder-smooth	✓	✓	$\times$	$\mathcal{O}(T^{-\frac{\mathfrak{p}-1}{\mathfrak{p}}})$
Ours (Theorem 3.5)	N.C.	Hölder-smooth	✓	✓	$\times$	$\mathcal{O}\Big(\frac{\log T}{T^{(\mathfrak{p}-1)/\mathfrak{p}}}\Big)$
Ours (Theorem 3.6)	N.C.	Hölder-smooth	✓	✓	✓	$\mathcal{O}\Big(T^{-\frac{\mathfrak{p}-1}{2\mathfrak{p}}}\Big)$

in expectation. Inspired by these recent developments, we decided to investigate the convergence rate of vanilla SGD with momentum under heavy-tailed noise.

Our contribution can be summarized as follows:

1. We provide the first theoretical guarantee that vanilla SGD with momentum converges under heavy-tailed noise for strongly convex, convex, and nonconvex functions. Specifically, for nonconvex objectives, we establish a convergence rate of $\mathcal{O}\left(T^{-\frac{\mathfrak{p}-1}{2\mathfrak{p}}}\right)$ when the tail index $\mathfrak{p} \in (1,2]$ is known, and $\mathcal{O}\left(T^{-\frac{\mathfrak{p}-1}{4}}\right)$ when it is unknown (see Theorem 3.6). While these rates are slightly inferior to those of normalized SGD with momentum [Liu and Zhou, 2025], our findings reveal how well vanilla SGD with momentum performs under heavy-tailed noise, serving as a fundamental baseline for future algorithmic improvements.
2. We establish that vanilla SGD converges in expectation under heavy-tailed noise across all three classes of objective functions. In contrast to existing work [Fatkhullin et al., 2025], our results do not require the assumption of bounded gradients and hold under significantly weaker conditions, offering a more general and robust theoretical framework.
3. Through experiments on synthetic functions, we demonstrate that the condition $\nu + 1 \leq \mathfrak{p}$ is crucial for stable convergence of vanilla SGD with momentum. Here, $\nu \in (0,1]$ denotes the Hölder continuity parameter (Assumption 2.1) and $\mathfrak{p} \in (1,2]$ is the tail index (Assumption 2.2). This condition highlights a vital alignment between our theory and empirical observations (Section 4). Notably, since standard L-smoothness ($\nu = 1$) fails to satisfy this condition when noise is strictly heavy-tailed

($\mathfrak{p} < 2$), our results suggest that Hölder smoothness is a more appropriate and versatile framework for analysis in such settings.

2 PRELIMINARIES

Notation Let $\mathbb{N}$ be the set of non-negative integers. For $m \in \mathbb{N} \setminus \{0\}$, define $[m] := \{1, 2, \dots, m\}$. Let $\mathbb{R}^d$ be a d-dimensional Euclidean space with inner product $\langle \cdot, \cdot \rangle$, which induces the norm $\| \cdot \|$. Let $(\boldsymbol{x}_t)_{t \geq 1} \subset \mathbb{R}^d$ denote the sequence of iterates generated by the recursion in Algorithm 1. Let ξ be a random variable, and let $\mathbb{E}_\xi[X]$ denote the expectation with respect to ξ of a random variable X. $\xi_{t,i}$ is a random variable generated from the i-th sampling at time t, and $\boldsymbol{\xi}_t := (\xi_{t,1}, \xi_{t,2}, \dots, \xi_{t,b})^\top$ is independent of $(\boldsymbol{x}_k)_{k=0}^t$, where $b \in [n]$ is the batch size. The independence of $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \dots$ allows us to define the total expectation $\mathbb{E}$ as $\mathbb{E} = \mathbb{E}_{\boldsymbol{\xi}_1} \mathbb{E}_{\boldsymbol{\xi}_2} \cdots \mathbb{E}_{\boldsymbol{\xi}_t}$. Let $\mathsf{G}_\xi(\boldsymbol{x})$ be the stochastic gradient of $f(\cdot)$ at $\boldsymbol{x} \in \mathbb{R}^d$. A mini-batch $\mathcal{S}_t$ consists of b samples at time t, and the mini-batch stochastic gradient of $f(\boldsymbol{x})$ for $\mathcal{S}_t$ is defined as $\nabla f_{\mathcal{S}_t}(\boldsymbol{x}) := \frac{1}{b} \sum_{i \in [b]} \mathsf{G}_{\xi_{t,i}}(\boldsymbol{x}) = \frac{1}{b} \sum_{i \in \mathcal{S}_t} \nabla f_i(\boldsymbol{x})$. We denote $\boldsymbol{x}^\star \in \operatorname{argmin}_{\boldsymbol{x} \in \mathbb{R}^d} f(\boldsymbol{x})$ and $f^\star := f(\boldsymbol{x}^\star)$ in the strongly convex and convex cases.

We will refer to SGD with momentum that employs neither normalization nor clipping as vanilla SGD with momentum. Furthermore, we will refer to vanilla SGD without momentum obtained by setting $\beta = 0$ in Algorithm 1 simply as vanilla SGD.

Algorithm 1 Vanilla SGD with Momentum

Require: $\boldsymbol{x}_1 \in \mathbb{R}^d, \eta_t > 0, b \in [n], \beta \in [0, 1), \boldsymbol{m}_0 := \boldsymbol{0}, T \in \mathbb{N}$
  for $t = 1$ to T **do**
    $\boldsymbol{m}_t := \beta \boldsymbol{m}_{t-1} + (1 - \beta) \nabla f_{\mathcal{S}_t}(\boldsymbol{x}_t)$
    $\boldsymbol{x}_{t+1} := \boldsymbol{x}_t - \eta_t \boldsymbol{m}_t$
  end for
  return $\boldsymbol{x}_{T+1}$

We make the following assumptions.

Assumption 2.1. *$\nabla f \colon \mathbb{R}^d \to \mathbb{R}^d$ is Hölder continuous; i.e., there exist $\nu \in (0, 1]$ and $L > 0$ such that, for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$,*

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

Assumption 2.2. *(i) For all $\boldsymbol{x} \in \mathbb{R}^d$ that does not depend on ξ,*

$$\mathbb{E}_\xi[\mathsf{G}_\xi(\boldsymbol{x})] = \nabla f(\boldsymbol{x}).$$

(ii) There exist $\sigma \geq 0$ and $\mathfrak{p} \in (1, 2]$ such that, for all $\boldsymbol{x} \in \mathbb{R}^d$ that does not depend on ξ,

$$\mathbb{E}_\xi[\|\mathsf{G}_\xi(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\|^\mathfrak{p}] \leq \sigma^\mathfrak{p}.$$

Assumption 2.1 forms the core of our analysis. Since it reduces to standard L-smoothness when $\nu = 1$, this assumption represents an extension of L-smoothness. Assumption 2.2(ii) is a standard condition for analyzing optimizers under heavy-tailed noise and is widely used in previous studies [Zhang et al., 2020, Cutkosky and Mehta, 2021, Sadiev et al., 2023, Nguyen et al., 2023a, Chezhegov et al., 2025]. When $\mathfrak{p} = 2$, it reduces to the standard bounded variance assumption [Nemirovski et al., 2009, Ghadimi and Lan, 2012, 2013]; when $\mathfrak{p} < 2$, the stochastic gradient may possess unbounded variance.

2.1 TOOLS FOR HEAVY-TAILED ANALYSIS

These lemmas are useful for analysis under heavy-tailed noise. For the sake of completeness, we include their proofs in Appendix A.

Lemma 2.1. *Suppose that Assumption 2.1 holds. Then, for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, the following inequality holds:*

$$f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{L}{\nu + 1} \|\boldsymbol{y} - \boldsymbol{x}\|^{\nu+1}.$$

Lemma 2.2. *Let $q \in (1, 2]$. For any $\boldsymbol{y} \in \mathbb{R}^d$ and any $\boldsymbol{x} \in \mathbb{R}^d \setminus \{\boldsymbol{0}\}$, the following inequality holds:*

$$\|\boldsymbol{x} \pm \boldsymbol{y}\|^q \leq \|\boldsymbol{x}\|^q \pm q \|\boldsymbol{x}\|^{q-2} \langle \boldsymbol{x}, \boldsymbol{y} \rangle + 2^{2-q} \|\boldsymbol{y}\|^q.$$

Lemma 2.3. *Suppose that Assumption 2.2 holds. Then, for all $t \in \mathbb{N}$ and all $\boldsymbol{x} \in \mathbb{R}^d$ that does not depend on $\boldsymbol{\xi}_t$, the following inequality holds:*

$$\mathbb{E}_{\boldsymbol{\xi}_t}[\|\nabla f_{\mathcal{S}_t}(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\|^\mathfrak{p}] \leq \frac{C_\mathfrak{p} \sigma^\mathfrak{p}}{b^{\mathfrak{p}-1}},$$

where $C_\mathfrak{p} \in [1, 2)$ is the constant given in Lemma A.1.

Lemma 2.4. *Let $\mathfrak{p} \in (1, 2]$ and $\nu \in (0, 1]$, suppose that $\nu + 1 \leq \mathfrak{p}$. Then, for all random vectors $\boldsymbol{X} \in \mathbb{R}^d$, the following inequality holds:*

$$\mathbb{E}_{\boldsymbol{X}}[\|\boldsymbol{X}\|^{\nu+1}] \leq (\mathbb{E}_{\boldsymbol{X}}[\|\boldsymbol{X}\|^\mathfrak{p}])^{\frac{\nu+1}{\mathfrak{p}}}.$$

Lemma 2.1 characterizes functions with Hölder continuous gradients. As it recovers the classical descent lemma when $\nu = 1$, it can be viewed as a generalized descent lemma (see, e.g., [Nesterov, 2015, Yashtini, 2016]). Lemma 2.2 has been extensively discussed in functional analysis [Beauzamy, 1982, Xu and Roach, 1991, Chidume, 2009, Rodomanov and Nesterov, 2020]; for $q = 2$, it yields the standard expansion of the squared Euclidean norm, $\|\boldsymbol{x} \pm \boldsymbol{y}\|^2 = \|\boldsymbol{x}\|^2 \pm 2\langle \boldsymbol{x}, \boldsymbol{y} \rangle + \|\boldsymbol{y}\|^2$. Lemma 2.3 is obtained using von Bahr-Esseen's inequality [von Bahr and Esseen, 1965] (Lemma A.1). Notably, it reduces to the standard result $\mathbb{E}_{\boldsymbol{\xi}_t}[\|\nabla f_{\mathcal{S}_t}(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\|^2] \leq \frac{\sigma^2}{b}$ when $\mathfrak{p} = 2$. Lemma 2.4 is crucial for our main results; it dictates the key constraint $\nu + 1 \leq \mathfrak{p}$ required for our convergence analysis.

3 MAIN RESULTS

3.1 INCOMPATIBILITY BETWEEN L-SMOOTHNESS AND HEAVY-TAILED NOISE

L-smoothness is a standard assumption in the analysis of first-order optimization algorithms ($\boldsymbol{x}_{t+1} := \boldsymbol{x}_t - \eta_t \boldsymbol{d}_t$). It allows the use of the classical descent lemma, which is a special case of Lemma 2.1 with $\nu = 1$:

$$\begin{aligned} &f(\boldsymbol{x}_{t+1}) \\ &\quad \leq f(\boldsymbol{x}_t) + \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}_{t+1} - \boldsymbol{x}_t \rangle + \frac{L}{2}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\|^2 \\ &\quad = f(\boldsymbol{x}_t) - \eta_t \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{d}_t \rangle + \frac{L\eta_t^2}{2}\|\boldsymbol{d}_t\|^2. \end{aligned}$$

Whether analyzing SGD ($\boldsymbol{d}_t := \nabla f_{\mathcal{S}_t}(\boldsymbol{x}_t)$) or any other stochastic variant, one cannot avoid the need to upper-bound $\mathbb{E}[\|\boldsymbol{d}_t\|^2]$. Under the standard bounded variance assumption, this is straightforward: $\mathbb{E}\left[\|\nabla f_{\mathcal{S}_t}(\boldsymbol{x}_t)\|^2\right] \leq \frac{\sigma^2}{b} + \mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right]$. However, when stochastic noise exhibits heavy-tailed behavior (Assumption 2.2) and the variance is unbounded, $\|\nabla f_{\mathcal{S}_t}(\boldsymbol{x}_t)\|^2$ cannot be bounded, causing the standard analysis to collapse. Consequently, convergence for vanilla SGD or SGD with momentum under heavy-tailed noise has been considered intractable under L-smoothness, leading prior studies to rely on clipping or normalization (see Section 1).

To circumvent this problem, we utilize the Hölder continuity of the gradient (Assumption 2.1). Applying Lemma 2.1 yields:

$$f(\boldsymbol{x}_{t+1}) \leq f(\boldsymbol{x}_t) - \eta_t \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{d}_t \rangle + \frac{L\eta_t^{\nu+1}}{\nu+1}\|\boldsymbol{d}_t\|^{\nu+1}.$$

Crucially, this shift of the framework requires an upper-bound estimate for $\|\nabla f_{\mathcal{S}_t}(\boldsymbol{x}_t)\|^{\nu+1}$ rather than the squared norm. By leveraging Lemmas 2.2–2.4, we obtain the following key auxiliary lemma, which serves as the technical foundation for our main results.

Lemma 3.1. *Suppose that Assumptions 2.1 and 2.2 hold. Then, for all $\boldsymbol{x} \in \mathbb{R}^d$ that does not depend on $\boldsymbol{\xi}_t$, and all $t \in \mathbb{N}$, the following holds:*

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{\xi}_t}\left[\|\nabla f_{\mathcal{S}_t}(\boldsymbol{x})\|^{\nu+1}\right] \\ &\quad \leq 2^{1-\nu}\left(\frac{C_{\mathfrak{p}}\sigma^{\mathfrak{p}}}{b^{\mathfrak{p}-1}}\right)^{\frac{\nu+1}{\mathfrak{p}}} + \|\nabla f(\boldsymbol{x})\|^{\nu+1} \\ &\quad \leq 2^{1-\nu}\left(\frac{C_{\mathfrak{p}}\sigma^{\mathfrak{p}}}{b^{\mathfrak{p}-1}}\right)^{\frac{\nu+1}{\mathfrak{p}}} + \frac{\nu+1}{2}\|\nabla f(\boldsymbol{x})\|^2 + \frac{1-\nu}{2}. \end{aligned}$$

Notably, Lemma 3.1 recovers the standard bound $\mathbb{E}_{\boldsymbol{\xi}_t}\left[\|\nabla f_{\mathcal{S}_t}(\boldsymbol{x})\|^2\right] \leq \frac{\sigma^2}{b} + \|\nabla f(\boldsymbol{x})\|^2$ when $\nu = 1$ and $\mathfrak{p} = 2$. This generalization demonstrates that existing analyses under L-smoothness and bounded variance can be extended to the heavy-tailed regime by adopting the Hölder smooth framework and the tools presented in Section 2.1.

Due to space constraints, we will define the following constant frequently appearing in our theorems:

$$E_0 := 2^{1-\nu}\left(\frac{C_{\mathfrak{p}}\sigma^{\mathfrak{p}}}{b^{\mathfrak{p}-1}}\right)^{\frac{\nu+1}{\mathfrak{p}}} + \frac{1-\nu}{2}. \tag{1}$$

3.2 STRONGLY CONVEX CASE

We introduce a key property that holds under Assumption 2.1 and μ-strong convexity. Specifically, for all $\boldsymbol{x}, \boldsymbol{y}$ in the domain where f is μ-strongly convex and Hölder smooth, we have:

$$\begin{aligned} \mu\|\boldsymbol{x} - \boldsymbol{y}\|^2 &\leq \langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \\ &\leq \|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|\|\boldsymbol{x} - \boldsymbol{y}\| \\ &\leq L\|\boldsymbol{x} - \boldsymbol{y}\|^{\nu+1}. \end{aligned}$$

That is, we have $\|\boldsymbol{x} - \boldsymbol{y}\|^{1-\nu} \leq \frac{L}{\mu}$ for all $\boldsymbol{x}, \boldsymbol{y}$ in the domain where f is μ-strongly convex and Hölder smooth ($\boldsymbol{x} \neq \boldsymbol{y}$). This property ensures that, for $\nu < 1$, the coexistence of μ-strong convexity and Hölder smoothness forces the domain to have a bounded diameter of at most $(L/\mu)^{\frac{1}{1-\nu}}$. In particular, no function defined on all of $\mathbb{R}^d$ can satisfy both properties with finite L. We use this fact in the proof of Theorem 3.2 (see Lemma B.1). When $\nu = 1$, this reduces to the standard condition $\mu \leq L$.

Theorem 3.1 (**Convergence Analysis of Vanilla SGD for Strongly Convex Functions**). *Let the function $f\colon \mathbb{R}^d \to \mathbb{R}$ be μ-strongly convex. Suppose that Assumptions 2.1 and 2.2 hold. With learning rate $\eta_t := \frac{2\nu}{\mu t}$, if $\nu + 1 \leq \mathfrak{p}$ and $\eta_t^\nu \leq \frac{1}{L}$ for all $t \in [T]$, then:*

$$\mathbb{E}\left[f(\boldsymbol{x}_T) - f^\star\right] \leq \frac{E_1}{T^\nu} = \mathcal{O}\left(\frac{1}{T^\nu}\right).$$

In particular, when $\nu + 1 = \mathfrak{p}$,

$$\mathbb{E}\left[f(\boldsymbol{x}_T) - f^\star\right] = \mathcal{O}\left(\frac{1}{T^{\mathfrak{p}-1}}\right),$$

where $E_1 := \max\left\{f(\boldsymbol{x}_1) - f^\star, \frac{E_0 L}{\nu(\nu+1)}\left(\frac{2\nu}{\mu}\right)^{\nu+1}\right\}$ and E_0 defined in Eq. (1) are non-negative constants.

The proof of Theorem 3.1 is in Appendix B.2. Theorem 3.1 shows that the convergence rate deteriorates when ν or $(\mathfrak{p} - 1)$ is sufficiently small. For $\nu = 1$ and $\mathfrak{p} = 2$, our result recovers the standard $\mathcal{O}(1/T)$ rate for vanilla SGD [Nemirovski et al., 2009, Rakhlin et al., 2012]. Unlike Fatkhullin et al. [2025], who require bounded gradients to achieve similar rates, our analysis holds without such restrictive assumptions.

Theorem 3.2 (**Convergence Analysis of Vanilla SGD with Momentum for Strongly Convex Functions**). *Let the function $f : \mathbb{R}^d \to \mathbb{R}$ be μ-strongly convex. Suppose that Assumptions 2.1 and 2.2 hold. With learning rate $\eta_t := \eta$, if*

$\nu+1 \leq \mathfrak{p}$ *and* $\eta \leq \min\left\{\frac{2(1-\beta)}{E_2}, \left(\frac{1}{8L}\right)^{\frac{1}{\nu}}, \dots\right\}$*, then:*

$$\mathbb{E}\left[f(\boldsymbol{x}_T) - f^\star\right] \leq \frac{f(\boldsymbol{x}_1) - f^\star}{2}\left(1 - \frac{\eta E_2}{4}\right)^T + \frac{E_3}{1-\beta} = \mathcal{O}\left(\left(1 - \frac{\eta E_2}{4}\right)^T + \eta^{\nu+1}\right).$$

In particular, when $\nu + 1 = \mathfrak{p}$*,*

$$\mathbb{E}\left[f(\boldsymbol{x}_T) - f^\star\right] = \mathcal{O}\left(\left(1 - \frac{\eta E_2}{4}\right)^T + \eta^{\mathfrak{p}}\right),$$

where $E_2 := \frac{2\mu(\nu+1)}{\nu\left(1+L^{\frac{1}{\nu}}\right)+1}$, $E_3 := \frac{LE_0}{\nu+1} + \frac{(1+2L)E_0E_2}{4(\nu+1)}\left(\frac{\beta}{1-\beta}\right)^{\nu+1}\eta + \frac{2L^2\nu(1-\beta)E_0}{\nu+1}\left(\frac{\beta}{1-\beta}\right)^{2\nu}\eta^\nu$*, and* E_0 *defined in Eq.* (1) *are non-negative constants.*

The proof of Theorem 3.2 is in Appendix B.3. Under standard assumptions ($\nu = 1, \mathfrak{p} = 2$), vanilla SGD with momentum is known to achieve an $\mathcal{O}((1-\mu\eta)^t + \eta)$ convergence rate for strongly convex functions [Liu et al., 2020]. Theorem 3.2 recovers this classical result as a special case. To the best of our knowledge, Theorem 3.2 provides the first convergence analysis in expectation for vanilla SGD with momentum under heavy-tailed noise.

3.3 CONVEX CASE

For the convex setting, we introduce an additional assumption regarding the boundedness of the iterates and the auxiliary sequence $\boldsymbol{z}_t$, where $\boldsymbol{z}_t := \boldsymbol{x}_t - \frac{\eta\beta}{1-\beta}\boldsymbol{m}_{t-1}$ is defined for the analysis of SGD with momentum.

Assumption 3.1. *There exist constants* $D_1, D_2 > 0$ *such that, for all* $t \in \mathbb{N}$*:*
(i) $\|\boldsymbol{x}_t - \boldsymbol{x}^\star\| \leq D_1$*, (ii)* $\|\boldsymbol{z}_t - \boldsymbol{x}^\star\| \leq D_2$*.*

While Assumption 3.1 is relatively strong, it is a standard way to derive performance upper bounds in both convex and nonconvex optimization [Nemirovski et al., 2009, Kingma and Ba, 2015, Reddi et al., 2018, Zhuang et al., 2020]. Note that while we do not explicitly assume boundedness of the gradient even in the convex case, combining Assumption 3.1(i) and Assumption 2.1 inherently ensures that the gradient remains bounded.

Theorem 3.3 (**Convergence Analysis of Vanilla SGD for Convex Functions**)**.** *Let the function* $f : \mathbb{R}^d \to \mathbb{R}$ *be convex. Suppose that Assumptions 2.1, 2.2, and 3.1(i) hold. With learning rate* $\eta_t := \eta_{\max} t^{-\frac{1}{\nu+1}}$*, if* $\nu + 1 \leq \mathfrak{p}$*, then:*

$$\min_{1\leq t\leq T}\mathbb{E}\left[f(\boldsymbol{x}_t) - f^\star\right] \leq \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^{\nu+1}}{(\nu+1)D_1^{\nu-1}}\frac{1}{\sum_{t=1}^T \eta_t} + E_4\frac{\sum_{t=1}^T \eta_t^{\nu+1}}{\sum_{t=1}^T \eta_t} = \mathcal{O}\left(\frac{\log T}{T^{\frac{\nu}{\nu+1}}}\right).$$

In particular, when $\nu + 1 = \mathfrak{p}$*,*

$$\min_{1\leq t\leq T}\mathbb{E}\left[f(\boldsymbol{x}_t) - f^\star\right] = \mathcal{O}\left(\frac{\log T}{T^{\frac{\mathfrak{p}-1}{\mathfrak{p}}}}\right),$$

where $E_4 := 2^{1-\nu}\left(E_0 + \frac{\nu+1}{2}L^2D_1^{2\nu}\right)$ *and* E_0 *defined in Eq.* (1) *are non-negative constants.*

The proof of Theorem 3.3 is in Appendix B.4. Under conventional assumptions ($\nu = 1, \mathfrak{p} = 2$), vanilla SGD is known to achieve a convergence rate of $\mathcal{O}\left(\left(\sum \eta_t\right)^{-1} + \sum \eta_t^2 / \sum \eta_t\right)$ for convex functions [Nemirovski et al., 2009, Gower et al., 2021]. Theorem 3.3 recovers this standard result as a special case. Under Assumptions 2.1 and 2.2, [Fatkhullin et al., 2025, Theorem 3.2] previously derived a similar convergence bound, so the novelty of Theorem 3.3 is incremental.

Theorem 3.4 (**Convergence Analysis of Vanilla SGD with Momentum for Convex Functions**)**.** *Let the function* $f : \mathbb{R}^d \to \mathbb{R}$ *be convex. Suppose that Assumptions 2.1, 2.2, 3.1(i), and 3.1(ii) hold. With learning rate* $\eta_t := \eta$*, if* $\nu+1 \leq \mathfrak{p}$*, then:*

$$\min_{1\leq t\leq T}\mathbb{E}\left[f(\boldsymbol{x}_t) - f^\star\right] \leq \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^{\nu+1}}{\eta(\nu+1)D_2^{\nu-1}T} + E_6\eta + \frac{2^{1-\nu}E_5}{(\nu+1)D_2^{\nu-1}}\eta^\nu = \mathcal{O}\left(\frac{1}{\eta T} + \eta^\nu\right).$$

In particular, when $\nu + 1 = \mathfrak{p}$*,*

$$\min_{1\leq t\leq T}\mathbb{E}\left[f(\boldsymbol{x}_t) - f^\star\right] = \mathcal{O}\left(\frac{1}{\eta T} + \eta^{\mathfrak{p}-1}\right),$$

where $E_5 := 2^{1-\nu}\left(\frac{C_{\mathfrak{p}}\sigma^{\mathfrak{p}}}{b^{\mathfrak{p}-1}}\right)^{\frac{\nu+1}{\mathfrak{p}}} + L^{\nu+1}D_1^{\nu(\nu+1)}$, $E_6 := \frac{\beta\left(E_5 + \nu L^{\frac{\nu+1}{\nu}}D_1^{\nu+1}\right)}{(1-\beta)(\nu+1)}$ *are non-negative constants.*

The proof of Theorem 3.4 is in Appendix B.5. Under standard assumptions ($\nu = 1, \mathfrak{p} = 2$), vanilla SGD with momentum is known to achieve a convergence rate of $\mathcal{O}(1/(\eta T) + \eta)$ for convex functions [Sebbouh et al., 2021]. Theorem 3.4 recovers this conventional result as a special case. To the best of our knowledge, Theorem 3.4 provides the first convergence analysis in expectation for vanilla SGD with momentum on convex functions under heavy-tailed noise.

3.4 NONCONVEX CASE

Theorem 3.5 (Convergence Analysis of Vanilla SGD for Nonconvex Functions). *Let the function $f : \mathbb{R}^d \to \mathbb{R}$ be nonconvex. Suppose that Assumptions 2.1 and 2.2 hold. With learning rate $\eta_t := \eta_{\max} t^{-\frac{1}{\nu+1}}$, if $\nu + 1 \leq \mathfrak{p}$ and $\eta_t^\nu < \frac{2}{L}$ for all $t \in [T]$, then:*

$$\begin{aligned}
&\min_{1\leq t\leq T} \mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right] \\
&\quad\leq \frac{2(f(\boldsymbol{x}_1) - f^\star)}{(2 - L\eta_{\max}^\nu)} \frac{1}{\sum_{t=1}^T \eta_t} + E_7 \frac{\sum_{t=1}^T \eta_t^{\nu+1}}{\sum_{t=1}^T \eta_t} \\
&\quad= \mathcal{O}\left(\frac{\log T}{T^{\frac{\nu}{\nu+1}}}\right).
\end{aligned}$$

In particular, when $\nu + 1 = \mathfrak{p}$,

$$\min_{1\leq t\leq T} \mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right] = \mathcal{O}\left(\frac{\log T}{T^{\frac{\mathfrak{p}-1}{\mathfrak{p}}}}\right),$$

where $E_7 := \frac{2E_0 L}{(\nu+1)(2-L\eta_{\max}^\nu)}$ and E_0 defined in Eq. (1) are non-negative constants.

The proof of Theorem 3.5 is in Appendix B.6. Under standard assumptions ($\nu = 1, \mathfrak{p} = 2$), vanilla SGD is known to achieve a convergence rate of $\mathcal{O}\big((\sum \eta_t)^{-1} + \sum \eta_t^2 / \sum \eta_t\big)$ for nonconvex functions [Sebbouh et al., 2021, Khaled and Richtárik, 2023]. Theorem 3.5 recovers this classical result as a special case. While Fatkhullin et al. [2025, Theorem 5.4] previously derived a similar bound under Assumptions 2.1 and 2.2, their analysis necessitates assuming bounded gradients. In contrast, our framework eliminates this requirement, providing a more general and robust convergence guarantee for the nonconvex setting.

Theorem 3.6 (Convergence Analysis of Vanilla SGD with Momentum for Nonconvex Functions). *Let the function $f : \mathbb{R}^d \to \mathbb{R}$ be nonconvex. Suppose that Assumptions 2.1 and 2.2 hold. With learning rate $\eta_t := \eta$, if $\nu + 1 \leq \mathfrak{p}$ and $\eta \leq \min\left\{\left(\frac{1}{4L}\right)^{\frac{1}{\nu}}, \frac{1-\beta}{\beta}\left(\frac{1}{4\nu L^2}\right)^{\frac{1}{2\nu}}\right\}$, then:*

$$\begin{aligned}
&\min_{1\leq t\leq T} \mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right] \\
&\quad\leq \frac{4(f(\boldsymbol{x}_1) - f^\star)}{\eta T} + \frac{LE_0\eta^\nu}{4(\nu+1)} + E_8\eta^{2\nu} \\
&\quad= \mathcal{O}\left(\frac{1}{\eta T} + \eta^\nu\right).
\end{aligned}$$

In particular, when $\nu + 1 = \mathfrak{p}$,

$$\min_{1\leq t\leq T} \mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right] = \mathcal{O}\left(\frac{1}{\eta T} + \eta^{\mathfrak{p}-1}\right),$$

where $E_8 := \frac{2L}{\nu+1}\left(\frac{\beta}{1-\beta}\right)^{2\nu}\{L(1-\nu) + 2E_0\nu\}$ and E_0 defined in Eq. (1) are non-negative constants.

The proof of Theorem 3.6 is in Appendix B.7. Under standard assumptions ($\nu = 1, \mathfrak{p} = 2$), vanilla SGD with momentum is known to achieve a convergence rate of $\mathcal{O}(1/(\eta T) + \eta)$ for nonconvex functions [Yan et al., 2018, Mai and Johansson, 2020, Liu et al., 2020]. Our analytical framework builds upon the approach established by Liu et al. [2020], and Theorem 3.6 recovers their results as a special case when $\nu = 1$ or $\mathfrak{p} = 2$. To the best of our knowledge, Theorem 3.6 represents the first convergence analysis in expectation for vanilla SGD with momentum on nonconvex functions under heavy-tailed noise.

Comparison with [Liu and Zhou, 2025] Liu and Zhou [2025] established that normalized SGD with momentum achieves the following convergence rates for nonconvex functions under heavy-tailed noise:

$$\begin{aligned}
&\min_{1\leq t\leq T} \mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|\right] = \mathcal{O}\left(T^{-\frac{\mathfrak{p}-1}{3\mathfrak{p}-2}}\right) \text{ (with known } \mathfrak{p}), \\
&\min_{1\leq t\leq T} \mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|\right] = \mathcal{O}\left(T^{-\frac{\mathfrak{p}-1}{2\mathfrak{p}}}\right) \text{ (with unknown } \mathfrak{p}).
\end{aligned}$$

In contrast, our Theorem 3.6 implies that vanilla SGD with momentum achieves:

$$\begin{aligned}
&\min_{1\leq t\leq T} \mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|\right] = \mathcal{O}\left(T^{-\frac{\mathfrak{p}-1}{2\mathfrak{p}}}\right) \text{ (with } \eta = T^{-1/\mathfrak{p}}), \\
&\min_{1\leq t\leq T} \mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|\right] = \mathcal{O}\left(T^{-\frac{\mathfrak{p}-1}{4}}\right) \text{ (with } \eta = T^{-1/2}).
\end{aligned}$$

Since $\frac{\mathfrak{p}-1}{3\mathfrak{p}-2} \geq \frac{\mathfrak{p}-1}{2\mathfrak{p}} \geq \frac{\mathfrak{p}-1}{4}$ for $\mathfrak{p} \in (1, 2]$, these results illustrate the efficiency of normalization in mitigating heavy-tailed noise, especially considering that the rate $\mathcal{O}\left(T^{-\frac{\mathfrak{p}-1}{3\mathfrak{p}-2}}\right)$ is optimal [Zhang et al., 2020]. While vanilla SGD with momentum exhibits sub-optimal rates compared to its normalized counterparts, our work fills a critical gap by characterizing the fundamental convergence behavior of the most widely used momentum-based optimizer without explicit gradient control. Our findings reveal that vanilla SGD with momentum maintains sublinear convergence even without normalization or clipping, providing a necessary theoretical baseline that more complex algorithms should aim to surpass.

Analytical framework As demonstrated, our analysis extends classical results established under L-smoothness and bounded variance to more general settings. These results are obtained by integrating existing analytical frameworks with the technical tools developed in Section 2.1. Specifically, for vanilla SGD with momentum, we employ an auxiliary sequence and a Lyapunov function to facilitate the analysis, following established methodologies (e.g., [Ghadimi et al., 2015, Gadat et al., 2018, Mai and Johansson, 2020, Liu et al., 2020, Defazio, 2021, Qiu et al., 2024, Kondo and Iiduka, 2025]). Furthermore, for convex and nonconvex settings, we adopt the weighting scheme introduced by Stich [2019] to streamline the convergence proofs. Notably, our

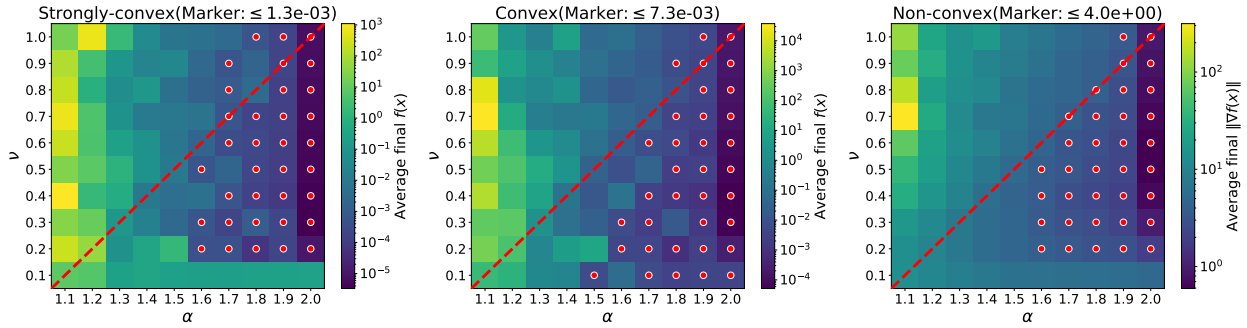

Figure 1: Convergence performance across various (ν, α) pairs. The heatmaps show the average final objective values (for strongly convex and convex cases) and gradient norms (for nonconvex cases) after 4,000 steps of vanilla SGD with momentum. The top 35 performing combinations are indicated by red markers. Below the red dotted line, the theoretical condition $\nu + 1 \leq \alpha$ is satisfied, ensuring convergence guarantees.

framework is sufficiently general to potentially establish convergence guarantees under heavy-tailed noise for other popular optimizers, such as Adam [Kingma and Ba, 2015] and Muon [Jordan et al., 2024].

Summary of theoretical findings Our theoretical analysis provides three key insights. First, vanilla SGD with momentum guarantees convergence in expectation without requiring explicit gradient control, such as clipping or normalization. Second, while convergence is guaranteed, the rate is inherently slower than those achieved under classical L-smoothness and bounded variance, or those obtained by heavy-tailed-specific methods (e.g., clipping-based algorithms). Third, the condition $\nu + 1 \leq \mathfrak{p}$ is essential for ensuring these convergence guarantees. In the following section, we provide numerical experiments to empirically validate these findings, with a particular focus on the necessity of this third condition.

4 NUMERICAL EXPERIMENTS

This section presents numerical experiments to empirically validate the theoretical findings established in Section 3. We evaluate the performance of vanilla SGD and SGD with momentum on a series of synthetic strongly convex, convex, and nonconvex objective functions to observe and compare their convergence behaviors under heavy-tailed noise.

Problem setup We prepared objective functions defined as follows:

$$
\begin{aligned}
f_{\text{strongly convex}}(\boldsymbol{x}) &= \frac{1}{2}\sum_{i=1}^{d} x_i^2 + \sum_{i=1}^{d} \frac{1}{\nu+1}|x_i|^{\nu+1}, \\
f_{\text{convex}}(\boldsymbol{x}) &= \sum_{i=1}^{d} \frac{1}{\nu+1}|x_i|^{\nu+1}, \\
f_{\text{nonconvex}}(\boldsymbol{x}) &= \sum_{i=1}^{d} \left(\frac{1}{\nu+1}|x_i|^{\nu+1} - \frac{1}{2}\cos(2\pi x_i) + \frac{1}{2} \right),
\end{aligned}
\tag{2}
$$

where $\boldsymbol{x} := (x_1, x_2, \ldots, x_d)^\top \in \mathbb{R}^d$. For any $\nu \in (0, 1]$, each of these functions satisfies Assumption 2.1 with a Hölder continuous gradient. In all experiments, the dimension is set to $d = 500$. The learning rate was set to $\eta_t := t^{-\frac{1}{\nu+1}}$ for convex and nonconvex cases, and $\eta_t := \frac{2\nu}{t}$ for strongly convex cases.

Heavy-tailed noise setup The minibatch stochastic gradient $\nabla f_{\mathcal{S}_t}(\boldsymbol{x}_t)$ in Algorithm 1 was simulated by injecting synthetic heavy-tailed noise $\boldsymbol{\omega}_t \in \mathbb{R}^d$ into the full gradient: $\nabla f_{\mathcal{S}_t}(\boldsymbol{x}_t) = \nabla f(\boldsymbol{x}_t) + \lambda \boldsymbol{\omega}_t$, where $\lambda > 0$ scales the noise intensity. Following Fatkhullin et al. [2025], each component of the synthetic noise was defined as:

$$
(\boldsymbol{\omega}_t)_i := s_i u_i^{-\frac{1}{\alpha}},
$$

where $s_i \sim \text{Unif}(\{-1, 1\})$, $u_i \sim \text{Unif}(0, 1)$, and $\alpha \in (1, 2)$ is the tail index. In this two-sided Pareto distribution, the $\mathfrak{p}$-th moment is finite for all $\mathfrak{p} \in (1, \alpha)$ and infinite for $\mathfrak{p} \geq \alpha$. Consequently, Assumption 2.2(ii) is satisfied for any $\mathfrak{p} < \alpha$. Furthermore, as $\mathbb{E}_{s_i}[s_i] = 0$, it follows from the independence of s_i and u_i that $\mathbb{E}_{(s_i,u_i)}[(\omega_t)_i] = \mathbb{E}_{s_i}[s_i]\mathbb{E}_{u_i}[u_i^{-1/\alpha}] = 0$, ensuring that Assumption 2.2(i) holds for any $\alpha \in (1, 2)$. For the boundary case $\alpha = 2$, we set $\boldsymbol{\omega}_t$ to standard Gaussian noise. The noise scale λ was set to 0.01 for strongly convex and convex functions, and 0.1 for the nonconvex setting.

Dependence of convergence on ν and α All theorems in Section 3 necessitate the condition $\nu + 1 \leq \alpha$ in order to leverage the moment bounds in Lemma 2.4. Consequently, our theory does not guarantee the convergence of vanilla SGD or SGD with momentum when this condition is violated. To empirically investigate the interplay between ν and α, we conducted experiments across 100 combinations of $\nu \in \{0.1, \ldots, 1.0\}$ and $\alpha \in \{1.1, \ldots, 2.0\}$. For each pair, SGD with momentum was executed for 4,000 steps, and we recorded the final objective values (for strongly convex and convex) or gradient norms (for nonconvex). Figure 1 presents heatmaps averaged over 20 independent trials,

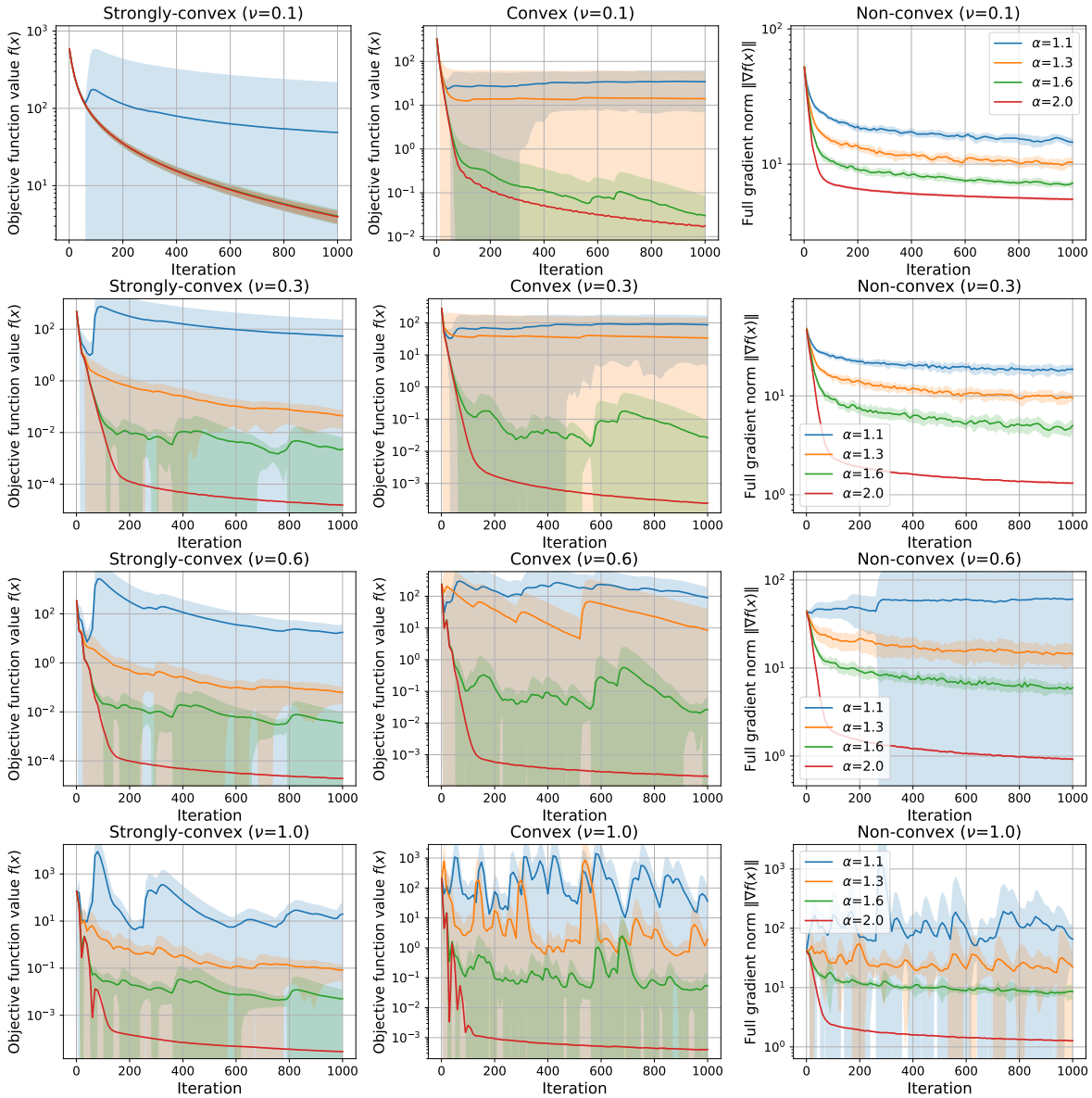

Figure 2: Convergence trajectories of vanilla SGD with momentum. The plots illustrate the evolution of objective values (for strongly convex and convex cases) and gradient norms (for nonconvex cases) over 1,000 steps for the functions defined in Eq. (2). Solid lines represent the average of 20 independent trials, while the lightly shaded regions indicate the range between the minimum and maximum values.

where the top 35 performing combinations are highlighted with red markers. Notably, the condition $\nu + 1 \leq \alpha$ corresponds to the lower-right triangular region of each map. As shown in the figure, the high-performance markers are predominantly concentrated within this region, empirically validating that $\nu + 1 \leq \alpha$ is a near-necessary requirement for the stable convergence of vanilla SGD with momentum. Furthermore, we observe that for small values of ν or α, the optimization performance remains poor even when the convergence condition is satisfied. This observation aligns with our theoretical rates; for instance, the nonconvex convergence rate $\mathcal{O}\big(1/(\eta T) + \eta^{\nu}\big)$ deteriorates when ν or $(\alpha - 1)$ is sufficiently small. Thus, these numerical results provide strong empirical support for our theoretical results.

Detailed observation of convergence behavior Figure 2 illustrates the evolution of objective values and gradient norms for fixed ν and varying α. Solid lines represent the average of 20 independent trials, with shaded regions indicating the range between extrema. In all cases, a smaller α not only degrades average performance but also amplifies variance, visually confirming our theoretical insight: the divergence of lower-order moments manifests as extreme instability in optimization trajectories. Crucially, under highly

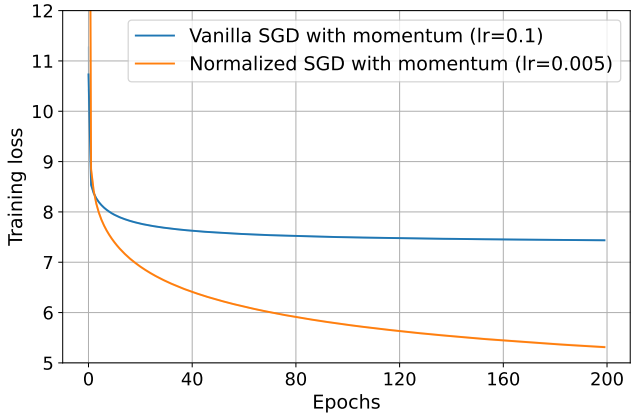

Figure 3: Loss function value for training versus the number of epochs in training Transformer-XL on the WikiText-2 dataset. The solid lines represent the mean value, and the shaded areas represent the maximum and minimum values over three runs.

heavy-tailed noise (small α), larger ν values lead to more erratic loss curves, whereas smaller ν maintains relative stability. This aligns with the requirement $\nu + 1 \leq \alpha$, as a smaller ν, representing broader Hölder smoothness, widens the guaranteed convergence range of α. These results underscore the inherent incompatibility between extreme heavy-tailed noise ($\alpha < 2$) and classical L-smoothness ($\nu = 1$), justifying the adoption of more flexible smoothness classes, such as Hölder or (L_0, L_1)-smoothness, in stochastic optimization theory. (See Appendix C for corresponding results for vanilla SGD).

Experiments on WikiText-2 dataset with Transformer-XL Finally, we conducted more realistic experiments on the WikiText-2 dataset [Merity et al., 2017] with Transformer-XL, where heavy-tailed gradient noise has been observed [Kunstner et al., 2023]. We evaluated a 6-layer, 8-head Transformer-XL model with approximately 26.9M parameters. The embedding, head, and inner dimensions were set to 256, 32, and 1024, respectively, with tied input/output embeddings. Training was performed for 200 epochs with a batch size of 20, and a target sequence and memory length of 128. We compared vanilla SGD with momentum and normalized SGD with momentum under constant learning rates. For each optimizer, we searched learning rates over $\{0.001, 0.005, 0.01, 0.05, 0.1\}$ and selected the best-performing one on the basis of the final training loss. The best learning rates were 0.1 for vanilla SGD with momentum and 0.005 for normalized SGD with momentum. The momentum parameter was fixed to 0.9 for both optimizers. Figure 3 confirms that (1) vanilla SGD with momentum converges without gradient clipping or normalization and (2) normalized SGD with momentum achieves a lower loss faster, consistent with our theory. We note that validating $\nu + 1 \leq \mathfrak{p}$ on real tasks would not be feasible since ν cannot be reliably estimated in practice. This is why the verification of this condition was conducted on synthetic functions where both ν and $\mathfrak{p}$ can be fully controlled.

5 CONCLUSION

In this study, we established a comprehensive convergence analysis for vanilla SGD and SGD with momentum under heavy-tailed noise, encompassing strongly convex, convex, and nonconvex functions. A key departure from traditional theory is our shift from classical L-smoothness and bounded variance to a more general framework of Hölder continuous gradients. This approach naturally extends existing results while proving that convergence in expectation is guaranteed for SGD with momentum, even without explicit gradient control such as clipping or normalization. Our framework identifies the critical condition $\nu + 1 \leq \mathfrak{p}$, capturing the intrinsic interplay between function smoothness and the noise tail index. Numerical experiments empirically validate this condition, demonstrating that Hölder continuity offers a more robust and realistic foundation for heavy-tailed optimization than standard L-smoothness.

Acknowledgements

We are sincerely grateful to Program Chairs, Area Chairs, and four anonymous reviewers for helping us improve the original manuscript. This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI Grants, Numbers 24K14846 and 26KJ2026, awarded to Hideaki Iiduka and Naoki Sato, respectively.

References

Kwangjun Ahn, Xiang Cheng, Minhak Song, Chulhee Yun, Ali Jadbabaie, and Suvrit Sra. Linear attention is (maybe) all you need (to understand transformer optimization). In *International Conference on Learning Representations*, 2024.

Barak Battash, Lior Wolf, and Ofir Lindenbaum. Revisiting the noise model of stochastic gradient descent. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238, pages 4780–4788, 2024.

Bernard Beauzamy. *Introduction to Banach Spaces and their Geometry*, volume 68 of *North-Holland Mathematics Studies*. North Holland, 1982.

Savelii Chezhegov, Klyukin Yaroslav, Andrei Semenov, Aleksandr Beznosikov, Alexander Gasnikov, Samuel Horváth, Martin Takáč, and Eduard Gorbunov. Clipping improves Adam-norm and AdaGrad-norm when the noise is heavy-tailed. In *Proceedings of the 42nd International*

Conference on Machine Learning, volume 267, pages 10269–10333, 2025.

Charles Chidume. *Geometric Properties of Banach Spaces and Nonlinear Iterations*. Number 1 in Lecture Notes in Mathematics. Springer London, 2009.

Ashok Cutkosky and Harsh 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.

Aaron Defazio. Momentum via primal averaging: Theoretical insights and learning rate schedules for non-convex optimization. *arXiv*, https://arxiv.org/abs/2010.00406, 2021.

Ilyas Fatkhullin, Florian Hübler, and Guanghui Lan. Can SGD handle heavy-tailed noise? In *NeurIPS 2025 Workshop: Optimization for Machine Learning*, 2025.

Sébastien Gadat, Fabien Panloup, and Sofiane Saadane. Stochastic heavy ball. *Electronic Journal of Statistics*, 12: 461–529, 2018.

Saurabh Garg, Joshua Zhanson, Emilio Parisotto, Adarsh Prasad, Zico Kolter, Zachary Lipton, Sivaraman Balakrishnan, Ruslan Salakhutdinov, and Pradeep Ravikumar. On proximal policy optimization's heavy-tailed gradients. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 3610–3619, 2021.

Euhanna Ghadimi, Hamid Reza Feyzmahdavian, and Mikael Johansson. Global convergence of the heavy-ball method for convex optimization. In *2015 European Control Conference (ECC)*, pages 310–315, 2015.

Saeed Ghadimi and Guanghui Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization I: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012.

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

Robert M. Gower, Othmane Sebbouh, and Nicolas Loizou. SGD for structured nonconvex functions: Learning rates, minibatching and interpolation. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130, pages 1315–1323, 2021.

Chuan He and Zhaosong Lu. Accelerated stochastic first-order method for convex optimization under heavy-tailed noise. *arXiv*, https://arxiv.org/abs/2510.11676, 2025.

Florian Hübler, Ilyas Fatkhullin, and Niao He. From gradient clipping to normalization for heavy tailed SGD. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258, pages 2413–2421, 2025.

Keller Jordan, Yuchen Jin, Vlado Boza, You Jiacheng, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks. https://kellerjordan.github.io/posts/muon/, 2024.

Ahmed Khaled and Peter Richtárik. Better theory for SGD in the nonconvex world. *Transactions on Machine Learning Research*, 2023.

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

Yuichi Kondo and Hideaki Iiduka. Accelerating SGDM via learning rate and batch size schedules: A Lyapunov-based analysis. *arXiv*, https://arxiv.org/abs/2508.03105, 2025.

Frederik Kunstner, Jacques Chen, Jonathan Wilder Lavington, and Mark Schmidt. Noise is not the main factor behind the gap between SGD and Adam on Transformers, but sign descent might be. In *International Conference on Representation Learning*, 2023.

Langqi Liu, Yibo Wang, and Lijun Zhang. High-probability bound for non-smooth non-convex stochastic optimization with heavy tails. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 32122–32138, 2024.

Yanli Liu, Yuan Gao, and Wotao Yin. An improved analysis of stochastic gradient descent with momentum. In *Advances in Neural Information Processing Systems*, volume 33, pages 18261–18271, 2020.

Zijian Liu and Zhengyuan Zhou. Nonconvex stochastic optimization under heavy-tailed noises: Optimal convergence without gradient clipping. In *International Conference on Learning Representations*, pages 92529–92554, 2025.

Zijian Liu, Jiawei Zhang, and Zhengyuan Zhou. Breaking the lower bound with (little) structure: Acceleration in non-convex stochastic optimization with heavy-tailed noise. In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195, pages 2266–2290, 2023.

Vien Mai and Mikael Johansson. Convergence of a stochastic gradient method with momentum for non-smooth non-convex optimization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 6630–6639, 2020.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

Yu Nesterov. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 152(1): 381–404, 2015.

Ta Duy Nguyen, Thien H Nguyen, Alina Ene, and Huy 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, 2023a.

Ta Duy Nguyen, Thien Hang Nguyen, Alina Ene, and Huy Le Nguyen. High probability convergence of clipped-SGD under heavy-tailed noise. *arXiv*, `https://arxiv.org/abs/2302.05437`, 2023b.

Iosif Pinelis. Multidimensional probability inequalities via spherical symmetry. *arXiv*, `https://arxiv.org/abs/2210.04391`, 2022.

Boris T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

Junwen Qiu, Bohao Ma, and Andre Milzarek. Convergence of SGD with momentum in the nonconvex case: A time window-based analysis. *arXiv*, `https://arxiv.org/abs/2405.16954`, 2024.

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1571–1578, 2012.

Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of Adam and beyond. In *International Conference on Learning Representations*, 2018.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22:400–407, 1951.

Anton Rodomanov and Yurii Nesterov. Smoothness parameter of power of Euclidean norm. *Journal of Optimization Theory and Applications*, 185(2):303–326, 2020.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by backpropagating errors. *Nature*, 323(6088):533–536, 1986.

Abdurakhmon Sadiev, Marina Danilova, Eduard Gorbunov, Samuel Horváth, Gauthier Gidel, Pavel Dvurechensky, Alexander Gasnikov, and Peter Richtárik. High-probability bounds for stochastic optimization and variational inequalities: the case of unbounded variance. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 29563–29648, 2023.

Othmane Sebbouh, Robert M. Gower, and Aaron Defazio. Almost sure convergence rates for stochastic gradient descent and stochastic heavy ball. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134, pages 3935–3971, 2021.

Umut Simsekli, Levent Sagun, and Mert Gürbüzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 5827–5837, 2019.

Sebastian U. Stich. Unified optimal analysis of the (stochastic) gradient method. *arXiv*, `http://arxiv.org/abs/1907.04232`, 2019.

Bengt von Bahr and Carl-Gustav Esseen. Inequalities for the rth absolute moment of a sum of random variables, $1 \leq r \leq 2$. *Annals of Mathematical Statistics*, 36(1): 299–303, 1965.

Zong-Ben Xu and G.F Roach. Characteristic inequalities of uniformly convex and uniformly smooth Banach spaces. *Journal of Mathematical Analysis and Applications*, 157 (1):189–210, 1991.

Yan Yan, Tianbao Yang, Zhe Li, Qihang Lin, and Yi Yang. Unified convergence analysis of stochastic momentum methods for convex and non-convex optimization. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 2955–2961, 2018.

Maryam Yashtini. On the global convergence rate of the gradient descent method for functions with Hölder continuous gradients. *Optimization Letters*, 10(6):1361–1370, 2016.

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

Juntang Zhuang, Tommy Tang, Yifan Ding, Sekhar Tatikonda, Nicha C. Dvornek, Xenophon Papademetris, and James S. Duncan. AdaBelief optimizer: Adapting stepsizes by the belief in observed gradients. In *Advances in Neural Information Processing Systems*, volume 33, pages 18795–18806, 2020.

Vanilla SGD with Momentum Survives Heavy-Tailed Noise: Convergence Analysis without Gradient Clipping or Normalization (Supplementary Material)

Ryusei Yamada[1 *] **Naoki Sato**[1 *] **Hideaki Iiduka**[1]

[1]Meiji University, Japan
[*]Equal contribution

CONTENTS

Notation. Throughout the appendix, t and T retain the same meaning as in Algorithm 1: $T \in \mathbb{N}$ denotes the total number of iterations and t indexes the corresponding iterates $(\boldsymbol{x}_t)_{t\geq 1}$ and $(\boldsymbol{m}_t)_{t\geq 0}$ generated by Algorithm 1.

A PROOFS OF LEMMAS IN SECTION 2

A.1 PROOF OF LEMMA 2.1

Lemma 2.1 (restated). *Suppose that Assumption 2.1 holds. Then, for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, the following inequality holds.*

$$f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{L}{\nu + 1} \|\boldsymbol{y} - \boldsymbol{x}\|^{\nu+1}.$$

Proof. For all $t \geq 0$, let us define the function $h\colon \mathbb{R} \to \mathbb{R}$ as $h(t) := f(t\boldsymbol{y} + (1-t)\boldsymbol{x}) = f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}))$. Then, we have

$$h'(t) = \langle \nabla f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})), \boldsymbol{y} - \boldsymbol{x} \rangle.$$

From the fundamental theorem of calculus,

$$f(\boldsymbol{y}) - f(\boldsymbol{x}) = h(1) - h(0) = \int_0^1 h'(t) d_t = \int_0^1 \langle \nabla f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})), \boldsymbol{y} - \boldsymbol{x} \rangle dt,$$

which implies

$$\begin{aligned} f(\boldsymbol{y}) - f(\boldsymbol{x}) &= \int_0^1 \langle \nabla f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})), \boldsymbol{y} - \boldsymbol{x} \rangle dt \\ &= \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \int_0^1 \langle \nabla f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})) - \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle dt. \end{aligned}$$

From the Cauchy-Schwarz inequality and Assumption 2.1,

$$\begin{aligned} f(\boldsymbol{y}) - f(\boldsymbol{x}) &\leq \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \int_0^1 \|\nabla f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})) - \nabla f(\boldsymbol{x})\| \|\boldsymbol{y} - \boldsymbol{x}\| dt \\ &\leq \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + L \|\boldsymbol{y} - \boldsymbol{x}\|^{\nu+1} \int_0^1 t^\nu dt. \end{aligned}$$

Here, we have

$$\int_0^1 t^\nu dt = \left[\frac{t^{\nu+1}}{\nu+1} \right]_0^1 = \frac{1}{\nu+1}.$$

Therefore, we find that

$$f(\boldsymbol{y}) - f(\boldsymbol{x}) \leq \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{L}{\nu+1} \|\boldsymbol{y} - \boldsymbol{x}\|^{\nu+1}.$$

This completes the proof. □

A.2 PROOF OF LEMMA 2.2

Lemma 2.2 (restated). *Let $q \in (1, 2]$. For any $\boldsymbol{y} \in \mathbb{R}^d$ and any $\boldsymbol{x} \in \mathbb{R}^d \setminus \{\boldsymbol{0}\}$, the following inequality holds:*

$$\|\boldsymbol{x} \pm \boldsymbol{y}\|^q \leq \|\boldsymbol{x}\|^q \pm q\|\boldsymbol{x}\|^{q-2} \langle \boldsymbol{x}, \boldsymbol{y} \rangle + 2^{2-q} \|\boldsymbol{y}\|^q.$$

Proof. The proof is trivial for $q = 2$. Thus, consider the case of $q \in (1, 2)$. Let us define

$$F(\boldsymbol{y}) := \|\boldsymbol{x}\|^q + q\|\boldsymbol{x}\|^{q-2} \langle \boldsymbol{x}, \boldsymbol{y} \rangle + 2^{2-q} \|\boldsymbol{y}\|^q - \|\boldsymbol{x} + \boldsymbol{y}\|^q.$$

Accordingly, we have

$$\nabla F(\boldsymbol{y}) = q\|\boldsymbol{x}\|^{q-2} \boldsymbol{x} + 2^{2-q} q \|\boldsymbol{y}\|^{q-2} \boldsymbol{y} - q\|\boldsymbol{x} + \boldsymbol{y}\|^{q-2} (\boldsymbol{x} + \boldsymbol{y}).$$

When $\nabla F(\boldsymbol{y}) = \boldsymbol{0}$, the following holds:

$$\left(\|\boldsymbol{x}\|^{q-2} - \|\boldsymbol{x} + \boldsymbol{y}\|^{q-2} \right) \boldsymbol{x} + \left(2^{2-q} \|\boldsymbol{y}\|^{q-2} - \|\boldsymbol{x} + \boldsymbol{y}\|^{q-2} \right) \boldsymbol{y} = \boldsymbol{0}.$$

Therefore, when $\nabla F(\boldsymbol{y}) = \boldsymbol{0}$, $\boldsymbol{x}$ and $\boldsymbol{y}$ are parallel. That is, it suffices to show that $F(\boldsymbol{y}) \geq 0$ holds when $\boldsymbol{x}$ and $\boldsymbol{y}$ are parallel. Then, let $\boldsymbol{y} = t\boldsymbol{x}$ ($t \in \mathbb{R}$). In this case, the inequality to be shown is:

$$|t + 1|^q \leq 1 + qt + 2^{2-q} |t|^q.$$

Here, let us define $g(t) := 1 + qt + 2^{2-q}|t|^q - |t+1|^q$. We will show that $g(t) \geq 0$ for all $t \in \mathbb{R}$. Note that $g(0) = 0$.

Case (i): $t \geq 0$ **i.e.,** $|t| = t$ **and** $|t+1| = t+1$**.** In this case, we have

$$g(t) = 1 + qt + 2^{2-q}t^q - (t+1)^q.$$

Therefore,

$$g'(t) = q\left\{1 + 2^{2-q}t^{q-1} - (t+1)^{q-1}\right\}.$$

Since $f(x) = x^{q-1}$ is a concave function and satisfies $f(0) \geq 0$, and given that any such concave function satisfies $f(x+y) \leq f(x) + f(y)$, we have

$$\begin{aligned} g'(t) &\geq q\left\{1 + 2^{2-q}t^{q-1} - (t^{q-1}+1)\right\} \\ &= q\left(2^{2-q} - 1\right)t^{q-1} \\ &> 0, \end{aligned}$$

where we have used $2^{2-q} > 1$ in the last inequality. Therefore, $g(t)$ is monotone increasing for $t \geq 0$. Since $g(0) = 0$, it follows that $g(t) \geq 0$ for $t \geq 0$.

Case (ii): $-1 \leq t < 0$ **i.e.,** $|t| = -t$ **and** $|t+1| = t+1$**.** In this case, we have

$$g(t) = 1 + qt + 2^{2-q}(-t)^q - (t+1)^q.$$

Therefore,

$$g'(t) = q\left\{1 - 2^{2-q}(-t)^{q-1} - (t+1)^{q-1}\right\}.$$

Here, let us define $u := -t$ ($u \in (0, 1]$). Then,

$$g'(t) = q\{1 - \underbrace{(2^{2-q}u^{q-1} + (1-u)^{q-1})}_{=:\psi(u)}\}.$$

Since $\psi(u)$ is the sum of concave functions, $\psi(u)$ itself is also concave. Therefore, the minimum point of $\psi(u)$ on an interval is an endpoint. From $\lim_{u\to 0}\psi(u) = 1$ and $\psi(1) = 2^{2-q} > 1$, we have $\psi(u) > 1$ and $g'(t) < 0$. Therefore, $g(t)$ is monotone decreasing for $t \in [-1, 0)$. Since $g(-1) = 1 - q + 2^{2-q} > 0$ and $g(0) = 0$, it follows that $g(t) \geq 0$ for $t \in [-1, 0)$.

Case (iii): $t < -1$ **i.e.,** $|t| = -t$ **and** $|t+1| = -(t+1)$**.** In this case, we have

$$g(t) = 1 + qt + 2^{2-q}(-t)^q - (-t-1)^q.$$

Therefore,

$$g'(t) = q\left\{1 - 2^{2-q}(-t)^{q-1} + (-t-1)^{q-1}\right\}.$$

Let us define $v := -t$ ($v > 1$). Here, we have $(-t)^{q-1} = v^{q-1}$ and $(-t-1)^{q-1} = (v-1)^{q-1}$. Accordingly,

$$g'(t) = q\{1 - 2^{2-q}v^{q-1} + (v-1)^{q-1}\}.$$

From $0 < q - 1 < 1$ and Jensen's inequality,

$$\frac{1 + (v-1)^{q-1}}{2} \leq \left\{\frac{1 + (v-1)}{2}\right\}^{q-1} = \frac{v^{q-1}}{2^{q-1}}.$$

Then, we have $1 + (v-1)^{q-1} \leq 2^{2-q}v^{q-1}$, which implies $g'(t) \leq 0$. Therefore, $g(t)$ is monotone decreasing for $t < -1$. Since $g(-1) = 1 - q + 2^{2-q} > 0$, it follows that $g(t) \geq 0$ for $t < -1$.

Therefore, $g(t) \geq 0$ holds for any $t \in \mathbb{R}$, and the proof is complete. □

A.3 PROOF OF LEMMA 2.3

Lemma A.1 (Theorem 3.1 in [Pinelis, 2022]). *Let $\mathfrak{p} \in (1, 2]$, and let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n \in \mathbb{R}^d$ be a sequence of independent random variables with $\mathbb{E}_{\boldsymbol{X}_i}[\boldsymbol{X}_i] = \boldsymbol{0}$ and $\mathbb{E}_{\boldsymbol{X}_i}[\|\boldsymbol{X}_i\|^{\mathfrak{p}}] < \infty$ for all $i \in [n]$. Then,*

$$\mathbb{E}_{(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)}\left[\left\|\sum_{i\in[n]} \boldsymbol{X}_i\right\|^{\mathfrak{p}}\right] \leq C_{\mathfrak{p}} \sum_{i\in[n]} \mathbb{E}_{\boldsymbol{X}_i}\left[\|\boldsymbol{X}_i\|^{\mathfrak{p}}\right],$$

where $C_{\mathfrak{p}}$ is a constant that depends only on $\mathfrak{p}$ and continuously and strictly decreasing in $\mathfrak{p} \in (1, 2]$ from $\lim_{\mathfrak{p}\to 1} C_{\mathfrak{p}} = 2$ to $C_2 = 1$.

Lemma 2.3 (restated). *Suppose that Assumption 2.2 holds. Then, for all $\boldsymbol{x} \in \mathbb{R}^d$ that does not depend on $\boldsymbol{\xi}_t$, the following inequality holds:*

$$\mathbb{E}_{\boldsymbol{\xi}_t}[\|\nabla f_{\mathcal{S}_t}(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\|^{\mathfrak{p}}] \leq \frac{C_{\mathfrak{p}}\sigma^{\mathfrak{p}}}{b^{\mathfrak{p}-1}},$$

where $C_{\mathfrak{p}} \in [1, 2)$ is the constant given in Lemma A.1.

Proof. From Assumption 2.2 and Lemma A.1, we have

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\xi}_t}\left[\|\nabla f_{S_t}(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\|^{\mathfrak{p}}\right] &= \mathbb{E}_{\boldsymbol{\xi}_t}\left[\left\|\frac{1}{b}\sum_{i\in[b]} \mathsf{G}_{\xi_{t,i}}(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\right\|^{\mathfrak{p}}\right] \\
&= \mathbb{E}_{\boldsymbol{\xi}_t}\left[\left\|\frac{1}{b}\sum_{i\in[b]} \left(\mathsf{G}_{\xi_{t,i}}(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\right)\right\|^{\mathfrak{p}}\right] \\
&= \frac{1}{b^{\mathfrak{p}}}\mathbb{E}_{\boldsymbol{\xi}_t}\left[\left\|\sum_{i\in[b]} \left(\mathsf{G}_{\xi_{t,i}}(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\right)\right\|^{\mathfrak{p}}\right] \\
&\leq \frac{C_{\mathfrak{p}}}{b^{\mathfrak{p}}}\sum_{i\in[b]} \mathbb{E}_{\boldsymbol{\xi}_t}\left[\left\|\mathsf{G}_{\xi_{t,i}}(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\right\|^{\mathfrak{p}}\right] \\
&= \frac{C_{\mathfrak{p}}\sigma^{\mathfrak{p}}}{b^{\mathfrak{p}-1}}.
\end{aligned}$$

This completes the proof. □

A.4 PROOF OF LEMMA 2.4

Lemma 2.4 (restated). *Let $\mathfrak{p} \in (1, 2]$ and $\nu \in (0, 1]$, suppose that $\nu + 1 \leq \mathfrak{p}$. Then, for all random vectors $\boldsymbol{X} \in \mathbb{R}^d$, the following inequality holds:*

$$\mathbb{E}_{\boldsymbol{X}}[\|\boldsymbol{X}\|^{\nu+1}] \leq \left(\mathbb{E}_{\boldsymbol{X}}[\|\boldsymbol{X}\|^{\mathfrak{p}}]\right)^{\frac{\nu+1}{\mathfrak{p}}}.$$

Proof. Consider the function $g(x) = x^k$ defined for $x \geq 0$, where $k = \frac{\mathfrak{p}}{\nu+1}$. Since we assumed $\nu + 1 \leq \mathfrak{p}$, it follows that $k \geq 1$. $g(x)$ is a convex function. From Jensen's inequality $g(\mathbb{E}_Y[Y]) \leq \mathbb{E}_Y[g(Y)]$, we have

$$\left(\mathbb{E}_{\boldsymbol{X}}[\|\boldsymbol{X}\|^{\nu+1}]\right)^{\frac{\mathfrak{p}}{\nu+1}} \leq \mathbb{E}_{\boldsymbol{X}}\left[\left(\|\boldsymbol{X}\|^{\nu+1}\right)^{\frac{\mathfrak{p}}{\nu+1}}\right] = \mathbb{E}_{\boldsymbol{X}}\left[\|\boldsymbol{X}\|^{\mathfrak{p}}\right],$$

where we have chosen $Y = \|\boldsymbol{X}\|^{\nu+1}$. Therefore, we have

$$\mathbb{E}_{\boldsymbol{X}}[\|\boldsymbol{X}\|^{\nu+1}] \leq \left(\mathbb{E}_{\boldsymbol{X}}\left[\|\boldsymbol{X}\|^{\mathfrak{p}}\right]\right)^{\frac{\nu+1}{\mathfrak{p}}}.$$

This completes the proof. □

B PROOFS OF LEMMAS AND THEOREMS IN SECTION 3

B.1 PROOF OF LEMMA 3.1

Lemma 3.1 (restated). *Suppose that Assumptions 2.1 and 2.2 hold. Then, for all $\boldsymbol{x} \in \mathbb{R}^d$ that does not depend on $\boldsymbol{\xi}_t$, the following holds:*

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\xi}_t}\left[\|\nabla f_{\mathcal{S}_t}(\boldsymbol{x})\|^{\nu+1}\right] &\leq 2^{1-\nu}\left(\frac{C_{\mathfrak{p}}\sigma^{\mathfrak{p}}}{b^{\mathfrak{p}-1}}\right)^{\frac{\nu+1}{\mathfrak{p}}} + \|\nabla f(\boldsymbol{x})\|^{\nu+1} \\ &\leq 2^{1-\nu}\left(\frac{C_{\mathfrak{p}}\sigma^{\mathfrak{p}}}{b^{\mathfrak{p}-1}}\right)^{\frac{\nu+1}{\mathfrak{p}}} + \frac{\nu+1}{2}\|\nabla f(\boldsymbol{x})\|^2 + \frac{1-\nu}{2}.\end{aligned}$$

Proof. From Lemma 2.2, we have

$$\begin{aligned}\|\nabla f_{\mathcal{S}_t}(\boldsymbol{x})\|^{\nu+1} &= \|\nabla f_{\mathcal{S}_t}(\boldsymbol{x}) - \nabla f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})\|^{\nu+1} \\ &\leq \|\nabla f(\boldsymbol{x})\|^{\nu+1} - (\nu+1)\|\nabla f(\boldsymbol{x})\|^{\nu-1}\langle \nabla f_{\mathcal{S}_t}(\boldsymbol{x}) - \nabla f(\boldsymbol{x}), \nabla f(\boldsymbol{x})\rangle \\ &\quad + 2^{1-\nu}\|\nabla f_{\mathcal{S}_t}(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\|^{\nu+1}.\end{aligned}$$

By taking the expectation,

$$\mathbb{E}_{\boldsymbol{\xi}_t}\left[\|\nabla f_{\mathcal{S}_t}(\boldsymbol{x})\|^{\nu+1}\right] \leq \mathbb{E}_{\boldsymbol{\xi}_t}\left[\|\nabla f(\boldsymbol{x})\|^{\nu+1}\right] + 2^{1-\nu}\mathbb{E}_{\boldsymbol{\xi}_t}\left[\|\nabla f_{\mathcal{S}_t}(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\|^{\nu+1}\right].$$

Here, from Lemmas 2.4 and 2.3, we obtain

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\xi}_t}\left[\|\nabla f_{\mathcal{S}_t}(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\|^{\nu+1}\right] &\leq \left(\mathbb{E}_{\boldsymbol{\xi}_t}\left[\|\nabla f_{\mathcal{S}_t}(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\|^{\mathfrak{p}}\right]\right)^{\frac{\nu+1}{\mathfrak{p}}} \\ &\leq \left(\frac{C_{\mathfrak{p}}\sigma^{\mathfrak{p}}}{b^{\mathfrak{p}-1}}\right)^{\frac{\nu+1}{\mathfrak{p}}}.\end{aligned}$$

Therefore, we have

$$\mathbb{E}_{\boldsymbol{\xi}_t}\left[\|\nabla f_{\mathcal{S}_t}(\boldsymbol{x})\|^{\nu+1}\right] \leq \mathbb{E}_{\boldsymbol{\xi}_t}\left[\|\nabla f(\boldsymbol{x})\|^{\nu+1}\right] + 2^{1-\nu}\left(\frac{C_{\mathfrak{p}}\sigma^{\mathfrak{p}}}{b^{\mathfrak{p}-1}}\right)^{\frac{\nu+1}{\mathfrak{p}}}.$$

In addition, from Young's inequality,

$$\|\nabla f(\boldsymbol{x})\|^{\nu+1} \cdot 1 \leq \frac{\nu+1}{2}\|\nabla f(\boldsymbol{x})\|^2 + \frac{1-\nu}{2}.$$

Therefore,

$$\mathbb{E}_{\boldsymbol{\xi}_t}\left[\|\nabla f_{\mathcal{S}_t}(\boldsymbol{x})\|^{\nu+1}\right] \leq 2^{1-\nu}\left(\frac{C_{\mathfrak{p}}\sigma^{\mathfrak{p}}}{b^{\mathfrak{p}-1}}\right)^{\frac{\nu+1}{\mathfrak{p}}} + \frac{\nu+1}{2}\|\nabla f(\boldsymbol{x})\|^2 + \frac{1-\nu}{2}.$$

This completes the proof. □

B.2 PROOF OF THEOREM 3.1

Theorem 3.1 (restated). *Let the function $f\colon \mathbb{R}^d \to \mathbb{R}$ be μ-strongly convex. Suppose that Assumptions 2.1 and 2.2 hold. With learning rate $\eta_t := \frac{2\nu}{\mu t}$, if $\nu + 1 \leq \mathfrak{p}$ and $\eta_t^\nu \leq \frac{1}{L}$ for all $t \in [T]$, then:*

$$\mathbb{E}\left[f(\boldsymbol{x}_T) - f^\star\right] \leq \frac{E_1}{T^\nu} = \mathcal{O}\left(\frac{1}{T^\nu}\right).$$

In particular, when $\nu + 1 = \mathfrak{p}$,

$$\mathbb{E}\left[f(\boldsymbol{x}_T) - f^\star\right] = \mathcal{O}\left(\frac{1}{T^{\mathfrak{p}-1}}\right),$$

where $E_1 := \max\left\{f(\boldsymbol{x}_1) - f^\star, \frac{E_0 L}{\nu(\nu+1)}\left(\frac{2\nu}{\mu}\right)^{\nu+1}\right\}$ and E_0 defined in Eq. (1) are non-negative constants.

Proof. From Lemma 2.1, we have

$$f(\boldsymbol{x}_{t+1}) \leq f(\boldsymbol{x}_t) + \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}_{t+1} - \boldsymbol{x}_t \rangle + \frac{L}{\nu+1}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\|^{\nu+1}$$
$$= f(\boldsymbol{x}_t) - \eta_t \langle \nabla f(\boldsymbol{x}_t), \nabla f_{\mathcal{S}_t}(\boldsymbol{x}_t) \rangle + \frac{L\eta_t^{\nu+1}}{\nu+1}\|\nabla f_{\mathcal{S}_t}(\boldsymbol{x}_t)\|^{\nu+1}.$$

By taking the expectation, we have

$$\mathbb{E}\left[f(\boldsymbol{x}_{t+1})\right] \leq \mathbb{E}\left[f(\boldsymbol{x}_t)\right] - \eta_t \mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right] + \frac{L\eta_t^{\nu+1}}{\nu+1}\mathbb{E}\left[\|\nabla f_{\mathcal{S}_t}(\boldsymbol{x}_t)\|^{\nu+1}\right].$$

Furthermore, from Lemma 3.1 and $\eta_t^\nu \leq \frac{1}{L}$,

$$\mathbb{E}\left[f(\boldsymbol{x}_{t+1})\right] \leq \mathbb{E}\left[f(\boldsymbol{x}_t)\right] - \eta_t \mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right] + \frac{L\eta_t^{\nu+1}}{\nu+1}\left\{2^{1-\nu}\left(\frac{C_{\mathfrak{p}}\sigma^{\mathfrak{p}}}{b^{\mathfrak{p}-1}}\right)^{\frac{\nu+1}{\mathfrak{p}}} + \frac{\nu+1}{2}\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right] + \frac{1-\nu}{2}\right\}$$
$$= \mathbb{E}\left[f(\boldsymbol{x}_t)\right] - \eta_t\left(1 - \frac{L\eta_t^\nu}{2}\right)\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right] + \frac{L\eta_t^{\nu+1}}{\nu+1}\left\{2^{1-\nu}\left(\frac{C_{\mathfrak{p}}\sigma^{\mathfrak{p}}}{b^{\mathfrak{p}-1}}\right)^{\frac{\nu+1}{\mathfrak{p}}} + \frac{1-\nu}{2}\right\}$$
$$\leq \mathbb{E}\left[f(\boldsymbol{x}_t)\right] - \frac{\eta_t}{2}\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right] + \frac{L}{\nu+1}\underbrace{\left\{2^{1-\nu}\left(\frac{C_{\mathfrak{p}}\sigma^{\mathfrak{p}}}{b^{\mathfrak{p}-1}}\right)^{\frac{\nu+1}{\mathfrak{p}}} + \frac{1-\nu}{2}\right\}}_{=:E_0}\eta_t^{\nu+1}.$$

Here, since f is μ-strongly convex, for all $\boldsymbol{x} \in \mathbb{R}^d$, we have

$$f(\boldsymbol{x}) - f^\star \leq \langle \nabla f(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{x}^\star \rangle - \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{x}^\star\|^2$$
$$\leq \frac{1}{2\mu}\|\nabla f(\boldsymbol{x})\|^2 + \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{x}^\star\|^2 - \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{x}^\star\|^2$$
$$= \frac{1}{2\mu}\|\nabla f(\boldsymbol{x})\|^2,$$

where we have used Young's inequality in the second inequality. Therefore, we have

$$\mathbb{E}\left[f(\boldsymbol{x}_{t+1})\right] \leq \mathbb{E}\left[f(\boldsymbol{x}_t)\right] - \mu\eta_t \mathbb{E}\left[f(\boldsymbol{x_t}) - f^\star\right] + \frac{E_0 L}{\nu+1}\eta_t^{\nu+1}.$$

Subtracting $f^\star$ from both sides yields

$$\mathbb{E}\left[f(\boldsymbol{x}_{t+1})\right] - f^\star \leq (1 - \mu\eta_t)\mathbb{E}\left[f(\boldsymbol{x}_t) - f^\star\right] + \frac{E_0 L}{\nu+1}\eta_t^{\nu+1}.$$

Using this inequality, we prove by induction that

$$\mathbb{E}\left[f(\boldsymbol{x}_t) - f^\star\right] \leq \frac{E_1}{t^\nu}, \text{ where } E_1 := \max\left\{f(\boldsymbol{x}_1) - f^\star, \frac{E_0 L}{\nu(\nu+1)}\left(\frac{2\nu}{\mu}\right)^{\nu+1}\right\}.$$

When $t = 1$, from the definition of E_1, we have $\mathbb{E}\left[f(\boldsymbol{x}_1) - f^\star\right] \leq E_1$. Assume that $\mathbb{E}\left[f(\boldsymbol{x}_t) - f^\star\right] \leq \frac{E_1}{t^\nu}$ holds for some $t \in \mathbb{N}$. Then, from $\eta_t := \frac{2\nu}{\mu t}$, we have

$$\mathbb{E}\left[f(\boldsymbol{x}_{t+1}) - f^\star\right] \leq (1 - \mu\eta_t)\frac{E_1}{t^\nu} + \frac{E_0 L}{\nu+1}\eta_t^{\nu+1}$$
$$= \left(1 - \frac{2\nu}{t}\right)\frac{E_1}{t^\nu} + \frac{E_0 L}{\nu+1}\left(\frac{2\nu}{\mu}\right)^{\nu+1}\frac{1}{t^{\nu+1}}.$$

From the definition of E_1, we have $\frac{E_0 L}{\nu+1}\left(\frac{2\nu}{\mu}\right)^{\nu+1} \leq \nu E_1$. Then,

$$\mathbb{E}\left[f(\boldsymbol{x}_{t+1}) - f^\star\right] \leq \frac{E_1}{t^\nu} - \frac{2\nu E_1}{t^{\nu+1}} + \frac{\nu E_1}{t^{\nu+1}}$$
$$= E_1\left(\frac{1}{t^\nu} - \frac{\nu}{t^{\nu+1}}\right).$$

Here, since the function $g(t) = t^{-\nu}$ ($\nu \in (0, 1]$) is convex, we have $g(t+1) \geq g(t) + g'(t)$, i.e., $\frac{1}{(t+1)^\nu} \geq \frac{1}{t^\nu} - \frac{\nu}{t^{\nu+1}}$. Hence,

$$\mathbb{E}\left[f(\boldsymbol{x}_{t+1}) - f^\star\right] \leq \frac{E_1}{(t+1)^\nu}.$$

Therefore, for all $t \in \mathbb{N}$, $\mathbb{E}\left[f(\boldsymbol{x}_t) - f^\star\right] \leq \frac{E_1}{t^\nu}$ holds. This completes the proof. □

B.3 PROOF OF THEOREM 3.2

Lemma B.1. *Suppose that Assumptions 2.1 and 2.2 hold and f is μ-strongly convex. Let $\Phi_t := f(\boldsymbol{z}_t) - f^\star + A\|\boldsymbol{m}_{t-1}\|^{\nu+1}$. Then, the following inequality holds:*

$$\|\nabla f(\boldsymbol{x}_t)\|^2 \geq E_2\Phi_t - E_2 A\|\boldsymbol{m}_{t-1}\|^{\nu+1} - \frac{(1+2L)E_2}{2(\nu+1)}\left(\frac{\eta\beta}{1-\beta}\right)^{\nu+1}\|\boldsymbol{m}_{t-1}\|^{\nu+1},$$

where $E_2 := \frac{2\mu(\nu+1)}{\nu\left(1+L^{\frac{1}{\nu}}\right)+1}$ is a non-negative constant.

Proof. From Lemma 2.1, we have

$$f(\boldsymbol{z}_t) \leq f(\boldsymbol{x}_t) + \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{z}_t - \boldsymbol{x}_t\rangle + \frac{L}{\nu+1}\|\boldsymbol{z}_t - \boldsymbol{x}_t\|^{\nu+1}.$$

Rearranging the terms and subtracting $f^\star$ from both sides gives

$$f(\boldsymbol{x}_t) - f^\star \geq f(\boldsymbol{z}_t) - f^\star - \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{z}_t - \boldsymbol{x}_t\rangle - \frac{L}{\nu+1}\|\boldsymbol{z}_t - \boldsymbol{x}_t\|^{\nu+1}.$$

Furthermore, multiplying both sides by 2μ yields

$$2\mu(f(\boldsymbol{x}_t) - f^\star) \geq 2\mu(f(\boldsymbol{z}_t) - f^\star) - 2\mu\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{z}_t - \boldsymbol{x}_t\rangle - \frac{2\mu L}{\nu+1}\|\boldsymbol{z}_t - \boldsymbol{x}_t\|^{\nu+1}.$$

Here, from the μ-strong convexity of f and Assumption 2.1, we have, for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$,

$$\mu\|\boldsymbol{x} - \boldsymbol{y}\|^2 \leq \langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y}\rangle \leq \|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|\|\boldsymbol{x} - \boldsymbol{y}\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|^{\nu+1}.$$

That is, we have $\|\boldsymbol{x} - \boldsymbol{y}\|^{1-\nu} \leq \frac{L}{\mu}$ for all $\nu \in (0, 1)$. Therefore, from Assumption 2.1,

$$\|\nabla f(\boldsymbol{x}_t)\|^{\frac{\nu+1}{\nu}} = \|\nabla f(\boldsymbol{x}_t)\|^2\|\nabla f(\boldsymbol{x}_t)\|^{\frac{1-\nu}{\nu}} \leq \|\nabla f(\boldsymbol{x}_t)\|^2 L^{\frac{1-\nu}{\nu}}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^{1-\nu} \leq \frac{L^{\frac{1}{\nu}}}{\mu}\|\nabla f(\boldsymbol{x}_t)\|^2.$$

Note that we arrive at the same result when $\nu = 1$, from $\mu \leq L$. Hence, from the Cauchy-Schwarz and Young's inequality,

$$\begin{aligned}-2\mu\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{z}_t - \boldsymbol{x}_t\rangle &\geq -2\mu\|\nabla f(\boldsymbol{x}_t)\|\|\boldsymbol{z}_t - \boldsymbol{x}_t\| \geq -\mu\left(\frac{\nu}{\nu+1}\|\nabla f(\boldsymbol{x}_t)\|^{\frac{\nu+1}{\nu}} + \frac{1}{\nu+1}\|\boldsymbol{z}_t - \boldsymbol{x}_t\|^{\nu+1}\right)\\ &\geq -\frac{\nu L^{\frac{1}{\nu}}}{\nu+1}\|\nabla f(\boldsymbol{x}_t)\|^2 - \frac{\mu}{\nu+1}\|\boldsymbol{z}_t - \boldsymbol{x}_t\|^{\nu+1}.\end{aligned}$$

Substituting this into the above inequality, we find that

$$2\mu(f(\boldsymbol{x}_t) - f^\star) \geq 2\mu(f(\boldsymbol{z}_t) - f^\star) - \frac{\nu L^{\frac{1}{\nu}}}{\nu+1}\|\nabla f(\boldsymbol{x}_t)\|^2 - \frac{\mu(1+2L)}{\nu+1}\|\boldsymbol{z}_t - \boldsymbol{x}_t\|^{\nu+1}.$$

Since f is μ-strongly convex, the inequality $\|\nabla f(\boldsymbol{x}_t)\|^2 \geq 2\mu(f(\boldsymbol{x}_t) - f^\star)$ holds. Then, we have

$$\|\nabla f(\boldsymbol{x}_t)\|^2 \geq 2\mu(f(\boldsymbol{z}_t) - f^\star) - \frac{\nu L^{\frac{1}{\nu}}}{\nu+1}\|\nabla f(\boldsymbol{x}_t)\|^2 - \frac{\mu(1+2L)}{\nu+1}\|\boldsymbol{z}_t - \boldsymbol{x}_t\|^{\nu+1}.$$

Isolating $\|\nabla f(\boldsymbol{x}_t)\|^2$ and organizing the terms lead to

$$\|\nabla f(\boldsymbol{x}_t)\|^2 \geq \frac{2\mu(\nu+1)}{\nu\left(1+L^{\frac{1}{\nu}}\right)+1}(f(\boldsymbol{z}_t)-f^\star) - \frac{\mu(1+2L)}{\nu\left(1+L^{\frac{1}{\nu}}\right)+1}\|\boldsymbol{z}_t-\boldsymbol{x}_t\|^{\nu+1}.$$

From the definition of Φ_t and $\boldsymbol{z}_t - \boldsymbol{x}_t = -\frac{\eta\beta}{1-\beta}\boldsymbol{m}_{t-1}$,

$$\begin{aligned}\|\nabla f(\boldsymbol{x}_t)\|^2 &\geq \underbrace{\frac{2\mu(\nu+1)}{\nu\left(1+L^{\frac{1}{\nu}}\right)+1}}_{=:E_2}\Phi_t - \frac{2\mu(\nu+1)A}{\nu\left(1+L^{\frac{1}{\nu}}\right)+1}\|\boldsymbol{m}_{t-1}\|^{\nu+1} - \frac{\mu(1+2L)}{\nu\left(1+L^{\frac{1}{\nu}}\right)+1}\left(\frac{\eta\beta}{1-\beta}\right)^{\nu+1}\|\boldsymbol{m}_{t-1}\|^{\nu+1} \\ &\geq E_2\Phi_t - E_2A\|\boldsymbol{m}_{t-1}\|^{\nu+1} - \frac{(1+2L)E_2}{2(\nu+1)}\left(\frac{\eta\beta}{1-\beta}\right)^{\nu+1}\|\boldsymbol{m}_{t-1}\|^{\nu+1}.\end{aligned}$$

This completes the proof. □

Lemma B.2. *Suppose that Assumption 2.2 holds. Then,*

$$\begin{aligned}&\mathbb{E}\left[\|\boldsymbol{m}_t\|^{\nu+1}\right] - \mathbb{E}\left[\|\boldsymbol{m}_{t-1}\|^{\nu+1}\right] \\ &\quad\leq -(1-\beta)\mathbb{E}\left[\|\boldsymbol{m}_{t-1}\|^{\nu+1}\right] + \frac{(1-\beta)(\nu+1)}{2}\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right] + \frac{(1-\beta)(1-\nu)}{2} + 2^{1-\nu}(1-\beta)^{\nu+1}\left(\frac{C_{\mathfrak{p}}\sigma^{\mathfrak{p}}}{b^{\mathfrak{p}-1}}\right)^{\frac{\nu+1}{\mathfrak{p}}}.\end{aligned}$$

Proof. From the definition of $\boldsymbol{m}_t$ and Lemma 2.2, we have

$$\begin{aligned}\|\boldsymbol{m}_t\|^{\nu+1} &= \|\beta\boldsymbol{m}_{t-1} + (1-\beta)\nabla f_{\mathcal{S}_t}(\boldsymbol{x}_t)\|^{\nu+1} \\ &= \|\left(\beta\boldsymbol{m}_{t-1} + (1-\beta)\nabla f(\boldsymbol{x}_t)\right) + (1-\beta)(\nabla f_{\mathcal{S}_t}(\boldsymbol{x}_t) - \nabla f(\boldsymbol{x}_t))\|^{\nu+1} \\ &= \|\beta\boldsymbol{m}_{t-1} + (1-\beta)\nabla f(\boldsymbol{x}_t)\|^{\nu+1} + 2^{1-\nu}(1-\beta)^{\nu+1}\|\nabla f_{\mathcal{S}_t}(\boldsymbol{x}_t) - \nabla f(\boldsymbol{x}_t)\|^{\nu+1} \\ &\quad + (\nu+1)(1-\beta)\|\beta\boldsymbol{m}_{t-1} + (1-\beta)\nabla f(\boldsymbol{x}_t)\|^{\nu-1}\langle\beta\boldsymbol{m}_{t-1} + (1-\beta)\nabla f(\boldsymbol{x}_t), \nabla f_{\mathcal{S}_t}(\boldsymbol{x}_t) - \nabla f(\boldsymbol{x}_t)\rangle.\end{aligned}$$

From Assumption 2.2, we find, by taking the expectation, that

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{m}_t\|^{\nu+1}\right] &= \mathbb{E}\left[\|\beta\boldsymbol{m}_{t-1} + (1-\beta)\nabla f(\boldsymbol{x}_t)\|^{\nu+1}\right] + 2^{1-\nu}(1-\beta)^{\nu+1}\left(\frac{C_{\mathfrak{p}}\sigma^{\mathfrak{p}}}{b^{\mathfrak{p}-1}}\right)^{\frac{\nu+1}{\mathfrak{p}}} \\ &\leq \beta\mathbb{E}\left[\|\boldsymbol{m}_{t-1}\|^{\nu+1}\right] + (1-\beta)\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^{\nu+1}\right] + 2^{1-\nu}(1-\beta)^{\nu+1}\left(\frac{C_{\mathfrak{p}}\sigma^{\mathfrak{p}}}{b^{\mathfrak{p}-1}}\right)^{\frac{\nu+1}{\mathfrak{p}}}.\end{aligned}$$

Subtracting $\mathbb{E}\left[\|\boldsymbol{m}_{t-1}\|^{\nu+1}\right]$ from both sides gives

$$\begin{aligned}&\mathbb{E}\left[\|\boldsymbol{m}_t\|^{\nu+1}\right] - \mathbb{E}\left[\|\boldsymbol{m}_{t-1}\|^{\nu+1}\right] \\ &\quad\leq -(1-\beta)\mathbb{E}\left[\|\boldsymbol{m}_{t-1}\|^{\nu+1}\right] + (1-\beta)\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^{\nu+1}\right] + 2^{1-\nu}(1-\beta)^{\nu+1}\left(\frac{C_{\mathfrak{p}}\sigma^{\mathfrak{p}}}{b^{\mathfrak{p}-1}}\right)^{\frac{\nu+1}{\mathfrak{p}}}.\end{aligned}$$

From Young's inequality, we have

$$\|\nabla f(\boldsymbol{x}_t)\|^{\nu+1}\cdot 1 \leq \frac{\nu+1}{2}\|\nabla f(\boldsymbol{x}_t)\|^2 + \frac{1-\nu}{2}.$$

Therefore, we have

$$\begin{aligned}&\mathbb{E}\left[\|\boldsymbol{m}_t\|^{\nu+1}\right] - \mathbb{E}\left[\|\boldsymbol{m}_{t-1}\|^{\nu+1}\right] \\ &\quad\leq -(1-\beta)\mathbb{E}\left[\|\boldsymbol{m}_{t-1}\|^{\nu+1}\right] + \frac{(1-\beta)(\nu+1)}{2}\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right] + \frac{(1-\beta)(1-\nu)}{2} + 2^{1-\nu}(1-\beta)^{\nu+1}\left(\frac{C_{\mathfrak{p}}\sigma^{\mathfrak{p}}}{b^{\mathfrak{p}-1}}\right)^{\frac{\nu+1}{\mathfrak{p}}}.\end{aligned}$$

This completes the proof. □

Lemma B.3. *Let $\boldsymbol{z}_t := \boldsymbol{x}_t - \frac{\beta\eta}{1-\beta}\boldsymbol{m}_{t-1}$. Suppose that Assumptions 2.1 and 2.2 hold. Then,*

$$\mathbb{E}[f(\boldsymbol{z}_{t+1})] \leq \mathbb{E}[f(\boldsymbol{z}_t)] - \left(\frac{\eta}{2} - \frac{L\eta^{\nu+1}}{2}\right)\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] + \frac{\eta L^2\nu}{\nu+1}\left(\frac{\eta\beta}{1-\beta}\right)^{2\nu}\mathbb{E}\left[\|\boldsymbol{m}_{t-1}\|^{\nu+1}\right] + \frac{\eta(1-\nu)L^2}{2(\nu+1)}\left(\frac{\eta\beta}{1-\beta}\right)^{2\nu} + \frac{L\eta^{\nu+1}}{\nu+1}\left\{2^{1-\nu}\left(\frac{C_{\mathfrak{p}}\sigma^{\mathfrak{p}}}{b^{\mathfrak{p}-1}}\right)^{\frac{\nu+1}{\mathfrak{p}}} + \frac{1-\nu}{2}\right\}.$$

Proof. From the definition of $\boldsymbol{z}_t$, we have

$$\begin{aligned}\boldsymbol{z}_{t+1} &= \boldsymbol{x}_{t+1} - \frac{\eta\beta}{1-\beta}\boldsymbol{m}_t = (\boldsymbol{x}_t - \eta\boldsymbol{m}_t) - \frac{\eta\beta}{1-\beta}\boldsymbol{m}_t = \boldsymbol{x}_t - \frac{\eta}{1-\beta}\boldsymbol{m}_t \\ &= \boldsymbol{x}_t - \frac{\eta}{1-\beta}\left\{\beta\boldsymbol{m}_{t-1} + (1-\beta)\nabla f_{\mathcal{S}_t}(\boldsymbol{x}_t)\right\} = \boldsymbol{z}_t - \eta\nabla f_{\mathcal{S}_t}(\boldsymbol{x}_t).\end{aligned}$$

From Assumption 2.1,

$$\begin{aligned}f(\boldsymbol{z}_{t+1}) &\leq f(\boldsymbol{z}_t) + \langle\nabla f(\boldsymbol{z}_t), \boldsymbol{z}_{t+1} - \boldsymbol{z}_t\rangle + \frac{L}{\nu+1}\|\boldsymbol{z}_{t+1} - \boldsymbol{z}_t\|^{\nu+1} \\ &= f(\boldsymbol{z}_t) - \eta\langle\nabla f(\boldsymbol{z}_t), \nabla f_{\mathcal{S}_t}(\boldsymbol{x}_t)\rangle + \frac{L\eta^{\nu+1}}{\nu+1}\|\nabla f_{\mathcal{S}_t}(\boldsymbol{x}_t)\|^{\nu+1}.\end{aligned}$$

By taking the expectation, we have

$$\mathbb{E}[f(\boldsymbol{z}_{t+1})] \leq \mathbb{E}[f(\boldsymbol{z}_t)] - \eta\mathbb{E}[\langle\nabla f(\boldsymbol{z}_t), \nabla f(\boldsymbol{x}_t)\rangle] + \frac{L\eta^{\nu+1}}{\nu+1}\mathbb{E}[\|\nabla f_{\mathcal{S}_t}(\boldsymbol{x}_t)\|^{\nu+1}].$$

From $\|\boldsymbol{x}-\boldsymbol{y}\|^2 = \|\boldsymbol{x}\|^2 - 2\langle\boldsymbol{x},\boldsymbol{y}\rangle + \|\boldsymbol{y}\|^2$ and Lemma 3.1,

$$\begin{aligned}\mathbb{E}[f(\boldsymbol{z}_{t+1})] &\leq \mathbb{E}[f(\boldsymbol{z}_t)] - \frac{\eta}{2}(\mathbb{E}[\|\nabla f(\boldsymbol{z}_t)\|^2] + \mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] - \mathbb{E}[\|\nabla f(\boldsymbol{z}_t) - \nabla f(\boldsymbol{x}_t)\|^2]) \\ &\quad + \frac{L\eta^{\nu+1}}{\nu+1}\left(2^{1-\nu}\left(\frac{C_{\mathfrak{p}}\sigma^{\mathfrak{p}}}{b^{\mathfrak{p}-1}}\right)^{\frac{\nu+1}{\mathfrak{p}}} + \frac{\nu+1}{2}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] + \frac{1-\nu}{2}\right) \\ &\leq \mathbb{E}[f(\boldsymbol{z}_t)] - \left(\frac{\eta}{2} - \frac{L\eta^{\nu+1}}{2}\right)\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] + \frac{\eta}{2}\mathbb{E}[\|\nabla f(\boldsymbol{z}_t) - \nabla f(\boldsymbol{x}_t)\|^2] \\ &\quad + \frac{L\eta^{\nu+1}}{\nu+1}\left\{2^{1-\nu}\left(\frac{C_{\mathfrak{p}}\sigma^{\mathfrak{p}}}{b^{\mathfrak{p}-1}}\right)^{\frac{\nu+1}{\mathfrak{p}}} + \frac{1-\nu}{2}\right\}.\end{aligned}$$

Here, from Assumption 2.1, the definition of $\boldsymbol{z}_t$, and Young's inequality, we have

$$\begin{aligned}\frac{\eta}{2}\mathbb{E}[\|\nabla f(\boldsymbol{z}_t) - \nabla f(\boldsymbol{x}_t)\|^2] &\leq \frac{\eta}{2}L^2\mathbb{E}[\|\boldsymbol{z}_t - \boldsymbol{x}_t\|^{2\nu}] \\ &= \frac{\eta}{2}L^2\left(\frac{\eta\beta}{1-\beta}\right)^{2\nu}\mathbb{E}\left[\|\boldsymbol{m}_{t-1}\|^{2\nu}\right] \\ &\leq \frac{\eta}{2}L^2\left(\frac{\eta\beta}{1-\beta}\right)^{2\nu}\left\{\frac{2\nu}{\nu+1}\mathbb{E}\left[\|\boldsymbol{m}_{t-1}\|^{\nu+1}\right] + \frac{1-\nu}{\nu+1}\right\} \\ &= \frac{\eta L^2\nu}{\nu+1}\left(\frac{\eta\beta}{1-\beta}\right)^{2\nu}\mathbb{E}\left[\|\boldsymbol{m}_{t-1}\|^{\nu+1}\right] + \frac{\eta(1-\nu)L^2}{2(\nu+1)}\left(\frac{\eta\beta}{1-\beta}\right)^{2\nu}.\end{aligned}$$

Therefore,

$$\mathbb{E}[f(\boldsymbol{z}_{t+1})] \leq \mathbb{E}[f(\boldsymbol{z}_t)] - \left(\frac{\eta}{2} - \frac{L\eta^{\nu+1}}{2}\right)\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] + \frac{\eta L^2\nu}{\nu+1}\left(\frac{\eta\beta}{1-\beta}\right)^{2\nu}\mathbb{E}\left[\|\boldsymbol{m}_{t-1}\|^{\nu+1}\right] + \frac{\eta(1-\nu)L^2}{2(\nu+1)}\left(\frac{\eta\beta}{1-\beta}\right)^{2\nu} + \frac{L\eta^{\nu+1}}{\nu+1}\left\{2^{1-\nu}\left(\frac{C_{\mathfrak{p}}\sigma^{\mathfrak{p}}}{b^{\mathfrak{p}-1}}\right)^{\frac{\nu+1}{\mathfrak{p}}} + \frac{1-\nu}{2}\right\}.$$

This completes the proof. □

Theorem 3.2 (restated). *Let the function $f\colon \mathbb{R}^d \to \mathbb{R}$ be μ-strongly convex. Suppose that Assumptions 2.1 and 2.2 hold. With learning rate $\eta_t := \eta$, if $\nu + 1 \le \mathfrak{p}$ and $\eta \le \min\left\{\frac{2(1-\beta)}{E_2}, \left(\frac{1}{8L}\right)^{\frac{1}{\nu}}, \left(\frac{(1-\beta)^{\nu+1}}{2E_2(1+2L)\beta^{\nu+1}}\right)^{\frac{1}{\nu+1}}, \left(\frac{(1-\beta)^{2\nu}}{2L^2\nu\beta^{2\nu}}\right)^{\frac{1}{2\nu}}\right\}$, then:*

$$\mathbb{E}\left[f(\boldsymbol{x}_T) - f^\star\right] \le \frac{f(\boldsymbol{x}_1) - f^\star}{2}\left(1 - \frac{\eta E_2}{4}\right)^T + \frac{E_3}{1-\beta} = \mathcal{O}\left(\left(1 - \frac{\eta E_2}{4}\right)^T + \eta^{\nu+1}\right).$$

In particular, when $\nu + 1 = \mathfrak{p}$,

$$\mathbb{E}\left[f(\boldsymbol{x}_T) - f^\star\right] = \mathcal{O}\left(\left(1 - \frac{\eta E_2}{4}\right)^T + \eta^{\mathfrak{p}}\right),$$

where $E_2 := \frac{2\mu(\nu+1)}{\nu\left(1+L^{\frac{1}{\nu}}\right)+1}$, $E_3 := \frac{LE_0}{\nu+1} + \frac{(1+2L)E_0E_2}{4(\nu+1)}\left(\frac{\beta}{1-\beta}\right)^{\nu+1}\eta + \frac{2L^2\nu(1-\beta)E_0}{\nu+1}\left(\frac{\beta}{1-\beta}\right)^{2\nu}\eta^\nu$, and E_0 defined in Eq. (1) are non-negative constants.

Proof. For all $t \in \mathbb{N}$, let $\Phi_t := f(\boldsymbol{z}_t) - f^\star + A\|m_{t-1}\|^{\nu+1}$. Then,

$$\mathbb{E}[\Phi_{t+1}] - \mathbb{E}[\Phi_t] = \mathbb{E}[f(z_{t+1})] - \mathbb{E}[f(\boldsymbol{z}_t)] + A(\mathbb{E}[\|m_t\|^{\nu+1}] - \mathbb{E}[\|m_{t-1}\|^{\nu+1}]).$$

From Lemma B.3, we have

$$\begin{aligned}\mathbb{E}[f(\boldsymbol{z}_{t+1})] - \mathbb{E}[f(\boldsymbol{z}_t)] \le &-\left(\frac{\eta}{2} - \frac{L\eta^{\nu+1}}{2}\right)\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] + \frac{\eta L^2\nu}{\nu+1}\left(\frac{\eta\beta}{1-\beta}\right)^{2\nu}\mathbb{E}\left[\|\boldsymbol{m}_{t-1}\|^{\nu+1}\right] \\ &+ \frac{\eta(1-\nu)L^2}{2(\nu+1)}\left(\frac{\eta\beta}{1-\beta}\right)^{2\nu} + \frac{L\eta^{\nu+1}}{\nu+1}\underbrace{\left\{2^{1-\nu}\left(\frac{C_{\mathfrak{p}}\sigma^{\mathfrak{p}}}{b^{\mathfrak{p}-1}}\right)^{\frac{\nu+1}{\mathfrak{p}}} + \frac{1-\nu}{2}\right\}}_{=:E_0}.\end{aligned}$$

Furthermore, from Lemma B.2 and Young's inequality,

$$\begin{aligned}A(\mathbb{E}[\|m_t\|^{\nu+1}] - \mathbb{E}[\|m_{t-1}\|^{\nu+1}]) \le &-A(1-\beta)\mathbb{E}[\|m_{t-1}\|^{\nu+1}] + \frac{A(1-\beta)(\nu+1)}{2}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] \\ &+ \frac{A(1-\beta)(1-\nu)}{2} + 2^{1-\nu}(1-\beta)^{\nu+1}A\left(\frac{C_{\mathfrak{p}}\sigma^{\mathfrak{p}}}{b^{\mathfrak{p}-1}}\right)^{\frac{\nu+1}{\mathfrak{p}}}.\end{aligned}$$

Combining these results,

$$\begin{aligned}\mathbb{E}[\Phi_{t+1}] - \mathbb{E}[\Phi_t] \le &-\left(\frac{\eta}{2} - \frac{L\eta^{\nu+1}}{2}\right)\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] + \frac{\eta L^2\nu}{\nu+1}\left(\frac{\eta\beta}{1-\beta}\right)^{2\nu}\mathbb{E}\left[\|\boldsymbol{m}_{t-1}\|^{\nu+1}\right] \\ &+ \frac{\eta(1-\nu)L^2}{2(\nu+1)}\left(\frac{\eta\beta}{1-\beta}\right)^{2\nu} + \frac{L\eta^{\nu+1}}{\nu+1}E_0 \\ &- A(1-\beta)\mathbb{E}[\|m_{t-1}\|^{\nu+1}] + \frac{A(1-\beta)(\nu+1)}{2}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] \\ &+ \frac{A(1-\beta)(1-\nu)}{2} + 2^{1-\nu}(1-\beta)^{\nu+1}A\left(\frac{C_{\mathfrak{p}}\sigma^{\mathfrak{p}}}{b^{\mathfrak{p}-1}}\right)^{\frac{\nu+1}{\mathfrak{p}}}.\end{aligned}$$

Next, we split the gradient term $-\frac{\eta}{2}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2]$ into two $-\frac{\eta}{4}$ parts and apply the result of Lemma B.1 to the first part,

$$
\begin{aligned}
&\mathbb{E}[\Phi_{t+1}] - \mathbb{E}[\Phi_t] \\
&\leq -\frac{\eta}{4}\left\{E_2\mathbb{E}\left[\Phi_t\right] - E_2 A\mathbb{E}\left[\|\boldsymbol{m}_{t-1}\|^{\nu+1}\right] - \frac{(1+2L)E_2}{2(\nu+1)}\left(\frac{\eta\beta}{1-\beta}\right)^{\nu+1}\mathbb{E}\left[\|\boldsymbol{m}_{t-1}\|^{\nu+1}\right]\right\} \\
&\quad - \left(\frac{\eta}{4} - \frac{L\eta^{\nu+1}}{2} - \frac{A(1-\beta)(1+\nu)}{2}\right)\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] + \left\{\frac{\eta L^2\nu}{\nu+1}\left(\frac{\eta\beta}{1-\beta}\right)^{2\nu} - A(1-\beta)\right\}\mathbb{E}[\|m_{t-1}\|^{\nu+1}] \\
&\quad + \frac{L\eta^{\nu+1}}{\nu+1}E_0 + \frac{\eta(1-\nu)L^2}{2(\nu+1)}\left(\frac{\eta\beta}{1-\beta}\right)^{2\nu} + A\left\{2^{1-\nu}(1-\beta)^{\nu+1}\left(\frac{C_{\mathfrak{p}}\sigma^{\mathfrak{p}}}{b^{\mathfrak{p}-1}}\right)^{\frac{\nu+1}{\mathfrak{p}}} + \frac{(1-\beta)(1-\nu)}{2}\right\} \\
&\leq -\frac{\eta E_2}{4}\mathbb{E}\left[\Phi_t\right] + \left\{\frac{\eta E_2}{4}A + \frac{\eta(1+2L)E_2}{8(\nu+1)}\left(\frac{\eta\beta}{1-\beta}\right)^{\nu+1} + \frac{\eta L^2\nu}{\nu+1}\left(\frac{\eta\beta}{1-\beta}\right)^{2\nu} - A(1-\beta)\right\}\mathbb{E}[\|m_{t-1}\|^{\nu+1}] \\
&\quad - \left(\frac{\eta}{4} - \frac{L\eta^{\nu+1}}{2} - \frac{A(1-\beta)(1+\nu)}{2}\right)\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] + \frac{L\eta^{\nu+1}}{\nu+1}E_0 \\
&\quad + A\left\{2^{1-\nu}(1-\beta)^{\nu+1}\left(\frac{C_{\mathfrak{p}}\sigma^{\mathfrak{p}}}{b^{\mathfrak{p}-1}}\right)^{\frac{\nu+1}{\mathfrak{p}}} + \frac{(1-\beta)(1-\nu)}{2}\right\},
\end{aligned}
$$

where $E_2 := \frac{2\mu(\nu+1)}{\nu\left(1+L\frac{1}{\nu}\right)+1}$. To eliminate the terms involving $\mathbb{E}[\|m_{t-1}\|^{\nu+1}]$, we define the coefficient A as follows:

$$
A = \frac{\frac{\eta(1+2L)E_2}{8(\nu+1)}\left(\frac{\eta\beta}{1-\beta}\right)^{\nu+1} + \frac{\eta L^2\nu}{\nu+1}\left(\frac{\eta\beta}{1-\beta}\right)^{2\nu}}{(1-\beta) - \frac{\eta E_2}{4}}.
$$

Since $\eta \leq \frac{2(1-\beta)}{E_2}$, $A > 0$ always holds and we have

$$
A \leq \frac{\eta(1+2L)E_2}{4(1-\beta)(\nu+1)}\left(\frac{\eta\beta}{1-\beta}\right)^{\nu+1} + \frac{2\eta L^2\nu}{\nu+1}\left(\frac{\eta\beta}{1-\beta}\right)^{2\nu}.
$$

Let us show that the coefficient of $\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right]$ is negative.

$$
\begin{aligned}
\frac{\eta}{4} - \frac{L\eta^{\nu+1}}{2} - \frac{A(1-\beta)(1+\nu)}{2} &\geq \frac{\eta}{4} - \frac{L\eta^{\nu+1}}{2} - \frac{\eta}{8}(1+2L)E_2\left(\frac{\eta\beta}{1-\beta}\right)^{\nu+1} - (1-\beta)\eta L^2\nu\left(\frac{\eta\beta}{1-\beta}\right)^{2\nu} \\
&= \frac{\eta}{8}\left\{2 - 4L\eta^{\nu} - (1+2L)E_2\left(\frac{\beta}{1-\beta}\right)^{\nu+1}\eta^{\nu+1} - (1-\beta)L^2\nu\left(\frac{\beta}{1-\beta}\right)^{2\nu}\eta^{2\nu}\right\}.
\end{aligned}
$$

Since $\eta \leq \min\left\{\left(\frac{1}{8L}\right)^{\frac{1}{\nu}}, \left[\frac{(1-\beta)^{\nu+1}}{2E_2(1+2L)\beta^{\nu+1}}\right]^{\frac{1}{\nu+1}}, \left[\frac{(1-\beta)^{2\nu}}{2L^2\nu\beta^{2\nu}}\right]^{\frac{1}{2\nu}}\right\}$, we have

$$
\eta^{\nu} \leq \frac{1}{8L},\ \eta^{\nu+1} \leq \frac{1}{2E_2(1+2L)}\left(\frac{1-\beta}{\beta}\right)^{\nu+1},\ \text{and}\ \eta^{2\nu} \leq \frac{1}{2L^2\nu(1-\beta)}\left(\frac{1-\beta}{\beta}\right)^{2\nu}.
$$

Hence,

$$
\frac{\eta}{4} - \frac{L\eta^{\nu+1}}{2} - \frac{A(1-\beta)(1+\nu)}{2} \geq \frac{\eta}{8}\left(2 - \frac{1}{2} - \frac{1}{2} - \frac{1}{2}\right) = \frac{\eta}{16}.
$$

Therefore, we have

$$\begin{aligned}\mathbb{E}[\Phi_{t+1}] &\leq \left(1-\frac{\eta E_2}{4}\right)\mathbb{E}[\Phi_t] + \frac{L\eta^{\nu+1}}{\nu+1}E_0 + A\left\{2^{1-\nu}(1-\beta)^{\nu+1}\left(\frac{C_{\mathfrak{p}}\sigma^{\mathfrak{p}}}{b^{\mathfrak{p}-1}}\right)^{\frac{\nu+1}{\mathfrak{p}}} + \frac{(1-\beta)(1-\nu)}{2}\right\} \\ &\leq \left(1-\frac{\eta E_2}{4}\right)\mathbb{E}[\Phi_t] + \frac{L\eta^{\nu+1}}{\nu+1}E_0 + A(1-\beta)E_0 \\ &\leq \underbrace{\left(1-\frac{\eta E_2}{4}\right)}_{=:\rho}\mathbb{E}[\Phi_t] + \underbrace{\left\{\frac{LE_0}{\nu+1} + \frac{(1+2L)E_0E_2}{4(\nu+1)}\left(\frac{\beta}{1-\beta}\right)^{\nu+1}\eta + \frac{2L^2\nu(1-\beta)E_0}{\nu+1}\left(\frac{\beta}{1-\beta}\right)^{2\nu}\eta^{\nu}\right\}}_{=:E_3}\eta^{\nu+1}.\end{aligned}$$

Since $\eta \leq \frac{2(1-\beta)}{E_2}$, we have $|\rho| < 1$. Then,

$$\mathbb{E}\left[\Phi_{t+1}\right] \leq \rho^t \mathbb{E}\left[\Phi_1\right] + \frac{E_3}{1-\rho} = \rho^{t+1}\cdot\frac{f(\boldsymbol{x}_1)-f^\star}{\rho} + \frac{E_3\eta^{\nu+1}}{1-\rho}.$$

Hence,

$$\mathbb{E}\left[f(\boldsymbol{z}_t)-f^\star\right] \leq \rho^t\cdot\frac{f(\boldsymbol{x}_1)-f^\star}{\rho} + \frac{E_3\eta^{\nu+1}}{1-\rho}.$$

Here, we can expand $\boldsymbol{m}_{t-1} = (1-\beta)\sum_{k=1}^{t-1}\beta^{t-1-k}\nabla f_{\mathcal{S}_k}(\boldsymbol{x}_k)$. Then, by the definition of $\boldsymbol{z}_t$,

$$\boldsymbol{x}_t = \boldsymbol{z}_t + \frac{\eta\beta}{1-\beta}\boldsymbol{m}_{t-1} = \boldsymbol{z}_t + \eta\sum_{k=1}^{t-1}\beta^{t-k}\nabla f_{\mathcal{S}_k}(\boldsymbol{x}_k).$$

Substituting $\eta\nabla f_{\mathcal{S}_k}(\boldsymbol{x}_k) = \boldsymbol{z}_k - \boldsymbol{z}_{k+1}$ gives

$$\begin{aligned}\boldsymbol{x}_t &= \boldsymbol{z}_t + \sum_{k=1}^{t-1}\beta^{t-k}(\boldsymbol{z}_k - \boldsymbol{z}_{k+1}) \\ &= \boldsymbol{z}_t + \left(\sum_{k=1}^{t-1}\beta^{t-k}\boldsymbol{z}_k - \sum_{k=1}^{t-1}\beta^{t-k}\boldsymbol{z}_{k+1}\right) \\ &= \boldsymbol{z}_t + \beta^{t-1}\boldsymbol{z}_1 + \sum_{k=2}^{t-1}\beta^{t-k}\boldsymbol{z}_k - \left(\sum_{k=2}^{t-1}\beta^{t-k+1}\boldsymbol{z}_k + \beta\boldsymbol{z}_t\right) \\ &= (1-\beta)\boldsymbol{z}_t + \sum_{k=2}^{t-1}\beta^{t-k}(1-\beta)\boldsymbol{z}_k + \beta^{t-1}\boldsymbol{z}_1.\end{aligned}$$

Finally, we have

$$\begin{aligned}(1-\beta) + (1-\beta)\sum_{k=2}^{t-1}\beta^{t-k} + \beta^{t-1} &= (1-\beta) + (1-\beta)\frac{\beta(1-\beta^{t-2})}{1-\beta} + \beta^{t-1} \\ &= 1-\beta+\beta-\beta^{t-1}+\beta^{t-1} \\ &= 1.\end{aligned}$$

Since $\beta \in [0,1)$, $\boldsymbol{x}_t$ can be expressed as a convex combination with the auxiliary sequence $\{\boldsymbol{z}_k\}_{k=1}^t$. From the convexity of f, we have

$$\mathbb{E}\left[f(\boldsymbol{x}_t)-f^\star\right] = \mathbb{E}\left[f\left(\sum_{k=1}^t w_k\boldsymbol{z}_k\right)-f^\star\right] \leq \sum_{k=1}^t w_k\mathbb{E}\left[f(\boldsymbol{z}_k)-f^\star\right] \leq \frac{f(\boldsymbol{x}_1)-f^\star}{\rho}\sum_{k=1}^t w_k\rho^k + \frac{E_3\eta^{\nu+1}}{1-\rho},$$

where $w_1 := \beta^{t-1}$, $w_t := 1-\beta$, and $w_k := (1-\beta)\beta^{t-k}$ $(1 < k < t)$. Since $\eta \leq \frac{2(1-\beta)}{E_2}$, we have $\beta < \rho$. Then,

$$\sum_{k=1}^t w_k\rho^k = \beta^{t-1}\rho + (1-\beta)\sum_{k=2}^t\beta^{t-k}\rho^k = \beta^{t-1}\rho + (1-\beta)\rho^t\sum_{j=0}^{t-2}\left(\frac{\beta}{\rho}\right)^j \leq \frac{1-\beta}{\rho-\beta}\rho^{t+1}.$$

Finally, we have

$$\begin{aligned}\mathbb{E}\left[f(\boldsymbol{x}_t)-f^\star\right] &\leq \frac{(1-\beta)(f(\boldsymbol{x}_1)-f^\star)}{\rho-\beta}\rho^t + \frac{E_3\eta^{\nu+1}}{1-\beta} \\ &\leq \frac{f(\boldsymbol{x}_1)-f^\star}{2}\left(1-\frac{\eta E_2}{4}\right)^t + \frac{E_3\eta^{\nu+1}}{1-\beta} \\ &= \mathcal{O}\left(\left(1-\frac{\eta E_2}{4}\right)^t + \eta^{\nu+1}\right),\end{aligned}$$

where $E_2 := \frac{2\mu(\nu+1)}{\nu\left(1+L^{\frac{1}{\nu}}\right)+1}$ and $E_3 := \frac{LE_0}{\nu+1} + \frac{(1+2L)E_0E_2}{4(\nu+1)}\left(\frac{\beta}{1-\beta}\right)^{\nu+1}\eta + \frac{2L^2\nu(1-\beta)E_0}{\nu+1}\left(\frac{\beta}{1-\beta}\right)^{2\nu}\eta^\nu$. This completes the proof. □

B.4 PROOF OF THEOREM 3.3

Lemma B.4. *Let $\eta_t := \eta_{\max}t^{-\frac{1}{\nu+1}}$ be the learning rate schedule. Then,*

$$\frac{1}{\sum_{t=1}^T \eta_t} \leq \frac{\frac{\nu}{\nu+1}}{\eta_{\max}\left\{(T+1)^{\frac{\nu}{\nu+1}}-1\right\}} = \mathcal{O}\left(\frac{1}{T^{\frac{\nu}{\nu+1}}}\right),$$

and

$$\frac{\sum_{t=1}^T \eta_t^{\nu+1}}{\sum_{t=1}^T \eta_t} \leq \frac{\frac{\nu}{\nu+1}\left(1+\log T\right)}{(T+1)^{\frac{\nu}{\nu+1}}-1} = \mathcal{O}\left(\frac{\log T}{T^{\frac{\nu}{\nu+1}}}\right).$$

Proof. Here, we have

$$\sum_{t=1}^T \eta_t = \eta_{\max}\sum_{t=1}^T t^{-\frac{1}{\nu+1}} \geq \eta_{\max}\int_1^{T+1} t^{-\frac{1}{\nu+1}} = \eta_{\max}\frac{(T+1)^{1-\frac{1}{\nu+1}}-1}{1-\frac{1}{\nu+1}} = \eta_{\max}\frac{(T+1)^{\frac{\nu}{\nu+1}}-1}{\frac{\nu}{\nu+1}}.$$

Therefore,

$$\frac{1}{\sum_{t=1}^T \eta_t} \leq \frac{\frac{\nu}{\nu+1}}{\eta_{\max}\left\{(T+1)^{\frac{\nu}{\nu+1}}-1\right\}}.$$

In addition,

$$\sum_{t=1}^T \eta_t^{\nu+1} = \eta_{\max}\sum_{t=1}^T t^{-1} \leq \eta_{\max}\left(1+\log T\right).$$

This completes the proof. □

Theorem 3.3 (restated). *Let the function $f\colon \mathbb{R}^d \to \mathbb{R}$ be convex. Suppose that Assumptions 2.1, 2.2, and 3.1(i) hold. With learning rate $\eta_t := \eta_{\max}t^{-\frac{1}{\nu+1}}$, if $\nu+1 \leq \mathfrak{p}$, then:*

$$\min_{1\leq t\leq T}\mathbb{E}\left[f(\boldsymbol{x}_t)-f^\star\right] \leq \frac{\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|^{\nu+1}}{(\nu+1)D_1^{\nu-1}}\frac{1}{\sum_{t=1}^T\eta_t} + E_4\frac{\sum_{t=1}^T\eta_t^{\nu+1}}{\sum_{t=1}^T\eta_t} = \mathcal{O}\left(\frac{\log T}{T^{\frac{\nu}{\nu+1}}}\right).$$

In particular, when $\nu+1=\mathfrak{p}$,

$$\min_{1\leq t\leq T}\mathbb{E}\left[f(\boldsymbol{x}_t)-f^\star\right] = \mathcal{O}\left(\frac{\log T}{T^{\frac{\mathfrak{p}-1}{\mathfrak{p}}}}\right),$$

where $E_4 := 2^{1-\nu}\left(E_0 + \frac{\nu+1}{2}L^2D_1^{2\nu}\right)$ and E_0 defined in Eq. (1) are non-negative constants.

Proof. From Lemma 2.2 and Assumption 3.1(i), we have

$$\begin{aligned}\|\boldsymbol{x}_{t+1}-\boldsymbol{x}^\star\|^{\nu+1} &= \|\boldsymbol{x}_t-\eta_t\nabla f_{\mathcal{S}_t}(\boldsymbol{x}_t)-\boldsymbol{x}^\star\|^{\nu+1}\\ &\le \|\boldsymbol{x}_t-\boldsymbol{x}^\star\|^{\nu+1}-\eta_t(\nu+1)\|\boldsymbol{x}_t-\boldsymbol{x}^\star\|^{\nu-1}\langle \boldsymbol{x}_t-\boldsymbol{x}^\star,\nabla f_{\mathcal{S}_t}(\boldsymbol{x}_t)\rangle + 2^{1-\nu}\|\nabla f_{\mathcal{S}_t}(\boldsymbol{x}_t)\|^{\nu+1}.\end{aligned}$$

Taking the conditional expectation of $\boldsymbol{x}_t$, we have, from the convexity of f and Assumption 3.1(i),

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{x}_{t+1}-\boldsymbol{x}^\star\|^{\nu+1}|\boldsymbol{x}_t\right] &\le \|\boldsymbol{x}_t-\boldsymbol{x}^\star\|^{\nu+1}-\eta_t(\nu+1)\|\boldsymbol{x}_t-\boldsymbol{x}^\star\|^{\nu-1}\langle \boldsymbol{x}_t-\boldsymbol{x}^\star,\nabla f(\boldsymbol{x}_t)\rangle + 2^{1-\nu}\mathbb{E}\left[\|\nabla f_{\mathcal{S}_t}(\boldsymbol{x}_t)\|^{\nu+1}|\boldsymbol{x}_t\right]\\ &\le \|\boldsymbol{x}_t-\boldsymbol{x}^\star\|^{\nu+1}-\eta_t(\nu+1)\|\boldsymbol{x}_t-\boldsymbol{x}^\star\|^{\nu-1}(f(\boldsymbol{x}_t)-f^\star) + 2^{1-\nu}\mathbb{E}\left[\|\nabla f_{\mathcal{S}_t}(\boldsymbol{x}_t)\|^{\nu+1}|\boldsymbol{x}_t\right]\\ &\le \|\boldsymbol{x}_t-\boldsymbol{x}^\star\|^{\nu+1}-\eta_t(\nu+1)D_1^{\nu-1}(f(\boldsymbol{x}_t)-f^\star) + 2^{1-\nu}\mathbb{E}\left[\|\nabla f_{\mathcal{S}_t}(\boldsymbol{x}_t)\|^{\nu+1}|\boldsymbol{x}_t\right].\end{aligned}$$

By taking the total expectation, we have

$$\mathbb{E}\left[\|\boldsymbol{x}_{t+1}-\boldsymbol{x}^\star\|^{\nu+1}\right] \le \mathbb{E}\left[\|\boldsymbol{x}_t-\boldsymbol{x}^\star\|^{\nu+1}\right]-\eta_t(\nu+1)D_1^{\nu-1}(f(\boldsymbol{x}_t)-f^\star) + 2^{1-\nu}\eta_t^{\nu+1}\mathbb{E}\left[\|\nabla f_{\mathcal{S}_t}(\boldsymbol{x}_t)\|^{\nu+1}\right].$$

From Lemma 3.1, we obtain

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{x}_{t+1}-\boldsymbol{x}^\star\|^{\nu+1}\right] &\le \mathbb{E}\left[\|\boldsymbol{x}_t-\boldsymbol{x}^\star\|^{\nu+1}\right]-\eta_t(\nu+1)D_1^{\nu-1}\mathbb{E}\left[f(\boldsymbol{x}_t)-f^\star\right]\\ &\quad + 2^{1-\nu}\left\{2^{1-\nu}\left(\frac{C_{\mathfrak{p}}\sigma^{\mathfrak{p}}}{b^{\mathfrak{p}-1}}\right)^{\frac{\nu+1}{\mathfrak{p}}} + \frac{\nu+1}{2}\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right] + \frac{1-\nu}{2}\right\}\eta_t^{\nu+1}.\end{aligned}$$

Here, from Assumptions 2.1 and 3.1(i), we have $\|\nabla f(\boldsymbol{x}_t)-\nabla f(\boldsymbol{x}^\star)\| \le L\|\boldsymbol{x}_t-\boldsymbol{x}^\star\|^\nu = LD_1^\nu$, i.e., $\|\nabla f(\boldsymbol{x}_t)\|^2 \le L^2D_1^{2\nu}$. Hence,

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{x}_{t+1}-\boldsymbol{x}^\star\|^{\nu+1}\right] &\le \mathbb{E}\left[\|\boldsymbol{x}_t-\boldsymbol{x}^\star\|^{\nu+1}\right]-\eta_t(\nu+1)D_1^{\nu-1}\mathbb{E}\left[f(\boldsymbol{x}_t)-f^\star\right]\\ &\quad + \underbrace{2^{1-\nu}\left\{2^{1-\nu}\left(\frac{C_{\mathfrak{p}}\sigma^{\mathfrak{p}}}{b^{\mathfrak{p}-1}}\right)^{\frac{\nu+1}{\mathfrak{p}}} + \frac{\nu+1}{2}L^2D_1^{2\nu} + \frac{1-\nu}{2}\right\}}_{=:E_4}\eta_t^{\nu+1}\\ &\le \mathbb{E}\left[\|\boldsymbol{x}_t-\boldsymbol{x}^\star\|^{\nu+1}\right]-\eta_t(\nu+1)D_1^{\nu-1}\mathbb{E}\left[f(\boldsymbol{x}_t)-f^\star\right]+E_4\eta_t^{\nu+1}.\end{aligned}$$

Rearranging the terms,

$$\eta_t\mathbb{E}\left[f(\boldsymbol{x}_t)-f^\star\right] \le \frac{1}{(\nu+1)D_1^{\nu-1}}\left(\mathbb{E}\left[\|\boldsymbol{x}_t-\boldsymbol{x}^\star\|^{\nu+1}\right]-\mathbb{E}\left[\|\boldsymbol{x}_{t+1}-\boldsymbol{x}^\star\|^{\nu+1}\right]\right)+\frac{E_4}{(\nu+1)D_1^{\nu-1}}\eta_t^{\nu+1},$$

and summing over $t=1$ to $t=T$, we have

$$\sum_{t=1}^T\eta_t\mathbb{E}\left[f(\boldsymbol{x}_t)-f^\star\right] \le \frac{\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|^{\nu+1}}{(\nu+1)D_1^{\nu-1}}+\frac{E_4}{(\nu+1)D_1^{\nu-1}}\sum_{t=1}^T\eta_t^{\nu+1}.$$

Hence,

$$\min_{1\le t\le T}\mathbb{E}\left[f(\boldsymbol{x}_t)-f^\star\right] \le \frac{\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|^{\nu+1}}{(\nu+1)D_1^{\nu-1}}\frac{1}{\sum_{t=1}^T\eta_t}+\frac{E_4}{(\nu+1)D_1^{\nu-1}}\frac{\sum_{t=1}^T\eta_t^{\nu+1}}{\sum_{t=1}^T\eta_t}.$$

Since $\eta_t := \eta_{\max}t^{-\frac{1}{\nu+1}}$, from Lemma B.4, this completes the proof. □

B.5 PROOF OF THEOREM 3.4

Theorem 3.4 (restated). *Let the function $f\colon\mathbb{R}^d\to\mathbb{R}$ be convex. Suppose that Assumptions 2.1, 2.2, 3.1(i), and 3.1(ii) hold. With learning rate $\eta_t := \eta$, if $\nu+1\le\mathfrak{p}$, then:*

$$\min_{1\le t\le T}f(\boldsymbol{x}_t)-f^\star \le \frac{\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|^{\nu+1}}{\eta(\nu+1)D_2^{\nu-1}T}+E_6\eta+\frac{2^{1-\nu}E_5}{(\nu+1)D_2^{\nu-1}}\eta^\nu = \mathcal{O}\left(\frac{1}{\eta T}+\eta^\nu\right).$$

In particular, when $\nu + 1 = \mathfrak{p}$*,*

$$\min_{1 \le t \le T} f(\boldsymbol{x}_t) - f^\star = \mathcal{O}\left(\frac{1}{\eta T} + \eta^{\mathfrak{p}-1}\right),$$

where $E_5 := 2^{1-\nu}\left(\frac{C_{\mathfrak{p}}\sigma^{\mathfrak{p}}}{b^{\mathfrak{p}-1}}\right)^{\frac{\nu+1}{\mathfrak{p}}} + L^{\nu+1}D_1^{\nu(\nu+1)}$*,* $E_6 := \frac{\beta\left(E_5 + \nu L^{\frac{\nu+1}{\nu}} D_1^{\nu+1}\right)}{(1-\beta)(\nu+1)}$ *are non-negative constants.*

Proof. For all $t \in \mathbb{N}$, let $w_t := \|\boldsymbol{z}_t - \boldsymbol{x}^\star\|^{\nu-1}$. From Lemma 2.2,

$$\begin{aligned}
\|\boldsymbol{z}_{t+1} - \boldsymbol{x}^\star\|^{\nu+1} &= \|\boldsymbol{z}_t - \boldsymbol{x}^\star - \eta\nabla f_{\mathcal{S}_t}(\boldsymbol{x}_t)\|^{\nu+1} \\
&\le \|\boldsymbol{z}_t - \boldsymbol{x}^\star\|^{\nu+1} - \eta(\nu+1)w_t\langle \boldsymbol{z}_t - \boldsymbol{x}^\star, \nabla f_{\mathcal{S}_t}(\boldsymbol{x}_t)\rangle + 2^{1-\nu}\eta^{\nu+1}\|\nabla f_{\mathcal{S}_t}(\boldsymbol{x}_t)\|^{\nu+1} \\
&= \|\boldsymbol{z}_t - \boldsymbol{x}^\star\|^{\nu+1} - \eta(\nu+1)w_t\langle \boldsymbol{x}_t - \boldsymbol{x}^\star, \nabla f_{\mathcal{S}_t}(\boldsymbol{x}_t)\rangle \\
&\quad + \eta(\nu+1)w_t\langle \boldsymbol{x}_t - \boldsymbol{z}_t, \nabla f_{\mathcal{S}_t}(\boldsymbol{x}_t)\rangle + 2^{1-\nu}\eta^{\nu+1}\|\nabla f_{\mathcal{S}_t}(\boldsymbol{x}_t)\|^{\nu+1} \\
&= \|\boldsymbol{z}_t - \boldsymbol{x}^\star\|^{\nu+1} - \eta(\nu+1)w_t\langle \boldsymbol{x}_t - \boldsymbol{x}^\star, \nabla f_{\mathcal{S}_t}(\boldsymbol{x}_t)\rangle \\
&\quad + \frac{\eta^2\beta(\nu+1)}{1-\beta}w_t\langle \boldsymbol{m}_{t-1}, \nabla f_{\mathcal{S}_t}(\boldsymbol{x}_t)\rangle + 2^{1-\nu}\eta^{\nu+1}\|\nabla f_{\mathcal{S}_t}(\boldsymbol{x}_t)\|^{\nu+1}.
\end{aligned}$$

Taking the conditional expectation of $\boldsymbol{x}_t$, we have, from the convexity of f,

$$\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{z}_{t+1} - \boldsymbol{x}^\star\|^{\nu+1}|\boldsymbol{x}_t\right] &\le \|\boldsymbol{z}_t - \boldsymbol{x}^\star\|^{\nu+1} - \eta(\nu+1)w_t(f(\boldsymbol{x}_t) - f^\star) \\
&\quad + \frac{\eta^2\beta(\nu+1)}{1-\beta}w_t\langle \boldsymbol{m}_{t-1}, \nabla f(\boldsymbol{x}_t)\rangle + 2^{1-\nu}\eta^{\nu+1}\mathbb{E}\left[\|\nabla f_{\mathcal{S}_t}(\boldsymbol{x}_t)\|^{\nu+1}|\boldsymbol{x}_t\right].
\end{aligned}$$

Moreover, from the Cauchy-Schwarz inequality, Young's inequality, and Assumptions 2.1 and 3.1(i),

$$\langle \boldsymbol{m}_{t-1}, \nabla f(\boldsymbol{x}_t)\rangle \le \|\boldsymbol{m}_{t-1}\|\|\nabla f(\boldsymbol{x}_t)\| \le \frac{\|\boldsymbol{m}_{t-1}\|^{\nu+1}}{\nu+1} + \frac{\nu\|\nabla f(\boldsymbol{x}_t)\|^{\frac{\nu+1}{\nu}}}{\nu+1} \le \frac{\|\boldsymbol{m}_{t-1}\|^{\nu+1}}{\nu+1} + \frac{\nu L^{\frac{\nu+1}{\nu}} D_1^{\nu+1}}{\nu+1}.$$

In addition, from Lemma 3.1 and the convexity of $\|\cdot\|^{\nu+1}$,

$$\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{m}_{t-1}\|^{\nu+1}\right] &\le \beta\mathbb{E}\left[\|\boldsymbol{m}_{t-1}\|^{\nu+1}\right] + (1-\beta)\mathbb{E}\left[\|\nabla f_{\mathcal{S}_t}(\boldsymbol{x}_t)\|^{\nu+1}\right] \\
&\le \beta\mathbb{E}\left[\|\boldsymbol{m}_{t-1}\|^{\nu+1}\right] + (1-\beta)\left\{2^{1-\nu}\left(\frac{C_{\mathfrak{p}}\sigma^{\mathfrak{p}}}{b^{\mathfrak{p}-1}}\right)^{\frac{\nu+1}{\mathfrak{p}}} + \mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^{\nu+1}\right]\right\} \\
&\le \beta\mathbb{E}\left[\|\boldsymbol{m}_{t-1}\|^{\nu+1}\right] + (1-\beta)\left\{2^{1-\nu}\left(\frac{C_{\mathfrak{p}}\sigma^{\mathfrak{p}}}{b^{\mathfrak{p}-1}}\right)^{\frac{\nu+1}{\mathfrak{p}}} + L^{\nu+1}D_1^{\nu(\nu+1)}\right\} \\
&\le \underbrace{2^{1-\nu}\left(\frac{C_{\mathfrak{p}}\sigma^{\mathfrak{p}}}{b^{\mathfrak{p}-1}}\right)^{\frac{\nu+1}{\mathfrak{p}}} + L^{\nu+1}D_1^{\nu(\nu+1)}}_{=:E_5}.
\end{aligned}$$

Therefore, by taking the expectation, we have

$$\mathbb{E}\left[\|\boldsymbol{z}_{t+1} - \boldsymbol{x}^\star\|^{\nu+1}\right] \le \mathbb{E}\left[\|\boldsymbol{z}_t - \boldsymbol{x}^\star\|^{\nu+1}\right] - \eta(\nu+1)w_t(f(\boldsymbol{x}_t) - f^\star) + \frac{\eta^2\beta w_t}{1-\beta}\left(E_5 + \nu L^{\frac{\nu+1}{\nu}} D_1^{\nu+1}\right) + 2^{1-\nu}\eta^{\nu+1}E_5.$$

Rearranging the terms,

$$w_t(f(\boldsymbol{x}_t) - f^\star) \le \frac{\mathbb{E}\left[\|\boldsymbol{z}_t - \boldsymbol{x}^\star\|^{\nu+1}\right] - \mathbb{E}\left[\|\boldsymbol{z}_{t+1} - \boldsymbol{x}^\star\|^{\nu+1}\right]}{\eta(\nu+1)} + \frac{\beta\left(E_5 + \nu L^{\frac{\nu+1}{\nu}} D_1^{\nu+1}\right)}{(1-\beta)(\nu+1)}\eta w_t + \frac{2^{1-\nu}E_5}{\nu+1}\eta^\nu,$$

and summing over $t = 1$ to $t = T$, we obtain

$$\sum_{t=1}^{T} w_t(f(\boldsymbol{x}_t) - f^\star) \le \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^{\nu+1}}{\eta(\nu+1)} + \frac{\beta\left(E_5 + \nu L^{\frac{\nu+1}{\nu}} D_1^{\nu+1}\right)}{(1-\beta)(\nu+1)}\eta\sum_{t=1}^{T} w_t + \frac{2^{1-\nu}E_5}{\nu+1}\eta^\nu T.$$

Hence,

$$\min_{1\le t\le T} f(\boldsymbol{x}_t) - f^\star \le \frac{\sum_{t=1}^T w_t(f(\boldsymbol{x}_t)-f^\star)}{\sum_{t=1}^T w_t} \le \frac{\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|^{\nu+1}}{\eta(\nu+1)\sum_{t=1}^T w_t} + \frac{\beta\left(E_5+\nu L^{\frac{\nu+1}{\nu}}D_1^{\nu+1}\right)}{(1-\beta)(\nu+1)}\eta + \frac{2^{1-\nu}E_5}{(\nu+1)\sum_{t=1}^T w_t}\eta^\nu T.$$

Here, from Assumption 3.1(ii), $w_t := \|\boldsymbol{z}_t-\boldsymbol{x}^\star\|^{\nu-1} \ge D_2^{\nu-1}$. Therefore,

$$\begin{aligned}\min_{1\le t\le T} f(\boldsymbol{x}_t) - f^\star &\le \frac{\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|^{\nu+1}}{\eta(\nu+1)D_2^{\nu-1}T} + \underbrace{\frac{\beta\left(E_5+\nu L^{\frac{\nu+1}{\nu}}D_1^{\nu+1}\right)}{(1-\beta)(\nu+1)}}_{=:E_6}\eta + \frac{2^{1-\nu}E_5}{(\nu+1)D_2^{\nu-1}}\eta^\nu \\ &= \mathcal{O}\left(\frac{1}{\eta T}+\eta^\nu\right).\end{aligned}$$

This completes the proof. □

B.6 PROOF OF THEOREM 3.5

Theorem 3.5 (restated). *Let the function $f\colon \mathbb{R}^d \to \mathbb{R}$ be nonconvex. Suppose that Assumptions 2.1 and 2.2 hold. With learning rate $\eta_t := \eta_{\max} t^{-\frac{1}{\nu+1}}$, if $\nu+1 \le \mathfrak{p}$ and $\eta_t^\nu < \frac{2}{L}$ for all $t\in[T]$, then:*

$$\min_{1\le t\le T} \mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right] \le \frac{2(f(\boldsymbol{x}_1)-f^\star)}{(2-L\eta_{\max}^\nu)}\frac{1}{\sum_{t=1}^T \eta_t} + E_7\frac{\sum_{t=1}^T \eta_t^{\nu+1}}{\sum_{t=1}^T \eta_t} = \mathcal{O}\left(\frac{\log T}{T^{\frac{\nu}{\nu+1}}}\right).$$

In particular, when $\nu+1=\mathfrak{p}$,

$$\min_{1\le t\le T} \mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right] = \mathcal{O}\left(\frac{\log T}{T^{\frac{\mathfrak{p}-1}{\mathfrak{p}}}}\right),$$

where $E_7 := \frac{2E_0L}{(\nu+1)(2-L\eta_{\max}^\nu)}$ and E_0 defined in Eq. (1) *are non-negative constants.*

Proof. From Lemma 2.1, we have

$$\begin{aligned}f(\boldsymbol{x}_{t+1}) &\le f(\boldsymbol{x}_t) + \langle\nabla f(\boldsymbol{x}_t), \boldsymbol{x}_{t+1}-\boldsymbol{x}_t\rangle + \frac{L}{\nu+1}\|\boldsymbol{x}_{t+1}-\boldsymbol{x}_t\|^{\nu+1} \\ &= f(\boldsymbol{x}_t) - \eta_t\langle\nabla f(\boldsymbol{x}_t), \nabla f_{\mathcal{S}_t}(\boldsymbol{x}_t)\rangle + \frac{L\eta_t^{\nu+1}}{\nu+1}\|\nabla f_{\mathcal{S}_t}(\boldsymbol{x}_t)\|^{\nu+1}.\end{aligned}$$

By taking the expectation, we have

$$\mathbb{E}\left[f(\boldsymbol{x}_{t+1})\right] \le \mathbb{E}\left[f(\boldsymbol{x}_t)\right] - \eta_t\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right] + \frac{L\eta_t^{\nu+1}}{\nu+1}\mathbb{E}\left[\|\nabla f_{\mathcal{S}_t}(\boldsymbol{x}_t)\|^{\nu+1}\right].$$

From Lemma 3.1,

$$\begin{aligned}\mathbb{E}\left[f(\boldsymbol{x}_{t+1})\right] &\le \mathbb{E}\left[f(\boldsymbol{x}_t)\right] - \eta_t\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right] + \frac{L\eta_t^{\nu+1}}{\nu+1}\left\{2^{1-\nu}\left(\frac{C_{\mathfrak{p}}\sigma^{\mathfrak{p}}}{b^{\mathfrak{p}-1}}\right)^{\frac{\nu+1}{\mathfrak{p}}} + \frac{\nu+1}{2}\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right] + \frac{1-\nu}{2}\right\} \\ &= \mathbb{E}\left[f(\boldsymbol{x}_t)\right] - \eta_t\left(1-\frac{L\eta_t^\nu}{2}\right)\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right] + \frac{L\eta_t^{\nu+1}}{\nu+1}\left\{2^{1-\nu}\left(\frac{C_{\mathfrak{p}}\sigma^{\mathfrak{p}}}{b^{\mathfrak{p}-1}}\right)^{\frac{\nu+1}{\mathfrak{p}}} + \frac{1-\nu}{2}\right\} \\ &\le \mathbb{E}\left[f(\boldsymbol{x}_t)\right] - \eta_t\left(1-\frac{L\eta_{\max}^\nu}{2}\right)\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right] + \frac{L\eta_t^{\nu+1}}{\nu+1}\left\{2^{1-\nu}\left(\frac{C_{\mathfrak{p}}\sigma^{\mathfrak{p}}}{b^{\mathfrak{p}-1}}\right)^{\frac{\nu+1}{\mathfrak{p}}} + \frac{1-\nu}{2}\right\}.\end{aligned}$$

Rearranging the terms,

$$\eta_t\left(1-\frac{L\eta_{\max}^\nu}{2}\right)\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right] \le \mathbb{E}\left[f(\boldsymbol{x}_t)\right] - \mathbb{E}\left[f(\boldsymbol{x}_{t+1})\right] + \frac{L\eta_t^{\nu+1}}{\nu+1}\left\{2^{1-\nu}\left(\frac{C_{\mathfrak{p}}\sigma^{\mathfrak{p}}}{b^{\mathfrak{p}-1}}\right)^{\frac{\nu+1}{\mathfrak{p}}} + \frac{1-\nu}{2}\right\},$$

and summing over $t = 1$ to $t = T$, we have

$$\left(1 - \frac{L\eta_{\max}^{\nu}}{2}\right) \sum_{t=1}^{T} \eta_t \mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right] \leq f(\boldsymbol{x}_1) - f^\star + \frac{L}{\nu+1} \underbrace{\left\{2^{1-\nu}\left(\frac{C_{\mathfrak{p}}\sigma^{\mathfrak{p}}}{b^{\mathfrak{p}-1}}\right)^{\frac{\nu+1}{\mathfrak{p}}} + \frac{1-\nu}{2}\right\}}_{=:E_0} \sum_{t=1}^{T} \eta_t^{\nu+1}.$$

Therefore,

$$\min_{1\leq t\leq T} \mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right] \leq \frac{2(f(\boldsymbol{x}_1) - f^\star)}{(2 - L\eta_{\max}^{\nu})} \frac{1}{\sum_{t=1}^{T}\eta_t} + \underbrace{\frac{2E_0 L}{(\nu+1)(2 - L\eta_{\max}^{\nu})}}_{=:E_7} \frac{\sum_{t=1}^{T}\eta_t^{\nu+1}}{\sum_{t=1}^{T}\eta_t}.$$

Since $\eta_t := \eta_{\max} t^{-\frac{1}{\nu+1}}$, we can use Lemma B.4 to complete the proof. □

B.7 PROOF OF THEOREM 3.6

Theorem 3.6 (restated). *Let the function $f\colon \mathbb{R}^d \to \mathbb{R}$ be nonconvex. Suppose that Assumptions 2.1 and 2.2 hold. With learning rate $\eta_t := \eta$, if $\nu + 1 \leq \mathfrak{p}$ and $\eta \leq \min\left\{\left(\frac{1}{4L}\right)^{\frac{1}{\nu}}, \frac{1-\beta}{\beta}\left(\frac{1}{4\nu L^2}\right)^{\frac{1}{2\nu}}\right\}$, then:*

$$\min_{1\leq t\leq T} \mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right] \leq \frac{4(f(\boldsymbol{x}_1) - f^\star)}{\eta T} + \frac{LE_0\eta^{\nu}}{4(\nu+1)} + E_8\eta^{2\nu} = \mathcal{O}\left(\frac{1}{\eta T} + \eta^{\nu}\right).$$

In particular, when $\nu + 1 = \mathfrak{p}$,

$$\min_{1\leq t\leq T} \mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right] = \mathcal{O}\left(\frac{1}{\eta T} + \eta^{\mathfrak{p}-1}\right),$$

where $E_8 := \frac{2L}{\nu+1}\left(\frac{\beta}{1-\beta}\right)^{2\nu}\{L(1-\nu) + 2E_0\nu\}$ and E_0 defined in Eq. (1) are non-negative constants.

Proof. For all $t \in \mathbb{N}$, let us define $\Phi_t := f(\boldsymbol{z}_t) + \lambda\|\boldsymbol{m}_{t-1}\|^{\nu+1}$, $\lambda := \frac{4\eta L\nu}{(1-\beta)(\nu+1)}\left(\frac{\eta\beta}{1-\beta}\right)^{2\nu}$. Then, we have

$$\mathbb{E}[\Phi_{t+1}] - \mathbb{E}[\Phi_t] = \mathbb{E}[f(\boldsymbol{z}_{t+1})] - \mathbb{E}[f(\boldsymbol{z}_t)] + \lambda\left(\mathbb{E}[\|\boldsymbol{m}_t\|^{\nu+1}] - \mathbb{E}[\|\boldsymbol{m}_{t-1}\|^{\nu+1}]\right).$$

Therefore, from Lemmas B.2 and B.3,

$$\begin{aligned}
&\mathbb{E}[\Phi_{t+1}] - \mathbb{E}[\Phi_t] \\
&\leq \mathbb{E}[f(\boldsymbol{z}_t)] - \left(\frac{\eta}{2} - \frac{L\eta^{\nu+1}}{2}\right)\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] + \frac{\eta L^2\nu}{\nu+1}\left(\frac{\eta\beta}{1-\beta}\right)^{2\nu}\mathbb{E}\left[\|\boldsymbol{m}_{t-1}\|^{\nu+1}\right] \\
&\quad + \frac{\eta(1-\nu)L^2}{2(\nu+1)}\left(\frac{\eta\beta}{1-\beta}\right)^{2\nu} + \frac{L\eta^{\nu+1}}{\nu+1}\underbrace{\left\{2^{1-\nu}\left(\frac{C_{\mathfrak{p}}\sigma^{\mathfrak{p}}}{b^{\mathfrak{p}-1}}\right)^{\frac{\nu+1}{\mathfrak{p}}} + \frac{1-\nu}{2}\right\}}_{=:E_0} \\
&\quad + \lambda\left\{-(1-\beta)\mathbb{E}[\|\boldsymbol{m}_{t-1}\|^{\nu+1}] + \frac{(1-\beta)(1+\nu)}{2}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] + \frac{(1-\beta)(1-\nu)}{2} + 2^{1-\nu}(1-\beta)^{\nu+1}\left(\frac{C_{\mathfrak{p}}\sigma^{\mathfrak{p}}}{b^{\mathfrak{p}-1}}\right)^{\frac{\nu+1}{\mathfrak{p}}}\right\} \\
&= -\frac{\eta}{2}\left(1 - L\eta^{\nu} - L^2\nu\left(\frac{\eta\beta}{1-\beta}\right)^{2\nu}\right)\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] + \left(\frac{\eta L^2\nu}{\nu+1}\left(\frac{\eta\beta}{1-\beta}\right)^{2\nu} - \lambda(1-\beta)\right)\mathbb{E}[\|\boldsymbol{m}_{t-1}\|^{\nu+1}] \\
&\quad + \frac{LE_0\eta^{\nu+1}}{\nu+1} + \frac{\eta(1-\nu)L^2}{2(\nu+1)}\left(\frac{\eta\beta}{1-\beta}\right)^{2\nu} + \lambda\left\{\frac{(1-\beta)(1-\nu)}{2} + 2^{1-\nu}(1-\beta)^{\nu+1}\left(\frac{C_{\mathfrak{p}}\sigma^{\mathfrak{p}}}{b^{\mathfrak{p}-1}}\right)^{\frac{\nu+1}{\mathfrak{p}}}\right\} \\
&\leq -\frac{\eta}{2}\left(1 - L\eta^{\nu} - L^2\nu\left(\frac{\eta\beta}{1-\beta}\right)^{2\nu}\right)\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] + \frac{LE_0\eta^{\nu+1}}{\nu+1} \\
&\quad + \frac{\eta(1-\nu)L^2}{2(\nu+1)}\left(\frac{\eta\beta}{1-\beta}\right)^{2\nu} + \frac{4\eta L\nu}{(\nu+1)}\left(\frac{\eta\beta}{1-\beta}\right)^{2\nu}E_0,
\end{aligned}$$

where we have used the definition of λ and $(1-\beta)^{\nu+1} < 1-\beta$ in the last inequality. Here, from the condition on η,

$$\frac{\eta}{2}\left(1 - L\eta^{\nu} - L^2\nu\left(\frac{\eta\beta}{1-\beta}\right)^{2\nu}\right) \geq \frac{\eta}{2}\cdot\frac{1}{2} = \frac{\eta}{4}.$$

Hence, we have

$$\frac{\eta}{4}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] \leq \mathbb{E}[\Phi_t] - \mathbb{E}[\Phi_{t+1}] + \frac{LE_0\eta^{\nu+1}}{\nu+1} + \frac{\eta(1-\nu)L^2}{2(\nu+1)}\left(\frac{\eta\beta}{1-\beta}\right)^{2\nu} + \frac{4\eta L\nu}{(\nu+1)}\left(\frac{\eta\beta}{1-\beta}\right)^{2\nu}E_0.$$

Summing over $t=1$ to $t=T$, we have

$$\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] \leq \frac{4(f(\boldsymbol{x}_1)-f^\star)}{\eta} + \frac{LE_0\eta^{\nu}}{4(\nu+1)}T + \frac{2(1-\nu)L^2}{(\nu+1)}\left(\frac{\eta\beta}{1-\beta}\right)^{2\nu}T + \frac{4L\nu}{(\nu+1)}\left(\frac{\eta\beta}{1-\beta}\right)^{2\nu}E_0T.$$

Therefore,

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] &\leq \frac{4(f(\boldsymbol{x}_1)-f^\star)}{\eta T} + \frac{LE_0}{4(\nu+1)}\eta^{\nu} + \frac{2(1-\nu)L^2}{(\nu+1)}\left(\frac{\beta}{1-\beta}\right)^{2\nu}\eta^{2\nu} + \frac{4L\nu}{(\nu+1)}\left(\frac{\beta}{1-\beta}\right)^{2\nu}E_0\eta^{2\nu} \\
&= \frac{4(f(\boldsymbol{x}_1)-f^\star)}{\eta T} + \frac{LE_0}{4(\nu+1)}\eta^{\nu} + \underbrace{\frac{2L}{\nu+1}\left(\frac{\beta}{1-\beta}\right)^{2\nu}\{L(1-\nu)+2E_0\nu\}}_{=:E_8}\eta^{2\nu} \\
&= \mathcal{O}\left(\frac{1}{\eta T} + \eta^{\nu}\right).
\end{aligned}$$

Since the minimum value is smaller than the average value, we have

$$\min_{1\leq t\leq T}\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right] \leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right].$$

This completes the proof. □

C ADDITIONAL EXPERIMENTS

In this section, we present numerical experimental results for vanilla SGD. These results parallel the observations for vanilla SGD with momentum discussed in Section 4. All results presented in this work are fully reproducible via the source code available at: `https://github.com/iiduka-researches/vanillaSGDM_uai2026`. The code is designed to run directly on Google Colab for ease of use.

Figure 4 displays the convergence performance of vanilla SGD across various combinations of ν and α. Similar to the SGD with momentum case, the high-performance configurations (marked in red) are concentrated in the lower-right region where the theoretical condition $\nu+1 \leq \alpha$ is satisfied. This further confirms that our stability boundary is a fundamental requirement for the vanilla stochastic gradient descent family, regardless of the momentum parameter.

Figure 5 illustrates the convergence behavior of vanilla SGD over 1,000 steps for fixed ν. The plots exhibit the same characteristics evident in Figure 2: a smaller α leads to both degraded average performance and significantly higher variance across trials. Consistent with our theory, trajectories for larger ν become increasingly erratic as α decreases, whereas smaller ν values provide a more robust optimization path.

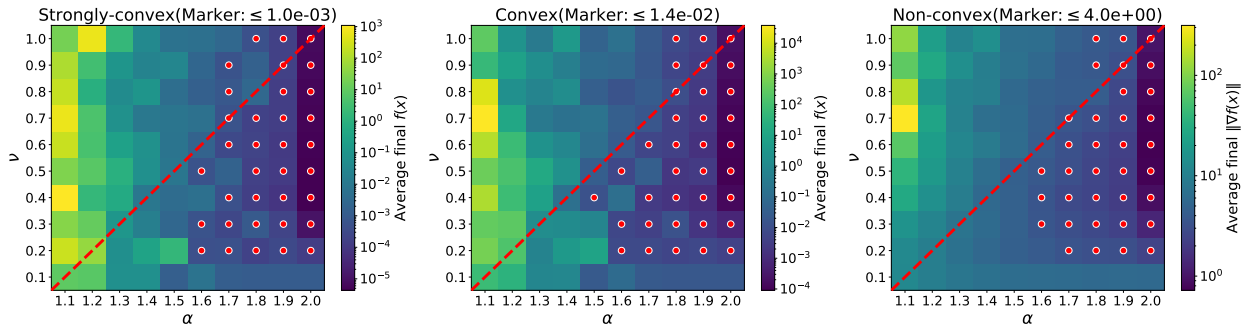

Figure 4: Convergence performance across various (ν, α) pairs. The heatmaps show the average final objective values (for strongly convex and convex cases) and gradient norms (for nonconvex cases) after 4,000 steps of vanilla SGD. The top 35 performing combinations are indicated by red markers. Below the red dotted line, the theoretical condition $\nu + 1 \leq \alpha$ is satisfied, ensuring convergence.

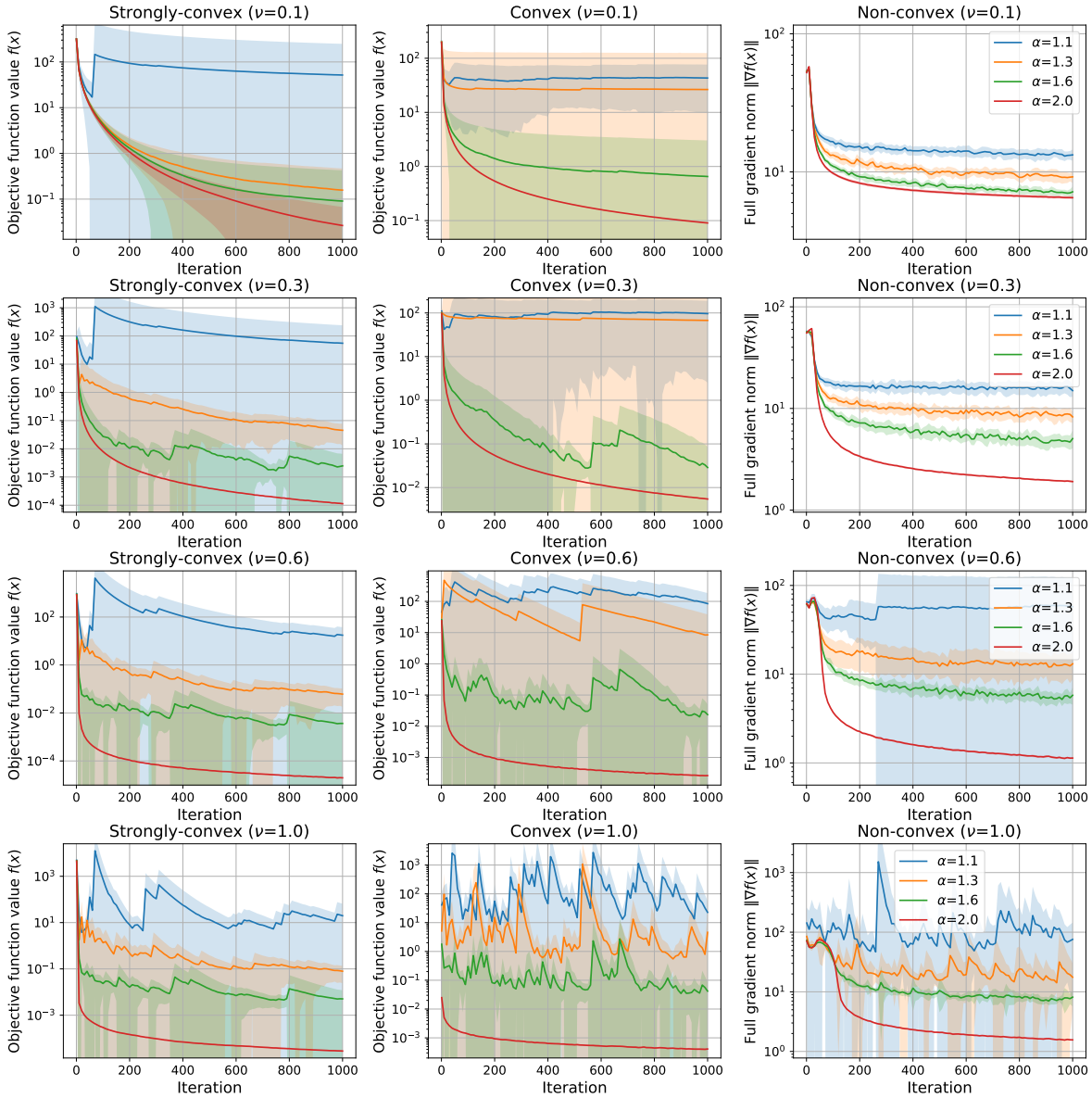

Figure 5: Convergence trajectories of vanilla SGD. The plots illustrate the evolution of objective values (for strongly convex and convex cases) and gradient norms (for nonconvex cases) over 1,000 steps for the functions defined in Eq. (2). Solid lines represent the average of 20 independent trials, while the lightly shaded regions indicate the range between the minimum and maximum values.