Tight Lower Bounds and Optimal Algorithms for Stochastic Nonconvex Optimization with Heavy-Tailed Noise

Adrien Fradin[1,2] **Abdurakhmon Sadiev**[1] **Laurent Condat**[1] **Peter Richtárik**[1]
[1]KAUST*
Thuwal, Saudi Arabia
[2]École Polytechnique
Paris, France

Abstract

We study stochastic nonconvex optimization under heavy-tailed noise. In this setting, the stochastic gradients only have bounded p-th central moment (p-BCM) for some $p \in (1, 2]$. Building on the foundational work of Arjevani et al. (2022) in stochastic optimization, we establish tight sample complexity lower bounds for all first-order methods under *relaxed* mean-squared smoothness (q-WAS) and δ-similarity ((q, δ)-S) assumptions, allowing any exponent $q \in [1, 2]$ instead of the standard $q = 2$. These results substantially broaden the scope of existing lower bounds. To complement them, we show that Normalized Stochastic Gradient Descent with Momentum Variance Reduction (NSGD-MVR), a known algorithm, matches these bounds in expectation. Beyond expectation guarantees, we introduce a new algorithm, Double-Clipped NSGD-MVR, which allows the derivation of high-probability convergence rates under weaker assumptions than in previous works. Finally, for second-order methods with stochastic Hessians satisfying bounded q-th central moment assumptions for some exponent $q \in [1, 2]$ (allowing $q \neq p$), we establish sharper lower bounds than previous works while improving over Sadiev et al. (2025) (where only $p = q$ is considered) and yielding stronger convergence exponents. Together, these results provide a nearly complete complexity characterization of stochastic nonconvex optimization in heavy-tailed regimes.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

1 INTRODUCTION

We consider the stochastic optimization problem

$$\min_{x \in \mathbb{R}^d} F(x), \qquad F(x) := \mathbb{E}_{\xi \sim \mathcal{D}} \left[f(x, \xi) \right], \tag{1}$$

where $d \geq 1$ is the dimension, $F \colon \mathbb{R}^d \to \mathbb{R}$ is a smooth possibly nonconvex objective function, and ξ is a random variable drawn from an unknown distribution $\mathcal{D}$. In the nonconvex setting, our goal is to find an ε–stationary point (Nesterov, 2018); that is, a vector $\bar{x}$ such that $\mathbb{E}[\|\nabla F(\bar{x})\|] \leq \varepsilon$. Problems of the form (1) are pervasive in machine learning where $f(x, \xi)$ denotes the loss of a model with weights x on a data sample $\xi \sim \mathcal{D}$, and $\mathcal{D}$ is the distribution of the training samples from the dataset (Bottou et al., 2018). While Gradient Descent (GD) is known to achieve the optimal rate $\mathcal{O}(\varepsilon^{-2})$ for finding an ε–stationary point (Carmon et al., 2020), it requires access to exact gradients $\nabla F(\cdot)$ which is infeasible in practice. Therefore, reliance on noisy gradients has become a gold standard approach, giving rise to stochastic gradient methods like Stochastic Gradient Descent (SGD) (Robbins and Monro, 1951). Under the standard σ_1^2–bounded variance assumption, SGD is provably optimal (Ghadimi and Lan, 2013), matching the lower bound $\Omega(L_1 \Delta / \varepsilon^2 + L_1 \Delta \sigma_1^2 / \varepsilon^4)$ (Arjevani et al., 2022) for L_1–smooth functions with $\Delta := F(x^0) - F^{\inf}$.

However, empirical observations in modern machine learning, e.g., image classification (Şimşekli et al., 2019a,b; Battash et al., 2024), large language models (Zhang et al., 2020; Ahn et al., 2024) and reinforcement learning (Garg et al., 2021), have shed light on the importance of *heavy-tailed* noise as a more realistic setting than the standard bounded variance assumption. That is, noisy gradients only have bounded p–th central moment, $p \in (1, 2]$. Such a relaxed moment assumption introduces challenges on both the algorithmic and theoretical sides, as SGD may fail to converge when $p < 2$. In the heavy-tailed regime, Zhang et al.

*King Abdullah University of Science and Technology

(2020) establish a lower bound of $\Omega(T^{-(p-1)/(3p-2)})$ on the convergence rate of first-order methods, where T denotes the number of iterations. Yet, their analysis is confined to the standard stochastic setting, where algorithms access smooth (potentially nonconvex) functions via unbiased stochastic gradients. Crucially, they did not address more structured assumptions on the gradient noise such as *mean-squared smoothness* or δ–*similarity*, which play a central role to the design and analysis of *variance-reduced* methods, but they remain largely unexplored in the heavy-tailed regime.

In this work, we aim to: (1) **develop lower complexity bounds for first-order variance-reduced methods in the heavy-tailed regime**, thereby closing an important open question from Liu et al. (2023), while advancing beyond the results of Tao et al. (2025); (2) **extend existing high-probability analysis to these new assumptions**, and finally, for second-order methods; (3) **characterize how the tail indices p (of gradients) and q (of Hessians) jointly affect convergence in the heavy-tailed regime**.

Beyond the development of lower bounds, a natural question arises:

Does there exist algorithm(s) which provably match (in expectation) the lower bounds developed in (1)*?*

Here, we answer this question affirmatively by revisiting a classical technique: gradient normalization. Remarkably, our results show that this simple method remains effective even in the heavy-tailed regime.

1.1 Our Contributions

We revisit the analysis of Normalized SGD (NSGD) with Momentum Variance Reduction (MVR) in the heavy-tailed regime. Our contributions are:

• **Tight lower bounds.** We obtain novel and tight lower bounds on the sample complexity of any first-order methods in the p–BCM noise model, under *relaxed* mean-squared smoothness and δ-similarity assumptions, allowing any exponent $q \in [1, 2]$ instead of $q = 2$. Our bound, $\Omega(\varepsilon^{-(p(2q+1)-2q)/q(p-1)})$, recovers the known $\Omega(\varepsilon^{-3})$ when $p = q = 2$ and achieves an improvement over the minimax complexity for first-order methods (Zhang et al., 2020; Tao et al., 2025; Hübler et al., 2025), thereby demonstrating the effectiveness of variance reduction in the heavy-tailed regime. Moreover, our result *continuously interpolates* between variance-reduced and non-variance-reduced regimes, as in the limit $q \to 1$ we recover the known $\Omega(\varepsilon^{-(3p-2)/(p-1)})$ complexity. *Our proofs are based on a generalization of Arjevani et al. (2022, Lemma 10) to the p–BCM noise assumption, which resolves an open question from Liu et al. (2023).*

• **Optimal first-order methods.** Revisiting NSGD-MVR, we show that, with a suitable choice of parameters, the method matches the lower bounds (up to constant factors) with respect to all parameters, thereby establishing the optimality of our analysis and the robustness of gradient normalization in the context of heavy-tailed noise. We further provide convergence rates with *unknown* tail indices p and q, recovering the best known rate of $\mathcal{O}(\varepsilon^{-2p/(p-1)})$ (Liu and Zhou, 2025).

• **High-probability upper bounds.** We propose a clipped variant of NSGD-MVR leading to a *new* algorithm: Double-Clipped NSGD-MVR (D-Clip-NSGD-MVR), which provably achieves high-probability convergence rates under weaker assumptions than in previous works (Liu et al., 2023).

• **Extension to second-order methods.** Lastly, for second-order methods using stochastic Hessians with bounded q-th central moment, for some exponent $q \in [1, 2]$, we establish sharper lower bounds than previous works (Arjevani et al., 2020a; Sadiev et al., 2025), extend previous analysis to the more general setting where $p \neq q$ (unlike Sadiev et al. (2025)), and derive upper bounds with stronger convergence exponents, thereby offering a more complete characterization of the limits of stochastic second-order optimization under heavy-tailed noise. Overall, our results reveal a striking parallel in the rates of variance-reduced methods and second-order optimization.

1.2 Related Works

Lower bounds in the heavy-tailed regime: In nonconvex optimization, Zhang et al. (2020); Liu and Zhou (2025) establish a lower bound of $\Omega\big(\frac{L_1\Delta}{\varepsilon^2} + \frac{L_1\Delta}{\varepsilon^2}\big(\frac{\sigma_1}{\varepsilon}\big)^{p/(p-1)}\big)$ under p–BCM noise and standard L_1–smoothness, recovering the well-known $\Omega(\varepsilon^{-4})$ lower bound when $p = 2$. Another line of work focuses on algorithm-dependent lower bounds: NSGD (Yang et al., 2023) and recently SGD (Fatkhullin et al., 2025), highlighting the ineffectiveness of SGD under heavy-tailed noise with only bounded p-th (non-central) moments. Beyond first-order methods, Sadiev et al. (2025) obtain near-optimal lower bounds for all second-order methods, extending the work of Arjevani et al. (2020a) to p–BCM noise.

Convergence under heavy-tailed noise: A substantial body of work investigates upper bounds for stochastic optimization with heavy-tailed noise, both in expectation and high probability. For smooth and nonconvex objectives, clipped or normalized variants of SGD achieve the rate $\mathcal{O}(\varepsilon^{-(3p-2)/(p-1)})$ up to logarithmic factors (Zhang et al., 2020; Cutkosky and Mehta, 2021; Liu et al., 2023; Sadiev et al., 2023). Among

Table 1: Sample complexities of stochastic methods for finding an ε-stationary point (in expectation or with high probability). Stochastic gradients satisfy p-BCM with $p \in (1, 2]$ (Assumption 2.3). The column "q" specifies the relevant moment assumptions: Assumption 2.6 (q-WAS), Assumption 2.7 ((q, δ)-S), or Assumption 2.5 (q-BCM); $q = \infty$ corresponds to bounded noise. The column "**HP?**" indicates whether a high-probability guarantee with only polylogarithmic dependence on $1/\beta$ is available. Lower bounds are stated *up to constant factors*. For the algorithm names: NSGD: Normalized SGD, Acc: Accelerated, SFOM: Stochastic First-Order Method, RM: Recursive Momentum, MVR: Momentum Variance Reduction, Hess: Hessian-corrected momentum, Clip / D-Clip: Clipping / Double-Clipped.

Setup	Algorithm	Sample Complexity	q	HP?
q-WAS	AccNSGD Liu et al. (2023)	$\left(\frac{\sqrt{\bar{L}\Delta}+\sigma_1}{\varepsilon}\right)^{2+\frac{1}{p-1}}$	∞ [1]	✓
	NSFOM with RM He et al. (2025)	$\left(\frac{\Delta+\sigma_1^p+L_1+L_1^p+\bar{L}^p}{\varepsilon}\right)^{2+\frac{1}{p-1}}$ [2]	$p=q$	✗
	NSGD-MVR Theorem 4.1	$\frac{\bar{L}\Delta}{\varepsilon^2}+\frac{\bar{L}\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{q(p-1)}}+\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}$	$(1,2]$	✗
	D-Clip-NSGD-MVR Theorem 6.1	$\left(\frac{\sqrt{\bar{L}\Delta}+\sigma_1}{\varepsilon}\right)^{2+\max\left\{\frac{1}{p-1},\frac{p}{q(p-1)}\right\}}$	$(1,2]$	✓
	Lower Bound Theorem 3.1	$\frac{\bar{L}\Delta}{\varepsilon^2}+\frac{\bar{L}\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{q(p-1)}}+\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}$	$(1,2]$	-
(q,δ)-S	NSGD-MVR Theorem 4.2	$\frac{(L_1+\delta)\Delta}{\varepsilon^2}+\frac{\delta\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{q(p-1)}}+\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}$	$(1,2]$	✗
	D-Clip-NSGD-MVR Theorem F.1	$\left(\frac{\sqrt{L_1\Delta}+\sigma_1}{\varepsilon}\right)^{2+\max\left\{\frac{1}{p-1},\frac{p}{q(p-1)}\right\}}+\left(\frac{\sqrt{\delta\Delta}}{\varepsilon}\right)^{2+\frac{p}{q(p-1)}}$	$(1,2]$	✓
	Lower Bound Theorem 3.2	$\min\left\{\frac{L_1\Delta}{\varepsilon^2}+\frac{L_1\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}},\frac{(L_1+\delta)\Delta}{\varepsilon^2}+\frac{\delta\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{q(p-1)}}+\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}\right\}$	$(1,2]$	-
q-BCM	NSGDHess Sadiev et al. (2025)	$\frac{(L_1+\sigma_2)\Delta}{\varepsilon^2}+\left(\frac{(L_1+\sigma_2)\Delta}{\varepsilon^2}+\frac{\sigma_1}{\varepsilon}\right)\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{1}{p-1}}$	$p=q$	✗
	Clip NSGDM-Hess Sadiev et al. (2025)	$\left(\frac{\sqrt{(L_1+\sigma_2)\Delta}+\sigma_1}{\varepsilon}\right)^{2+\frac{1}{p-1}}$	$p=q$	✓
	NSGD-Hess Theorem 5.1	$\frac{\sigma_2\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{q(p-1)}}+\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}+\frac{\sqrt{L_2}\Delta\sigma_1^{1/4}}{\varepsilon^{7/4}}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{1}{4(p-1)}}$ [3]	$(1,2]$	✗
	Clip NSGD-Hess Theorem F.2	$\left(\frac{\sqrt{L_1\Delta}+\sigma_1}{\varepsilon}\right)^{2+\max\left\{\frac{1}{p-1},\frac{p}{q(p-1)}\right\}}+\left(\frac{\sqrt{\sigma_2\Delta}}{\varepsilon}\right)^{2+\frac{p}{q(p-1)}}$ [3]	$(1,2]$	✓
	Lower Bound Theorem 3.3	$\min\left\{\frac{\sigma_2\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{q(p-1)}},\frac{L_1\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}},\frac{\sqrt{L_2}\Delta}{\varepsilon^{3/2}}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}\right\}$ [3]	$(1,2]$	-

[1] Liu et al. (2023) provide analysis under stronger assumptions, which implies the bounded stochastic Hessian.
[2] He et al. (2025) establish a near-optimal bound in terms of ε for the case $p = q$.
[3] For simplicity we present only the stochastic part of the complexity.

these, Zhang et al. (2020) obtained the sharpest in-expectation guarantees, later shown to be tight by Hübler et al. (2025) via minibatch-NSGD, while Nguyen et al. (2023) established improved high-probability guarantees, avoiding extra $\mathcal{O}(\log T)$ factors. Later works extended known high-probability analyses to settings with higher-order smoothness (Sadiev et al., 2025) and to more general nonlinear SGD-type methods (Armacki et al., 2024, 2025b). In parallel, Tao et al. (2025) studied clip-SGD under stronger smoothness assumptions, obtaining a rate of $\mathcal{O}(\varepsilon^{-(2p-1)/(p-1)})$, while Liu and Zhou (2025) derived in-expectation bounds for NSGD-Mom under a general (σ_0, σ_1)–affine p–BCM model, albeit without matching lower bounds.

Gradient clipping and normalization: Gradient clipping and normalization are two closely related techniques that have become central in modern optimization. Gradient clipping, originally popularized to stabilize training across various machine learning applications (Pascanu et al., 2013; Schulman et al., 2017), has been analyzed extensively, providing robustness under relaxed moment assumptions (Polyak and Tsypkin, 1979; Jakovetić et al., 2023), high-order smoothness (Sadiev et al., 2025), and enabling high-probability convergence guarantees with only logarithmic dependence on the failure probability. These results hold in both convex (Nazin et al., 2019; Gorbunov et al., 2020; Davis et al., 2021; Gorbunov et al., 2024b; Liu and Zhou, 2023; Gorbunov et al., 2024a; Puchkin et al., 2024; Armacki et al., 2025b,a) and nonconvex settings (Zhang et al., 2020; Cutkosky and Mehta, 2021; Sadiev et al., 2023; Nguyen et al., 2023;

Liu et al., 2023; Sadiev et al., 2025). Importantly, high-probability guarantees are valuable both theoretically and practically, as they capture the behavior of individual runs rather than merely characterizing the average-case performance.

Normalized gradient methods date back to Nesterov's pioneering work (Nesterov, 1984) and were later extended to the smooth and stochastic regimes (Kiwiel, 2001; Hazan et al., 2015; Levy, 2017; Nacson et al., 2019). In deep learning, normalization addresses exploding and vanishing gradients (You et al., 2017, 2020), though rigorous nonconvex guarantees were first obtained by Cutkosky and Mehta (2020), showing how Polyak's momentum ensures convergence without large batches. Later works explored NSGD's strengths and limits, including saddle-point escape (Levy, 2016), lower bounds (Yang et al., 2023) and parameter-agnostic convergence (Hübler et al., 2024). Recently, extensions to heavy-tailed noise have been investigated (Cutkosky and Mehta, 2021; Liu et al., 2023; Tao et al., 2025), though typically under stronger smoothness assumptions.

Despite strong empirical performance and complementarity with clipping, the theory of gradient normalization under heavy-tailed noise remains incomplete, motivating our contributions.

Variance reduction for stochastic optimization: Variance reduction originated as a tool to accelerate convergence in convex finite-sum optimization, with seminal works such as Roux et al. (2012); Johnson and Zhang (2013); Shalev-Shwartz and Zhang (2013); Mairal (2013); Defazio et al. (2014) introducing various algorithms, e.g., SAG, SVRG, SAGA. Building on these foundations and the key advances of Allen-Zhu (2017), a sequence of works (Lan et al., 2019; Zhou et al., 2019; Song et al., 2020; Kovalev et al., 2020) developed algorithms attaining near-optimal or optimal rates under various regimes. In nonconvex optimization, variance reduction can also improve convergence: in the general stochastic setting (1), several works (Fang et al., 2018; Cutkosky and Orabona, 2019; Tran-Dinh et al., 2019; Liu et al., 2020; Li et al., 2021) established an $\mathcal{O}(\varepsilon^{-3})$ convergence rate in expectation, which improves upon the classical $\Theta(\varepsilon^{-4})$ of SGD and matches the minimax lower bound $\Omega(\varepsilon^{-3})$ under mean-squared smoothness (Arjevani et al., 2022). More recently, there has been growing interest in extending variance reduction techniques to the heavy-tailed regime (Liu et al., 2023; Tao et al., 2025; He et al., 2025), where normalization and clipping have emerged as effective mechanisms to guarantee robustness under weaker moment assumptions.

In this work, we characterize the fundamental limits of first-order "variance-reduced" methods in the heavy-tailed regime and provide optimal methods which attain these limits.

2 NOTATION AND ASSUMPTIONS

We review here the basic notation and assumptions needed in this paper (see Appendix A for more details).

Notations. For integer $n > 0$, $[n] := \{1, 2, \dots, n\}$. We let $d \geq 1$ be the dimension, $\langle \cdot, \cdot \rangle$ the standard dot product on $\mathbb{R}^d$, $\|\cdot\|$ the ℓ^2–norm and $\|\cdot\|_{\text{op}}$ the canonical spectral/operator norm. Here, $\nabla f(\cdot, \cdot)$ and $\nabla^2 f(\cdot, \cdot)$ denote the stochastic gradient and Hessian oracles; $a \wedge b := \min\{a, b\}$ and $a \vee b := \max\{a, b\}$. We use the standard $\mathcal{O}(\cdot)$ and $\Omega(\cdot)$ for complexity notation. To avoid confusion, we use subscript 1 for gradient parameters (e.g., L_1, σ_1) and 2 for Hessian ones.

In this work, we make the following assumptions.

Assumption 2.1 (Lower Boundedness)**.** *The objective F is lower bounded: $F^{\inf} := \inf_{x \in \mathbb{R}^d} F(x) > -\infty$.*

We then let $\Delta := F(x^0) - F^{\inf}$ be the initial suboptimality where x^0 is the starting point.

Assumption 2.2 (L_1–Lipschitz Gradients)**.** *The objective F is differentiable over $\mathbb{R}^d$ and its gradient is L_1–Lipschitz for some $L_1 \geq 0$, i.e., for all $x, y \in \mathbb{R}^d$,*

$$\|\nabla F(x) - \nabla F(y)\| \leq L_1 \|x - y\|.$$

Assumption 2.3 (p–BCM for Gradients)**.** *We have access to unbiased stochastic gradients $\nabla f(x, \xi)$, which have p-bounded central moment for some $p \in (1, 2]$, i.e., there exists $\sigma_1 > 0$ such that for all $x \in \mathbb{R}^d$,*

$$\mathbb{E}_{\xi \sim \mathcal{D}}[\nabla f(x, \xi)] = \nabla F(x)$$

and

$$\mathbb{E}_{\xi \sim \mathcal{D}}[\|\nabla f(x, \xi) - \nabla F(x)\|^p] \leq \sigma_1^p.$$

Assumption 2.4 (L_2–Lipschitz Hessians)**.** *The objective F is twice continuously differentiable over $\mathbb{R}^d$ and we have access to stochastic Hessian-vector products $\nabla^2 f(x, \xi) \cdot v$ for any $v \in \mathbb{R}^d$. Moreover, the Hessian of F is L_2–Lipschitz, i.e., for all $x, y \in \mathbb{R}^d$,*

$$\left\|\nabla^2 F(x) - \nabla^2 F(y)\right\|_{\text{op}} \leq L_2 \|x - y\|.$$

Assumption 2.5 (q–BCM for Hessians)**.** *The stochastic Hessians $\nabla^2 f(x, \xi)$ are unbiased and have q-bounded central moment for some $q \in [1, 2]$, i.e., there exists $\sigma_2 > 0$ such that for all $x \in \mathbb{R}^d$*

$$\mathbb{E}_{\xi \sim \mathcal{D}}\left[\nabla^2 f(x, \xi)\right] = \nabla^2 F(x),$$

and

$$\mathbb{E}_{\xi \sim \mathcal{D}}\left[\left\|\nabla^2 f(x, \xi) - \nabla^2 F(x)\right\|_{\text{op}}^q\right] \leq \sigma_2^q.$$

Assumption 2.6 (q-Weak Average Smoothness (He et al., 2025))**.** *For some $q \in [1,2]$, there exists a finite constant $\bar{L} \geq 0$ such that, for all $x, y \in \mathbb{R}^d$ we have*

$$\mathbb{E}_{\xi\sim\mathcal{D}}\left[\|\nabla f(x,\xi) - \nabla f(y,\xi)\|^q\right] \leq \bar{L}^q \|x-y\|^q .$$

The next assumption has been introduced in Arjevani et al. (2020a) in the special case $q = 2$.

Assumption 2.7 ((q,δ)–Similarity)**.** *For some exponent $q \in [1,2]$, there exists a finite constant $\delta \geq 0$ such that, for all $x, y \in \mathbb{R}^d$ we have*

$$\mathbb{E}_{\xi\sim\mathcal{D}}\left[\|[\nabla f(x,\xi) - \nabla f(y,\xi)] - [\nabla F(x) - \nabla F(y)]\|^q\right] \leq \delta^q \|x-y\|^q .$$

It is worth noting that Assumption 2.6 is in fact equivalent to Assumptions 2.2 and 2.7. Specifically, the latter two assumptions imply Assumption 2.6 with $\bar{L}^q = 2^{q-1}(L_1^q + \delta^q)$, while conversely, if Assumption 2.6 holds then both Assumptions 2.2 and 2.7 hold with $L_1 = \bar{L}$ and $\delta = \bar{L}$. Hence, $L_1 \leq \bar{L}$, $\delta \leq \bar{L}$, and $\bar{L}^q \leq 2^{q-1}(L_1^q + \delta^q) \leq 2^q \bar{L}^q$.

3 LOWER COMPLEXITY BOUNDS

To establish our lower bounds, we analyze the distributional complexity of finding an ε-stationary point. We focus on the broad class of *zero-respecting* algorithms, denoted by $\mathcal{A}_{\mathtt{zr}}$, which only query coordinates previously discovered to have non-zero gradients (Carmon et al., 2020). In our optimization protocol, at each round, an algorithm $\mathtt{A} \in \mathcal{A}_{\mathtt{zr}}$ makes a batch of $K \geq 1$ *multi-point queries* to a stochastic oracle $\mathtt{O}_F$. Crucially, these K queries are evaluated at the *same* random seed ξ, rather than using independent seeds. Because our hard instance construction guarantees that if a drawn seed yields zero, none of the K simultaneous queries can reveal a new coordinate, the parameter K does not appear explicitly in our final lower bound expressions. Consequently, our bounds hold regardless of whether $K = 1$ or $K \geq 2$. (For context, our variance-reduced methods in Algorithms 1 and 3 utilize $K = 2$, while Algorithm 2 utilizes $K = 1$).

Formally, following Arjevani et al. (2022), we define the minimax sample complexity as:

$$\mathfrak{m}_\varepsilon^{\mathtt{zr}}(K,\Delta\ldots) := \sup_{F\in\mathcal{F}(\Delta\ldots)} \sup_{(\mathbb{P}_\xi,\mathtt{O}_F)\in\mathcal{O}(F\ldots)} \inf_{\mathtt{A}\in\mathcal{A}_{\mathtt{zr}}} \inf\left\{T \geq 1 : \mathbb{E}_{\mathtt{O}_F,\mathtt{A}}\left[\left\|\nabla F\left(\mathbf{x}^{(T)}_{\mathtt{A}[\mathtt{O}_F]}\right)\right\|\right] \leq \varepsilon\right\},$$

where the expectation is taken over the randomness in the oracle $\mathtt{O}_F$ and in the algorithm $\mathtt{A} \in \mathcal{A}_{\mathtt{zr}}$, if any.

To derive these bounds, we rely on a technique developed by Carmon et al. (2020); Arjevani et al. (2022), introducing a "worst-case" nonconvex function $F_T : \mathbb{R}^T \to \mathbb{R}$ that is hard for any zero-respecting algorithm to optimize. Our main theoretical novelty here is Lemma C.1, an important result lying at the core of all our lower bounds. By generalizing Arjevani et al. (2022, Lemma 10) to the p–BCM assumption, this lemma yields a tight improvement from $\Omega(\sigma_1^2/\varepsilon^2)$ to $\Omega((\sigma_1/\varepsilon)^{p/(p-1)})$. This handles a critical edge case in Theorems 3.1 to 3.3 where the parameters would otherwise force the dimension T of the "hard instance" F_T to be too small.

While the key complexity definitions are provided above, we defer the formal definitions of the function and oracle classes $(\mathcal{F}, \mathcal{O})$, the exact construction of F_T, and further motivations behind our generalization to Remark C.4 and Appendix B. The full proofs for the following three lower bounds can be found in Appendices C.3 to C.5.

3.1 Under q-Weak Average Smoothness

We first establish a lower bound on the oracle complexity of any zero-respecting algorithm under Assumption 2.6, an extension of the mean-squared smoothness (MSS) assumption. This result generalizes one of the main contributions of Arjevani et al. (2022), who obtained optimal lower bound in the $p = q = 2$ setting.

Theorem 3.1. *Given $\Delta, \bar{L} > 0$, $\sigma_1 \geq 0$ and $0 < \varepsilon \leq c_1\sqrt{\bar{L}\Delta}$ for some universal constant $c_1 > 0$. Then, for any algorithm $A \in \mathcal{A}_{\mathtt{zr}}$, there exists a function $f \in \mathcal{F}(\Delta)$, an oracle and a distribution $(O,\mathcal{D}) \in \mathcal{O}(f, \bar{L}^q, \sigma_1^p)$ satisfying Assumptions 2.1, 2.3 and 2.6 such that*

$$\mathfrak{m}_\varepsilon^{\mathtt{zr}}\left(K,\Delta,\bar{L}^q,\sigma_1^p\right) \geq \Omega(1)\cdot\left(\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}} + \frac{\bar{L}\Delta}{\varepsilon^2} + \frac{\bar{L}\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{q(p-1)}}\right).$$

In the special case $p = q = 2$, our result recovers the optimal complexity lower bound $\Omega(\sigma_1^2/\varepsilon^2 + \bar{L}\Delta/\varepsilon^2 + \bar{L}\Delta\sigma_1/\varepsilon^3)$ derived by Arjevani et al. (2022). We further establish the tightness of this bound in Theorem 4.1, where we prove a matching upper bound.

3.2 Under (q,δ)–Similarity

The previous bound captures limits of first-order methods under Assumption 2.6, but this assumption couples smoothness and noise together into $\bar{L}$. Assuming separately (q,δ)–similarity with L_1-smoothness decouples these effects, yielding sharper lower bounds.

Theorem 3.2. *Given $\Delta, L_1, \delta > 0$, $\sigma_1 \geq 0$ and $0 < \varepsilon \leq c_1\sqrt{L_1\Delta}$ for some universal constant $c_1 > 0$. Then, for any algorithm $A \in \mathcal{A}_{\mathtt{zr}}$, there exists a function $f \in \mathcal{F}(\Delta, L_1)$, an oracle and a distribution $(O,\mathcal{D}) \in \mathcal{O}(f, \delta^q, \sigma_1^p)$ satisfying Assumptions 2.1*

to 2.3 and 2.7 such that

$$\mathfrak{m}_{\varepsilon}^{\mathtt{zr}}\left(K,\Delta,L_1,\delta^q,\sigma_1^p\right) \geq \Omega(1)\cdot\min\left\{\frac{L_1\Delta}{\varepsilon^2}+\frac{L_1\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}, \left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}+\frac{(L_1+\delta)\Delta}{\varepsilon^2}+\frac{\delta\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{q(p-1)}}\right\}.$$

Expressing the rate in terms of L_1 and δ is more precise than using $\bar{L}$ alone; in particular, when $\delta \ll \bar{L}$, we show in Theorem 4.2 that NSGD-MVR can achieve improved complexity, even matching the second term of the above $\min\{\ldots\}$.

3.3 Under Bounded Central Moments

Compared to the previous results, which focused on first-order methods under q-weak average smoothness and (q,δ)–similarity, we consider here the setting where the oracle's noise is controlled only via bounded central moments, allowing distinct exponents p and q for the stochastic gradients and Hessians.

Theorem 3.3. *Given $\Delta, L_1, L_2 > 0$, $\sigma_1, \sigma_2 \geq 0$ and $0 < \varepsilon \leq c_1 \min\{\sqrt{L_1\Delta}, L_2^{1/3}\Delta^{2/3}\}$ for some universal constant $c_1 > 0$. Then, for any algorithm $A \in \mathcal{A}_{zr}$, there exists a function $f \in \mathcal{F}(\Delta, L_1, L_2)$, an oracle and a distribution $(O,\mathcal{D}) \in \mathcal{O}(f, \sigma_1^p, \sigma_2^q)$ satisfying Assumptions 2.1 to 2.5 such that*[1]

$$\mathfrak{m}_{\varepsilon}^{\mathtt{zr}}\left(K,\Delta,L_1,L_2,\sigma_1^p,\sigma_2^q\right) \geq \Omega(1)\cdot\min\left\{\frac{L_1\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}, \frac{L_2^{1/2}\Delta}{\varepsilon^{3/2}}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}, \left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}+\frac{\sigma_2\Delta}{\varepsilon^2}+\frac{\sigma_2\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{q(p-1)}}\right\}.$$

Remark 3.1. In the noiseless setting, i.e., $\sigma_1 = \sigma_2 = 0$, the complexity of Theorem 3.3 reduces to

$$\min\left\{\frac{L_1\Delta}{\varepsilon^2}, \frac{L_2^{1/2}\Delta}{\varepsilon^{3/2}}\right\}, \tag{2}$$

which is matched by the combination of gradient descent (GD) and cubic regularized Newton method (Nesterov and Polyak, 2006), hence is optimal. This is slightly better than the $\mathcal{O}(L_1^{1/2}L_2^{1/4}\Delta\varepsilon^{-7/4})$ bound achieved in Carmon et al. (2017), since

$$\min\left\{\frac{L_1\Delta}{\varepsilon^2}, \frac{L_2^{1/2}\Delta}{\varepsilon^{3/2}}\right\} \leq \sqrt{\frac{L_1\Delta}{\varepsilon^2}\cdot\frac{L_2^{1/2}\Delta}{\varepsilon^{3/2}}} = \frac{L_1^{1/2}L_2^{1/4}\Delta}{\varepsilon^{7/4}},$$

but achieving (2) requires full Hessian access in order to get rid of the dependency in L_1 (Arjevani et al., 2020a; Carmon et al., 2021).

[1] For clarity, we omit the optimization term in this lower bound. The full lower bound can be found in the proof of Theorem 3.3 in Appendix C.5.

4 OPTIMAL METHOD UNDER FIRST-ORDER BCM

To establish the optimal rate, we consider the well-known Normalized SGD algorithm with the momentum variance reduction (MVR) technique (Cutkosky and Orabona, 2019), and present its pseudo-code in Algorithm 1.

Algorithm 1: NSGD-MVR (Normalized SGD with MVR)

Initialization:
- $x_0 \in \mathbb{R}^d$, the starting point
- $T > 0$, the number of iterations
- $g_0 \in \mathbb{R}^d$, an initial vector
- $\gamma > 0$, the stepsize
- $\alpha \in (0,1]$, the momentum parameter for MVR

$x_1 \leftarrow x_0 - \gamma\frac{g_0}{\|g_0\|}$

For $t = 1, 2, \ldots, T-1$ **do**

// Apply MVR.

$g_t \leftarrow (1-\alpha)\left(g_{t-1} - \nabla f(x_{t-1},\xi_t)\right) + \nabla f(x_t,\xi_t)$

// Do one descent step.

$x_{t+1} \leftarrow x_t - \gamma\frac{g_t}{\|g_t\|}$

Output: x_T

4.1 Convergence Analysis

4.1.1 Case of Known Tail Indices p and q

Theorem 4.1. *Under Assumptions 2.1, 2.3 and 2.6, let the initial gradient estimate g_0 be given by*

$$g_0 = \frac{1}{B_{\text{init}}}\sum_{j=0}^{B_{\text{init}}-1}\nabla f(x_0,\xi_{0,j}),$$

where $B_{\text{init}} = \max\left\{1, \left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}\right\}$, let the stepsize $\gamma = \sqrt{\frac{\Delta\alpha^{1/q}}{\bar{L}T}}$, the momentum parameter $\alpha = \min\{1, \alpha_{\text{eff}}\}$ where

$$\alpha_{\text{eff}} = \max\left\{\left(\frac{\varepsilon}{\sigma_1 T}\right)^{\frac{p}{2p-1}}, \left(\frac{\bar{L}\Delta}{\sigma_1^2 T}\right)^{\frac{pq}{p(2q+1)-2q}}\right\}. \tag{3}$$

Then, Algorithm 1 guarantees to find an ε–stationary point with total sample complexity

$$\mathcal{O}\left(\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}+\frac{\bar{L}\Delta}{\varepsilon^2}+\frac{\bar{L}\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{q(p-1)}}\right).$$

The proof is deferred to Appendix D.

Using Assumption 2.7 combined with Assumption 2.2, we obtain the following refined result.

Theorem 4.2. *Under Assumptions 2.1 to 2.3 and 2.7, let the initial gradient estimate g_0 be given by*

$$g_0 = \frac{1}{B_{\text{init}}} \sum_{j=0}^{B_{\text{init}}-1} \nabla f\left(x_0, \xi_{0,j}\right),$$

with $B_{\text{init}} = \max\left\{1, \left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}\right\}$, *stepsize* $\gamma = \min\left\{\sqrt{\frac{\Delta}{L_1 T}}, \sqrt{\frac{\Delta \alpha^{1/q}}{\delta T}}\right\}$, *momentum parameter* $\alpha = \min\left\{1, \alpha_{\text{eff}}\right\}$, *where*

$$\alpha_{\text{eff}} = \max\left\{\left(\frac{\varepsilon}{\sigma_1 T}\right)^{\frac{p}{2p-1}}, \left(\frac{\delta\Delta}{\sigma_1^2 T}\right)^{\frac{pq}{p(2q+1)-2q}}\right\}. \quad (4)$$

Then Algorithm 1 is guaranteed to find an ε-stationary point with total sample complexity

$$\mathcal{O}\left(\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}} + \frac{(L_1+\delta)\Delta}{\varepsilon^2} + \frac{\delta\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{q(p-1)}}\right).$$

The proof is deferred to Appendix D. Essentially, the constant $L_1 + \delta$ is of the same order as the constant $\bar{L}$ from Assumption 2.6.

By (Hübler et al., 2024, Theorem 2), this implies that the lower bound from Theorem 3.2 is *tight*: it is matched, up to some constant factors, by the combination of NSGD-Mom (Normalized SGD with Momentum) with NSGD-MVR (Algorithm 1). Additionally, it is worth noting that Normalized Minibatch SGD, like NSGD-Mom, also attains the first term in the min $\{\ldots\}$ of Theorem 3.2 (Hübler et al., 2025, Corollary 3).

4.1.2 Case of Unknown Tail Indices p and q

Theorem 4.3. *Under Assumptions 2.1, 2.3 and 2.6, assume $g_0 = \nabla f(x_0, \xi_0)$, $\gamma = \sqrt{\frac{\Delta\alpha}{LT}}$, $\alpha = T^{-\frac{1}{2}} \in (0, 1]$. Then in Algorithm 1 we have*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla F(x_t)\|\right] = \mathcal{O}\left(\frac{\sigma_1}{T^{\frac{p-1}{2p}}} + \frac{\sqrt{\bar{L}\Delta}}{T^{1/4}}\right). \quad (5)$$

Our Theorem 4.3 makes the pessimistic assumption $q = 1$, under which we recover the usual $\mathcal{O}(T^{-\frac{p-1}{2p}})$ rate with unknown tail index p (Hübler et al., 2025; Liu and Zhou, 2025). The bound (5) can be slightly refined under Assumptions 2.2 and 2.7.

Theorem 4.4. *Under Assumptions 2.1 to 2.3 and 2.7, assume $g_0 = \nabla f(x_0, \xi_0)$, $\gamma = \min\left\{\sqrt{\frac{\Delta}{L_1 T}}, \sqrt{\frac{\Delta\alpha}{\delta T}}\right\}$, $\alpha = T^{-\frac{1}{2}} \in (0, 1]$. Then in Algorithm 1 we have*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla F(x_t)\|\right] = \mathcal{O}\left(\frac{\sigma_1}{T^{\frac{p-1}{2p}}} + \frac{\sqrt{\delta\Delta}}{T^{1/4}} + \frac{\sqrt{L_1\Delta}}{T^{1/2}}\right).$$

The proofs are deferred to Appendix D.

4.2 Discussion of the Obtained Results

As we can see, NSGD-MVR achieves the optimal rate: our upper bounds in Theorems 4.1 and 4.2 match the lower bounds in Theorems 3.1 and 3.2. For the standard case $q = 2$, the stochastic terms are $\frac{\bar{L}\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{p/2(p-1)}$ and $\frac{\delta\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{p/2(p-1)}$, under Assumption 2.6 ($q$-WAS) and Assumption 2.7 ((q,δ)-S), respectively. Compared to Liu et al. (2023), this yields faster rates under weaker assumptions.

In the case $p = q$, these stochastic terms coincide with those for NSGDHess in Sadiev et al. (2025) under Assumption 2.5: $\frac{\sigma_2\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{1/(p-1)}$. This relation follows since Assumption 2.7 implies Assumption 2.5, and under certain conditions $\delta \leq \sigma_2$. Moreover, for empirical risk minimization with $p = q = 2$, the two assumptions are equivalent (Arjevani et al., 2020a).

These insights motivate the study of a more general framework for second-order stochastic optimization.

5 NEAR-OPTIMAL METHOD UNDER BCM GRADIENTS AND BCM HESSIANS

Unlike Algorithm 1, we consider a method that employs Hessian–vector products instead of differences of stochastic gradients. This modification leads to the Hessian-corrected momentum (Hess) technique (Tran and Cutkosky, 2022; Salehkaleybar et al., 2022).

Algorithm 2: NSGD-Hess (Normalized SGD with Hessian-corrected Momentum)

Initialization:
- $x_0 \in \mathbb{R}^d$, the starting point
- $T > 0$, the number of iterations
- $g_0 \in \mathbb{R}^d$, an initial vector
- $\gamma > 0$, the stepsize
- $\alpha \in (0, 1]$, the momentum parameter for MVR

$x_1 \leftarrow x_0 - \gamma \frac{g_0}{\|g_0\|}$

For $t = 1, 2, \ldots, T-1$ **do**
- Sample $q_t \sim \mathcal{U}([0, 1])$
- $\hat{x}_t \leftarrow q_t x_t + (1 - q_t) x_{t-1}$
- // $\xi_t, \hat{\xi}_t \sim \mathcal{D}$ are independent.
- // Apply Hessian-vector products
- $g_t \leftarrow (1-\alpha)\left(g_{t-1} + \nabla^2 f(\hat{x}_t, \hat{\xi}_t)(x_t - x_{t-1})\right) + \alpha \nabla f\left(x_t, \xi_t\right)$
- // Do one descent step.
- $x_{t+1} \leftarrow x_t - \gamma \frac{g_t}{\|g_t\|}$

Output: x_T

Remark 5.1. Algorithm 2 is not new and has been considered in Sadiev et al. (2025). Here, we provide a refined analysis and improve over Sadiev et al. (2025, Theorem 2). Note that Algorithm 2 is *zero-respecting* (see Definition B.4) since $\operatorname{supp}(\hat{x}_t) \subseteq \operatorname{supp}(x_{t-1}) \cup \operatorname{supp}(x_t)$, where $\operatorname{supp}(x) := \{i \in [d] : x_i \neq 0\}$ is the support of $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$.

Additionally, the use of $\hat{x}_t$ taken uniformly at random on the line $[x_{t-1}, x_t]$ allows to use Lemma G.1, which provides a better overall sample complexity than the one derived in Tran and Cutkosky (2022).

Theorem 5.1. *Under Assumptions 2.1 to 2.5, let the initial gradient estimate g_0 be given by*

$$g_0 = \frac{1}{B_{\text{init}}} \sum_{j=0}^{B_{\text{init}}-1} \nabla f(x_0, \xi_{0,j}),$$

where $B_{\text{init}} = \max\left\{1, \left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}\right\}$, let the stepsize

$$\gamma = \min\left\{\sqrt{\frac{\Delta}{L_1 T}}, \sqrt{\frac{\Delta \alpha^{1/q}}{\sigma_2 T}}, \sqrt[3]{\frac{\Delta \alpha^{1/2}}{L_2 T}}\right\},$$

the momentum parameter $\alpha = \min\{1, \alpha_{\text{eff}}\}$ where $\alpha_{\text{eff}} =$

$$\max\left\{\left(\frac{\varepsilon}{\sigma_1 T}\right)^{\frac{p}{2p-1}}, \left(\frac{\sigma_2 \Delta}{\sigma_1^2 T}\right)^{\frac{pq}{p(2q+1)-2q}}, \left(\frac{L_2^{1/2}\Delta}{\sigma_1^{3/2} T}\right)^{\frac{4p}{7p-6}}\right\}.$$

Then, Algorithm 2 guarantees to find an ε–stationary point with the total sample complexity[2]

$$\mathcal{O}\left(\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}} + \frac{\Delta\sigma_2}{\varepsilon^2} + \frac{\Delta\sigma_2}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{q(p-1)}} + \frac{L_2^{1/2}\Delta\sigma_1^{1/4}}{\varepsilon^{7/4}}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{1}{4(p-1)}}\right).$$

The proof is deferred to Appendix E.

6 GRADIENT CLIPPING FOR HIGH-PROBABILITY CONVERGENCE

In this section, we conduct a high-probability analysis under heavy-tailed noise. Since we work under Assumption 2.3, we incorporate gradient clipping in Algorithm 1; the pseudo-code of the new method is provided in Algorithm 3. Formally, a clipping operator (clipping for short) is defined as

$$\texttt{clip}(v, \lambda) = \min\left\{1, \frac{\lambda}{\|v\|}\right\} v \quad \text{for any } v \neq 0 \text{ in } \mathbb{R}^d,$$

where $\lambda > 0$ is called clipping level/threshold. The proofs are deferred to Appendix F.

Algorithm 3: D-clip-NSGD-MVR (Double Clipped Normalized SGD with MVR)

Initialization:
- $x_0 \in \mathbb{R}^d$, the starting point
- $T > 0$, the number of iterations
- $g_0 \in \mathbb{R}^d$, an initial vector
- $\gamma > 0$, the stepsize
- $\alpha \in (0, 1]$, the momentum parameter for MVR
- $\lambda_1, \lambda_2 > 0$, the clipping thresholds

$x_1 \leftarrow x_0 - \gamma \frac{g_0}{\|g_0\|}$

For $t = 1, 2, \ldots, T-1$ **do**

// Apply MVR with clip operator.

$g_t \leftarrow (1-\alpha)\big(g_{t-1} + \texttt{clip}(\nabla f(x_t, \xi_t) - \nabla f(x_{t-1}, \xi_t), \lambda_1)\big) + \alpha\, \texttt{clip}(\nabla f(x_t, \xi_t), \lambda_2)$

// Do one descent step.

$x_{t+1} \leftarrow x_t - \gamma \frac{g_t}{\|g_t\|}$

Output: x_T

Algorithm 3 modifies Algorithm 1 by applying clipping not only to the stochastic gradient — to control heavy-tailed noise — but also to the gradient difference term. This additional clipping step is the primary feature distinguishing our method from Accelerated NSGD with clipping and momentum (Liu et al., 2023). Crucially, this modification enables us to establish high-probability guarantees under assumptions weaker than those in Liu et al. (2023). While their analysis relies on individual smoothness (i.e., assuming $f(\cdot, \xi)$ is $\bar{L}$-smooth almost surely), our analysis holds under q-WAS (Assumption 2.6). Individual smoothness is a significantly stronger condition; notably, under that assumption alone, Lei et al. (2019) showed it is possible to attain an $\mathcal{O}(\varepsilon^{-4})$ rate independent of the heavy-tail index p.

Theorem 6.1. *Under Assumptions 2.1, 2.3 and 2.6, let $T \geq 1$ and $\beta \in (0, 1]$ be such that $\log \frac{8T}{\beta} \geq 1$. Let $x_0 \in \mathbb{R}^d$ and define $\Delta_1 := F(x_0) - F^{\inf}$. Suppose that Algorithm 3 is run with $g_0 = 0$, momentum parameter $\alpha = \max\{T^{-\frac{p}{2p-1}}, T^{-\frac{pq}{p(2q+1)-2q}}\}$, clipping thresholds $\lambda_1 = 2\gamma\bar{L}\alpha^{-\frac{1}{q}}$ and $\lambda_2 = \max\{4\sqrt{\bar{L}\Delta_1}, \sigma_1\alpha^{-\frac{1}{p}}\}$,*

[2] For clarity, we omit the optimization terms in this upper bound. The full upper bound can be found in the proof of Theorem 5.1 in Appendix E.2.

and stepsize

$$\gamma = \mathcal{O}\left(\min\left\{\sqrt{\frac{\Delta_1}{\bar{L}T}}, \alpha\sqrt{\frac{\Delta_1}{\bar{L}}}, \frac{1}{\alpha T \log\frac{T}{\beta}}\sqrt{\frac{\Delta_1}{\bar{L}}}, \frac{\Delta_1}{\sigma_1 \alpha^{\frac{p-1}{p}} T \log\frac{T}{\beta}}, \sqrt{\frac{\Delta_1 \alpha^{\frac{1}{q}}}{\bar{L}T\log\frac{T}{\beta}}}\right\}\right).$$

Then, with probability at least $1-\beta$*, the output of Algorithm 3 satisfies*

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F(x_t)\| \leq \frac{2\Delta_1}{\gamma T},$$

and, by our choice of parameters, the gradient norm converges with high probability at the rate

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F(x_t)\| = \mathcal{O}\left(\left(\frac{\sqrt{\bar{L}\Delta_1}+\sigma_1}{T^{\frac{p-1}{2p-1}\wedge\frac{q(p-1)}{p(2q+1)-2q}}}\right)\log\frac{T}{\beta}\right).$$

The detailed proof is provided in Appendix F.

Notably, in the case where $p = q$, we establish the same rate $\widetilde{\mathcal{O}}\left(\varepsilon^{\frac{2p-1}{p-1}}\right)$ as Liu et al. (2023) did for Algorithm 2, but under weaker assumptions. Moreover, our lower bounds (see Theorem 3.1) indicate that this rate is optimal in terms of the dependence on ε.

Another interesting observation is that when $p \leq q$, the high-probability rate remains identical to the $p = q$ case. In contrast, for in-expectation guarantees, we observe a faster rate (see Theorem 4.1). Determining whether this gap is fundamental or a limitation of the current proof technique remains an open question, and calls for a tighter high-probability analysis for these methods.

Furthermore, Appendix F establishes high-probability convergence guarantees for Algorithm 3 under Assumption 2.7, and for Algorithm 4 under Assumptions 2.4 and 2.5.

Interpretation of Clipping Thresholds and Practical Implementation. While deriving optimal high-probability bounds in heavy-tailed settings remains an open challenge, our double-clipping mechanism provides a significant theoretical advantage: it establishes convergence guarantees without assuming bounded noise—a restrictive assumption often required by prior methods lacking this specific mechanism (e.g., Liu et al. (2023)).

Beyond the theoretical guarantees, our analysis offers a crucial insight for practical implementation. Under Assumption 2.3 (heavy-tailed noise, exponent p) and Assumptions 2.6 and 2.7 (q-WAS and (q,δ)-S, exponent q), our theoretical clipping thresholds scale as $\lambda_1 \sim \mathcal{O}(\gamma\alpha^{-1/q})$ for the gradient differences, and $\lambda_2 \sim \mathcal{O}(\alpha^{-1/p})$ for the raw stochastic gradients. In practice, we can rewrite the gradient difference clipping step by factoring out the stepsize γ:

$$\begin{aligned}&\texttt{clip}\big(\nabla f(x_t,\xi_t) - \nabla f(x_{t-1},\xi_t), \lambda_1\big) = \\ &= \gamma \cdot \texttt{clip}\left(\frac{\nabla f(x_t,\xi_t) - \nabla f(x_{t-1},\xi_t)}{\gamma}, \bar{\lambda}_1\right),\end{aligned}$$

where we define the rescaled threshold $\bar{\lambda}_1 := \gamma^{-1}\lambda_1$. This reveals that $\bar{\lambda}_1$ scales as $\mathcal{O}(\alpha^{-1/q})$. Consequently, in the standard case where $p = q$, both thresholds share the exact same scaling: $\bar{\lambda}_1 \sim \lambda_2 \sim \mathcal{O}(\alpha^{-1/p})$. This is a highly valuable insight for practitioners, as it implies that we do not need to tune two independent clipping hyperparameters. Instead, we can simply tune a single clipping parameter $\lambda := \bar{\lambda}_1 = \lambda_2$, significantly simplifying the deployment of D-clip-NSGD-MVR.

Acknowledgments

The work of Adrien Fradin was performed during a summer research internship in the Optimization and Machine Learning Lab at KAUST led by Peter Richtárik.

This work was supported by funding from King Abdullah University of Science and Technology (KAUST):

i) KAUST Baseline Research Scheme,

ii) Center of Excellence for Generative AI, under award no. 5940,

iii) Competitive Research Grant (CRG) Program, under award no. 6460,

iv) SDAIA-KAUST Center of Excellence in Data Science and Artificial Intelligence (SDAIA-KAUST AI).

References

Ahn, K., Cheng, X., Song, M., Yun, C., Jadbabaie, A., and Sra, S. (2024). Linear attention is (maybe) all you need (to understand transformer optimization). In *The Twelfth International Conference on Learning Representations*.

Allen-Zhu, Z. (2017). Katyusha: the first direct acceleration of stochastic gradient methods. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2017, page 1200–1205, New York, NY, USA. Association for Computing Machinery.

Andreas, W. (2014). Moments and absolute moments of the normal distribution. preprint arXiv:1209.4340.

Arjevani, Y., Carmon, Y., Duchi, J. C., Foster, D. J., Sekhari, A., and Sridharan, K. (2020a). Second-order information in non-convex stochastic optimization: Power and limitations. In Abernethy, J. and Agarwal, S., editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 242–299. PMLR.

Arjevani, Y., Carmon, Y., Duchi, J. C., Foster, D. J., Srebro, N., and Woodworth, B. (2022). Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, pages 1–50.

Arjevani, Y., Shamir, O., and Srebro, N. (2020b). A tight convergence analysis for stochastic gradient descent with delayed updates. In *Algorithmic Learning Theory*, pages 111–132. PMLR.

Armacki, A., Bajovic, D., Jakovetic, D., and Kar, S. (2025a). Optimal high-probability convergence of nonlinear SGD under heavy-tailed noise via symmetrization. preprint arXiv:2507.09093.

Armacki, A., Yu, S., Sharma, P., Joshi, G., Bajovic, D., Jakovetic, D., and Kar, S. (2024). Nonlinear stochastic gradient descent and heavy-tailed noise: A unified framework and high-probability guarantees. *CoRR*, abs/2410.13954.

Armacki, A., Yu, S., Sharma, P., Joshi, G., Bajovic, D., Jakovetic, D., and Kar, S. (2025b). High-probability convergence bounds for online nonlinear stochastic gradient descent under heavy-tailed noise. In Li, Y., Mandt, S., Agrawal, S., and Khan, E., editors, *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 1774–1782. PMLR.

Battash, B., Wolf, L., and Lindenbaum, O. (2024). Revisiting the noise model of stochastic gradient descent. In Dasgupta, S., Mandt, S., and Li, Y., editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 4780–4788. PMLR.

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

Carmon, Y., Duchi, J. C., Hinder, O., and Sidford, A. (2017). "Convex until proven guilty": dimension-free acceleration of gradient descent on non-convex functions. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 654–663. JMLR.org.

Carmon, Y., Duchi, J. C., Hinder, O., and Sidford, A. (2020). Lower bounds for finding stationary points I. *Mathematical Programming*, 184(1):71–120.

Carmon, Y., Duchi, J. C., Hinder, O., and Sidford, A. (2021). Lower bounds for finding stationary points II: first-order methods. *Mathematical Programming*, 185(1):315–355.

Chen, W. (1993). Landau-Kolmogorov inequality on a finite interval. *Bulletin of the Australian Mathematical Society*, 48(3):485–494.

Cutkosky, A. and Mehta, H. (2020). Momentum improves normalized SGD. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2260–2268. PMLR.

Cutkosky, A. and Mehta, H. (2021). High-probability bounds for non-convex stochastic optimization with heavy tails. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 4883–4895. Curran Associates, Inc.

Cutkosky, A. and Orabona, F. (2019). *Momentum-based variance reduction in non-convex SGD*. Curran Associates Inc., Red Hook, NY, USA.

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

Defazio, A., Bach, F., and Lacoste-Julien, S. (2014). SAGA: a fast incremental gradient method with support for non-strongly convex composite objectives. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'14, page 1646–1654, Cambridge, MA, USA. MIT Press.

DLMF (2025). NIST Digital Library of Mathematical Functions. https://dlmf.nist.gov/, Release 1.2.4 of 2025-03-15. Edited by F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain.

Fabry, C. (1987). An elementary proof of Gorny's inequality. *Proceedings of the Royal Society of Edinburgh: Section A Mathematics*, 105(1):345–349.

Fang, C., Li, C. J., Lin, Z., and Zhang, T. (2018). SPIDER: Near-optimal non-convex optimization via stochastic path integrated differential estimator. In *NeurIPS Information Processing Systems*.

Fatkhullin, I., Hübler, F., and Lan, G. (2025). Can SGD handle heavy-tailed noise? preprint arXiv:2508.04860.

Garg, S., Zhanson, J., Parisotto, E., Prasad, A., Kolter, Z., Lipton, Z., Balakrishnan, S., Salakhut-

dinov, R., and Ravikumar, P. (2021). On proximal policy optimization's heavy-tailed gradients. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 3610–3619. PMLR.

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

Gorbunov, E., Danilova, M., and Gasnikov, A. (2020). Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Gorbunov, E., Danilova, M., Shibaev, I., Dvurechensky, P., and Gasnikov, A. (2024a). High-probability complexity bounds for non-smooth stochastic convex optimization with heavy-tailed noise. *Journal of Optimization Theory and Applications*, 203(3):2679–2738.

Gorbunov, E., Sadiev, A., Danilova, M., Horváth, S., Gidel, G., Dvurechensky, P., Gasnikov, A., and Richtárik, P. (2024b). High-probability convergence for composite and distributed stochastic minimization and variational inequalities with heavy-tailed noise. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Hazan, E., Levy, K. Y., and Shalev-Shwartz, S. (2015). Beyond convexity: stochastic quasi-convex optimization. In *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 1594–1602, Cambridge, MA, USA. MIT Press.

He, C., Lu, Z., Sun, D., and Deng, Z. (2025). Complexity of normalized stochastic first-order methods with momentum under heavy-tailed noise.

Hübler, F., Fatkhullin, I., and He, N. (2024). From gradient clipping to normalization for heavy tailed SGD. In *OPT 2024: Optimization for Machine Learning*.

Hübler, F., Fatkhullin, I., and He, N. (2025). From gradient clipping to normalization for heavy tailed sgd. In Li, Y., Mandt, S., Agrawal, S., and Khan, E., editors, *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 2413–2421. PMLR.

Hübler, F., Yang, J., Li, X., and He, N. (2024). Parameter-agnostic optimization under relaxed smoothness. In Dasgupta, S., Mandt, S., and Li, Y., editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 4861–4869. PMLR.

Islamov, R., As, Y., and Fatkhullin, I. (2025). Safe-EF: Error feedback for non-smooth constrained optimization. In *Forty-second International Conference on Machine Learning*.

Jakovetić, D., Bajović, D., Sahu, A. K., Kar, S., Milošević, N., and Stamenković, D. (2023). Nonlinear gradient mappings and stochastic optimization: A general framework with applications to heavy-tail noise. *SIAM Journal on Optimization*, 33(2):394–423.

Johnson, R. and Zhang, T. (2013). Accelerating stochastic gradient descent using predictive variance reduction. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'13, page 315–323, Red Hook, NY, USA. Curran Associates Inc.

Kiwiel, K. C. (2001). Convergence and efficiency of subgradient methods for quasiconvex minimization. *Mathematical Programming*, 90(1):1–25.

Kovalev, D., Horváth, S., and Richtárik, P. (2020). Don't jump through hoops and remove those loops: SVRG and Katyusha are better without the outer loop. In *Algorithmic learning theory*, pages 451–467. PMLR.

Lan, G., Li, Z., and Zhou, Y. (2019). A unified variance-reduced accelerated gradient method for convex optimization. *Advances in Neural Information Processing Systems*, 32.

Lei, Y., Hu, T., Li, G., and Tang, K. (2019). Stochastic gradient descent for nonconvex learning without bounded gradient assumptions. *IEEE transactions on neural networks and learning systems*, 31(10):4394–4400.

Levy, K. (2017). Online to offline conversions, universality and adaptive minibatch sizes. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Levy, K. Y. (2016). The power of normalization: Faster evasion of saddle points. preprint arXiv:1611.04831.

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

Liu, D., Nguyen, L. M., and Tran-Dinh, Q. (2020). An optimal hybrid variance-reduced algorithm for

stochastic composite nonconvex optimization. *arXiv preprint arXiv:2008.09055*.

Liu, Z., Zhang, J., and Zhou, Z. (2023). Breaking the lower bound with (little) structure: Acceleration in non-convex stochastic optimization with heavy-tailed noise. In Neu, G. and Rosasco, L., editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 2266–2290. PMLR.

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

Liu, Z. and Zhou, Z. (2025). Nonconvex stochastic optimization under heavy-tailed noises: Optimal convergence without gradient clipping. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*. Available at https://openreview.net/forum?id=NKotdPUc3L.

Mairal, J. (2013). Optimization with first-order surrogate functions. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, page III–783–III–791. JMLR.org.

Mitrinović, D. S., Pečarić, J. E., and Fink, A. M. (1991). *Landau-Kolmogorov and Related Inequalities*, pages 1–65. Springer Netherlands, Dordrecht.

Nacson, M. S., Lee, J., Gunasekar, S., Savarese, P. H. P., Srebro, N., and Soudry, D. (2019). Convergence of gradient descent on separable data. In Chaudhuri, K. and Sugiyama, M., editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 3420–3428. PMLR.

Nazin, A. V., Nemirovsky, A. S., Tsybakov, A. B., and Juditsky, A. B. (2019). Algorithms of robust stochastic optimization based on mirror descent method. *Automation and Remote Control*, 80(9):1607–1627.

Nesterov, Y. (2018). *Lectures on convex optimization*, volume 137. Springer.

Nesterov, Y. and Polyak, B. T. (2006). Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205.

Nesterov, Y. E. (1984). Minimization methods for nonsmooth convex and quasiconvex functions. *Matekon*, 29(3):519–531.

Nguyen, T. D., Nguyen, T. H., Ene, A., and Nguyen, H. (2023). Improved convergence in high probability of clipped gradient methods with heavy tailed noise. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 24191–24222. Curran Associates, Inc.

Pascanu, R., Mikolov, T., and Bengio, Y. (2013). On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, page III–1310–III–1318. JMLR.org.

Polyak, B. T. and Tsypkin, Y. Z. (1979). Adaptive estimation algorithms: convergence, optimality, stability. *Automat. i Telemekh.*, 40(3):378–389. Translated from original in Russian.

Puchkin, N., Gorbunov, E., Kutuzov, N., and Gasnikov, A. (2024). Breaking the heavy-tailed noise barrier in stochastic optimization problems. In Dasgupta, S., Mandt, S., and Li, Y., editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 856–864. PMLR.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *Ann. Math. Statistics*, 22:400–407.

Roux, N. L., Schmidt, M., and Bach, F. (2012). A stochastic gradient method with an exponential convergence rate for finite training sets. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'12, page 2663–2671, Red Hook, NY, USA. Curran Associates Inc.

Sadiev, A., Danilova, M., Gorbunov, E., Horváth, S., Gidel, G., Dvurechensky, P., Gasnikov, A., and Richtárik, P. (2023). High-probability bounds for stochastic optimization and variational inequalities: the case of unbounded variance. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J., editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29563–29648. PMLR.

Sadiev, A., Richtárik, P., and Fatkhullin, I. (2025). Second-order optimization under heavy-tailed noise: Hessian clipping and sample complexity limits. *arXiv preprint arXiv:2510.10690*.

Salehkaleybar, S., Khorasani, S., Kiyavash, N., He, N., and Thiran, P. (2022). Adaptive momentum-based policy gradient with second-order information. *arXiv preprint arXiv:2205.08253*.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. preprint arXiv:1707.06347.

Shalev-Shwartz, S. and Zhang, T. (2013). Stochastic dual coordinate ascent methods for regularized loss. *J. Mach. Learn. Res.*, 14(1):567–599.

Song, C., Jiang, Y., and Ma, Y. (2020). Variance reduction via accelerated dual averaging for finite-sum optimization. *Advances in Neural Information Processing Systems*, 33:833–844.

Sun, Z. and Wei, E. (2025). Improved lower bounds for first-order stochastic non-convex optimization under markov sampling. In *Forty-second International Conference on Machine Learning*.

Tao, S., Xinwang, L., and Kun, Y. (2025). Revisiting gradient normalization and clipping for nonconvex sgd under heavy-tailed noise: Necessity, sufficiency, and acceleration. *Journal of Machine Learning Research*, 26:1–42.

Tran, H. and Cutkosky, A. (2022). Better SGD using second-order momentum. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS'22, Red Hook, NY, USA. Curran Associates Inc.

Tran-Dinh, Q., Pham, N. H., Phan, D. T., and Nguyen, L. M. (2019). Hybrid stochastic gradient descent algorithms for stochastic nonconvex optimization. *arXiv preprint arXiv:1905.05920*.

Tyurin, A., Gruntkowska, K., and Richtárik, P. (2025). Freya PAGE: first optimal time complexity for large-scale nonconvex finite-sum optimization with heterogeneous asynchronous computations. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS'24, Red Hook, NY, USA. Curran Associates Inc.

Tyurin, A. and Richtárik, P. (2024). On the optimal time complexities in decentralized stochastic asynchronous optimization. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C., editors, *Advances in Neural Information Processing Systems*, volume 37, pages 122652–122705. Curran Associates, Inc.

Yang, J., Li, X., Fatkhullin, I., and He, N. (2023). Two sides of one coin: the limits of untuned SGD and the power of adaptive methods. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS'23, Red Hook, NY, USA. Curran Associates Inc.

You, Y., Gitman, I., and Ginsburg, B. (2017). Large batch training of convolutional networks. preprint arXiv:1708.03888.

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

Zhang, J., Karimireddy, S. P., Veit, A., Kim, S., Reddi, S., Kumar, S., and Sra, S. (2020). Why are adaptive methods good for attention models? In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Zhou, K., Ding, Q., Shang, F., Cheng, J., Li, D., and Luo, Z.-Q. (2019). Direct acceleration of SAGA using sampled negative momentum. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1602–1610. PMLR.

Şimşekli, U., Gürbüzbalaban, M., Nguyen, T. H., Richard, G., and Sagun, L. (2019a). On the heavy-tailed theory of stochastic gradient descent for deep neural networks. preprint arXiv:1912.00018.

Şimşekli, U., Sagun, L., and Gürbüzbalaban, M. (2019b). A tail-index analysis of stochastic gradient noise in deep neural networks. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5827–5837. PMLR.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Yes, this is the main focus of the whole work.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries.
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes, our assumptions are clearly stated.
   (b) Complete proofs of all theoretical results. Yes.
   (c) Clear explanations of any assumptions. Yes.
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Not Applicable, our paper is theoretical.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Not Applicable, our paper is theoretical.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Not Applicable, our paper is theoretical.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). Not Applicable, our paper is theoretical.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. Not Applicable.
   (b) The license information of the assets, if applicable. Not Applicable.
   (c) New assets either in the supplemental material or as a URL, if applicable. Not Applicable.
   (d) Information about consent from data providers/curators. Not Applicable.
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable.
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. Not Applicable.
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable.
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable.

Tight Lower Bounds and Optimal Algorithms for Stochastic Nonconvex Optimization with Heavy-Tailed Noise: Supplementary Material

Contents

A NOTATION

Following Arjevani et al. (2020a), a q^{th}-order ($q \geq 0$) tensor $T \in \mathbb{R}^{d\times\cdots\times d} = \mathbb{R}^{\otimes^q d}$ is a q-dimensional array of real numbers, where $\mathbb{R}^{\otimes^q d}$ denotes the q-fold tensor product of $\mathbb{R}^d$. By convention, a 0^{th}-order tensor corresponds to a scalar (i.e., an element of $\mathbb{R}$), a 1^{st}-order tensor corresponds to a vector in $\mathbb{R}^d$, and a 2^{nd}-order tensor corresponds to a matrix in $\mathbb{R}^{d\times d}$. If $q \geq 1$, we denote $T = (T_1, \ldots, T_d)$ where each T_i is the $(q-1)^{\text{th}}$-order subtensor of T obtained by fixing the first index to i. Formally, for all $i \in [d]$, we define $[T_i]_{j_1,\ldots,jq-1} = T_{i,j_1,\ldots,j_{q-1}}$ where $j_1, \ldots, j_{q-1} \in [d]$. This recursive definition allows us to view any tensor as an ordered collection of its subtensors along a given mode. When $q = 2$, i.e., when T is a matrix, we write $[T]_{i,\cdot}$ for its i^{th} row to avoid confusion with the i^{th} coordinate of a vector. Similarly, $[T]_{\cdot,j}$ denotes its j^{th} column. Throughout this work, we only consider tensors of order 1 and 2: first-order tensors correspond to gradients, and second-order tensors correspond to Hessians.

For any integer $n > 0$, we define the index set $[n] := \{1, 2, \ldots, n\}$. Let $d \geq 1$ denote the ambient dimension. We use $\langle\cdot,\cdot\rangle$ to represent the standard Euclidean inner product on $\mathbb{R}^d$, i.e., $\langle x, y\rangle = \sum_{i=1}^d x_i y_i$, and $\|\cdot\|$ to denote the associated ℓ^2–norm, $\|x\| = \sqrt{\langle x, x\rangle}$. For a matrix $A \in \mathbb{R}^{d\times d}$, we denote by $\|A\|_{\text{op}}$ its operator (spectral) norm, defined as $\|A\|_{\text{op}} = \sup_{\|x\|=1} \|Ax\|$. We write $\nabla f(\cdot,\cdot)$ and $\nabla^2 f(\cdot,\cdot)$ for the stochastic gradient and Hessian oracles, respectively. For any two real numbers a, b, we use $a \wedge b := \min\{a, b\}$ and $a \vee b := \max\{a, b\}$. We adopt the standard asymptotic notations $\mathcal{O}(\cdot)$ and $\Omega(\cdot)$ to denote upper and lower bounds on growth rates. To prevent ambiguity, constants related to first-order (gradient) quantities are indexed by the subscript 1 (e.g., L_1, σ_1), whereas those related to second-order (Hessian) quantities carry the subscript 2 (e.g., L_2, σ_2).

We denote by $\mathbb{P}(E)$ the probability of an event E and by $\mathbb{E}[X]$ the expectation of a random variable X. Conditional probability and conditional expectation are written, respectively, as $\mathbb{P}(E \mid \mathcal{F})$ and $\mathbb{E}[X \mid \mathcal{F}]$, where $\mathcal{F}$ denotes a σ–algebra or a conditioning event. The notation $\mathbb{E}_\xi[\cdot]$ (or equivalently $\mathbb{E}_{\xi\sim\mathbb{P}}[\cdot]$) indicates that the expectation is taken with respect to the randomness of ξ, and explicitly that ξ is distributed according to the probability law $\mathbb{P}$. This notation is used to clarify the source of randomness when multiple random variables are involved or when the distribution of ξ is not immediately clear from the context.

B ADDITIONAL DEFINITIONS

We present in this section the formal setup (function, oracle and algorithm/protocol classes) we considered and under which our (dimension-free) lower and upper bounds are derived.

B.1 The Setup

B.1.1 Function Class

The lower bounds developed in this work apply to algorithms that find an ε–stationary point of (nonconvex) functions. All functions considered here are defined from $\mathbb{R}^d$ to $\mathbb{R}$ where $d \geq 1$ is an integer. Depending on the assumptions (Assumptions 2.4, 2.6 and 2.7), the class of functions may vary. Formally, we define

$$\mathcal{F}(\Delta) := \left\{ F \in \mathcal{C}^1(\mathbb{R}^d, \mathbb{R}) \,:\, F(0) - \inf_{x\in\mathbb{R}^d} F(x) \leq \Delta \right\}, \tag{6}$$

which is the class of continuously differentiable and Δ–bounded functions. If the considered function further satisfies some (standard) regularity conditions (see Assumptions 2.2 and 2.4) we let

$$\mathcal{F}(\Delta, L_1) := \left\{ F \in \mathcal{C}^1(\mathbb{R}^d, \mathbb{R}) \,:\, F(0) - \inf_{x\in\mathbb{R}^d} F(x) \leq \Delta \text{ and } \|\nabla F(x) - \nabla F(y)\| \leq L_1 \|x - y\| \text{ for all } x, y \in \mathbb{R}^d \right\},$$

and

$$\mathcal{F}(\Delta, L_1, L_2) := \Bigg\{ F \in \mathcal{C}^2(\mathbb{R}^d, \mathbb{R}) \,:\, F(0) - \inf_{x\in\mathbb{R}^d} F(x) \leq \Delta \text{ and } \begin{cases} \|\nabla F(x) - \nabla F(y)\| \leq L_1 \|x - y\| \\ \left\|\nabla^2 F(x) - \nabla^2 F(y)\right\|_{\text{op}} \leq L_2 \|x - y\| \end{cases} \text{for all } x, y \in \mathbb{R}^d \Bigg\},$$

when access to second-order information is possible. The Lipchitz constants L_1 and L_2 appearing above are measured with respect to the canonical (Euclidean) ℓ^2–norm over $\mathbb{R}^d$ for the gradients and points while, for the Hessian we use the operator norm $\|\cdot\|_{\mathrm{op}}$ induced by $\|\cdot\|$. We recall in Definition H.1 the definition of $\|\cdot\|_{\mathrm{op}}$.

The dimension $d \geq 1$ appearing in the above definitions will be made explicit in the proofs (see Appendix C) and it may depends on some problem specific parameters, e.g., ε, Δ and the smoothness constants.

Note that the function class (6) as stated is too broad and lacks regularity, e.g., Lipchitz continuous gradients. Nonetheless in some cases, appropriate assumptions on the stochastic gradients (for instance, *mean-squared smoothness*, see Assumption 2.6) are enough to ensure the underlying function itself is smooth, i.e., belongs to $\mathcal{F}(\Delta, L)$ for some constant $L \geq 0$ (for the exact derivation of this fact, see (65)).

B.1.2 Oracle Class

Definition B.1 (Stochastic p^{th}-order Oracles (Arjevani et al., 2020a))**.** Given integers $d, p \geq 1$ and a function $F \in \mathcal{C}^p(\mathbb{R}^d, \mathbb{R})$, we define $\mathcal{O}_p(F)$ as the class stochastic p^{th}-order oracles, i.e., the pairs $(\mathbb{P}_\xi, \mathtt{O}_F^p)$ where $\mathbb{P}_\xi$ is a distribution on a measurable set $\mathcal{Z}$ and $\mathtt{O}_F^p$ is an unbiased mapping defined as

$$\mathtt{O}_F^p \colon (x, \xi) \mapsto (F(x), \nabla f(x, \xi), \ldots, \nabla^p f(x, \xi))\,;$$

that is, for every $r \in [p]$ we have

$$\mathbb{E}_{\xi \sim \mathbb{P}_\xi}\left[\nabla^r f(x, \xi)\right] = \nabla^r F(x).$$

Furthermore, if some of the derivative estimator (here, only the gradient estimator $\nabla f(x, \xi)$ or the Hessian estimator $\nabla^2 f(x, \xi)$ are used) satisfies additional properties or assumption depending on some parameters σ_1 (Assumption 2.3), σ_2 (Assumption 2.5), $\bar{L}$ (Assumption 2.6), δ (Assumption 2.7) and so on, the oracle class will be denoted by

$$\mathcal{O}_p(F, \sigma_1, \sigma_2 \ldots), \tag{7}$$

with all parameters listed inside the parenthesis, in an arbitrary order. As each assumption defines unambiguously its own parameters, we keep the notation (7) for simplicity as it avoids any risk of ambiguity. We also superscript some parameters when needed (e.g., δ^q, σ_1^p) to indicate the various exponents (p and q) involved in the assumptions. Notably, this has nothing to do with exponentiation.

B.1.3 Optimization Protocol and Algorithm Class

First, let us introduce some important definitions.

Definition B.2 (Support of a Vector/Tensor)**.** Let $d \geq 1$ be an integer, the support of a vector $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ is the set

$$\operatorname{supp}(x) := \{i \in [d] \,:\, x_i \neq 0\}\,,$$

i.e., the set of all indices $i \in [d]$ for which x has a nonzero i^{th} coordinate.

For a given p^{th}-order tensor $T = (T_1, \ldots, T_d) \in \mathbb{R}^{d \times \cdots \times d} = \mathbb{R}^{\otimes^p d}$, its support is defined as

$$\operatorname{supp}(T) := \{i \in [d] \,:\, T_i \neq 0\}\,,$$

where T_i is the i^{th} sub-tensor (which is a $(p-1)^{\text{th}}$-order tensor), e.g., the i^{th} row of the matrix T if $p = 2$.

Definition B.3 (Progress, "prog")**.** Let $d \geq 1$ be an integer, for any $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ and any $\alpha \in [0, +\infty)$ we define

$$\operatorname{prog}_\alpha(x) := \max\{i \in [d] \,:\, |x_i| > \alpha\}\,,$$

and $\operatorname{prog}_\alpha(x) = 0$ if $|x_i| \leq \alpha$ for all $i \in [d]$.

If $\alpha = 0$, for a given p^{th}-order tensor $T = (T_1, \ldots, T_d) \in \mathbb{R}^{d \times \cdots \times d} = \mathbb{R}^{\otimes^p d}$, its "prog" is defined as

$$\operatorname{prog}_0(T) := \max\{i \in [d] \,:\, T_i \neq 0\}\,,$$

or 0 if no there is not such index $i \in [d]$ for which $T_i \neq 0$.

Notably, if $\alpha = 0$ then $\mathrm{prog}_0(x)$ is the largest index at which $x \in \mathbb{R}^d$ has a nonzero coordinate. prog_0 will be used to capture the rate at which new coordinates are "discovered". Initially, all coordinates are set to 0 and, as we progressively acquire information from the queries to the oracle, the union of the support of the oracle responses grows. The growth rate is quantified using prog and controlled thanks to the notion os *zero-chain*, which we recall formally below in Definitions B.5 and B.6.

We recall below some technical notions from the paper of Arjevani et al. (2022).

Optimization Protocol: the lower bound guarantees obtained in this work apply to algorithms interacting with an oracle over several rounds, where in each round they may issue a batch of $K \geq 1$ queries ("*multi-point*" queries). More formally, at round $i \geq 1$, the algorithm queries the oracle at a batch $\mathbf{x}^{(i)} = (x_1^{(i)}, \ldots, x_K^{(i)})$ of K points in $\mathbb{R}^d$ and for each point $x_j^{(i)}$, $j \in [K]$, the oracle performs an independent draw $\xi^{(i)} \sim \mathbb{P}_\xi$ and replies with

$$\mathtt{O}_F^{p,K}\left(\mathbf{x}^{(i)}, \xi^{(i)}\right) := \left(\mathtt{O}_F^p\left(x_1^{(i)}, \xi^{(i)}\right), \ldots, \mathtt{O}_F^p\left(x_K^{(i)}, \xi^{(i)}\right)\right),$$

where the randomness is shared across all queries of the batch: the same seed $\xi^{(i)}$ is used. For instance, in Algorithm 1 we use a batch a $K = 2$ queries.

Optimization Algorithms: a (randomized) algorithm $\mathtt{A}$ consists of a distribution $\mathbb{P}_r$ (over a measurable set $\mathcal{R}$), a random seed $r \sim \mathbb{P}_r$ drawn at the very beginning of the protocol, and a sequence of measurable mappings $\{\mathtt{A}^{(i)}\}_{i\geq 1}$ such that $\mathtt{A}^{(i)}$ takes all the previous $i-1$ oracle responses and use the random seed $r \in \mathcal{R}$ to produce the next i^{th} query. Formally, a randomized algorithm $\mathtt{A}$ produces the sequence of iterates $\{\mathbf{x}^{(i)}_{\mathtt{A}[\mathtt{O}_F^{p,K}]}\}_{i\geq 1}$ where

$$\mathbf{x}^{(i)}_{\mathtt{A}[\mathtt{O}_F^{p,K}]} := \mathtt{A}^{(i)}\left(\left[\mathtt{O}_F^{p,K}\left(\mathbf{x}^{(1)}_{\mathtt{A}[\mathtt{O}_F^{p,K}]}, \xi^{(1)}\right), \ldots, \mathtt{O}_F^{p,K}\left(\mathbf{x}^{(i-1)}_{\mathtt{A}[\mathtt{O}_F^{p,K}]}, \xi^{(i-1)}\right)\right], r\right).$$

We define $\mathcal{A}_{\mathtt{rand}}(K)$ as the class of all algorithms that follow the aforementioned protocol with a batch size of K queries per round.

Definition B.4 (Zero-Respecting Algorithm (Arjevani et al., 2022, 2020a, Definition 1))**.** A stochastic p^{th}-order algorithm $\mathtt{A}$ is *zero-respecting* if for any function $F \in \mathcal{C}^p(\mathbb{R}^d, \mathbb{R})$ and any p^{th}-order oracle $\mathtt{O}_F^{p,K}$, the iterates $\{\mathbf{x}^{(i)}_{\mathtt{A}[\mathtt{O}_F^{p,K}]}\}_{i\geq 1}$ satisfies, for any

$$\mathrm{supp}\left(\mathbf{x}^{(i)}_{\mathtt{A}[\mathtt{O}_F^{p,K}]}\right) \subseteq \bigcup_{j=1}^{i-1} \mathrm{supp}\left(\mathtt{O}_F^{p,K}\left(\mathbf{x}^{(j)}_{\mathtt{A}[\mathtt{O}_F^{p,K}]}, \xi^{(j)}\right)\right),$$

We define $\mathcal{A}_{\mathtt{zr}}(K) \subseteq \mathcal{A}_{\mathtt{rand}}(K)$ as the class of all *zero-respecting* algorithms.

In other words, a zero-respecting algorithm cannot modify coordinates where no information is known; that is, its queries at each round have support in the union of the supports of all previous oracle responses.

In the rest of the paper, we drop the K in all our notation for simplicity and because it does not appear in any of the lower bound complexity we derive.

Proof Strategy: Zero-Chains The main motivation for building the hard instance (11) is the notion of zero-chain, which in some sense, are functions for which it is "hard" to uncover new, nonzero, coordinates.

Definition B.5 (Deterministic p^{th}-order Zero-Chain (Carmon et al., 2020, Definition 3))**.** Given an integer $p \geq 1$ and a function $F \in \mathcal{C}^p(\mathbb{R}^d, \mathbb{R})$, we say that F is a p^{th}*-order zero-chain* if, for every $x \in \mathbb{R}^d$ and every $i \in [d]$ we have

$$\mathrm{supp}\,(x) \subseteq \{1, 2, \ldots, i-1\} \quad \text{implies} \quad \bigcup_{r=1}^{p} \mathrm{supp}\left(\nabla^r F(x)\right) \subseteq \{1, 2, \ldots, i\}. \tag{8}$$

We say that F is a *zero-chain* if it is a p^{th}-order zero-chain for every integer $p \geq 1$.

In other word, Definition B.5 tells us that given $x \in \mathbb{R}^d$, we can "discover" at most one new coordinate when accessing the gradient of F at x (or any high-order derivatives of F). When dealing with stochastic estimators instead, we can extend the previous Definition B.5 as follows:

Definition B.6 (Probability-θ p^{th}-order Zero-Chain (Arjevani et al., 2020a, Definition 2)). Given an integer $p \geq 1$, $\theta \in (0,1]$, a function $F \in \mathcal{C}^p(\mathbb{R}^d, \mathbb{R})$ and derivative estimators $\nabla f(x,\xi)$, ..., $\nabla^p f(x,\xi)$ of F, we say that F is a probability-θ p^{th}-*order zero-chain* if we have

$$\mathbb{P}\left(\exists x_0 \mid \operatorname{prog}_0\left(\nabla f(x_0,\xi),\ldots,\nabla^p f(x_0,\xi)\right) > \operatorname{prog}_\alpha(x_0)+1\right) = 0, \tag{9}$$

and

$$\mathbb{P}\left(\exists x_0 \mid \operatorname{prog}_0\left(\nabla f(x_0,\xi),\ldots,\nabla^p f(x_0,\xi)\right) = \operatorname{prog}_\alpha(x_0)+1\right) \leq \theta, \tag{10}$$

We say that F is a *probability-θ zero-chain* if it is a probability-θ p^{th}-order zero-chain for every integer $p \geq 1$.

In the above Definition B.6, condition (9) is the analogue of condition (8) when we have only access to noisy derivatives estimators of F. In addition, condition (10) tells us that we have a "small" chance to discover a new coordinate upon querying the oracle, i.e., the added noise behaves "adversarially" and can slow down the discovery of new coordinates (in expectation).

An important properties of probability-θ zero-chain is the following lemma.

Lemma B.1 (Arjevani et al. (2020a, Lemma 16)). *Let $p \geq 1$, $\theta \in (0,1]$, a function $F \in \mathcal{C}^p(\mathbb{R}^d,\mathbb{R})$ and unbiased derivative estimators $\nabla f(x,\xi)$, ..., $\nabla^p f(x,\xi)$ of F which form a probability-θ zero-chain and let $\mathtt{O}_F^p$ be an oracle such that $\mathtt{O}_F^p(x,\xi) = (F(x), \nabla f(x,\xi),\ldots,\nabla^p f(x,\xi))$.*

Let $\{\mathbf{x}^{(i)}_{\mathtt{A}[\mathtt{O}_F^p]}\}_{i\geq 1}$ be the queries produced by any zero-respecting algorithm $\mathtt{A} \in \mathcal{A}_{\mathtt{zr}}$ interacting with $\mathtt{O}_F^p$. Then, with probability at least $1-\delta$, we have

$$\operatorname{prog}_0\left(\mathbf{x}^{(t)}_{\mathtt{A}[\mathtt{O}_F^p]}\right) < T \quad \textit{for all} \quad t \leq \frac{T - \log(\frac{1}{\delta})}{2\theta}.$$

The proof can be found in Arjevani et al. (2020a, Lemma 16) and Arjevani et al. (2020b, Lemma 1).

B.1.4 Complexity Measures

As in Arjevani et al. (2022, 2020a), we develop lower bounds on the *distributional complexity* for finding an ε–stationary point, which in turn, implies lower bounds on the *minimax complexity* for finding such stationary points. Formally, following Arjevani et al. (2022)

$$\mathfrak{m}^{\mathtt{zr}}_\varepsilon(K,\Delta\ldots) := \sup_{F\in\mathcal{F}(\Delta\ldots)} \sup_{(\mathbb{P}_\xi,\mathtt{O}_F)\in\mathcal{O}(F\ldots)} \inf_{\mathtt{A}\in\mathcal{A}_{\mathtt{zr}}} \inf\left\{T\geq 1 \,:\, \mathbb{E}_{\mathtt{O}_F,\mathtt{A}}\left[\left\|\nabla F\left(\mathbf{x}^{(T)}_{\mathtt{A}[\mathtt{O}_F]}\right)\right\|\right] \leq \varepsilon\right\},$$

where the expectation is taken over the randomness in the oracle $\mathtt{O}_F$ and in the algorithm $\mathtt{A} \in \mathcal{A}_{\mathtt{zr}}$, if any.

As in the definition of the oracle class, complexity measures may depend on various parameters depending on the assumptions considered. The convention is to list all involved parameters (in an arbitrary order), and there is no ambiguity in doing so.

B.2 The Worst-Case Function

In this section, we recall the "worst-case" function introduced in Carmon et al. (2020); Arjevani et al. (2022) and which is used to prove our lower bounds. This function (or some variations, depending on the targeted class of functions) is at the core of many lower bounds in stochastic nonvonvex optimization (Arjevani et al., 2022; Tyurin and Richtárik, 2024; Tyurin et al., 2025; Islamov et al., 2025; Sun and Wei, 2025; Sadiev et al., 2025). Formally, given an integer $T \geq 1$ which denotes the dimension in which the hard instance $F_T\colon \mathbb{R}^T \to \mathbb{R}$ lies, we define

$$F_T\colon x \mapsto -\Psi(1)\Phi(x_1) + \sum_{i=2}^{T}\left[\Psi(-x_{i-1})\Phi(-x_i) - \Psi(x_{i-1})\Phi(x_i)\right], \tag{11}$$

where the inner components Ψ and Φ are defined, for any $t \in \mathbb{R}$, as

$$\Psi(t) := \begin{cases} 0, & \text{if } t \leq \frac{1}{2}; \\ \exp\left(1 - \frac{1}{(2t-1)^2}\right), & \text{if } t > \frac{1}{2}; \end{cases} \quad \text{and} \quad \Phi(t) := \sqrt{e}\int_{-\infty}^{t} e^{-\frac{s^2}{2}}\,\mathrm{d}s.$$

The particular design (11) of F_T is such that it is a (deterministic) zero-chain (see Definition B.5) and the gradients of F_T are large unless all coordinates are large (i.e., $\operatorname{prog}_1(x) \geq T$); these two keys properties are recalled below (5. and 6.). Several other properties of the function F_T are stated in the next lemma.

Lemma B.2 (Properties of the Hard Instance (Carmon et al., 2020; Arjevani et al., 2022)). *For any integer $T \geq 1$, the function F_T satisfies:*

1. $F_T(0) - \inf_{x \in \mathbb{R}^T} F_T(x) \leq \Delta_0 T$ *where* $\Delta_0 := 12$,
2. *The gradient of F_T is ℓ_1–Lipschitz continuous with* $\ell_1 := 152$,
3. *For all $x \in \mathbb{R}^T$, we have* $\|\nabla F_T(x)\|_\infty \leq \gamma_\infty$ *where* $\gamma_\infty := 23$. *Notably, this shows that* $\|\nabla F_T(x)\| \leq 23\sqrt{T}$,
4. *There exists an universal constant $0 < c < +\infty$ such that for every integer $p \geq 1$, the p–th order derivatives of F_T are ℓ_p–Lipschitz continuous for some* $\ell_p \leq \exp\left(\frac{5}{2} p \log p + cp\right)$,
5. *For all $x \in \mathbb{R}^T$ we have* $\operatorname{prog}_0(\nabla F_T(x)) \leq \operatorname{prog}_{\frac{1}{2}}(x) + 1$,
6. *For all $x \in \mathbb{R}^T$, if* $\operatorname{prog}_0(x) < T$ *then* $\|\nabla F_T(x)\| > 1$.

In the proofs of our lower bounds, we mostly focus on properties 1., 2., 3. and 4. which, after a specific rescaling, are enough to ensure the hard instance F_T belongs to the right class of functions. The two remaining properties 5. and 6. are not explicitly used in this paper but their role, however, should not be underestimated: they guarantee that the inner mechanics of the hard instance of Carmon et al. (2020); Arjevani et al. (2022) remain the same, thereby ensuring that any zero-respecting algorithm on F_T behaves as intended and that their lower-bound strategy carries over.

B.3 The Stochastic Gradient and Hessian Estimator

We recall below from Arjevani et al. (2022, 2020a) the gradient and Hessian estimators used:

Definition B.7 (Gradient Estimator). Given an integer $T \geq 1$ and $\theta \in (0, 1]$, the gradient estimator $g_T \colon \mathbb{R}^T \times \mathcal{Z} \to \mathbb{R}^T$ of the hard instance F_T in (11) is defined coordinate-wise as

$$[g_T(x, \xi)]_i := [\nabla F_T(x)]_i \cdot \left(1 + \mathbb{I}\left\{i > \operatorname{prog}_{\frac{1}{4}}(x)\right\}\left(\frac{\xi}{\theta} - 1\right)\right),$$

where $i \in [T]$, $x \in \mathbb{R}^T$ and $\xi \sim \operatorname{Ber}(\theta)$, i.e., we take as distribution $\mathbb{P}_\xi = \operatorname{Ber}(\theta)$ and measurable set $\mathcal{Z} = \{0, 1\}$.

Definition B.8 (Hessian Estimator). Given an integer $T \geq 1$ and $\theta \in (0, 1]$, the Hessian estimator $\nabla g_T \colon \mathbb{R}^T \times \mathcal{Z} \to \mathbb{R}^T$ of the hard instance F_T in (11) is defined row-wise as

$$[\nabla g_T(x, \xi)]_{i,\cdot} := \left(1 + \mathbb{I}\left\{i > \operatorname{prog}_{\frac{1}{4}}(x)\right\}\left(\frac{\xi}{\theta} - 1\right)\right) \cdot \left[\nabla^2 F_T(x)\right]_{i,\cdot},$$

where $i \in [T]$, $x \in \mathbb{R}^T$ and $\xi \sim \operatorname{Ber}(\theta)$ (as before).

Note that, for all $x \in \mathbb{R}^T$ and all $i > \operatorname{prog}_{\frac{1}{4}}(x) + 1$ we have $[\nabla F_T(x)]_i = 0$ by Lemma B.2 (property 5.) so, only the specific coordinate at $i = \operatorname{prog}_{\frac{1}{4}}(x) + 1$ is noisy.

In the case of Assumptions 2.6 and 2.7, i.e., additional regularity conditions on the gradient estimator, we define, as in Arjevani et al. (2022) the "smoothed" indicator Θ_i of $\mathbb{I}\left\{i > \operatorname{prog}_{\frac{1}{4}}(\cdot)\right\}$ as:

$$\Theta_i \colon x \mapsto \Gamma\left(1 - \left(\sum_{k=i}^{T} \Gamma\left(|x_k|\right)^2\right)^{\frac{1}{2}}\right) = \Gamma\left(1 - \|(\Gamma(|x_i|), \ldots, \Gamma(|x_T|))\|\right),$$

where $\Gamma \colon \mathbb{R} \to \mathbb{R}$ is any $\mathcal{C}^\infty$, non-decreasing Lipschitz continuous function with $\Gamma(t) = 0$ for all $t \leq \frac{1}{4}$ and $\Gamma(t) = 1$ for all $t \geq \frac{1}{2}$. For instance, we can take, as in Arjevani et al. (2022)

$$\Gamma(t) := \frac{\int_{\frac{1}{4}}^{t} \Lambda(s)\,\mathrm{d}s}{\int_{\frac{1}{4}}^{\frac{1}{2}} \Lambda(s)\,\mathrm{d}s}, \quad \text{where} \quad \Lambda(t) := \begin{cases} 0, & \text{if } t \leq \frac{1}{4} \text{ or } t \geq \frac{1}{2}; \\ \exp\left(-\frac{1}{100\left(t - \frac{1}{4}\right)\left(\frac{1}{2} - t\right)}\right), & \text{if } \frac{1}{4} < t < \frac{1}{2}; \end{cases}$$

for any $t \in \mathbb{R}$. Notably, the smoothed indicator Θ_i satisfies for all $x \in \mathbb{R}^T$

$$\mathbb{I}\left\{i > \text{prog}_{\frac{1}{4}}(x)\right\} \leq \Theta_i(x) \leq \mathbb{I}\left\{i > \text{prog}_{\frac{1}{2}}(x)\right\},$$

hence

$$\Theta_i(x) = \begin{cases} 1, & \text{for all } i > \text{prog}_{\frac{1}{4}}(x); \\ 0, & \text{for all } i \leq \text{prog}_{\frac{1}{2}}(x). \end{cases}$$

Overall, this gives the following "smoothed" gradient estimator:

Definition B.9 ("Smoothed" Gradient Estimator)**.** Given an integer $T \geq 1$ and $\theta \in (0, 1]$, the gradient estimator $g_T \colon \mathbb{R}^T \times \mathcal{Z} \to \mathbb{R}^T$ of the hard instance F_T in (11) is defined coordinate-wise as

$$[g_T(x, \xi)]_i := [\nabla F_T(x)]_i \cdot \left(1 + \Theta_i(x)\left(\frac{\xi}{\theta} - 1\right)\right),$$

where $i \in [T]$, $x \in \mathbb{R}^T$ and $\xi \sim \text{Ber}(\theta)$ (as before).

The gradient and Hessian estimators defined above are all probability-θ zero-chain.

C PROOFS OF THE LOWER BOUNDS

C.1 A Lower Bound in the *Global Stochastic Model* Under Heavy-Tail Noise

In this part, we extend Arjevani et al. (2022, Lemma 11) in the setting where we only assume the variance of the gradient estimator to have bounded p–th moment for some $p \in (1, 2]$. This result holds in any dimension $d \geq 1$.

Lemma C.1. *Under Assumptions 2.1, 2.3 and 2.6 and as long as* $0 < \varepsilon \leq \frac{1}{8}\sqrt{\bar{L}\Delta}$, *the number of samples required to obtain an* ε*–stationary point in the global stochastic model defined above is*

$$\Omega(1) \cdot \left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}},$$

where $p \in (1, 2]$.

In particular, this lower bound does not depend on the exponent q from Assumption 2.6.

Remark C.1. More precisely, Lemma C.1 shows that

$$\mathfrak{m}_{\varepsilon}^{\mathtt{zr}}\left(K, \Delta, \bar{L}, \sigma_1^p\right) \geq \Omega(1) \cdot \max\left\{1, \left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}\right\}. \tag{12}$$

Remark C.2. It is also worth mentioning that the same lower bound holds under Assumptions 2.1 to 2.3 and 2.7 since, by our choice of function (14) for any $s \in \{-1, 1\}$ and any $x, y \in \mathbb{R}^d$ we have

$$\begin{aligned}
&\mathbb{E}_{\xi\sim\mathbb{P}^s}\left[\left\|\left[\nabla f_d^\star(x,\xi) - \nabla f_d^\star(y,\xi)\right] - \left[\nabla F_{d,s}^\star(x) - \nabla F_{d,s}^\star(y)\right]\right\|^q\right] \\
&\quad \overset{(15)}{=} \beta^q \mathbb{E}_{\xi\sim\mathbb{P}^s}\left[\left\|\left[\nabla f_d(\beta x,\xi) - \nabla f_d(\beta y,\xi)\right] - \left[\nabla F_{d,s}(\beta x) - \nabla F_{d,s}(\beta y)\right]\right\|^q\right] \\
&\quad \overset{(18)}{=} \beta^q \mathbb{E}_{\xi\sim\mathbb{P}^s}\left[\left\|\left[L(\beta x - \boldsymbol{\xi}) - L(\beta y - \boldsymbol{\xi})\right] - \left[L(\beta x - \theta_s) - L(\beta y - \theta_s)\right]\right\|^q\right] \\
&\quad = 0,
\end{aligned}$$

hence Assumption 2.7 is satisfied. Moreover, taking $L = \frac{2\Delta}{r^2}$ is enough to ensure that $F_{d,s}^\star$ has Δ–bounded sub-optimality (see (16)). On the other hand, for any $x, y \in \mathbb{R}^d$ we have

$$\begin{aligned}
\left\|\nabla F_{d,s}^\star(x) - \nabla F_{d,s}^\star(y)\right\| &\overset{(15)}{=} \beta \left\|\nabla F_{d,s}(\beta x) - \nabla F_{d,s}(\beta y)\right\| \\
&= \beta^2 L \left\|x - y\right\|,
\end{aligned} \tag{13}$$

and it's enough to take $0 < \beta \leq \sqrt{\frac{L_1}{L}}$ to ensure the function $F_{d,s}^\star$ has L_1–Lipschitz gradients. The rest of the proof (**steps 3**, **4** and **5**) is the same, using L_1 instead of $\bar{L}$.

Remark C.3. As a last remark before proving Lemma C.1, it is also worth noting that the stochastic function $f_d \colon \mathbb{R}^d \to \mathbb{R}$ used to prove the lower bound (12) still holds when one has access to high-order information (e.g., the Hessian), that is, under Assumptions 2.1 to 2.3 and 2.5, because by construction f_d is a quadratic function so, for any $s \in \{-1, 1\}$ and any $x, y \in \mathbb{R}^d$, we have

$$\begin{aligned}
\mathbb{E}_{\xi\sim\mathbb{P}^s}\left[\left\|\nabla^2 f_d^\star(x,\xi) - \nabla^2 F_{d,s}^\star(x)\right\|_{\mathrm{op}}^q\right] &= \beta^{2q}\mathbb{E}_{\xi\sim\mathbb{P}^s}\left[\left\|\nabla^2 f_d(\beta x,\xi) - \nabla^2 F_{d,s}(\beta x)\right\|_{\mathrm{op}}^q\right] \\
&= \beta^{2q}\mathbb{E}_{\xi\sim\mathbb{P}^s}\left[\left\|LI_d - LI_d\right\|_{\mathrm{op}}^q\right] \\
&= 0,
\end{aligned}$$

where $I_d \in \mathbb{R}^{d\times d}$ is the identity matrix. Hence, Assumption 2.5 is satisfied. Additionally, since $\nabla^2 F_{d,s}^\star(x) = L\beta^2 I_d$ for any $x \in \mathbb{R}^d$ then the Hessian of $F_{d,s}^\star$ is L_2–Lipschitz continuous for any constant $L_2 \geq 0$. Combining this observation with the arguments in the previous remark (see (13)), we obtain that $F_{d,s}^\star \in \mathcal{F}(\Delta, L_1, L_2)$ for any $d \geq 1$ and any $s \in \{-1, 1\}$, as long as we choose

$$L = \frac{2\Delta}{r^2} \quad \text{and} \quad 0 < \beta \leq \sqrt{\frac{L_1}{L}}.$$

The rest of the proof (**steps 3**, **4** and **5**) is the same.

Proof. Let the accuracy parameter $\varepsilon > 0$, initial sub-optimality $\Delta \geq 0$, the mean-squared smoothness parameter $\bar{L}$, and the variance parameter $\sigma_1 \geq 0$ be fixed, and $0 < L \leq \bar{L}$ to be specified later. Our proof follows the same steps as in Arjevani et al. (2020b) except that, taking inspiration from Hübler et al. (2025) and the gradient oracle from Arjevani et al. (2022) as recalled in Definition B.7, we take two Bernoulli distributions instead of normal distributions.

Let us fix $d \geq 1$ an integer and consider the following family of functions $f_d \colon \mathbb{R}^d \times \{-1, 0, 1\} \to \mathbb{R}$ defined as

$$f_d(x, \xi) := \frac{L}{2}\left(\|x\|^2 - 2x_1\xi + r^2\right), \tag{14}$$

where $r \in (0, 1]$ is a fixed parameter (to be specified later), $(x, \xi) \in \mathbb{R}^d \times \{-1, 0, 1\}$ and $x = (x_1, \ldots, x_d)$ are its coordinates. Then, for $\beta > 0$, we rescale the function f_d as

$$f_d^\star(x, \xi) \colon x \mapsto f_d(\beta x, \xi). \tag{15}$$

Next, we define the two probability distributions $\mathbb{P}^1 = \mathrm{Ber}(r)$ and $\mathbb{P}^{-1} = -\mathrm{Ber}(r)$ (whose support is $\{-1, 0\}$) and we let $\theta_s := (rs, 0, \ldots, 0) \in \mathbb{R}^d$ for $s \in \{-1, 1\}$. Additionally, for any $s \in \{-1, 1\}$, we define the function $F_{d,s}^\star \colon \mathbb{R}^d \to \mathbb{R}$ for all $x \in \mathbb{R}^d$ as

$$F_{d,s}^\star(x) := \mathbb{E}_{\xi \sim \mathbb{P}^s}\left[f_d^\star(x, \xi)\right] \overset{(a)}{=} \frac{L}{2}\left(\|\beta x\|^2 - 2\beta x_1 rs + r^2\right) = \frac{L}{2}\|\beta x - \theta_s\|^2 =: F_{d,s}(\beta x),$$

where (a) follows from the fact that, when $s = 1$ then $\xi \sim \mathrm{Ber}(r)$ and $\mathbb{E}[\xi] = r = rs$, while when $s = -1$ then $\xi \sim -\mathrm{Ber}(r)$ and $\mathbb{E}[\xi] = -r = rs$.

- **Step 1:** *Ensuring $F_{d,s}^\star$ has Δ–bounded initial sub-optimality (Assumption 2.1).*

  To ensure $F_{d,s}^\star$ satisfies Assumption 2.1, let us compute the initial sub-optimality. Assuming the starting point is $x^0 = 0$, we have

  $$F_{d,s}^\star(0) - \inf_{x \in \mathbb{R}^d} F_{d,s}^\star(x) \overset{(15)}{=} F_{d,s}(0) = \frac{Lr^2}{2}, \tag{16}$$

  thus, it's enough to take $L = \frac{2\Delta}{r^2} < +\infty$ (since $r > 0$) so as to ensure the function $F_{d,s}^\star$ has Δ–bounded initial sub-optimality. In the next step, we will show that $F_{d,s}^\star$ also has $\bar{L}$–Lipschitz gradients, as a consequence of the mean-squared smoothness property.

- **Step 2:** *The Oracle Class and Assumptions 2.3 and 2.6.*

  Now, let us compute the p–th central moment, for any $x \in \mathbb{R}^d$ and any $s \in \{-1, 1\}$ we have

  $$\begin{aligned}
  2\mathbb{E}_{\xi \sim \mathbb{P}^s}\left[\left\|\nabla f_d^\star(x, \xi) - \nabla F_{d,s}^\star(x)\right\|^p\right] &\overset{(15)}{=} \beta^p\, \mathbb{E}_{\xi \sim \mathbb{P}^s}\left[\|\nabla f_d(\beta x, \xi) - \nabla F_{d,s}(\beta x)\|^p\right] \\
  &\overset{(a)}{=} \beta^p\, \mathbb{E}_{\xi \sim \mathbb{P}^s}\left[\|L(\beta x - \boldsymbol{\xi}) - L(\beta x - \theta_s)\|^p\right] \\
  &= (L\beta)^p\, \mathbb{E}_{\xi \sim \mathbb{P}^s}\left[|\xi - rs|^p\right] \\
  &= (L\beta r)^p\, \mathbb{E}_{\xi \sim \mathbb{P}^s}\left[\left|\frac{\xi}{r} - s\right|^p\right] \\
  &\overset{(b)}{=} (L\beta r)^p \left((1 - r)|s|^p + r\left|\frac{s}{r} - s\right|^p\right) \\
  &= (L\beta r)^p \left((1 - r) + r\left|\frac{1}{r} - 1\right|^p\right) \\
  &= (L\beta r)^p (1 - r)\left(1 + \left(\frac{1 - r}{r}\right)^{p-1}\right) \\
  &\overset{(c)}{=} (L\beta)^p\, r(1 - r)\left(r^{p-1} + (1 - r)^{p-1}\right) \\
  &\leq 2(L\beta)^p\, r(1 - r),
  \end{aligned} \tag{17}$$

where in (a) we let $\boldsymbol{\xi} = (\xi, 0, \ldots, 0) \in \mathbb{R}^d$ and the gradient of f_d and $F_{d,s}$ are given, for any $x \in \mathbb{R}^d$ by

$$\nabla f_d(x, \xi) = L(x - \boldsymbol{\xi}) \quad \text{and} \quad \nabla F_{d,s}(x) = L(x - \theta_s), \tag{18}$$

notably, $\mathbb{E}_{\xi \sim \mathbb{P}^s}\left[\nabla f_d^\star(x, \xi)\right] = \nabla F_{d,s}^\star(x)$. In (b) we use the fact that $|s| = 1$ while in (c) we use both $p > 1$ and $r \in (0, 1]$ to bound $r^{p-1} + (1-r)^{p-1}$ by 2. Hence, following (17) we need to guarantee

$$2\,(L\beta)^p\, r(1-r) \leq \sigma_1^p \quad \text{so} \quad 2^{\frac{1}{p}} L\beta\,(r(1-r))^{\frac{1}{p}} \leq \sigma_1. \tag{19}$$

Now, computing the mean-squared smoothness constant, we have, for $q \in (1, 2]$, $s \in \{-1, 1\}$ and any $x, y \in \mathbb{R}^d$

$$\begin{aligned}
\mathbb{E}_{\xi \sim \mathbb{P}^s}\left[\|\nabla f_d^\star(x, \xi) - \nabla f_d^\star(y, \xi)\|^q\right] &\overset{(15)}{=} \beta^q\, \mathbb{E}_{\xi \sim \mathbb{P}^s}\left[\|\nabla f_d(\beta x, \xi) - \nabla f_d(\beta y, \xi)\|^q\right] \\
&\overset{(18)}{=} (L\beta)^q\, \mathbb{E}_{\xi \sim \mathbb{P}^s}\left[\|\beta x - \beta y\|^q\right] \\
&= \left(L\beta^2\right)^q \|x - y\|^q,
\end{aligned}$$

and it suffices to ensure

$$\left(L\beta^2\right)^q \leq \bar{L}^q, \quad \text{that is,} \quad L\beta^2 \leq \bar{L}, \tag{20}$$

so as to satisfy Assumption 2.6.

- **Step 3:** *Choice of β, r.*

  It remains to choose $\beta > 0$, $r \in (0, 1]$ (universal constant) to satisfy the two inequalities (19) and (20). Additionally, let $c' > 0$, to be fixed later, be an universal constant such that $0 < \varepsilon \leq c'\sqrt{\bar{L}\Delta}$. From inequality (20) we have

  $$\beta \leq \sqrt{\frac{\bar{L}}{L}} = \frac{1}{L}\sqrt{\bar{L}L} = \frac{\sqrt{2\bar{L}\Delta}}{Lr},$$

  hence for any constant $c' > 0$, taking $\beta = \frac{\varepsilon\sqrt{2}}{c'Lr} > 0$ gives

  $$\beta = \frac{\varepsilon\sqrt{2}}{c'Lr} \leq \frac{c'\sqrt{\bar{L}\Delta}\sqrt{2}}{c'Lr} = \frac{\sqrt{2\bar{L}\Delta}}{Lr},$$

  as desired. Plugging back this value in the inequality (19) we need to have

  $$\frac{2^{\frac{1}{p}}\sqrt{2}}{c'}\varepsilon r^{\frac{1}{p}-1}(1-r)^{\frac{1}{p}} \leq \sigma_1, \quad \text{i.e.,} \quad \frac{2^{\frac{1}{p}}\sqrt{2}}{c'}\left(\frac{\varepsilon}{\sigma_1}\right)(1-r)^{\frac{1}{p}} \leq r^{\frac{p-1}{p}},$$

  which is equivalent to

  $$C^{\frac{p}{p-1}}\left(\frac{\varepsilon}{\sigma_1}\right)^{\frac{p}{p-1}}(1-r)^{\frac{1}{p-1}} \leq r,$$

  where we set $C := \frac{2^{\frac{1}{p}}\sqrt{2}}{c'}$. Thus, as $1 - r \leq 1$, it is enough to take

  $$r = \min\left\{1, \left(\frac{C\varepsilon}{\sigma_1}\right)^{\frac{p}{p-1}}\right\} > 0,$$

  and now all the inequalities are satisfied so do Assumptions 2.1, 2.3 and 2.6 hold.

- **Step 4:** *Transforming the Optimization Problem into a Function Identification Problem.*

  We now continue to follow the proof of Arjevani et al. (2022). First, let us randomized the selection of the instances $\{F_{d,s}^\star\}_{s \in \{-1,1\}}$ by drawing s uniformly in the set $\{-1, 1\}$. Hence, we let $S \sim \mathcal{U}(\{-1, 1\})$ and consider any algorithm $\mathtt{A}$ that takes as input the iid samples $\xi_1, \ldots, \xi_T \sim \mathbb{P}^S$, where T is the number of queries, and which returns the (random) iterate $\widehat{x} \in \mathbb{R}^d$. We now bound the expected norm of the gradient at

$\widehat{x}$. To do so, let us define the (random) quantity $\widehat{S} := \arg\min_{s\in\{-1,1\}} \|\nabla F^\star_{d,s}(\widehat{x})\|$, breaking ties arbitrarily. Then if $S \neq \widehat{S}$ we have, by definition of $\widehat{S}$,

$$\left\|\nabla F^\star_{d,S}(\widehat{x})\right\| \geq \left\|\nabla F^\star_{d,\widehat{S}}(\widehat{x})\right\|,$$

hence

$$\begin{aligned}
2\mathbb{E}\left[\left\|\nabla F^\star_{d,S}(\widehat{x})\right\|\right] &\geq \mathbb{E}\left[\left\|\nabla F^\star_{d,S}(\widehat{x})\right\|\right] + \mathbb{E}\left[\left\|\nabla F^\star_{d,\widehat{S}}(\widehat{x})\right\|\right] \\
&= \mathbb{E}\left[\left\|\nabla F^\star_{d,S}(\widehat{x})\right\| + \left\|\nabla F^\star_{d,\widehat{S}}(\widehat{x})\right\|\right] \\
&\overset{(a)}{\geq} \inf_{x\in\mathbb{R}^d}\left(\left\|\nabla F^\star_{d,1}(x)\right\| + \left\|\nabla F^\star_{d,-1}(x)\right\|\right) \\
&\overset{(18)}{=} L\beta \inf_{x\in\mathbb{R}^d}\left(\|\beta x - \theta_1\| + \|\beta x - \theta_{-1}\|\right) \\
&\overset{(b)}{\geq} L\beta\|\theta_1 - \theta_{-1}\| \\
&= 2L\beta r, \qquad (21)
\end{aligned}$$

where in (a) we use $S \neq \widehat{S}$ and that both belongs to $\{-1,1\}$. In (b) we use the triangle inequality. Hence,

$$\begin{aligned}
\mathbb{E}\left[\left\|\nabla F^\star_{d,S}(\widehat{x})\right\|\right] &\overset{\text{Lem. H.11}}{\geq} L\beta r\mathbb{P}\left(\left\|\nabla F^\star_{d,S}(\widehat{x})\right\| \geq L\beta r\right) \\
&\overset{(21)}{\geq} L\beta r\mathbb{P}\left(\widehat{S} \neq S\right),
\end{aligned}$$

since $L\beta r > 0$ and the last inequality follows from the lower bound in (21), implied by the event $\{\widehat{S} \neq S\}$.

- **Step 5:** *Lower Bounding the Misidentification Probability* $\mathbb{P}\left(\widehat{S} \neq S\right)$.

  Next, for $s \in \{-1,1\}$ let $\mathbb{P}^s_T = s\mathrm{Ber}^{\otimes T}(r)$ be the law of $(\xi_1,\dots,\xi_T)$ conditioned on the event $\{S = s\}$, then

$$\begin{aligned}
\mathbb{P}\left(\widehat{S} \neq S\right) &= 1 - \mathbb{P}\left(\widehat{S} = S\right) \\
&= 1 - \left(\mathbb{P}(S=1)\,\mathbb{P}\left(\widehat{S} = S \middle| S = 1\right) + \mathbb{P}(S=-1)\,\mathbb{P}\left(\widehat{S} = S \middle| S = -1\right)\right) \\
&= 1 - \frac{1}{2}\left(\mathbb{P}^1_T\left(\widehat{S} = 1\right) + \mathbb{P}^{-1}_T\left(\widehat{S} \neq 1\right)\right) \\
&\leq 1 - \frac{1}{2}\sup_{A\subseteq\mathbb{R}^T \text{ mesurable}}\left(\mathbb{P}^1_T(A) + \mathbb{P}^{-1}_T(A^\mathsf{c})\right) \\
&= \frac{1}{2} - \frac{1}{2}\sup_{A\subseteq\mathbb{R}^T \text{ mesurable}}\left(\mathbb{P}^1_T(A) - \mathbb{P}^{-1}_T(A)\right) \\
&\overset{(a)}{=} \frac{1}{2}\left(1 - \left\|\mathbb{P}^1_T - \mathbb{P}^{-1}_T\right\|_{\mathrm{TV}}\right),
\end{aligned}$$

where (a) follows form the definition of the total variation distance. Then we compute $\left\|\mathbb{P}^1_T - \mathbb{P}^{-1}_T\right\|_{\mathrm{TV}}$, to do so observe that the support of $\mathbb{P}^1_T$ is $\{0,1\}^T$ and the support of $\mathbb{P}^{-1}_T$ is $\{-1,0\}^T$ thus, if we let $\mathbf{0} := (0,\dots,0)$,

$$\begin{aligned}
\left\|\mathbb{P}^1_T - \mathbb{P}^{-1}_T\right\|_{\mathrm{TV}} &= \frac{1}{2}\sum_{x\in\{-1,0,1\}^T}\left|\mathbb{P}^1_T(x) - \mathbb{P}^{-1}_T(x)\right| \\
&\overset{(a)}{=} \frac{1}{2}\sum_{x\in\{-1,0\}^T\setminus\{\mathbf{0}\}}\mathbb{P}^{-1}_T(x) + \frac{1}{2}\sum_{x\in\{0,1\}^T\setminus\{\mathbf{0}\}}\mathbb{P}^1_T(x) \\
&= \frac{1}{2}\left(1 - (1-r)^T\right) + \frac{1}{2}\left(1 - (1-r)^T\right) \\
&= 1 - (1-r)^T,
\end{aligned}$$

where (a) follows from $\mathbb{P}_T^1(\mathbf{0}) = (1-r)^T = \mathbb{P}_T^{-1}(\mathbf{0})$ along with the fact that $\{-1,0\}^T \cap \{0,1\}^T = \{\mathbf{0}\}$. Hence, we obtain

$$\mathbb{P}\left(\widehat{S} \neq S\right) \geq \frac{1}{2}(1-r)^T,$$

and plugging this bound into (21) gives

$$\mathbb{E}\left[\left\|\nabla F_{d,S}^\star(\widehat{x})\right\|\right] \geq \frac{L\beta r}{2}(1-r)^T.$$

Also, note that

$$\max\left\{\mathbb{E}\left[\left\|\nabla F_{d,1}^\star(\widehat{x})\right\|\right], \mathbb{E}\left[\left\|\nabla F_{d,-1}^\star(\widehat{x})\right\|\right]\right\} \geq \frac{1}{2}\left(\mathbb{E}\left[\left\|\nabla F_{d,1}^\star(\widehat{x})\right\|\right] + \mathbb{E}\left[\left\|\nabla F_{d,-1}^\star(\widehat{x})\right\|\right]\right) = \mathbb{E}\left[\left\|\nabla F_{d,S}^\star(\widehat{x})\right\|\right].$$

It remains to lower bound adequately the quantity $\frac{L\beta r}{2}(1-r)^T$. For this, we consider two cases:

– if $\sigma_1 > C\varepsilon$ then $1 > \frac{C\varepsilon}{\sigma_1}$ so $1 > \left(\frac{C\varepsilon}{\sigma_1}\right)^{\frac{p}{p-1}}$ thus $r = \left(\frac{C\varepsilon}{\sigma_1}\right)^{\frac{p}{p-1}}$ so

$$\frac{L\beta r}{2}(1-r)^T = \frac{\varepsilon\sqrt{2}}{2c'}(1-r)^T,$$

and,

∗ assume we have $T \leq \frac{1}{2r}$ then

$$(1-r)^T \overset{(a)}{\geq} 1 - rT \geq \frac{1}{2},$$

where (a) follows from the Bernoulli's inequality (see Lemma H.12) thus

$$\max\left\{\mathbb{E}\left[\left\|\nabla F_{d,1}^\star(\widehat{x})\right\|\right], \mathbb{E}\left[\left\|\nabla F_{d,-1}^\star(\widehat{x})\right\|\right]\right\} \geq \frac{\varepsilon\sqrt{2}}{4c'},$$

and it's enough to take $c' = \frac{1}{8}$ to ensure

$$\max\left\{\mathbb{E}\left[\left\|\nabla F_{d,1}^\star(\widehat{x})\right\|\right], \mathbb{E}\left[\left\|\nabla F_{d,-1}^\star(\widehat{x})\right\|\right]\right\} \geq 2\varepsilon\sqrt{2} > 2\varepsilon$$

So, when $T \leq \frac{1}{2r}$ it is not possible to reach an ε–stationary point on both $F_{d,1}^\star$ and on $F_{d,-1}^\star$.

∗ Hence, we deduce that we must have

$$T > \frac{1}{2r} = \frac{1}{2}\left(\frac{\sigma_1}{C\varepsilon}\right)^{\frac{p}{p-1}} = \Omega(1)\cdot\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}} \geq \Omega(1),$$

as desired,

– if $\sigma_1 \leq C\varepsilon$ then $r = 1$ and

$$\left(\frac{\sigma_1}{C\varepsilon}\right)^{\frac{p}{p-1}} \leq 1,$$

and since for any $s \in \{-1,1\}$ we have

$$\left\|\nabla F_{d,s}^\star(0)\right\| \overset{(18)}{=} L\beta r = \frac{\varepsilon\sqrt{2}}{c'} = 8\varepsilon\sqrt{2} > 2\varepsilon,$$

so at least one query is required to reach an ε–stationary point, i.e., $T \geq 1$ from where

$$T \geq 1 \geq \left(\frac{\sigma_1}{C\varepsilon}\right)^{\frac{p}{p-1}} = \Omega(1)\cdot\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}},$$

which achieves the proof of the lemma.

□

Remark C.4. Let us expand on the proof strategy of the above lemma. First, we start by defining a stochastic function $f\colon (x,\xi) \mapsto \frac{L}{2}\left(\|x\|^2 - 2x_1\xi + r^2\right)$ and we then provide two probability distributions $\mathbb{P}^-$ and $\mathbb{P}^+$ for the random variable ξ which allows to define two deterministic functions

$$F_-\colon x \mapsto \mathbb{E}_{\xi\sim\mathbb{P}_-}\left[f(x,\xi)\right] \quad \text{and} \quad F_+\colon x \mapsto \mathbb{E}_{\xi\sim\mathbb{P}_+}\left[f(x,\xi)\right].$$

The strategy to establish a lower bound then is to randomized the initial choice of the function by choosing F_- or F_+ with probability $\frac{1}{2}$ and lower bound the optimization error on the norm of the gradient by the misidentification error, in other words, how many samples $\xi^{(1)}, \dots, \xi^{(T)}$ are needed in expectation to distinguish between the two distributions $\mathbb{P}^-$ and $\mathbb{P}^+$.

Contrary to Arjevani et al. (2022), instead of using normal distributions we use two Bernoulli distributions $\mathrm{Ber}(r)$ and $-\mathrm{Ber}(r)$ where, after tuning the parameters so as to satisfy the different assumptions, we obtain

$$r = \Theta\left(\min\left\{1, \left(\frac{\varepsilon}{\sigma_1}\right)^{\frac{p}{p-1}}\right\}\right).$$

The choice of Bernoulli distributions is actually inspired from the definition of the gradient estimator in the work of Arjevani et al. (2022) where a Bernoulli distribution is used to probabilistically hide new coordinates (see Definition B.7).

Hence with our Bernoulli distributions, upon querying the oracle, it produces repeated 0s until it ultimately returns -1 or $+1$. In the former situation, we are unable to distinguish between the distributions $\mathrm{Ber}(r)$ and $-\mathrm{Ber}(r)$ since the outcome 0 happens with equal probability while, in the later case, we can immediately tell which distribution the oracle has chosen initially. As the outcome ± 1 happens with probability r, we have to do in expectation $\frac{1}{r} = \Omega(1 + (\sigma_1/\varepsilon)^{p/p-1})$ queries to be able to distinguish between $\mathbb{P}^-$ and $\mathbb{P}^+$ and this leads to the claimed lower bound.

Remark C.5. Combining Arjevani et al. (2022, Lemma 11) and the above lemma, we deduce that under the general p–bounded central moment where $p > 1$ is any real number, we have

$$\mathfrak{m}_\varepsilon^{\mathtt{zr}}\left(K, \Delta, \bar{L}, \sigma_1^p\right) \geq \Omega(1) \cdot \left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}\vee 2}, \tag{22}$$

where $\vee$ denotes the maximum between the two exponent, and the hidden constant in $\Omega(1)$ may depends on p. Hence

- p–bounded central moment for gradient when $p = 1$ brings literally *no information* and we can't find an ε–stationary points in a finite number T of oracle queries,
- assuming high-order bounded moments ($p > 2$) does not bring additional information than in the bounded variance case ($p = 2$).

Proof of (22). The proof here applies not only to the setting of Lemma C.1, i.e., Assumptions 2.1, 2.3 and 2.6 but also to the settings discussed in Remarks C.2 and C.3.

First, the case of an exponent $p \in (1, 2]$ is already covered by Lemma C.1. For the case when $p > 2$, we reuse the exact same proof as in Arjevani et al. (2022, Lemma 11). More precisely, instead of considering Bernoulli distributions for $\mathbb{P}^{-1}$ and $\mathbb{P}^1$, we fall back to normal distributions, that is, $\mathbb{P}^s = \mathcal{N}(rs, \frac{\sigma_1^2}{c_p^2})$ for $s \in \{-1, 1\}$ and some well-chosen constant $c_p > 0$ (depending on the exponent p and $\bar{L}$) that we fix later. Taking $\beta = 1$ and $L = \bar{L}$, as in Arjevani et al. (2022), it remains to bound the p^{th} central moment of the stochastic gradient, we have for any $x \in \mathbb{R}^d$ and any $s \in \{-1, 1\}$, by (17)

$$\begin{aligned}
\mathbb{E}_{\xi\sim\mathbb{P}^s}\left[\left\|\nabla f_d^\star(x,\xi) - \nabla F_{d,s}^\star(x)\right\|^p\right] &\overset{(17)}{=} \bar{L}^p \mathbb{E}_{\xi\sim\mathbb{P}^s}\left[|\xi - rs|^p\right] \\
&\overset{\text{Lem. H.13}}{=} \bar{L}^p \left(\frac{\sigma_1\sqrt{2}}{c_p}\right)^p \frac{\Gamma\left(\frac{p+1}{2}\right)}{\sqrt{\pi}} \\
&\overset{(a)}{\leq} \sigma_1^p,
\end{aligned}$$

where in (a) we took

$$c_p = \bar{L}\sqrt{2}\left(\frac{\Gamma\left(\frac{p+1}{2}\right)}{\sqrt{\pi}}\right)^{\frac{1}{p}} = C_p\bar{L} > 0,$$

where $C_p = \sqrt{2}\left(\frac{\Gamma\left(\frac{p+1}{2}\right)}{\sqrt{\pi}}\right)^{\frac{1}{p}}$. The rest of the proof is similar to Arjevani et al. (2022, Lemma 10). □

C.2 Some Bounds on the Gradient Estimator $\bar{g}_T$

Lemma C.2 (Properties of the Gradient Estimator $\bar{g}_T$)**.** *The stochastic gradient estimator $\bar{g}_T$ is a probability–θ zero-chain, is unbiased with respect to ∇F_T and satisfies*

$$\mathbb{E}\left[\|\bar{g}_T(x,\xi) - \nabla F_T(x)\|^p\right] \leq \frac{2\gamma_\infty^p(1-\theta)}{\theta^{p-1}}, \quad \text{and} \quad \mathbb{E}\left[\|\bar{g}_T(x,\xi) - \bar{g}_T(y,\xi)\|^q\right] \leq \frac{\bar{\ell}_1^q}{\theta^{q-1}}\|x-y\|^q, \tag{23}$$

and

$$\mathbb{E}\left[\|[\bar{g}_T(x,\xi) - \bar{g}_T(y,\xi)] - [\nabla F_T(x) - \nabla F_T(y)]\|^q\right] \leq \frac{\bar{\delta}_1^{\,q}(1-\theta)}{\theta^{q-1}}\|x-q\|^q, \tag{24}$$

for all $x, y \in \mathbb{R}^T$, where $p, q \in (1,2]$, γ_∞ is defined in Lemma B.2, $\bar{\ell}_1 := 2\left(2^q\left(6^{2q}\gamma_\infty^q + \ell_1^q\right) + \ell_1^q\right)^{\frac{1}{q}} \geq \ell_1$ and $\bar{\delta}_1 := 4\left(6^{2q}\gamma_\infty^q + \ell_1^q\right)^{\frac{1}{q}}$.

The proof of Lemma C.2 is largely inspired from (Arjevani et al., 2022, Lemma 4). Notably, we rely on Lemma H.6 which states the following inequality

$$\|a+b\|^\alpha \leq 2^{\alpha-1}\left(\|a\|^\alpha + \|b\|^\alpha\right), \tag{25}$$

holds, for any vectors $a, b \in \mathbb{R}^d$ and any exponent $\alpha \geq 1$. This generalizes the well-known inequality $\|a+b\|^2 \leq 2(\|a\|^2 + \|b\|^2)$. We use inequality (25) in (27) as a "substitute" for the squared norm expansion (as used in Arjevani et al. (2022)) which does not hold anymore with exponent $q \in (1,2]$ instead of 2.

Proof. Following the proof of Arjevani et al. (2022, Lemma 4), for any ξ, the vector $\delta(x,\xi) := \bar{g}_T(x,\xi) - \nabla F_T(x)$ has at most one nonzero entry at coordinate $i_x = \operatorname{prog}_{\frac{1}{4}}(x) + 1$. Moreover, for any $i \in \{1,\dots,T\}$, the i^{th} entry $\delta_i(x,\xi)$ of $\delta(x,\xi)$ reads

$$\delta_i(x,\xi) = [\nabla F_T(x)]_i\,\Theta_i(x)\left(\frac{\xi}{\theta} - 1\right),$$

where the function Θ_i is defined in Definition B.9. Hence, we have

$$\begin{aligned}
\mathbb{E}\left[\|\bar{g}_T(x,\xi) - \nabla F_T(x)\|^p\right] &= \mathbb{E}\left[|\delta_{i_x}(x,\xi)|^p\right] \\
&= \left|[\nabla F_T(x)]_{i_x}\right|^p |\Theta_{i_x}(x)|^p\,\mathbb{E}\left[\left|\frac{\xi}{\theta} - 1\right|^p\right] \\
&\overset{\text{Lem. B.2}}{\leq} \gamma_\infty^p\,|\Theta_{i_x}(x)|^p\,\mathbb{E}\left[\left|\frac{\xi}{\theta} - 1\right|^p\right] \\
&\leq \gamma_\infty^p\,\mathbb{E}\left[\left|\frac{\xi}{\theta} - 1\right|^p\right] \\
&\overset{(a)}{=} \gamma_\infty^p\left((1-\theta) + \theta\left(\frac{1}{\theta} - 1\right)^p\right) \\
&= \gamma_\infty^p(1-\theta)\left(1 + \left(\frac{1-\theta}{\theta}\right)^{p-1}\right) \\
&\overset{(b)}{\leq} \frac{2\gamma_\infty^p(1-\theta)}{\theta^{p-1}},
\end{aligned} \tag{26}$$

where in (a) we use the fact that $\xi \sim \mathrm{Ber}(\theta)$. In (b) we use $\theta \in [0,1]$ and $p \geq 1$ to bound $(1-\theta)^{p-1} \leq 1$ and $1 \leq \theta^{-(p-1)}$. This establishes the first inequality from (23).

Now, for the second inequality in (23), we use the fact that $\delta(\cdot,\xi) := \bar{g}_T(\cdot,\xi) - \nabla F_T(\cdot)$ has at most one nonzero coordinate, then if we let $i_x = \text{prog}_{\frac{1}{4}}(x)+1$ and $i_y = \text{prog}_{\frac{1}{4}}(y)+1$ we obtain

$$\begin{aligned}
\mathbb{E}\left[\|\bar{g}_T(x,\xi) - \bar{g}_T(y,\xi)\|^q\right] &= \mathbb{E}\left[\|[\delta(x,\xi)-\delta(y,\xi)] + [\nabla F_T(x) - \nabla F_T(y)]\|^q\right] \\
&\overset{\text{Lem. H.6}}{\leq} 2^{q-1}\left(\mathbb{E}\left[\|\delta(x,\xi)-\delta(y,\xi)\|^q\right] + \|\nabla F_T(x) - \nabla F_T(y)\|^q\right) \\
&\leq 2^{q-1}\left(\mathbb{E}\left[|\delta_{i_x}(x,\xi)-\delta_{i_x}(y,\xi)|^q\right] + \mathbb{E}\left[\left|\delta_{i_y}(x,\xi)-\delta_{i_y}(y,\xi)\right|^q\right] + \|\nabla F_T(x) - \nabla F_T(y)\|^q\right),
\end{aligned} \tag{27}$$

and, for any integer $i \in \{1,\dots,T\}$ we have

$$\begin{aligned}
\mathbb{E}\left[|\delta_i(x,\xi)-\delta_i(y,\xi)|^q\right] &= \left|[\nabla F_T(x)]_i\,\Theta_i(x) - [\nabla F_T(y)]_i\,\Theta_i(y)\right|^q \mathbb{E}\left[\left|\frac{\xi}{\theta}-1\right|^q\right] \\
&\overset{(a)}{\leq} \left|[\nabla F_T(x)]_i\,\Theta_i(x) - [\nabla F_T(y)]_i\,\Theta_i(y)\right|^q \left(\frac{2}{\theta^{q-1}}\right) \\
&= \left|[\nabla F_T(x)]_i\,(\Theta_i(x)-\Theta_i(y)) + ([\nabla F_T(x)]_i - [\nabla F_T(y)]_i)\,\Theta_i(y)\right|^q \left(\frac{2}{\theta^{q-1}}\right) \\
&\overset{\text{Lem. H.6}}{\leq} 2^{q-1}\left(|[\nabla F_T(x)]_i|^q\,|\Theta_i(x)-\Theta_i(y)|^q + |[\nabla F_T(x)]_i - [\nabla F_T(y)]_i|^q\,|\Theta_i(y)|^q\right)\left(\frac{2}{\theta^{q-1}}\right) \\
&\overset{(b)}{\leq} 2^{q-1}\left(6^{2q}\,|[\nabla F_T(x)]_i|^q\,\|x-y\|^q + |[\nabla F_T(x)]_i - [\nabla F_T(y)]_i|^q\right)\left(\frac{2}{\theta^{q-1}}\right) \\
&\overset{\text{Lem. B.2}}{\leq} 2^{q-1}\left(6^{2q}\gamma_\infty^q\,\|x-y\|^q + \|\nabla F_T(x) - \nabla F_T(y)\|^q\right)\left(\frac{2}{\theta^{q-1}}\right),
\end{aligned} \tag{28}$$

where in (a) we use our previous bound from (26) (which we derived with exponent p instead of q). In (b) we use the fact that Θ_i is 6^2–Lipschitz (see Arjevani et al. (2022)) and $|\Theta_i(\cdot)| \leq 1$. Now, plugging back the bound (28) in (27) we obtain

$$\begin{aligned}
\mathbb{E}\left[\|\bar{g}_T(x,\xi) - \bar{g}_T(y,\xi)\|^q\right] &\overset{(28)+(27)}{\leq} 2^{q-1}\left(2^q\left(6^{2q}\gamma_\infty^q\,\|x-y\|^q + \|\nabla F_T(x) - \nabla F_T(y)\|^q\right)\left(\frac{2}{\theta^{q-1}}\right) + \|\nabla F_T(x) - \nabla F_T(y)\|^q\right) \\
&\overset{\text{Lem. B.2}}{\leq} 2^{q-1}\left(2^q\left(6^{2q}\gamma_\infty^q + \ell_1^q\right)\left(\frac{2}{\theta^{q-1}}\right) + \ell_1^q\right)\|x-y\|^q \\
&\overset{(a)}{\leq} 2^q\left(2^q\left(6^{2q}\gamma_\infty^q + \ell_1^q\right) + \ell_1^q\right)\frac{\|x-y\|^q}{\theta^{q-1}},
\end{aligned}$$

where in (a) we use $1 \leq {}^2/_{\theta^{q-1}}$ to factor it out. If we let $\bar{\ell}_1 := 2\left(2^q\left(6^{2q}\gamma_\infty^q + \ell_1^q\right) + \ell_1^q\right)^{\frac{1}{q}}$ we obtain

$$\mathbb{E}\left[\|\bar{g}_T(x,\xi) - \bar{g}_T(y,\xi)\|^q\right] \leq \frac{\bar{\ell}_1^q}{\theta^{q-1}}\|x-y\|^q,$$

as desired.

It remains to establish the third inequality (24). Combining the bounds (27) and (28) we have

$$\begin{aligned}
\mathbb{E}\left[\|[\bar{g}_T(x,\xi) - \bar{g}_T(y,\xi)] - [\nabla F_T(x) - \nabla F_T(y)]\|^q\right] &= \mathbb{E}\left[\|\delta(x,\xi)-\delta(y,\xi)\|^q\right] \\
&\overset{\text{Lem. H.6}}{\leq} 2^{q-1}\left(\mathbb{E}\left[|\delta_{i_x}(x,\xi)-\delta_{i_x}(y,\xi)|^q\right] + \mathbb{E}\left[\left|\delta_{i_y}(x,\xi)-\delta_{i_y}(y,\xi)\right|^q\right]\right) \\
&\overset{(26)+(28)}{\leq} 4^q\left(6^{2q}\gamma_\infty^q + \ell_1^q\right)\frac{1-\theta}{\theta^{q-1}}\|x-y\|^q \\
&= \frac{\bar{\delta}_1^{\,q}(1-\theta)}{\theta^{q-1}}\|x-q\|^q,
\end{aligned}$$

where we define $\bar{\delta}_1 := 4\left(6^{2q}\gamma_\infty^q + \ell_1^q\right)^{\frac{1}{q}}$. This achieves the proof of the lemma. □

C.3 Proof of Theorem 3.1

Theorem 3.1. *Given $\Delta, \bar{L} > 0$, $\sigma_1 \geq 0$ and $0 < \varepsilon \leq c_1\sqrt{\bar{L}\Delta}$ for some universal constant $c_1 > 0$. Then, for any algorithm $A \in \mathcal{A}_{zr}$, there exists a function $f \in \mathcal{F}(\Delta)$, an oracle and a distribution $(O, \mathcal{D}) \in \mathcal{O}\left(f, \bar{L}^q, \sigma_1^p\right)$ satisfying Assumptions 2.1, 2.3 and 2.6 such that*

$$\mathfrak{m}_\varepsilon^{\mathtt{zr}}\left(K, \Delta, \bar{L}^q, \sigma_1^p\right) \geq \Omega(1) \cdot \left(\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}} + \frac{\bar{L}\Delta}{\varepsilon^2} + \frac{\bar{L}\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{q(p-1)}}\right).$$

Proof. Let Δ_0, ℓ_1, γ_∞ and $\bar{\ell}_1$ be the numerical constants in Lemmas B.2 and C.2 respectively. Additionally, we let the accuracy parameter $\varepsilon > 0$, initial sub-optimality $\Delta \geq 0$, the q–weak average smoothness parameter $\bar{L}$, and the variance parameter $\sigma_1 \geq 0$ be fixed, and $0 < L \leq \bar{L}$ to be specified later. Then, for $\alpha, \beta > 0$ two positive real numbers, following Arjevani et al. (2022), we rescale the function F_T as

$$F_T^\star \colon x \mapsto \alpha F_T(\beta x). \tag{29}$$

- **Step 1:** *Ensuring $F_T^\star \in \mathcal{F}(\Delta, L)$.*

  To guarantee the rescaled function $F_T^\star$ belongs to the function class $\mathcal{F}(\Delta, L)$, let us compute the initial sub-optimality Δ and the smoothness constant L. Assuming the algorithm $\mathtt{A}$ starts at $x^0 = 0$ we have

$$F_T^\star(0) - \inf_{x\in\mathbb{R}^T} F_T^\star(x) \overset{(29)}{=} \alpha\left(F_T(0) - \inf_{x\in\mathbb{R}^T} F_T(x)\right) \overset{\text{Lem. B.2}}{\leq} \alpha\Delta_0 T,$$

  thus, it's enough to take $T = \left\lfloor \frac{\Delta}{\alpha\Delta_0} \right\rfloor$ so as to ensure $F_T^\star(0) - \inf_{x\in\mathbb{R}^T} F_T^\star(x) \leq \Delta$. Moreover, for any $x, y \in \mathbb{R}^T$,

$$\begin{aligned}\left\|\nabla F_T^\star(x) - \nabla F_T^\star(y)\right\| &= \alpha\beta\left\|\nabla F_T(\beta x) - \nabla F_T(\beta y)\right\| \\ &\overset{\text{Lem. B.2}}{\leq} \alpha\beta\ell_1\left\|\beta x - \beta y\right\| \\ &= \alpha\beta^2\ell_1\left\|x - y\right\|,\end{aligned}$$

  and it suffices to take $\alpha = \frac{L}{\beta^2\ell_1} > 0$ to ensure the function $F_T^\star$ has L–Lipschitz gradients. Consequently, we have $F_T^\star \in \mathcal{F}(\Delta, L)$, as desired.

- **Step 2:** *Analysis of the Protocol and Choice for β.*

  Following the proof of Arjevani et al. (2022, Theorem 1), according to Lemma B.2, for all points $x \in \mathbb{R}^T$ such that $\mathrm{prog}_0(x) < T$ we have $\mathrm{prog}_0(\beta x) = \mathrm{prog}_0(x) < T$ so

$$\left\|\nabla F_T^\star(x)\right\| \overset{(29)}{=} \frac{L}{\ell_1\beta}\left\|\nabla F_T(\beta x)\right\| \overset{\text{Lem. B.2}}{>} \frac{L}{\ell_1\beta}, \tag{30}$$

  and we need to guarantee that

$$\left\|\nabla F_T^\star(x)\right\| > 2\varepsilon,$$

  for all $x \in \mathbb{R}^T$ with $\mathrm{prog}_0(x) < T$ which, given (30), can be done if we set $\beta = \frac{L}{2\ell_1\varepsilon}$.

- **Step 3:** *The Oracle Class and Assumptions 2.3 and 2.6.*

  It remains to choose the parameter $\theta \in (0, 1]$ and constant L such that the gradient estimator $\bar{g}_T^\star$ of $\nabla F_T^\star$ satisfies Assumptions 2.3 and 2.6. Computing the p–th central moment of $\bar{g}_T^\star$ gives, for all $x \in \mathbb{R}^T$

$$\begin{aligned}\mathbb{E}\left[\left\|\bar{g}_T^\star(x,\xi) - \nabla F_T^\star(x)\right\|^p\right] &\overset{(29)}{=} (\alpha\beta)^p\,\mathbb{E}\left[\left\|\bar{g}_T(\beta x,\xi) - \nabla F_T(\beta x)\right\|^p\right] \\ &\overset{\text{Lem. C.2}}{\leq} \frac{2\left(\gamma_\infty\alpha\beta\right)^p(1-\theta)}{\theta^{p-1}} \\ &\overset{(a)}{\leq} \frac{\left(2\gamma_\infty\alpha\beta\right)^p}{\theta^{p-1}},\end{aligned} \tag{31}$$

  where in (a) we use the fact that $0 < \theta \leq 1$ and $p \geq 1$ so that $2 \leq 2^p$. From (31), so as to satisfy Assumption 2.3 it's enough to take $\theta = \min\left\{1, \bar{\theta}\right\}$ where

$$\bar{\theta}^{p-1} \geq \left(\frac{2\gamma_\infty \alpha\beta}{\sigma_1}\right)^p \overset{\text{(a)}}{=} \left(\frac{4\gamma_\infty \varepsilon}{\sigma_1}\right)^p, \quad \text{so} \quad \bar{\theta} \geq \left(\frac{4\gamma_\infty \varepsilon}{\sigma_1}\right)^{\frac{p}{p-1}}$$

where in (a) we use the value of α and β fixed earlier. Hence

$$\theta = \min\left\{1, \left(\frac{4\gamma_\infty \varepsilon}{\sigma_1}\right)^{\frac{p}{p-1}}\right\}. \tag{32}$$

Next, concerning the mean-squared smoothness assumption, we have

$$\begin{aligned}
\mathbb{E}\left[\|\bar{g}_T^\star(x,\xi) - \bar{g}_T^\star(y,\xi)\|^q\right] &\overset{(29)}{=} (\alpha\beta)^q \,\mathbb{E}\left[\|\bar{g}_T(\beta x,\xi) - \bar{g}_T(\beta y,\xi)\|^q\right] \\
&\overset{\text{Lem. C.2}}{\leq} \frac{\left(\alpha\beta\bar{\ell}_1\right)^q}{\theta^{q-1}} \|\beta x - \beta y\|^q \\
&= \left(\frac{\alpha\beta^2\bar{\ell}_1}{\theta^{\frac{q-1}{q}}}\right)^q \|x-y\|^q \\
&\overset{\text{(a)}}{=} \left(\frac{L\bar{\ell}_1}{\ell_1\theta^{\frac{q-1}{q}}}\right)^q \|x-y\|^q,
\end{aligned} \tag{33}$$

where in (a) we use $\alpha = \frac{L}{\beta^2\ell_1}$. Hence, from the upper bound (33) it suffices to take

$$L = \frac{\ell_1\bar{L}}{\bar{\ell}_1}\theta^{\frac{q-1}{q}} \leq \bar{L}\min\left\{1, \left(\frac{4\gamma_\infty\varepsilon}{\sigma_1}\right)^{\frac{p(q-1)}{q(p-1)}}\right\} \leq \bar{L},$$

since $\ell_1 = 152 \leq \bar{\ell}_1$ (see Lemma C.2). This proves Assumption 2.6 is satisfied by the gradient estimator $\bar{g}_T^\star$.

- **Step 4:** *Lower Bounding* $\mathfrak{m}_\varepsilon^{\mathtt{zr}}\left(K,\Delta,\bar{L}^q,\sigma_1^p\right)$.

  Continuing on **step 2**, by Lemma B.1 we know that with probability at least $\frac{1}{2}$ it holds that for all integer $0 \leq t \leq \frac{T-1}{2\theta}$ and all $k \in [K]$ we have

$$\left\|\nabla F_T^\star\left(x_{\mathtt{A}[\mathtt{0}_F]}^{(t,k)}\right)\right\| > 2\varepsilon, \quad \text{hence} \quad \mathbb{E}\left[\left\|\nabla F_T^\star\left(x_{\mathtt{A}[\mathtt{0}_F]}^{(t,k)}\right)\right\|\right] > \varepsilon,$$

  from where it follows that

$$\begin{aligned}
\mathfrak{m}_\varepsilon^{\mathtt{zr}}\left(K,\Delta,\bar{L}^q,\sigma_1^p\right) > \frac{T-1}{2\theta} = \frac{1}{2\theta}\left(\left\lfloor\frac{\Delta}{\alpha\Delta_0}\right\rfloor - 1\right) &= \frac{1}{2\theta}\left(\left\lfloor\frac{\beta^2\ell_1\Delta}{L\Delta_0}\right\rfloor - 1\right) \\
&= \frac{1}{2\theta}\left(\left\lfloor\frac{L\Delta}{4\Delta_0\ell_1\varepsilon^2}\right\rfloor - 1\right) \\
&= \frac{1}{2\theta}\left(\left\lfloor\frac{\bar{L}\Delta\theta^{\frac{q-1}{q}}}{4\Delta_0\bar{\ell}_1\varepsilon^2}\right\rfloor - 1\right),
\end{aligned} \tag{34}$$

  we then distinguish two cases:

  – if $\frac{\bar{L}\Delta\theta^{\frac{q-1}{q}}}{4\Delta_0\bar{\ell}_1\varepsilon^2} \geq 3$ then, using the inequality $\lfloor x\rfloor - 1 \geq \frac{x}{2}$, valid for all real number $x \geq 3$ we obtain

$$\begin{aligned}
\mathfrak{m}_\varepsilon^{\mathtt{zr}}\left(K,\Delta,\bar{L}^q,\sigma_1^p\right) &\overset{(34)}{>} \frac{1}{4\theta}\cdot\frac{\bar{L}\Delta\theta^{\frac{q-1}{q}}}{4\Delta_0\bar{\ell}_1\varepsilon^2} \\
&= \frac{1}{16\Delta_0\bar{\ell}_1}\cdot\frac{\bar{L}\Delta}{\varepsilon^2\theta^{\frac{1}{q}}} \\
&\overset{(32)}{\geq} \frac{1}{32\Delta_0\bar{\ell}_1}\cdot\frac{\bar{L}\Delta}{\varepsilon^2}\left[\left(\frac{\sigma_1}{4\gamma_\infty\varepsilon}\right)^{\frac{p}{q(p-1)}} + 1\right] \\
&= \Omega(1)\cdot\left(\frac{\bar{L}\Delta}{\varepsilon^2} + \frac{\bar{L}\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{q(p-1)}}\right),
\end{aligned}$$

    and combining the lower bound above with Lemma C.1, with the same choice for c' as provided below, gives the desired results.

– otherwise, if $\frac{\bar{L}\Delta\theta^{\frac{q-1}{q}}}{4\Delta_0\bar{\ell}_1\varepsilon^2} < 3$, choosing the universal constant $0 < c' = \left(12\Delta_0\bar{\ell}_1\right)^{-\frac{1}{2}} \leq \left(12\Delta_0\ell_1\right)^{-\frac{1}{2}} \approx 0.0067592 < \frac{1}{8}$ then the assumption

$$0 < \varepsilon \leq c'\sqrt{\bar{L}\Delta} = \sqrt{\frac{\bar{L}\Delta}{12\Delta_0\bar{\ell}_1}},$$

precludes the possibility that $\theta = 1$ so $\sigma_1 \geq 4\gamma_\infty\varepsilon$ hence $\sigma_1/\varepsilon \gtrsim 1$. Moreover, we have

$$3 > \frac{\bar{L}\Delta}{4\Delta_0\bar{\ell}_1\varepsilon^2}\left(\frac{4\gamma_\infty\varepsilon}{\sigma_1}\right)^{\frac{p(q-1)}{q(p-1)}},$$

and multiplying both sides by $\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}} > 0$ leads to

$$\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}} > \frac{\bar{L}\Delta\left(4\gamma_\infty\right)^{\frac{p(q-1)}{q(p-1)}}}{12\Delta_0\bar{\ell}_1\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}-\frac{p(q-1)}{q(p-1)}} = \frac{\bar{L}\Delta\left(4\gamma_\infty\right)^{\frac{p(q-1)}{q(p-1)}}}{12\Delta_0\bar{\ell}_1\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{q(p-1)}},$$

and using Lemma C.1, since $0 < \varepsilon \leq \frac{1}{8}\sqrt{\bar{L}\Delta}$ then there exists an universal constant $C_p > 0$ (depending only on p) such that

$$\mathfrak{m}_\varepsilon^{\mathtt{zr}}\left(K, \Delta, \bar{L}^q, \sigma_1^p\right) \geq C_p\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}} \geq \Omega(1)\cdot\frac{\bar{L}\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{q(p-1)}} \gtrsim \frac{\bar{L}\Delta}{\varepsilon^2}, \tag{35}$$

hence,

$$\mathfrak{m}_\varepsilon^{\mathtt{zr}}\left(K, \Delta, \bar{L}^q, \sigma_1^p\right) \geq \Omega(1)\cdot\left(\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}} + \frac{\bar{L}\Delta}{\varepsilon^2} + \frac{\bar{L}\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{q(p-1)}}\right),$$

which holds in both cases and concludes the proof of the theorem.

□

C.4 Proof of Theorem 3.2

Theorem 3.2. *Given $\Delta, L_1, \delta > 0$, $\sigma_1 \geq 0$ and $0 < \varepsilon \leq c_1\sqrt{L_1\Delta}$ for some universal constant $c_1 > 0$. Then, for any algorithm $A \in \mathcal{A}_{\mathtt{zr}}$, there exists a function $f \in \mathcal{F}(\Delta, L_1)$, an oracle and a distribution $(O, \mathcal{D}) \in \mathcal{O}(f, \delta^q, \sigma_1^p)$ satisfying Assumptions 2.1 to 2.3 and 2.7 such that*

$$\mathfrak{m}_\varepsilon^{\mathtt{zr}}\left(K, \Delta, L_1, \delta^q, \sigma_1^p\right) \geq \Omega(1)\cdot\min\left\{\frac{L_1\Delta}{\varepsilon^2} + \frac{L_1\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}, \left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}} + \frac{(L_1+\delta)\Delta}{\varepsilon^2} + \frac{\delta\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{q(p-1)}}\right\}.$$

Proof. The proof follows the same lines as the proof of Theorem 3.1. Let Δ_0, ℓ_1, γ_∞ and $\bar{\delta}_1$ be the numerical constants in Lemmas B.2 and C.2 respectively. Additionally, we let the accuracy parameter $\varepsilon > 0$, initial sub-optimality $\Delta \geq 0$, the smoothness constant $L_1 \geq 0$, the q–weak average smoothness parameter δ (Assumption 2.7), and the variance parameter $\sigma_1 \geq 0$ be fixed, and $0 < L \leq L_1$ to be specified later. Then, for $\alpha, \beta > 0$ two positive real numbers, following Arjevani et al. (2022), we rescale the function F_T as

$$F_T^\star \colon x \mapsto \alpha F_T(\beta x). \tag{36}$$

- **Step 1:** *Ensuring $F_T^\star \in \mathcal{F}(\Delta, L)$.*

  To guarantee the rescaled function $F_T^\star$ belongs to the function class $\mathcal{F}(\Delta, L)$, let us compute the initial sub-optimality Δ and the smoothness constant L. Assuming the algorithm $\mathtt{A}$ starts at $x^0 = 0$ we have

$$F_T^\star(0) - \inf_{x\in\mathbb{R}^T} F_T^\star(x) \overset{(29)}{=} \alpha\left(F_T(0) - \inf_{x\in\mathbb{R}^T} F_T(x)\right) \overset{\text{Lem. B.2}}{\leq} \alpha\Delta_0 T,$$

thus, it's enough to take $T = \left\lfloor \frac{\Delta}{\alpha \Delta_0} \right\rfloor$ so as to ensure $F_T^\star(0) - \inf_{x \in \mathbb{R}^T} F_T^\star(x) \leq \Delta$. Moreover, as done previously, for any $x, y \in \mathbb{R}^T$,

$$\begin{aligned} \|\nabla F_T^\star(x) - \nabla F_T^\star(y)\| &= \alpha\beta \|\nabla F_T(\beta x) - \nabla F_T(\beta y)\| \\ &\overset{\text{Lem. B.2}}{\leq} \alpha\beta\ell_1 \|\beta x - \beta y\| \\ &= \alpha\beta^2\ell_1 \|x - y\|, \end{aligned}$$

and it suffices to take $\alpha = \frac{L}{\beta^2 \ell_1} > 0$ to ensure the function $F_T^\star$ has L–Lipschitz gradients. Consequently, we have $F_T^\star \in \mathcal{F}(\Delta, L)$, as desired.

- **Step 2:** *Analysis of the Protocol and Choice for β.*

  Following the proof of Arjevani et al. (2022, Theorem 1), according to Lemma B.2, for all points $x \in \mathbb{R}^T$ such that $\text{prog}_0(x) < T$ we have $\text{prog}_0(\beta x) = \text{prog}_0(x) < T$ so

$$\|\nabla F_T^\star(x)\| \overset{(29)}{=} \frac{L}{\ell_1 \beta} \|\nabla F_T(\beta x)\| \overset{\text{Lem. B.2}}{>} \frac{L}{\ell_1 \beta}, \tag{37}$$

  and we need to guarantee that

$$\|\nabla F_T^\star(x)\| > 2\varepsilon, \tag{38}$$

  for all $x \in \mathbb{R}^T$ with $\text{prog}_0(x) < T$ which, given (30), can be done if we set $\beta = \frac{L}{2\ell_1 \varepsilon}$.

- **Step 3:** *The Oracle Class and Assumptions 2.3 and 2.6.*

  It remains to choose the parameter $\theta \in (0, 1]$ and constant L such that the gradient estimator $\bar{g}_T^\star$ of $\nabla F_T^\star$ satisfies Assumptions 2.3 and 2.7. Computing the p–th central moment of $\bar{g}_T^\star$ gives, for all $x \in \mathbb{R}^T$

$$\begin{aligned} \mathbb{E}\left[\|\bar{g}_T^\star(x, \xi) - \nabla F_T^\star(x)\|^p\right] &\overset{(29)}{=} (\alpha\beta)^p \, \mathbb{E}\left[\|\bar{g}_T(\beta x, \xi) - \nabla F_T(\beta x)\|^p\right] \\ &\overset{\text{Lem. C.2}}{\leq} \frac{2\left(\gamma_\infty \alpha\beta\right)^p (1 - \theta)}{\theta^{p-1}} \\ &\overset{(a)}{\leq} \frac{\left(2\gamma_\infty \alpha\beta\right)^p}{\theta^{p-1}}, \end{aligned} \tag{39}$$

  where in (a) we use the fact that $0 < \theta \leq 1$ and $p \geq 1$ so that $2 \leq 2^p$. From (31), so as to satisfy Assumption 2.3 it's enough to take $\theta = \min\left\{1, \bar{\theta}\right\}$ where

$$\bar{\theta}^{p-1} \geq \left(\frac{2\gamma_\infty \alpha\beta}{\sigma_1}\right)^p \overset{(a)}{=} \left(\frac{4\gamma_\infty \varepsilon}{\sigma_1}\right)^p, \quad \text{so} \quad \bar{\theta} \geq \left(\frac{4\gamma_\infty \varepsilon}{\sigma_1}\right)^{\frac{p}{p-1}}$$

  where in (a) we use the value of α and β fixed earlier. Hence

$$\theta = \min\left\{1, \left(\frac{4\gamma_\infty \varepsilon}{\sigma_1}\right)^{\frac{p}{p-1}}\right\}. \tag{40}$$

  Next, concerning the mean-squared smoothness assumption (Assumption 2.7), we have

$$\begin{aligned} &\mathbb{E}\left[\|[\bar{g}_T^\star(x, \xi) - \bar{g}_T^\star(y, \xi)] - [\nabla F_T^\star(x) - \nabla F_T^\star(y)]\|^q\right] \\ &\quad \overset{(29)}{=} (\alpha\beta)^q \, \mathbb{E}\left[\|[\bar{g}_T^\star(\beta x, \xi) - \bar{g}_T^\star(\beta y, \xi)] - [\nabla F_T^\star(\beta x) - \nabla F_T^\star(\beta y)]\|^q\right] \\ &\quad \overset{\text{Lem. C.2}}{\leq} \frac{\left(\alpha\beta\bar{\delta}_1\right)^q (1 - \theta)}{\theta^{q-1}} \|\beta x - \beta y\|^q \\ &\quad = \left(\frac{\alpha\beta^2\bar{\delta}_1}{\theta^{\frac{q-1}{q}}}\right)^q (1 - \theta) \|x - y\|^q \\ &\quad \overset{(a)}{=} \left(\frac{L\bar{\delta}_1}{\ell_1 \theta^{\frac{q-1}{q}}}\right)^q (1 - \theta) \|x - y\|^q, \end{aligned} \tag{41}$$

where in (a) we use $\alpha = \frac{L}{\beta^2 \ell_1}$. Hence, from the upper bound (33) it suffices to take

$$L \leq \frac{\ell_1 \delta}{\bar{\delta}_1} \theta^{\frac{q-1}{q}},$$

and since we must have $L \leq L_1$, we set

$$L = \min\left\{L_1, \frac{\ell_1 \delta}{\bar{\delta}_1} \theta^{\frac{q-1}{q}}\right\}. \tag{42}$$

This proves Assumption 2.7 is satisfied by the gradient estimator $\bar{g}_T^\star$.

- **Step 4:** *Lower Bounding* $\mathfrak{m}_\varepsilon^{\mathtt{zr}}(K, \Delta, L_1, \delta^q, \sigma_1^p)$.

  Continuing on **step 2**, by Lemma B.1 we know that with probability at least $\frac{1}{2}$ it holds that for all integer $0 \leq t \leq \frac{T-1}{2\theta}$ and all $k \in [K]$ we have

$$\left\|\nabla F_T^\star\left(x_{\mathtt{A}[\mathtt{0}_F]}^{(t,k)}\right)\right\| > 2\varepsilon, \quad \text{hence} \quad \mathbb{E}\left[\left\|\nabla F_T^\star\left(x_{\mathtt{A}[\mathtt{0}_F]}^{(t,k)}\right)\right\|\right] > \varepsilon,$$

  from where it follows that

$$\mathfrak{m}_\varepsilon^{\mathtt{zr}}(K, \Delta, L_1, \delta^q, \sigma_1^p) > \frac{T-1}{2\theta} = \frac{1}{2\theta}\left(\left\lfloor \frac{\Delta}{\alpha \Delta_0} \right\rfloor - 1\right) = \frac{1}{2\theta}\left(\left\lfloor \frac{\beta^2 \ell_1 \Delta}{L \Delta_0} \right\rfloor - 1\right) = \frac{1}{2\theta}\left(\left\lfloor \frac{L\Delta}{4\Delta_0 \ell_1 \varepsilon^2} \right\rfloor - 1\right) \tag{43}$$

  we then distinguish two cases:

  – if $\frac{L\Delta}{4\Delta_0 \ell_1 \varepsilon^2} \geq 3$ then, using the inequality $\lfloor x \rfloor - 1 \geq \frac{x}{2}$, valid for all real number $x \geq 3$ we obtain

$$\begin{aligned}
\mathfrak{m}_\varepsilon^{\mathtt{zr}}(K, \Delta, L_1, \delta^q, \sigma_1^p) &\overset{(43)}{>} \frac{1}{4\theta} \cdot \frac{L\Delta}{4\Delta_0 \ell_1 \varepsilon^2} \\
&\overset{(42)}{=} \frac{1}{16\Delta_0 \ell_1} \cdot \frac{\Delta}{\theta \varepsilon^2} \min\left\{L_1, \frac{\ell_1 \delta}{\bar{\delta}_1} \theta^{\frac{q-1}{q}}\right\} \\
&= \frac{1}{16\Delta_0 \ell_1} \cdot \min\left\{\frac{L_1 \Delta}{\theta \varepsilon^2}, \frac{\ell_1 \delta \Delta}{\bar{\delta}_1 \varepsilon^2} \theta^{-\frac{1}{q}}\right\} \\
&\overset{(40)}{\geq} \frac{1}{32\Delta_0 \ell_1} \cdot \min\left\{\frac{L_1 \Delta}{\varepsilon^2}\left(1 + \left(\frac{\sigma_1}{4\gamma_\infty \varepsilon}\right)^{\frac{p}{p-1}}\right), \frac{\ell_1 \delta \Delta}{\bar{\delta}_1 \varepsilon^2}\left(1 + \left(\frac{\sigma_1}{4\gamma_\infty \varepsilon}\right)^{\frac{p}{q(p-1)}}\right)\right\} \\
&= \Omega(1) \cdot \min\left\{\frac{L_1 \Delta}{\varepsilon^2} + \frac{L_1 \Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}, \frac{\delta \Delta}{\varepsilon^2} + \frac{\delta \Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{q(p-1)}}\right\},
\end{aligned} \tag{44}$$

    and combining the lower bound (44) with Lemma C.1, using the choice for c' provided below, gives:

$$\mathfrak{m}_\varepsilon^{\mathtt{zr}}(K, \Delta, L_1, \delta^q, \sigma_1^p) \geq \Omega(1) \cdot \min\left\{\frac{L_1 \Delta}{\varepsilon^2} + \frac{L_1 \Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}, \left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}} + \frac{\delta \Delta}{\varepsilon^2} + \frac{\delta \Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{q(p-1)}}\right\}, \tag{45}$$

    since from the assumption $0 < \varepsilon \leq c' \sqrt{L_1 \Delta}$ we have

$$\frac{L_1 \Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}} \gtrsim \left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}.$$

  – otherwise, if $\frac{L\Delta}{4\Delta_0 \ell_1 \varepsilon^2} < 3$, choosing the universal constant $0 < c' = (12\Delta_0 \ell_1)^{-\frac{1}{2}} \approx 0.0067592 < \frac{1}{8}$ then the assumption

$$0 < \varepsilon \leq c' \sqrt{L_1 \Delta} = \sqrt{\frac{L_1 \Delta}{12\Delta_0 \ell_1}},$$

    precludes the possibility that $L = L_1$. While we can still argue the same way as previously done in the proof of Theorem 3.1, we follows here a different strategy, thanks to Remark C.1. So, let us divide both sides of $\frac{L\Delta}{4\Delta_0 \ell_1 \varepsilon^2} < 3$ by our choice of $\theta > 0$ to obtain

$$\frac{1}{12\Delta_0 \ell_1} \cdot \frac{L\Delta}{\varepsilon^2} \theta^{-1} \leq \theta^{-1} = \max\left\{1, \left(\frac{\sigma_1}{4\gamma_\infty \varepsilon}\right)^{\frac{p}{p-1}}\right\} \leq \max\left\{1, \left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}\right\}, \tag{46}$$

where the last inequality follows from the fact that $\frac{p}{p-1} > 0$ and $4\gamma_\infty = 92 \geq 1$. Using the definition of L we obtain

$$\begin{aligned}\max\left\{1, \left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}\right\} &\overset{(46)+(42)}{\geq} \frac{1}{12\Delta_0\ell_1} \cdot \min\left\{\frac{L_1\Delta}{\theta\varepsilon^2}, \frac{\ell_1\delta\Delta}{\bar{\delta}_1\varepsilon^2}\theta^{-\frac{1}{q}}\right\} \\ &\overset{(40)}{\geq} \frac{1}{24\Delta_0\ell_1} \cdot \min\left\{\frac{L_1\Delta}{\varepsilon^2} + \frac{L_1\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{4\gamma_\infty\varepsilon}\right)^{\frac{p}{p-1}}, \frac{\ell_1\delta\Delta}{\bar{\delta}_1\varepsilon^2} + \frac{\ell_1\delta\Delta}{\bar{\delta}_1\varepsilon^2}\left(\frac{\sigma_1}{4\gamma_\infty\varepsilon}\right)^{\frac{p}{q(p-1)}}\right\} \\ &\geq C_{p,q} \min\left\{\frac{L_1\Delta}{\varepsilon^2} + \frac{L_1\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}, \frac{\delta\Delta}{\varepsilon^2} + \frac{\delta\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{q(p-1)}}\right\},\end{aligned}$$

where $C_{p,q} > 0$ is an universal constant depending only on p and q. Moreover, using Lemma C.1 (and more precisely Remark C.2), since $0 < \varepsilon \leq \frac{1}{8}\sqrt{L_1\Delta}$ then there exists an universal constant $C_p > 0$ (which depends only on p) such that

$$\mathfrak{m}_\varepsilon^{\mathtt{zr}}\left(K, \Delta, L_1, \delta^q, \sigma_1^p\right) \geq C_p \max\left\{1, \left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}\right\},$$

hence

$$\begin{aligned}\mathfrak{m}_\varepsilon^{\mathtt{zr}}\left(K, \Delta, L_1, \delta^q, \sigma_1^p\right) &\geq \Omega(1) \cdot \left(\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}} + \min\left\{\frac{L_1\Delta}{\varepsilon^2} + \frac{L_1\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}, \frac{\delta\Delta}{\varepsilon^2} + \frac{\delta\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{q(p-1)}}\right\}\right) \\ &\geq \Omega(1) \cdot \min\left\{\frac{L_1\Delta}{\varepsilon^2} + \frac{L_1\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}, \left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}} + \frac{\delta\Delta}{\varepsilon^2} + \frac{\delta\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{q(p-1)}}\right\}, \quad (47)\end{aligned}$$

as $0 < \varepsilon \leq c'\sqrt{L_1\Delta}$ implies

$$\frac{L_1\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}} \gtrsim \left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}},$$

and we can forget $\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}$ in the first term of the min.

- **Step 5:** *A Last Bound: the Case $\theta = 1$.*

  Observe that, if instead of taking θ as in (40), we choose directly $\theta = 1$ then, thanks to Lemma C.2 we immediately have

  $$\mathbb{E}\left[\|\bar{g}_T^\star(x,\xi) - \nabla F_T^\star(x)\|^p\right] = 0,$$

  and

  $$\mathbb{E}\left[\|[\bar{g}_T^\star(x,\xi) - \bar{g}_T^\star(y,\xi)] - [\nabla F_T^\star(x) - \nabla F_T^\star(y)]\|^q\right] = 0,$$

  so Assumption 2.3 and Assumption 2.7 are satisfied. Hence, if we set

  $$T = \left\lfloor \frac{\Delta}{\alpha\Delta_0} \right\rfloor, \quad \alpha = \frac{L_1}{\beta^2\ell_1}, \quad \text{and} \quad \beta = \frac{L_1}{2\ell_1\varepsilon},$$

  then $F_T^\star \in \mathcal{F}(\Delta, L_1)$ and the inequality (38) is satisfied. Hence, with probability at least $\frac{1}{2}$ it holds that for all integer $0 \leq t \leq \frac{T-1}{2\theta}$ and all $k \in [K]$ we have

  $$\left\|\nabla F_T^\star\left(x_{\mathtt{A}[\mathtt{0}_F]}^{(t,k)}\right)\right\| > 2\varepsilon, \quad \text{hence} \quad \mathbb{E}\left[\left\|\nabla F_T^\star\left(x_{\mathtt{A}[\mathtt{0}_F]}^{(t,k)}\right)\right\|\right] > \varepsilon,$$

  from where we obtain also

  $$\mathfrak{m}_\varepsilon^{\mathtt{zr}}\left(K, \Delta, L_1, \delta^q, \sigma_1^p\right) > \frac{T-1}{2\theta} = \frac{1}{2\theta}\left(\left\lfloor \frac{\Delta}{\alpha\Delta_0} \right\rfloor - 1\right) = \frac{1}{2\theta}\left(\left\lfloor \frac{\beta^2\ell_1\Delta}{L_1\Delta_0} \right\rfloor - 1\right) = \frac{1}{2\theta}\left(\left\lfloor \frac{L_1\Delta}{4\Delta_0\ell_1\varepsilon^2} \right\rfloor - 1\right) \quad (48)$$

  and, since we assume $0 < \varepsilon < c'\sqrt{L_1\Delta} = \sqrt{\frac{L_1\Delta}{12\Delta_0\ell_1}}$ this implies $\frac{L_1\Delta}{4\Delta_0\ell_1\varepsilon^2} \geq 3$ thus

  $$\begin{aligned}\mathfrak{m}_\varepsilon^{\mathtt{zr}}\left(K, \Delta, L_1, \delta^q, \sigma_1^p\right) &\overset{(48)}{>} \frac{1}{4\theta} \cdot \frac{L_1\Delta}{4\Delta_0\ell_1\varepsilon^2} \\ &= \Omega(1) \cdot \frac{L_1\Delta}{\varepsilon^2},\end{aligned}$$

and, combining this bound with (45) and (47) respectively leads to the desired result, i.e.,

$$\mathfrak{m}_\varepsilon^{\mathtt{zr}}\left(K,\Delta,L_1,\delta^q,\sigma_1^p\right) \geq \Omega(1)\cdot\min\left\{\frac{L_1\Delta}{\varepsilon^2}+\frac{L_1\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}},\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}+\frac{(L_1+\delta)\Delta}{\varepsilon^2}+\frac{\delta\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{q(p-1)}}\right\}.$$

□

C.5 Proof of Theorem 3.3

Lemma C.3 (Properties of the Gradient and Hessian Estimators g_T and ∇g_T)**.** *The stochastic gradient estimator g_T is a probability–θ zero-chain, is unbiased with respect to ∇F_T and satisfies*

$$\mathbb{E}\left[\|\nabla f_T(x,\xi)-\nabla F_T(x)\|^p\right] \leq \frac{2\gamma_\infty^p(1-\theta)}{\theta^{p-1}}, \quad \text{and} \quad \mathbb{E}\left[\left\|\nabla^2 f_T(x,\xi)-\nabla^2 F_T(x)\right\|_{\mathrm{op}}^q\right] \leq \frac{2\ell_1^q(1-\theta)}{\theta^{q-1}},$$

for all $x\in\mathbb{R}^T$, where $p\in(1,2]$, $q\in[1,2]$, γ_∞ and ℓ_1 are defined in Lemma B.2.

The proof of Lemma C.3 is very similar to Lemma C.2 (and is simpler since we can directly bound $\mathbb{I}\{i>\mathrm{prog}_{\frac{1}{4}}(x)\}$ by 1). For the bound on the difference $\nabla^2 f_T(x,\xi)-\nabla^2 F_T(x)$ we use Lemma B.2, that is, F_T has ℓ_1–Lipschitz gradients along with the fact that F_T is twice continuously differentiable which allows to bound the Hessian appropriately (see Lemma H.14). Moreover, by definition of the operator norm $\|\cdot\|_{\mathrm{op}}$ (Definition H.1) we have

$$\begin{aligned}
\left\|\left(\left[\nabla^2 F_T(x)\right]_{i,\cdot}\mathbb{I}\left\{i>\mathrm{prog}_{\frac{1}{4}}(x)\right\}\right)_{i\in[T]}\right\|_{\mathrm{op}} &\overset{\text{Def. H.1}}{=} \sup_{y\in\mathbb{R}^d,\|y\|=1}\left(\sum_{i=1}^T \mathbb{I}\left\{i>\mathrm{prog}_{\frac{1}{4}}(x)\right\}\left|\left[\nabla^2 F_T(x)\right]_{i,\cdot}{}^\top y\right|^2\right)^{\frac{1}{2}}\\
&\overset{(a)}{\leq} \sup_{y\in\mathbb{R}^d,\|y\|=1}\left(\sum_{i=1}^T \left|\left[\nabla^2 F_T(x)\right]_{i,\cdot}{}^\top y\right|^2\right)^{\frac{1}{2}}\\
&= \left\|\nabla^2 F_T(x)\right\|_{\mathrm{op}}\\
&\overset{\text{Lem. H.14}}{\leq} \ell_1,
\end{aligned}$$

where $\left[\nabla^2 F_T(x)\right]_{i,\cdot}$ denotes the i–th row of the Hessian of F_T. In (a) we use $\mathbb{I}\left\{i>\mathrm{prog}_{\frac{1}{4}}(x)\right\}\leq 1$.

Theorem 3.3. *Given $\Delta,L_1,L_2>0$, $\sigma_1,\sigma_2\geq 0$ and $0<\varepsilon\leq c_1\min\{\sqrt{L_1\Delta},L_2^{1/3}\Delta^{2/3}\}$ for some universal constant $c_1>0$. Then, for any algorithm $A\in\mathcal{A}_{\mathtt{zr}}$, there exists a function $f\in\mathcal{F}(\Delta,L_1,L_2)$, an oracle and a distribution $(O,\mathcal{D})\in\mathcal{O}(f,\sigma_1^p,\sigma_2^q)$ satisfying Assumptions 2.1 to 2.5 such that*

$$\begin{aligned}
\mathfrak{m}_\varepsilon^{\mathtt{zr}}\left(K,\Delta,L_1,L_2,\sigma_1^p,\sigma_2^q\right) \geq \Omega(1)\cdot\min\Bigg\{&\frac{L_1\Delta}{\varepsilon^2}+\frac{L_1\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}},\frac{L_2^{1/2}\Delta}{\varepsilon^{3/2}}+\frac{L_2^{1/2}\Delta}{\varepsilon^{3/2}}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}},\\
&\min\left\{\frac{L_1\Delta}{\varepsilon^2},\frac{L_2^{1/2}\Delta}{\varepsilon^{3/2}}\right\}+\frac{\Delta\sigma_2}{\varepsilon^2}+\frac{\Delta\sigma_2}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{q(p-1)}}+\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}\Bigg\}.
\end{aligned}$$

Proof. Let Δ_0, ℓ_1 and γ_∞ be the numerical constants in Lemma B.2 respectively. Additionally, we let the accuracy parameter $\varepsilon>0$, initial sub-optimality $\Delta\geq 0$, the Lipschitz constants $L_1,L_2\geq 0$ of the gradients and Hessians of F respectively, and the variance parameters $\sigma_1,\sigma_2\geq 0$ of the stochastic gradients and Hessians be fixed. Then, for $\alpha,\beta>0$ two positive real numbers, as in Arjevani et al. (2022), we rescale the function F_T as

$$F_T^\star\colon x\mapsto \alpha F_T(\beta x). \tag{49}$$

- **Step 1:** *Ensuring $F_T^\star\in\mathcal{F}(\Delta,L_1,L_2)$.*

  To guarantee the rescaled function $F_T^\star$ belongs to the function class $\mathcal{F}(\Delta,L_1,L_2)$, let us compute the initial sub-optimality Δ and check if F_T has L_1–Lipschitz gradients and L_2–Lipschitz Hessians. Assuming the algorithm $\mathtt{A}$ starts at $x^0=0$ we have

$$F_T^\star(0)-\inf_{x\in\mathbb{R}^T}F_T^\star(x) \overset{(49)}{=} \alpha\left(F_T(0)-\inf_{x\in\mathbb{R}^T}F_T(x)\right) \overset{\text{Lem. B.2}}{\leq} \alpha\Delta_0 T,$$

thus, it's enough to take $T = \left\lfloor \frac{\Delta}{\alpha\Delta_0} \right\rfloor$ so as to ensure $F_T^\star(0) - \inf_{x\in\mathbb{R}^T} F_T^\star(x) \leq \Delta$. Moreover, for any $x, y \in \mathbb{R}^T$,

$$\begin{aligned}\|\nabla F_T^\star(x) - \nabla F_T^\star(y)\| &= \alpha\beta \|\nabla F_T(\beta x) - \nabla F_T(\beta y)\| \\ &\overset{\text{Lem. B.2}}{\leq} \alpha\beta\ell_1 \|\beta x - \beta y\| \\ &= \alpha\beta^2\ell_1 \|x - y\|,\end{aligned}$$

and

$$\begin{aligned}\left\|\nabla^2 F_T^\star(x) - \nabla^2 F_T^\star(y)\right\|_{\text{op}} &= \alpha\beta^2 \left\|\nabla^2 F_T(\beta x) - \nabla^2 F_T(\beta y)\right\| \\ &\overset{\text{Lem. B.2}}{\leq} \alpha\beta^2\ell_2 \|\beta x - \beta y\| \\ &= \alpha\beta^3\ell_2 \|x - y\|,\end{aligned}$$

so it suffices to take $0 < \alpha\beta^2 \leq \frac{L_1}{\ell_1}$ and $0 < \alpha\beta^3 \leq \frac{L_2}{\ell_2}$ to ensure the function $F_T^\star$ has L_1–Lipschitz gradients and L_2–Lipschitz Hessians. Consequently, we have $F_T^\star \in \mathcal{F}(\Delta, L_1, L_2)$, as desired.

- **Step 2:** *Analysis of the Protocol and Choice for β.*

  Following the proof of Arjevani et al. (2022, Theorem 1), according to Lemma B.2, for all points $x \in \mathbb{R}^T$ such that $\text{prog}_0(x) < T$ we have $\text{prog}_0(\beta x) = \text{prog}_0(x) < T$ so

$$\|\nabla F_T^\star(x)\| \overset{(49)}{=} \alpha\beta \|\nabla F_T(\beta x)\| \overset{\text{Lem. B.2}}{>} \alpha\beta, \tag{50}$$

  and we need to guarantee that

$$\|\nabla F_T^\star(x)\| > 2\varepsilon, \tag{51}$$

  for all $x \in \mathbb{R}^T$ with $\text{prog}_0(x) < T$ which, given (50), can be done if we set $\alpha = \frac{2\varepsilon}{\beta}$.

- **Step 3:** *The Oracle Class and Assumptions 2.3 and 2.5.*

  It remains to choose the parameter $\theta \in (0, 1]$ and $\beta > 0$ such that $0 < \alpha\beta^2 = 2\varepsilon\beta \leq \frac{L_1}{\ell_1}$, $0 < \alpha\beta^3 = 2\varepsilon\beta^2 \leq \frac{L_2}{\ell_2}$ and the gradient estimator $g_T^\star$ of $\nabla F_T^\star$ and Hessian estimator $\nabla g_T^\star$ of $\nabla^2 F_T^\star$ satisfies Assumptions 2.3 and 2.5. Computing the p–th central moment of $g_T^\star$ gives, for all $x \in \mathbb{R}^T$

$$\begin{aligned}\mathbb{E}\left[\|g_T^\star(x,\xi) - \nabla F_T^\star(x)\|^p\right] &\overset{(29)}{=} (\alpha\beta)^p \,\mathbb{E}\left[\|g_T(\beta x,\xi) - \nabla F_T(\beta x)\|^p\right] \\ &\overset{\text{Lem. C.3}}{\leq} \frac{2\left(\gamma_\infty\alpha\beta\right)^p(1-\theta)}{\theta^{p-1}} \\ &\overset{\text{(a)}}{\leq} \frac{\left(2\gamma_\infty\alpha\beta\right)^p}{\theta^{p-1}} \\ &\overset{\text{(b)}}{=} \frac{\left(4\gamma_\infty\varepsilon\right)^p}{\theta^{p-1}},\end{aligned} \tag{52}$$

  where in (a) we use the fact that $0 < \theta \leq 1$ and $p \geq 1$ so that $2 \leq 2^p$ while in (b) we use $\alpha = \frac{2\varepsilon}{\beta}$. Moreover, following the sames lines as in (52) we have

$$\begin{aligned}\mathbb{E}\left[\left\|\nabla g_T^\star(x,\xi) - \nabla^2 F_T^\star(x)\right\|^q\right] &\overset{(29)}{=} \left(\alpha\beta^2\right)^q \mathbb{E}\left[\left\|\nabla g_T(\beta x,\xi) - \nabla^2 F_T(\beta x)\right\|^q\right] \\ &\overset{\text{Lem. C.3}}{\leq} \frac{2\left(\ell_1\alpha\beta^2\right)^q(1-\theta)}{\theta^{q-1}} \\ &\leq \frac{\left(2\ell_1\alpha\beta^2\right)^q}{\theta^{q-1}}.\end{aligned} \tag{53}$$

  From (52), so as to satisfy Assumptions 2.3 and 2.5 it's enough to take $\theta = \min\{1, \theta_1\}$ such that

$$\theta_1^{p-1} \geq \left(\frac{4\gamma_\infty\varepsilon}{\sigma_1}\right)^p, \quad \text{so} \quad \theta_1 \geq \left(\frac{4\gamma_\infty\varepsilon}{\sigma_1}\right)^{\frac{p}{p-1}}, \tag{54}$$

hence

$$\theta = \min\left\{1, \left(\frac{4\gamma_\infty \varepsilon}{\sigma_1}\right)^{\frac{p}{p-1}}\right\}, \tag{55}$$

while for (53) we need to have $0 < \alpha\beta^2 \leq \frac{\sigma_2 \theta^{\frac{q-1}{q}}}{2\ell_1}$ thus we fix β such that

$$\alpha\beta^2 = 2\varepsilon\beta \leq \min\left\{\frac{L_1}{\ell_1}, \frac{\sigma_2\theta^{\frac{q-1}{q}}}{2\ell_1}\right\} \quad \text{and} \quad \alpha\beta^3 = 2\varepsilon\beta^2 \leq \frac{L_2}{\ell_2},$$

that is to say

$$\beta = \min\left\{\frac{L_1}{2\ell_1\varepsilon}, \sqrt{\frac{L_2}{2\ell_2\varepsilon}}, \frac{\sigma_2\theta^{\frac{q-1}{q}}}{4\ell_1\varepsilon}\right\}. \tag{56}$$

- **Step 4:** *Lower Bounding* $\mathfrak{m}_\varepsilon^{\mathtt{zr}}(K, \Delta, L_1, L_2, \sigma_1^p, \sigma_2^q)$.

  Continuing on **step 2**, by Lemma B.1 we know that with probability at least $\frac{1}{2}$ it holds that for all integer $0 \leq t \leq \frac{T-1}{2\theta}$ and all $k \in [K]$ we have

$$\left\|\nabla F_T^\star\left(x_{\mathtt{A}[\mathtt{O}_F]}^{(t,k)}\right)\right\| > 2\varepsilon, \quad \text{hence} \quad \mathbb{E}\left[\left\|\nabla F_T^\star\left(x_{\mathtt{A}[\mathtt{O}_F]}^{(t,k)}\right)\right\|\right] > \varepsilon,$$

  from where it follows that

$$\mathfrak{m}_\varepsilon^{\mathtt{zr}}(K, \Delta, L_1, L_2, \sigma_1^p, \sigma_2^q) > \frac{T-1}{2\theta} = \frac{1}{2\theta}\left(\left\lfloor\frac{\Delta}{\alpha\Delta_0}\right\rfloor - 1\right) = \frac{1}{2\theta}\left(\left\lfloor\frac{\Delta\beta}{2\Delta_0\varepsilon}\right\rfloor - 1\right), \tag{57}$$

  we then distinguish two cases:

  – if $\frac{\Delta\beta}{2\Delta_0\varepsilon} \geq 3$ then, using the inequality $\lfloor x\rfloor - 1 \geq \frac{x}{2}$, valid for all real number $x \geq 3$ we obtain

$$\begin{aligned}
&\mathfrak{m}_\varepsilon^{\mathtt{zr}}(K, \Delta, L_1, L_2, \sigma_1^p, \sigma_2^q) \\
&\overset{(57)}{>} \frac{1}{4\theta}\cdot\frac{\Delta\beta}{2\Delta_0\varepsilon} \\
&= \frac{1}{16\Delta_0\ell_1}\cdot\frac{\Delta}{\varepsilon}\min\left\{\frac{L_1}{\varepsilon\theta}, \frac{1}{\theta}\sqrt{\frac{2L_2}{\ell_2\varepsilon}}, \frac{\sigma_2\theta^{-\frac{1}{q}}}{2\varepsilon}\right\} \\
&\overset{(55)}{\geq} \frac{1}{32\Delta_0\ell_1}\cdot\frac{\Delta}{\varepsilon}\min\left\{\frac{L_1}{\varepsilon}\left(1+\left(\frac{\sigma_1}{4\gamma_\infty\varepsilon}\right)^{\frac{p}{p-1}}\right), \sqrt{\frac{2L_2}{\ell_2\varepsilon}}\left(1+\left(\frac{\sigma_1}{4\gamma_\infty\varepsilon}\right)^{\frac{p}{p-1}}\right), \frac{\sigma_2}{2\varepsilon}\left(1+\left(\frac{\sigma_1}{4\gamma_\infty\varepsilon}\right)^{\frac{p}{q(p-1)}}\right)\right\} \\
&= \Omega(1)\cdot\min\left\{\frac{L_1\Delta}{\varepsilon^2}+\frac{L_1\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}, \frac{L_2^{1/2}\Delta}{\varepsilon^{3/2}}+\frac{L_2^{1/2}\Delta}{\varepsilon^{3/2}}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}, \frac{\Delta\sigma_2}{\varepsilon^2}+\frac{\Delta\sigma_2}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{q(p-1)}}\right\},
\end{aligned} \tag{58}$$

    and combining the lower bound (58) with Lemma C.1, using the choice for c' provided below, gives:

$$\begin{aligned}
&\mathfrak{m}_\varepsilon^{\mathtt{zr}}(K, \Delta, L_1, L_2, \sigma_1^p, \sigma_2^q) \\
&\geq \Omega(1)\cdot\min\left\{\frac{L_1\Delta}{\varepsilon^2}+\frac{L_1\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}, \frac{L_2^{1/2}\Delta}{\varepsilon^{3/2}}+\frac{L_2^{1/2}\Delta}{\varepsilon^{3/2}}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}, \left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}+\frac{\Delta\sigma_2}{\varepsilon^2}+\frac{\Delta\sigma_2}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{q(p-1)}}\right\},
\end{aligned} \tag{59}$$

    since from the assumption $0 < \varepsilon \leq c'\min\left\{\sqrt{L_1\Delta}, L_2^{1/3}\Delta^{2/3}\right\}$ we have

$$\frac{L_1\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}} \gtrsim \left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}} \quad \text{and} \quad \frac{L_2^{1/2}\Delta}{\varepsilon^{3/2}}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}} \gtrsim \left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}},$$

    and we can forget $\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}$ in the first two terms of the min.

– otherwise, if $\frac{\Delta\beta}{2\Delta_0\varepsilon} < 3$, we choose the universal constant c' as

$$0 < c' = \min\left\{(12\Delta_0\ell_1)^{-\frac{1}{2}}, \left(6\sqrt{2}\Delta_0\ell_2^{1/2}\right)^{-\frac{2}{3}}\right\} < \frac{1}{8}, \tag{60}$$

then, using Remark C.1, dividing both sides of $\frac{\Delta\beta}{2\Delta_0\varepsilon} < 3$ by our choice of $\theta > 0$ yields

$$\frac{1}{6\Delta_0}\cdot\frac{\Delta\beta}{\varepsilon}\theta^{-1} \leq \theta^{-1} = \max\left\{1, \left(\frac{\sigma_1}{4\gamma_\infty\varepsilon}\right)^{\frac{p}{p-1}}\right\} \leq \max\left\{1, \left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}\right\}, \tag{61}$$

where the last inequality follows from the fact that $\frac{p}{p-1} > 0$ and $4\gamma_\infty = 92 \geq 1$. Using the definition of β we obtain

$$\begin{aligned}
&\max\left\{1, \left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}\right\} \\
&\quad\overset{(61)+(56)}{\geq} \frac{1}{6\Delta_0}\cdot\min\left\{\frac{L_1\Delta}{2\ell_1\theta\varepsilon^2}, \frac{\Delta}{\theta\varepsilon}\sqrt{\frac{L_2}{2\ell_2\varepsilon}}, \frac{\Delta\sigma_2\theta^{-\frac{1}{q}}}{4\ell_1\varepsilon^2}\right\} \\
&\quad\overset{(55)}{\geq} \frac{1}{24\Delta_0\max\left\{\ell_1, \ell_2^{1/2}\right\}}\cdot\min\left\{\frac{L_1\Delta}{\varepsilon^2} + \frac{L_1\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{4\gamma_\infty\varepsilon}\right)^{\frac{p}{p-1}}, \right. \\
&\qquad\qquad \left. \frac{L_2^{1/2}\Delta}{\varepsilon^{3/2}} + \frac{L_2^{1/2}\Delta}{\varepsilon^{3/2}}\left(\frac{\sigma_1}{4\gamma_\infty\varepsilon}\right)^{\frac{p}{p-1}}, \frac{\Delta\sigma_2}{\varepsilon^2} + \frac{\Delta\sigma_2}{\varepsilon^2}\left(\frac{\sigma_1}{4\gamma_\infty\varepsilon}\right)^{\frac{p}{q(p-1)}}\right\} \\
&\quad\geq C_{p,q}\min\left\{\frac{L_1\Delta}{\varepsilon^2} + \frac{L_1\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}, \frac{L_2^{1/2}\Delta}{\varepsilon^{3/2}} + \frac{L_2^{1/2}\Delta}{\varepsilon^{3/2}}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}, \frac{\Delta\sigma_2}{\varepsilon^2} + \frac{\Delta\sigma_2}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{q(p-1)}}\right\},
\end{aligned}$$

where $C_{p,q} > 0$ is an universal constant depending only on p and q. Moreover, using Lemma C.1 (and more precisely Remark C.3), since $0 < \varepsilon \leq \frac{1}{8}\sqrt{L_1\Delta}$ then there exists an universal constant $C_p > 0$ (which depends only on p) such that

$$\mathfrak{m}_\varepsilon^{\mathtt{zr}}(K, \Delta, L_1, L_2, \sigma_1^p, \sigma_2^q) \geq C_p\max\left\{1, \left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}\right\},$$

hence

$$\begin{aligned}
&\mathfrak{m}_\varepsilon^{\mathtt{zr}}(K, \Delta, L_1, L_2, \sigma_1^p, \sigma_2^q) \\
&\quad\geq \Omega(1)\cdot\left(\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}} + \min\left\{\frac{L_1\Delta}{\varepsilon^2} + \frac{L_1\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}, \frac{L_2^{1/2}\Delta}{\varepsilon^{3/2}} + \frac{L_2^{1/2}\Delta}{\varepsilon^{3/2}}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}, \frac{\Delta\sigma_2}{\varepsilon^2} + \frac{\Delta\sigma_2}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{q(p-1)}}\right\}\right) \\
&\quad\geq \Omega(1)\cdot\min\left\{\frac{L_1\Delta}{\varepsilon^2} + \frac{L_1\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}, \frac{L_2^{1/2}\Delta}{\varepsilon^{3/2}} + \frac{L_2^{1/2}\Delta}{\varepsilon^{3/2}}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}, \left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}} + \frac{\Delta\sigma_2}{\varepsilon^2} + \frac{\Delta\sigma_2}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{q(p-1)}}\right\},
\end{aligned} \tag{62}$$

as $0 < \varepsilon \leq c'\min\left\{\sqrt{L_1\Delta}, L_2^{1/3}\Delta^{2/3}\right\}$ implies

$$\frac{L_1\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}} \gtrsim \left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}} \quad\text{and}\quad \frac{L_2^{1/2}\Delta}{\varepsilon^{3/2}}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}} \gtrsim \left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}},$$

and we can forget $\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}$ in the first two terms of the min.

- **Step 5:** *A Last Bound: the Case $\theta = 1$.*

  Observe that, if instead of taking θ as in (55), we choose directly $\theta = 1$ then, thanks to Lemma C.3 we immediately have

$$\mathbb{E}\left[\|g_T^\star(x,\xi) - \nabla F_T^\star(x)\|^p\right] = 0,$$

and

$$\mathbb{E}\left[\left\|\nabla g_T^\star(x,\xi) - \nabla^2 F_T^\star(x)\right\|^q\right] = 0,$$

so Assumption 2.3 and Assumption 2.5 are satisfied. Hence, if we set

$$T = \left\lfloor \frac{\Delta}{\alpha\Delta_0} \right\rfloor, \quad \alpha = \frac{2\varepsilon}{\beta}, \quad \text{and} \quad \beta = \min\left\{\frac{L_1}{2\ell_1\varepsilon}, \sqrt{\frac{L_2}{2\ell_2\varepsilon}}\right\},$$

then $F_T^\star \in \mathcal{F}(\Delta, L_1, L_2)$ and the inequality (51) is satisfied. Hence, with probability at least $\frac{1}{2}$ it holds that for all integer $0 \le t \le \frac{T-1}{2\theta}$ and all $k \in [K]$ we have

$$\left\|\nabla F_T^\star\left(x^{(t,k)}_{\mathtt{A}[\mathtt{0}_F]}\right)\right\| > 2\varepsilon, \quad \text{hence} \quad \mathbb{E}\left[\left\|\nabla F_T^\star\left(x^{(t,k)}_{\mathtt{A}[\mathtt{0}_F]}\right)\right\|\right] > \varepsilon,$$

from where we obtain also

$$\mathfrak{m}_\varepsilon^{\mathtt{zr}}\left(K, \Delta, L_1, L_2, \sigma_1^p, \sigma_2^q\right) > \frac{T-1}{2\theta} = \frac{1}{2\theta}\left(\left\lfloor \frac{\Delta}{\alpha\Delta_0}\right\rfloor - 1\right) = \frac{1}{2\theta}\left(\left\lfloor \frac{\Delta\beta}{2\Delta_0\varepsilon}\right\rfloor - 1\right), \tag{63}$$

and by our assumption on ε (see (60)), we assumed

$$0 < \varepsilon < c' \min\left\{\sqrt{L_1\Delta}, L_2^{1/3}\Delta^{2/3}\right\} \le \min\left\{\sqrt{\frac{L_1\Delta}{12\Delta_0\ell_1}}, \left(\frac{L_2^{1/2}\Delta}{6\sqrt{2}\Delta_0\ell_2^{1/2}}\right)^{\frac{2}{3}}\right\},$$

which is enough to imply the inequality

$$\min\left\{\frac{L_1\Delta}{4\Delta_0\ell_1\varepsilon^2}, \frac{L_2^{1/2}\Delta}{2\sqrt{2}\Delta_0\ell_2^{1/2}\varepsilon^{3/2}}\right\} = \frac{\Delta\beta}{2\Delta_0\varepsilon} \ge 3,$$

thus, we have

$$\begin{aligned}
\mathfrak{m}_\varepsilon^{\mathtt{zr}}\left(K, \Delta, L_1, L_2, \sigma_1^p, \sigma_2^q\right) &\overset{(63)}{>} \frac{1}{4\theta}\cdot\frac{\Delta\beta}{2\Delta_0\varepsilon^2} \\
&= \frac{1}{8\Delta_0}\cdot\frac{\Delta}{\varepsilon}\min\left\{\frac{L_1}{2\ell_1\varepsilon}, \sqrt{\frac{L_2}{2\ell_2\varepsilon}}\right\} \\
&= \Omega(1)\cdot\min\left\{\frac{L_1\Delta}{\varepsilon^2}, \frac{L_2^{1/2}\Delta}{\varepsilon^{3/2}}\right\},
\end{aligned}$$

and, combining this bound with (59) and (62) respectively leads to the desired result, i.e.,

$$\begin{aligned}
\mathfrak{m}_\varepsilon^{\mathtt{zr}}\left(K, \Delta, L_1, L_2, \sigma_1^p, \sigma_2^q\right) \ge \Omega(1)\cdot\min\Bigg\{&\frac{L_1\Delta}{\varepsilon^2} + \frac{L_1\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}, \frac{L_2^{1/2}\Delta}{\varepsilon^{3/2}} + \frac{L_2^{1/2}\Delta}{\varepsilon^{3/2}}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}, \\
&\min\left\{\frac{L_1\Delta}{\varepsilon^2}, \frac{L_2^{1/2}\Delta}{\varepsilon^{3/2}}\right\} + \frac{\Delta\sigma_2}{\varepsilon^2} + \frac{\Delta\sigma_2}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{q(p-1)}} + \left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}\Bigg\}.
\end{aligned}$$

□

D MISSING PROOFS IN SECTION 4

D.1 Auxiliary Lemmas

Lemma D.1 (A Descent Lemma). *Under Assumptions 2.3 and 2.6, for any choice of stepsize $\gamma > 0$, and for any $t \in \{0, \dots, T-1\}$ we have*

$$F(x_{t+1}) \leq F(x_t) + 2\gamma \|\hat{e}_t\| - \gamma \|\nabla F(x_t)\| + \frac{\gamma^2 \bar{L}}{2}, \tag{64}$$

where $\hat{e}_t := g_t - \nabla F(x_t)$ is the error term.

Proof. According to Assumption 2.6 we know that the function F is $\bar{L}$–smooth (Nesterov, 2018) since by Jensen's inequality (Lemma H.9) applied on the convex function $x \mapsto \|x\|^q$ (since $q \geq 1$)

$$\|\nabla F(x) - \nabla F(y)\|^q \overset{\text{Ass. 2.3}}{=} \|\mathbb{E}_{\xi\sim\mathcal{D}}\left[\nabla f(x,\xi) - \nabla f(y,\xi)\right]\|^q \overset{\text{Lem. H.9}}{\leq} \mathbb{E}_{\xi\sim\mathcal{D}}\left[\|\nabla f(x,\xi) - \nabla f(y,\xi)\|^q\right] \overset{\text{Ass. 2.6}}{\leq} \bar{L}^q \|x-y\|^q, \tag{65}$$

thus it holds

$$\begin{aligned}
F(x_{t+1}) &\leq F(x_t) + \langle \nabla F(x_t), x_{t+1} - x_t\rangle + \frac{\bar{L}}{2}\|x_{t+1} - x_t\|^2 \\
&\overset{(a)}{=} F(x_t) - \gamma \left\langle \nabla F(x_t), \frac{g_t}{\|g_t\|}\right\rangle + \frac{\gamma^2\bar{L}}{2} \\
&= F(x_t) - \gamma \left\langle \nabla F(x_t) - g_t, \frac{g_t}{\|g_t\|}\right\rangle - \gamma\|g_t\| + \frac{\gamma^2\bar{L}}{2} \\
&\overset{\text{Lem. H.3}}{\leq} F(x_t) + \gamma\|\nabla F(x_t) - g_t\| - \gamma\|g_t\| + \frac{\gamma^2\bar{L}}{2} \\
&\overset{(b)}{\leq} F(x_t) + 2\gamma\|\nabla F(x_t) - g_t\| - \gamma\|\nabla F(x_t)\| + \frac{\gamma^2\bar{L}}{2} \\
&= F(x_t) + 2\gamma\|\hat{e}_t\| - \gamma\|\nabla F(x_t)\| + \frac{\gamma^2\bar{L}}{2},
\end{aligned}$$

where in (a) we use the update rule $x_{t+1} = x_t - \gamma\frac{g_t}{\|g_t\|}$ while in (b) we use the triangle inequality. This establishes the desired claim. □

Lemma D.2 (Another Descent Lemma). *Under Assumptions 2.2 and 2.3, for any choice of stepsize $\gamma > 0$, and for any $t \in \{0, \dots, T-1\}$ we have*

$$F(x_{t+1}) \leq F(x_t) + 2\gamma \|\hat{e}_t\| - \gamma \|\nabla F(x_t)\| + \frac{\gamma^2 L_1}{2}, \tag{66}$$

where $\hat{e}_t := g_t - \nabla F(x_t)$ is the error term.

Proof. The proof is the same as in the previous descent lemma (Lemma D.1) where now F has L_1–Lipschitz continuous gradients (instead of $\bar{L}$). □

Lemma D.3 (Unrolling the Descent Lemma). *Under Assumptions 2.1, 2.3 and 2.6, for any choice of stepsize $\gamma > 0$ the iterates $\{x_t\}_{t\in\{0,\dots,T\}}$ produced by Algorithm 1 satisfy*

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F(x_t)\| \leq \frac{\Delta}{\gamma T} + \frac{2}{T}\sum_{t=0}^{T-1}\|\hat{e}_t\| + \frac{\gamma\bar{L}}{2}.$$

Proof. From the previous descent lemma (Lemma D.1), summing inequality (64) over $t \in \{0, \dots, T-1\}$ gives

$$\gamma\sum_{t=0}^{T-1}\|\nabla F(x_t)\| \leq F(x_0) - F(x_T) + 2\gamma\sum_{t=0}^{T-1}\|\hat{e}_t\| + \frac{\gamma^2\bar{L}T}{2}, \tag{67}$$

where we telescope the terms $F(x_t) - F(x_{t+1})$. Multiplying both sides of (67) by $1/\gamma T$ leads to

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F(x_t)\| \leq \frac{1}{\gamma T}\left(F(x_0) - F(x_T)\right) + \frac{2}{T}\sum_{t=0}^{T-1}\|\hat{e}_t\| + \frac{\gamma \bar{L}}{2},$$

and using Assumption 2.1 we obtain

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F(x_t)\| \leq \frac{\Delta}{\gamma T} + \frac{2}{T}\sum_{t=0}^{T-1}\|\hat{e}_t\| + \frac{\gamma \bar{L}}{2},$$

as desired. □

If we assume the function F has L_1–Lipschitz continuous gradients (Assumption 2.2) then Lemma D.2 holds and we can unroll it in the same way as we did above. For that reason, we only state the result and we omit the proof.

Lemma D.4 (Unrolling the Descent Lemma)**.** *Under Assumptions 2.1 to 2.3, for any choice of stepsize $\gamma > 0$ the iterates $\{x_t\}_{t\in\{0,\ldots,T\}}$ produced by Algorithm 1 satisfy*

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F(x_t)\| \leq \frac{\Delta}{\gamma T} + \frac{2}{T}\sum_{t=0}^{T-1}\|\hat{e}_t\| + \frac{\gamma L_1}{2}.$$

Lemma D.5 (Bounding the Error Term)**.** *Under Assumptions 2.3 and 2.6, for all $t \in \{0,\ldots,T-1\}$ we have*

$$\mathbb{E}\left[\|\hat{e}_t\|\right] \leq (1-\alpha)^t \mathbb{E}\left[\|\hat{e}_0\|\right] + 2\sigma_1 \alpha^{\frac{p-1}{p}} + 4\gamma\bar{L}\alpha^{-\frac{1}{q}},$$

where $\hat{e}_t := g_t - \nabla F(x_t)$.

Proof. By the update rule of the gradient estimator in Algorithm 1 (line 9) we have, for all $t \in \{1,\ldots,T-1\}$

$$\begin{aligned}
\hat{e}_t &:= g_t - \nabla F(x_t)\\
&= (1-\alpha)(g_{t-1} + \nabla f(x_t,\xi_t) - \nabla f(x_{t-1},\xi_t) + \alpha\nabla f(x_t,\xi_t) - \nabla F(x_t)\\
&= (1-\alpha)(g_{t-1} - \nabla F(x_{t-1})) + \alpha\left(\nabla f(x_t,\xi_t) - \nabla F(x_t)\right)\\
&\qquad\qquad - (1-\alpha)\left([\nabla F(x_t) - \nabla f(x_t,\xi_t)] - [\nabla F(x_{t-1}) - \nabla f(x_{t-1},\xi_t)]\right)\\
&= (1-\alpha)\hat{e}_{t-1} + \alpha e_t - (1-\alpha)\hat{S}_t,
\end{aligned} \tag{68}$$

where we let $e_t := \nabla f(x_t,\xi_t) - \nabla F(x_t)$ and $\hat{S}_t := [\nabla F(x_t) - \nabla f(x_t,\xi_t)] - [\nabla F(x_{t-1}) - \nabla f(x_{t-1},\xi_t)]$. It is worth noting that both $(e_t)_{t\geq 0}$ and $(\hat{S}_t)_{t\geq 0}$ are martingale difference sequence with respect to the filtration $(\mathcal{F}_t)_{t\geq 0}$ where $\mathcal{F}_t := \sigma(g_0,\xi_1,\ldots,\xi_t)$.

Then, unrolling the recursion (68) gives

$$\hat{e}_t = (1-\alpha)^t\hat{e}_0 + \alpha\sum_{j=0}^{t-1}(1-\alpha)^{t-j-1}e_{j+1} - \sum_{j=0}^{t-1}(1-\alpha)^{t-j}\hat{S}_{j+1},$$

and taking the norm followed by the total expectation yields

$$\mathbb{E}\left[\|\hat{e}_t\|\right] \leq (1-\alpha)^t\mathbb{E}\left[\|\hat{e}_0\|\right] + \alpha\mathbb{E}\left[\left\|\sum_{j=0}^{t-1}(1-\alpha)^{t-j-1}e_{j+1}\right\|\right] + (1-\alpha)\mathbb{E}\left[\left\|\sum_{j=0}^{t-1}(1-\alpha)^{t-j-1}\hat{S}_{j+1}\right\|\right].$$

We now need to upper bound the last two terms of the previous inequality. For the first term, using Jensen's

inequality (Lemma H.9) we have

$$\begin{aligned}
\mathbb{E}\left[\left\|\sum_{j=0}^{t-1}(1-\alpha)^{t-j-1}e_{j+1}\right\|\right] &\overset{\text{Lem. H.9}}{\leq} \left(\mathbb{E}\left[\left\|\sum_{j=0}^{t-1}(1-\alpha)^{t-j-1}e_{j+1}\right\|^p\right]\right)^{\frac{1}{p}} \\
&\overset{\text{Lem. G.1}}{\leq} \left(2\sum_{j=0}^{t-1}(1-\alpha)^{p(t-j-1)}\mathbb{E}\left[\|e_{j+1}\|^p\right]\right)^{\frac{1}{p}} \\
&\overset{\text{Ass. 2.3}}{\leq} 2\left(\sum_{j=0}^{t-1}(1-\alpha)^{p(t-j-1)}\sigma_1^p\right)^{\frac{1}{p}} \\
&\leq 2\sigma_1\left(\sum_{j=0}^{t-1}(1-\alpha)^{t-j-1}\right)^{\frac{1}{p}} \\
&\leq 2\sigma_1\left(\sum_{j\geq 0}(1-\alpha)^{j}\right)^{\frac{1}{p}} \\
&= 2\sigma_1\alpha^{-\frac{1}{p}}, \qquad (69)
\end{aligned}$$

while, for the last term we have

$$\begin{aligned}
\mathbb{E}\left[\left\|\sum_{j=0}^{t-1}(1-\alpha)^{t-j-1}\hat{S}_{j+1}\right\|\right] &\overset{\text{Lem. H.9}}{\leq} \left(\mathbb{E}\left[\left\|\sum_{j=0}^{t-1}(1-\alpha)^{t-j-1}\hat{S}_{j+1}\right\|^q\right]\right)^{\frac{1}{q}} \\
&\overset{\text{Lem. G.1}}{\leq} \left(2\sum_{j=0}^{t-1}(1-\alpha)^{q(t-j-1)}\mathbb{E}\left[\left\|\hat{S}_{j+1}\right\|^q\right]\right)^{\frac{1}{q}}, \qquad (70)
\end{aligned}$$

and, using Assumption 2.6 we obtain

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{S}_{j+1}\right\|^q\right] &= \mathbb{E}\left[\|[\nabla F(x_{j+1}) - \nabla F(x_j)] - [\nabla f(x_{j+1},\xi_{j+1}) - \nabla f(x_j,\xi_{j+1})]\|^q\right] \\
&\overset{\text{Lem. H.6}}{\leq} 2^{q-1}\mathbb{E}\left[\|\nabla F(x_{j+1}) - \nabla F(x_j)\|^q + \|\nabla f(x_{j+1},\xi_{j+1}) - \nabla f(x_j,\xi_{j+1})\|^q\right] \\
&\overset{\text{Ass. 2.6}}{\leq} 2^q\bar{L}^q\mathbb{E}\left[\|x_{j+1}-x_j\|^q\right] \\
&= 2^q\gamma^q\bar{L}^q, \qquad (71)
\end{aligned}$$

thus, using (71) and $2^{\frac{1}{q}} \leq 2$ we obtain

$$\begin{aligned}
\mathbb{E}\left[\left\|\sum_{j=0}^{t-1}(1-\alpha)^{t-j-1}\hat{S}_{j+1}\right\|\right] &\overset{(70)}{\leq} 4\gamma\bar{L}\left(\sum_{j=0}^{t-1}(1-\alpha)^{q(t-j-1)}\right)^{\frac{1}{q}} \\
&\leq 4\gamma\bar{L}\alpha^{-\frac{1}{q}}. \qquad (72)
\end{aligned}$$

Then, combining the bounds (69) and (72) we have

$$\mathbb{E}\left[\|\hat{e}_t\|\right] \leq (1-\alpha)^t\mathbb{E}\left[\|\hat{e}_0\|\right] + 2\sigma_1\alpha^{\frac{p-1}{p}} + 4\gamma\bar{L}\alpha^{-\frac{1}{q}},$$

which achieves the proof of the lemma. □

Remark D.1. In (71) we only use the q–weak average smoothness assumption (Assumption 2.6) to achieve the bound $2^q\gamma^q\bar{L}^q$. It is worth mentioning that this assumption can be replaced by the combinations of both $\bar{L}$–Lipschitz continuous gradients of F and

$$\mathbb{E}_{\xi\sim\mathcal{D}}\left[\|[\nabla f(x,\xi) - \nabla f(y,\xi)] - [\nabla F(x) - \nabla F(y)]\|^q\right] \leq \delta^q\|x-y\|^q,$$

for all $x, y \in \mathbb{R}^d$, where $\delta \geq 0$ is some fixed constant which can be much smaller than δ (see Assumption 2.7).

Corollary D.1 (Bounding the Error Term: a Refined Version)**.** *Under Assumptions 2.3 and 2.7, for all $t \in \{0, \ldots, T-1\}$ we have*

$$\mathbb{E}\left[\|\hat{e}_t\|\right] \leq (1-\alpha)^t \mathbb{E}\left[\|\hat{e}_0\|\right] + 2\sigma_1 \alpha^{\frac{p-1}{p}} + 2\gamma\delta\alpha^{-\frac{1}{q}}, \tag{73}$$

where $\hat{e}_t := g_t - \nabla F(x_t)$.

Proof. The first two term in the upper bound (73) are obtained exactly the same way as in Lemma D.5. For the last term, we start exactly as in (70) and, using Assumption 2.7 we have

$$\mathbb{E}\left[\left\|\hat{S}_{j+1}\right\|^q\right] = \mathbb{E}\left[\|[\nabla F(x_t) - \nabla F(x_{t-1})] - [\nabla f(x_t, \xi_t) - \nabla f(x_{t-1}, \xi_t)]\|^q\right] \overset{\text{Ass. 2.7}}{\leq} \delta^q \|x_t - x_{t-1}\|^q \leq \gamma^q \delta^q,$$

and plugging this new bound into (72) gives the inequality (73), as claimed. □

D.2 Proof of Theorem 4.1

Thanks to Lemmas D.3 and D.5 we can now establish the convergence analysis (in expectation) of Algorithm 1.

Theorem 4.1. *Under Assumptions 2.1, 2.3 and 2.6, let the initial gradient estimate g_0 be given by*

$$g_0 = \frac{1}{B_{\text{init}}} \sum_{j=0}^{B_{\text{init}}-1} \nabla f\left(x_0, \xi_{0,j}\right),$$

where $B_{\text{init}} = \max\left\{1, \left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}\right\}$, let the stepsize $\gamma = \sqrt{\frac{\Delta\alpha^{1/q}}{\bar{L}T}}$, the momentum parameter $\alpha = \min\{1, \alpha_{\text{eff}}\}$ where

$$\alpha_{\text{eff}} = \max\left\{\left(\frac{\varepsilon}{\sigma_1 T}\right)^{\frac{p}{2p-1}}, \left(\frac{\bar{L}\Delta}{\sigma_1^2 T}\right)^{\frac{pq}{p(2q+1)-2q}}\right\}. \tag{74}$$

Then, Algorithm 1 guarantees to find an ε–stationary point with the total sample complexity

$$\mathcal{O}\left(\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}} + \frac{\bar{L}\Delta}{\varepsilon^2} + \frac{\bar{L}\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{q(p-1)}}\right).$$

Proof. According to Lemma D.3 we have

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla F(x_t)\| \leq \frac{\Delta}{\gamma T} + \frac{2}{T}\sum_{t=0}^{T-1} \|\hat{e}_t\| + \frac{\gamma\bar{L}}{2},$$

and using Lemma D.5 this yields

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\|\hat{e}_t\|\right] \leq \frac{1}{T}\sum_{t=0}^{T-1} (1-\alpha)^t \mathbb{E}\left[\|\hat{e}_0\|\right] + 2\sigma_1\alpha^{\frac{p-1}{p}} + 4\gamma\bar{L}\alpha^{-\frac{1}{q}} \leq \frac{\mathbb{E}\left[\|\hat{e}_0\|\right]}{\alpha T} + 2\sigma_1\alpha^{\frac{p-1}{p}} + 4\gamma\bar{L}\alpha^{-\frac{1}{q}},$$

hence

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla F(x_t)\|\right] \leq \frac{\Delta}{\gamma T} + \frac{2\mathbb{E}\left[\|\hat{e}_0\|\right]}{\alpha T} + 4\sigma_1\alpha^{\frac{p-1}{p}} + 8\gamma\bar{L}\alpha^{-\frac{1}{q}} + \frac{\gamma\bar{L}}{2} \leq \frac{\Delta}{\gamma T} + \frac{2\mathbb{E}\left[\|\hat{e}_0\|\right]}{\alpha T} + 4\sigma_1\alpha^{\frac{p-1}{p}} + 9\gamma\bar{L}\alpha^{-\frac{1}{q}}, \tag{75}$$

since $0 < \alpha \leq 1$. Now, using $\gamma = \sqrt{\frac{\Delta\alpha^{1/q}}{\bar{L}T}}$ we have

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla F(x_t)\|\right] = \mathcal{O}\left(\alpha^{-\frac{1}{2q}}\sqrt{\frac{\bar{L}\Delta}{T}} + \frac{\mathbb{E}\left[\|\hat{e}_0\|\right]}{\alpha T} + \sigma_1\alpha^{\frac{p-1}{p}}\right). \tag{76}$$

Now, by our choice of g_0 we have

$$\begin{aligned}\mathbb{E}\left[\|\hat{e}_0\|\right] &\overset{\text{Lem. H.9}}{\leq} \left(\mathbb{E}\left[\|\hat{e}_0\|^p\right]\right)^{\frac{1}{p}} \\ &= \left(\mathbb{E}\left[\|g_0 - \nabla F(x_0)\|^p\right]\right)^{\frac{1}{p}} \\ &\overset{\text{Lem. G.1}}{\leq} \frac{2}{B_{\text{init}}}\left(\sum_{j=0}^{B_{\text{init}}-1}\mathbb{E}\left[\|\nabla f(x_0,\xi_{0,j}) - \nabla F(x_0)\|^p\right]\right)^{\frac{1}{p}} \\ &\overset{\text{Ass. 2.3}}{\leq} \frac{2}{B_{\text{init}}}\left(\sum_{j=0}^{B_{\text{init}}-1}\sigma_1^p\right)^{\frac{1}{p}} \\ &= \frac{2\sigma_1}{B_{\text{init}}^{\frac{p-1}{p}}}, \end{aligned} \tag{77}$$

and since $B_{\text{init}} = \max\left\{1, \left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}\right\}$ we have $\mathbb{E}\left[\|\hat{e}_0\|\right] \leq \sigma_1 \times \left(\frac{\varepsilon}{\sigma_1}\right) = \varepsilon$. This gives

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla F(x_t)\|\right] &\overset{(76)}{=} \mathcal{O}\left(\alpha^{-\frac{1}{2q}}\sqrt{\frac{\bar{L}\Delta}{T}} + \frac{\mathbb{E}\left[\|\hat{e}_0\|\right]}{\alpha T} + \sigma_1\alpha^{\frac{p-1}{p}}\right) \\ &= \mathcal{O}\left(\alpha^{-\frac{1}{2q}}\sqrt{\frac{\bar{L}\Delta}{T}} + \frac{\varepsilon}{\alpha T} + \sigma_1\alpha^{\frac{p-1}{p}}\right) \\ &= \mathcal{O}\left(\left[\sqrt{\frac{\bar{L}\Delta}{T}} + \alpha_{\text{eff}}^{-\frac{1}{2q}}\sqrt{\frac{\bar{L}\Delta}{T}}\right] + \frac{\varepsilon}{T}\alpha_{\text{eff}}^{-1} + \sigma_1\alpha_{\text{eff}}^{\frac{p-1}{p}}\right) \\ &\overset{(a)}{=} \mathcal{O}\left(\sqrt{\frac{\bar{L}\Delta}{T}} + \sigma_1\left(\frac{\bar{L}\Delta}{\sigma_1^2 T}\right)^{\frac{q(p-1)}{p(2q+1)-2q}} + \sigma_1\left(\frac{\varepsilon}{\sigma_1 T}\right)^{\frac{p-1}{2p-1}}\right), \end{aligned} \tag{78}$$

where in (a) we use the choice of α_{eff} from (74) since

$$\frac{\varepsilon}{T}\alpha_{\text{eff}}^{-1} \leq \frac{\varepsilon}{T}\left(\frac{\sigma_1 T}{\varepsilon}\right)^{\frac{p}{2p-1}} = \sigma_1\left(\frac{\varepsilon}{\sigma_1 T}\right)^{\frac{p-1}{2p-1}},$$

and

$$\sigma_1\alpha_{\text{eff}}^{\frac{p-1}{p}} \leq \sigma_1\left(\frac{\varepsilon}{\sigma_1 T}\right)^{\frac{p-1}{2p-1}} + \sigma_1\left(\frac{\bar{L}\Delta}{\sigma_1^2 T}\right)^{\frac{q(p-1)}{p(2q+1)-2q}}.$$

Finally, from the bound (78) we deduce that the sample complexity of Algorithm 1 is exactly

$$\mathcal{O}\left(\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}} + \frac{\bar{L}\Delta}{\varepsilon^2} + \frac{\bar{L}\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{q(p-1)}}\right),$$

as claimed, and it matches our lower bound from Theorem 3.1. □

D.3 Proof of Theorem 4.2

Theorem 4.2. *Under Assumptions 2.1 to 2.3 and 2.7, let the initial gradient estimate g_0 be given by*

$$g_0 = \frac{1}{B_{\text{init}}}\sum_{j=0}^{B_{\text{init}}-1}\nabla f\left(x_0,\xi_{0,j}\right),$$

where $B_{\text{init}} = \max\left\{1, \left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}\right\}$, let the stepsize $\gamma = \min\left\{\sqrt{\frac{\Delta}{L_1 T}}, \sqrt{\frac{\Delta\alpha^{1/q}}{\delta T}}\right\}$, the momentum parameter $\alpha = \min\left\{1, \alpha_{\text{eff}}\right\}$ where

$$\alpha_{\text{eff}} = \max\left\{\left(\frac{\varepsilon}{\sigma_1 T}\right)^{\frac{p}{2p-1}}, \left(\frac{\delta\Delta}{\sigma_1^2 T}\right)^{\frac{pq}{p(2q+1)-2q}}\right\}. \tag{79}$$

Then, Algorithm 1 guarantees to find an ε–stationary point with the total sample complexity

$$\mathcal{O}\left(\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}} + \frac{(L_1+\delta)\Delta}{\varepsilon^2} + \frac{\delta\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{q(p-1)}}\right).$$

Proof. As the function F has L_1–Lipschitz gradients by Assumption 2.2 then applying Lemma D.4 gives

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F(x_t)\| \leq \frac{\Delta}{\gamma T} + \frac{2}{T}\sum_{t=0}^{T-1}\|\hat{e}_t\| + \frac{\gamma L_1}{2},$$

and using Corollary D.1 yields

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\hat{e}_t\|\right] \leq \frac{1}{T}\sum_{t=0}^{T-1}(1-\alpha)^t\mathbb{E}\left[\|\hat{e}_0\|\right] + 2\sigma_1\alpha^{\frac{p-1}{p}} + 2\gamma\delta\alpha^{-\frac{1}{q}} \leq \frac{\mathbb{E}\left[\|\hat{e}_0\|\right]}{\alpha T} + 2\sigma_1\alpha^{\frac{p-1}{p}} + 2\gamma\delta\alpha^{-\frac{1}{q}},$$

hence

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla F(x_t)\|\right] \leq \frac{\Delta}{\gamma T} + \frac{2\mathbb{E}\left[\|\hat{e}_0\|\right]}{\alpha T} + 4\sigma_1\alpha^{\frac{p-1}{p}} + 4\gamma\delta\alpha^{-\frac{1}{q}} + \frac{\gamma L_1}{2}, \tag{80}$$

Now, using our choice of stepsize $\gamma = \min\left\{\sqrt{\frac{\Delta}{L_1 T}}, \sqrt{\frac{\Delta\alpha^{1/q}}{\delta T}}\right\}$ we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla F(x_t)\|\right] = \mathcal{O}\left(\sqrt{\frac{L_1\Delta}{T}} + \alpha^{-\frac{1}{2q}}\sqrt{\frac{\delta\Delta}{T}} + \frac{\mathbb{E}\left[\|\hat{e}_0\|\right]}{\alpha T} + \sigma_1\alpha^{\frac{p-1}{p}}\right). \tag{81}$$

Next, by our choice of g_0 we have, as in we did in the previous proof

$$\mathbb{E}\left[\|\hat{e}_0\|\right] \overset{(77)}{\leq} \frac{2\sigma_1}{B_{\text{init}}^{\frac{p-1}{p}}},$$

and since $B_{\text{init}} = \max\left\{1, \left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}\right\}$ we have $\mathbb{E}\left[\|\hat{e}_0\|\right] \leq \sigma_1 \times \left(\frac{\varepsilon}{\sigma_1}\right) = \varepsilon$. This gives

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla F(x_t)\|\right] &\overset{(81)}{=} \mathcal{O}\left(\sqrt{\frac{L_1\Delta}{T}} + \alpha^{-\frac{1}{2q}}\sqrt{\frac{\delta\Delta}{T}} + \frac{\mathbb{E}\left[\|\hat{e}_0\|\right]}{\alpha T} + \sigma_1\alpha^{\frac{p-1}{p}}\right) \\
&= \mathcal{O}\left(\sqrt{\frac{L_1\Delta}{T}} + \alpha^{-\frac{1}{2q}}\sqrt{\frac{\delta\Delta}{T}} + \frac{\varepsilon}{\alpha T} + \sigma_1\alpha^{\frac{p-1}{p}}\right) \\
&= \mathcal{O}\left(\sqrt{\frac{L_1\Delta}{T}} + \left[\sqrt{\frac{\delta\Delta}{T}} + \alpha_{\text{eff}}^{-\frac{1}{2q}}\sqrt{\frac{\delta\Delta}{T}}\right] + \frac{\varepsilon}{T}\alpha_{\text{eff}}^{-1} + \sigma_1\alpha_{\text{eff}}^{\frac{p-1}{p}}\right) \\
&\overset{(a)}{=} \mathcal{O}\left(\sqrt{\frac{L_1\Delta}{T}} + \sqrt{\frac{\delta\Delta}{T}} + \sigma_1\left(\frac{\delta\Delta}{\sigma_1^2 T}\right)^{\frac{q(p-1)}{p(2q+1)-2q}} + \sigma_1\left(\frac{\varepsilon}{\sigma_1 T}\right)^{\frac{p-1}{2p-1}}\right),
\end{aligned} \tag{82}$$

where in (a) we use the choice of α_{eff} from (79) since

$$\frac{\varepsilon}{T}\alpha_{\text{eff}}^{-1} \leq \frac{\varepsilon}{T}\left(\frac{\sigma_1 T}{\varepsilon}\right)^{\frac{p}{2p-1}} = \sigma_1\left(\frac{\varepsilon}{\sigma_1 T}\right)^{\frac{p-1}{2p-1}},$$

and

$$\sigma_1\alpha_{\text{eff}}^{\frac{p-1}{p}} \leq \sigma_1\left(\frac{\varepsilon}{\sigma_1 T}\right)^{\frac{p-1}{2p-1}} + \sigma_1\left(\frac{\delta\Delta}{\sigma_1^2 T}\right)^{\frac{q(p-1)}{p(2q+1)-2q}}.$$

Finally, from the bound (82) we deduce that the sample complexity of Algorithm 1 is exactly

$$\mathcal{O}\left(\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}} + \frac{(L_1+\delta)\Delta}{\varepsilon^2} + \frac{\delta\Delta}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{q(p-1)}}\right),$$

as claimed, and, combining Algorithm 1 with NSGD-Mom is enough to match our lower bound from Theorem 3.2. □

D.4 Proof of Theorem 4.3

Theorem 4.3. *Under Assumptions 2.1, 2.3 and 2.6, let the initial gradient estimate $g_0 = \nabla f(x_0, \xi_0)$, let the stepsize $\gamma = \sqrt{\frac{\Delta\alpha}{\bar{L}T}}$, the momentum parameter $\alpha = T^{-\frac{1}{2}} \in (0, 1]$. Then, Algorithm 1 guarantees the bound*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla F(x_t)\|\right] = \mathcal{O}\left(\frac{\sigma_1}{T^{\frac{p-1}{2p}}} + \frac{\sqrt{\bar{L}\Delta}}{T^{1/4}}\right).$$

Proof. Notice that the bound (75) holds for any choice of the parameters $\gamma > 0$, $\alpha \in (0, 1]$ and $g_0 \in \mathbb{R}^d$. Hence, for our particular choice $g_0 = \nabla f(x_0, \xi_0)$, $\gamma = \sqrt{\frac{\Delta\alpha}{\bar{L}T}}$ and $\alpha = T^{-\frac{1}{2}}$ we obtain

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla F(x_t)\|\right] &\overset{(75)}{=} \mathcal{O}\left(\frac{\Delta}{\gamma T} + \frac{\mathbb{E}\left[\|\hat{e}_0\|\right]}{\alpha T} + \sigma_1\alpha^{\frac{p-1}{p}} + \gamma\bar{L}\alpha^{-\frac{1}{q}}\right)\\
&= \mathcal{O}\left(\alpha^{-\frac{1}{2}}\sqrt{\frac{\bar{L}\Delta}{T}} + \frac{\mathbb{E}\left[\|\hat{e}_0\|\right]}{T^{1/2}} + \frac{\sigma_1}{T^{\frac{p-1}{2p}}} + \alpha^{-\frac{2-q}{2q}}\sqrt{\frac{\bar{L}\Delta}{T}}\right)\\
&\overset{(a)}{=} \mathcal{O}\left(\frac{\sqrt{\bar{L}\Delta}}{T^{1/4}} + \frac{\mathbb{E}\left[\|\hat{e}_0\|\right]}{T^{1/2}} + \frac{\sigma_1}{T^{\frac{p-1}{2p}}}\right)\\
&\overset{(b)}{=} \mathcal{O}\left(\frac{\sigma_1}{T^{\frac{p-1}{2p}}} + \frac{\sigma_1}{T^{1/2}} + \frac{\sqrt{\bar{L}\Delta}}{T^{1/4}}\right)\\
&= \mathcal{O}\left(\frac{\sigma_1}{T^{\frac{p-1}{2p}}} + \frac{\sqrt{\bar{L}\Delta}}{T^{1/4}}\right),
\end{aligned}$$

where in (a) we use $\alpha^{-\frac{1}{2}} \geq \alpha^{-\frac{2-q}{2q}}$ since $\alpha \in (0, 1]$ and $\frac{2-q}{2q} \leq \frac{1}{2}$. In (b) we use the definition of $\hat{e}_0$ and g_0, i.e.,

$$\mathbb{E}\left[\|\hat{e}_0\|\right] = \mathbb{E}\left[\|\nabla f(x_0, \xi_0) - \nabla F(x_0)\|\right] \overset{\text{Lem. H.9}}{\leq} \left(\mathbb{E}\left[\|\nabla f(x_0, \xi_0) - \nabla F(x_0)\|^p\right]\right)^{\frac{1}{p}} \overset{\text{Ass. 2.3}}{\leq} \sigma_1.$$

This achieves the proof of the theorem. □

D.5 Proof of Theorem 4.4

Theorem 4.4. *Under Assumptions 2.1 to 2.3 and 2.7, let the initial gradient estimate $g_0 = \nabla f(x_0, \xi_0)$, let the stepsize $\gamma = \min\left\{\sqrt{\frac{\Delta}{L_1 T}}, \sqrt{\frac{\Delta\alpha}{\delta T}}\right\}$, the momentum parameter $\alpha = T^{-\frac{1}{2}} \in (0, 1]$. Then, Algorithm 1 guarantees guarantees the bound*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla F(x_t)\|\right] = \mathcal{O}\left(\frac{\sigma_1}{T^{\frac{p-1}{2p}}} + \frac{\sqrt{\delta\Delta}}{T^{1/4}} + \frac{\sqrt{L_1\Delta}}{T^{1/2}}\right). \tag{83}$$

Proof. As in the previous theorem, note that the bound (80) holds for any choice of the parameters $\gamma > 0$, $\alpha \in (0, 1]$ and $g_0 \in \mathbb{R}^d$. Hence, for our particular choice $g_0 = \nabla f(x_0, \xi_0)$, $\gamma = \min\left\{\sqrt{\frac{\Delta}{L_1 T}}, \sqrt{\frac{\Delta\alpha}{\delta T}}\right\}$ and $\alpha = T^{-\frac{1}{2}}$

we obtain, as in the previous proof

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla F(x_t)\|\right] &\overset{(80)}{=} \mathcal{O}\left(\frac{\Delta}{\gamma T}+\frac{\mathbb{E}\left[\|\hat{e}_0\|\right]}{\alpha T}+\sigma_1\alpha^{\frac{p-1}{p}}+\gamma\delta\alpha^{-\frac{1}{q}}+\gamma L_1\right)\\
&= \mathcal{O}\left(\sqrt{\frac{L_1\Delta}{T}}+\alpha^{-\frac{1}{2}}\sqrt{\frac{\delta\Delta}{T}}+\frac{\mathbb{E}\left[\|\hat{e}_0\|\right]}{T^{1/2}}+\frac{\sigma_1}{T^{\frac{p-1}{2p}}}+\alpha^{-\frac{2-q}{2q}}\sqrt{\frac{\delta\Delta}{T}}\right)\\
&\overset{(a)}{=} \mathcal{O}\left(\sqrt{\frac{L_1\Delta}{T}}+\frac{\sqrt{\delta\Delta}}{T^{1/4}}+\frac{\mathbb{E}\left[\|\hat{e}_0\|\right]}{T^{1/2}}+\frac{\sigma_1}{T^{\frac{p-1}{2p}}}\right)\\
&\overset{(b)}{=} \mathcal{O}\left(\frac{\sigma_1}{T^{\frac{p-1}{2p}}}+\frac{\sigma_1}{T^{1/2}}+\frac{\sqrt{\delta\Delta}}{T^{1/4}}+\sqrt{\frac{L_1\Delta}{T}}\right)\\
&= \mathcal{O}\left(\frac{\sigma_1}{T^{\frac{p-1}{2p}}}+\frac{\sqrt{\delta\Delta}}{T^{1/4}}+\sqrt{\frac{L_1\Delta}{T}}\right),
\end{aligned}$$

where in (a) we use $\alpha^{-\frac{1}{2}} \geq \alpha^{-\frac{2-q}{2q}}$ since $\alpha \in (0,1]$ and $\frac{2-q}{2q} \leq \frac{1}{2}$. In (b) we use, as before, the bound (83). This achieves the proof of the theorem. □

E MISSING PROOFS IN SECTION 5

E.1 Auxiliary Lemmas

Some of the auxiliary lemmas needed in this section (Lemmas D.2 and D.4) have already been established in the previous section.

Lemma E.1 (Bounding the Error Term). *Under Assumptions 2.3 to 2.5, for all $t \in \{0, \dots, T-1\}$ we have*

$$\mathbb{E}\left[\|\hat{e}_t\|\right] \leq (1-\alpha)^t \mathbb{E}\left[\|\hat{e}_0\|\right] + 2\sigma_1 \alpha^{\frac{p-1}{p}} + 2\gamma\sigma_2 \alpha^{-\frac{1}{q}} + 4\gamma \min\left\{L_1, \gamma L_2\right\} \alpha^{-\frac{1}{2}},$$

where $\hat{e}_t := g_t - \nabla F(x_t)$.

Proof. By the update rule of the gradient estimator in Algorithm 2 (line 9) we have, for all $t \in \{1, \dots, T-1\}$

$$\begin{aligned}
\hat{e}_t &:= g_t - \nabla F(x_t) \\
&= (1-\alpha)(g_{t-1} + \nabla^2 f(\hat{x}_t, \hat{\xi}_t)(x_t - x_{t-1}) + \alpha \nabla f(x_t, \xi_t) - \nabla F(x_t) \\
&= (1-\alpha)(g_{t-1} - \nabla F(x_{t-1})) + \alpha \left(\nabla f(x_t, \xi_t) - \nabla F(x_t)\right) \\
&\qquad + (1-\alpha)\left(\nabla F(x_{t-1}) - \nabla F(x_t) - \nabla^2 f(\hat{x}_t, \hat{\xi}_t)(x_{t-1} - x_t)\right) \\
&= (1-\alpha)(g_{t-1} - \nabla F(x_{t-1})) + \alpha \left(\nabla f(x_t, \xi_t) - \nabla F(x_t)\right) \\
&\qquad + (1-\alpha)\left(\nabla F(x_{t-1}) - \nabla F(x_t) - \nabla^2 F(\hat{x}_t)(x_{t-1} - x_t)\right) \\
&\qquad + (1-\alpha)\left(\nabla^2 F(\hat{x}_t)(x_{t-1} - x_t) - \nabla^2 f(\hat{x}_t, \hat{\xi}_t)(x_{t-1} - x_t)\right) \\
&= (1-\alpha)\hat{e}_{t-1} + \alpha e_t + (1-\alpha)\hat{R}_t + (1-\alpha)\hat{S}_t, \qquad (84)
\end{aligned}$$

where we let $e_t := \nabla f(x_t, \xi_t) - \nabla F(x_t)$ and

$$\hat{R}_t := \nabla F(x_{t-1}) - \nabla F(x_t) - \nabla^2 F(\hat{x}_t)(x_{t-1} - x_t), \qquad (85)$$

$$\hat{S}_t := \nabla^2 F(\hat{x}_t)(x_{t-1} - x_t) - \nabla^2 f(\hat{x}_t, \hat{\xi}_t)(x_{t-1} - x_t). \qquad (86)$$

Notably, it is worth mentioning that $(e_t)_{t\geq 0}$, $(\hat{R}_t)_{t\geq 0}$ and $(\hat{S}_t)_{t\geq 0}$ are all martingale difference sequence with respect to the filtration $(\mathcal{F}_t)_{t\geq 0}$ where $\mathcal{F}_t := \sigma(g_0, (\xi_1, \hat{\xi}_1, q_1), \dots, (\xi_t, \hat{\xi}_t, q_t))$. Effectively, for $\hat{R}_t$ we have

$$\mathbb{E}_{q_t}\left[F(\hat{x}_t)(x_t - x_{t-1})\right] = \int_0^1 \nabla^2 F(sx_t + (1-s)x_{t-1})(x_t - x_{t-1})\,\mathrm{d}s = \nabla F(x_t) - \nabla F(x_{t-1}). \qquad (87)$$

Then, unrolling the recursion (84) gives

$$\hat{e}_t = (1-\alpha)^t \hat{e}_0 + \alpha \sum_{j=0}^{t-1} (1-\alpha)^{t-j-1} e_{j+1} + \sum_{j=0}^{t-1} (1-\alpha)^{t-j} \hat{R}_{j+1} + \sum_{j=0}^{t-1} (1-\alpha)^{t-j} \hat{S}_{j+1},$$

and taking the norm followed by the total expectation yields

$$\begin{aligned}
\mathbb{E}\left[\|\hat{e}_t\|\right] \leq (1-\alpha)^t \mathbb{E}\left[\|\hat{e}_0\|\right] + \alpha \mathbb{E}\left[\left\|\sum_{j=0}^{t-1} (1-\alpha)^{t-j-1} e_{j+1}\right\|\right] &+ (1-\alpha)\mathbb{E}\left[\left\|\sum_{j=0}^{t-1} (1-\alpha)^{t-j-1} \hat{R}_{j+1}\right\|\right] \\
&+ (1-\alpha)\mathbb{E}\left[\left\|\sum_{j=0}^{t-1} (1-\alpha)^{t-j-1} \hat{S}_{j+1}\right\|\right].
\end{aligned}$$

We now need to upper bound the last three terms of the previous inequality. For the first term, using Jensen's inequality (Lemma H.9) as we did before in Lemma D.5 (see (70)) we have

$$\mathbb{E}\left[\left\|\sum_{j=0}^{t-1} (1-\alpha)^{t-j-1} e_{j+1}\right\|\right] \overset{(70)}{\leq} 2\sigma_1 \alpha^{-\frac{1}{p}}. \qquad (88)$$

Then, for the second term we have

$$\mathbb{E}\left[\left\|\sum_{j=0}^{t-1}(1-\alpha)^{t-j-1}\hat{R}_{j+1}\right\|\right] \overset{\text{Lem. H.9}}{\leq} \left(\mathbb{E}\left[\left\|\sum_{j=0}^{t-1}(1-\alpha)^{t-j-1}\hat{R}_{j+1}\right\|^2\right]\right)^{\frac{1}{2}} \overset{\text{Lem. G.1}}{\leq} \left(2\sum_{j=0}^{t-1}(1-\alpha)^{2(t-j-1)}\mathbb{E}\left[\left\|\hat{R}_{j+1}\right\|^2\right]\right)^{\frac{1}{2}}, \tag{89}$$

and, the variance of $\nabla^2 F(\hat{x}_t)(x_t - x_{t-1})$ can be bounded in two different ways. First, by using Assumption 2.2 and Jensen's inequality we have

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{R}_t\right\|^2\right] &\overset{(85)}{=} \mathbb{E}\left[\left\|\nabla F(x_{t-1}) - \nabla F(x_t) - \nabla^2 F(\hat{x}_t)(x_{t-1} - x_t)\right\|^2\right] \\
&\overset{\text{Lem. H.10}}{\leq} 2\mathbb{E}\left[\left\|\nabla F(x_t) - \nabla F(x_{t-1})\right\|^2\right] + 2\mathbb{E}\left[\left\|\nabla^2 F(\hat{x}_t)(x_t - x_{t-1})\right\|^2\right] \\
&\overset{\text{Ass. 2.2}}{\leq} 2L_1^2\mathbb{E}\left[\left\|x_t - x_{t-1}\right\|^2\right] + 2\mathbb{E}\left[\left\|\nabla^2 F(\hat{x}_t)\right\|_{\text{op}}^2 \cdot \left\|x_t - x_{t-1}\right\|^2\right] \\
&\overset{\text{Lem. H.14}}{\leq} 4L_1^2\mathbb{E}\left[\left\|x_t - x_{t-1}\right\|^2\right] \\
&= 4\gamma^2 L_1^2,
\end{aligned} \tag{90}$$

and, using the connection between $\nabla^2 F(\hat{x}_t)(x_t - x_{t-1})$ and $\nabla F(x_t) - \nabla F(x_{t-1})$ as displayed in (87) we also have

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{R}_t\right\|^2\right] &\overset{(85)}{=} \mathbb{E}\left[\left\|\nabla F(x_{t-1}) - \nabla F(x_t) - \nabla^2 F(\hat{x}_t)(x_{t-1} - x_t)\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\int_0^1 \left(\nabla^2 F(x_{t-1} + s(x_t - x_{t-1})) - \nabla^2 F(\hat{x}_t)\right)(x_t - x_{t-1})\,\mathrm{d}s\right\|^2\right] \\
&\overset{(a)}{\leq} \mathbb{E}\left[\left(\int_0^1 \left\|\left(\nabla^2 F(x_{t-1} + s(x_t - x_{t-1})) - \nabla^2 F(\hat{x}_t)\right)(x_t - x_{t-1})\right\|\mathrm{d}s\right)^2\right] \\
&\leq \mathbb{E}\left[\left(\int_0^1 \left\|\nabla^2 F(x_{t-1} + s(x_t - x_{t-1})) - \nabla^2 F(\hat{x}_t)\right\|_{\text{op}} \cdot \left\|x_t - x_{t-1}\right\|\mathrm{d}s\right)^2\right] \\
&= \gamma^2\,\mathbb{E}\left[\left(\int_0^1 \left\|\nabla^2 F(x_{t-1} + s(x_t - x_{t-1})) - \nabla^2 F(\hat{x}_t)\right\|_{\text{op}}\mathrm{d}s\right)^2\right] \\
&\overset{\text{Ass. 2.4}}{\leq} \gamma^2 L_2^2\,\mathbb{E}\left[\left(\int_0^1 \left\|s(x_t - \hat{x}_t) + (1-s)(x_{t-1} - \hat{x}_t)\right\|\mathrm{d}s\right)^2\right] \\
&\overset{(b)}{\leq} \gamma^2 L_2^2\,\mathbb{E}\left[\left(\int_0^1 s\left\|x_t - \hat{x}_t\right\|\mathrm{d}s + \int_0^1 (1-s)\left\|x_{t-1} - \hat{x}_t\right\|\mathrm{d}s\right)^2\right] \\
&= \frac{\gamma^2 L_2^2}{4}\,\mathbb{E}\left[\left(\left\|x_t - \hat{x}_t\right\| + \left\|x_{t-1} - \hat{x}_t\right\|\right)^2\right] \\
&\overset{(c)}{=} \frac{\gamma^2 L_2^2}{4}\,\mathbb{E}\left[\left((1-q_t)\left\|x_t - x_{t-1}\right\| + q_t\left\|x_t - x_{t-1}\right\|\right)^2\right] \\
&= \frac{\gamma^4 L_2^2}{4},
\end{aligned} \tag{91}$$

where in (a) and (b) we use the triangle inequality. In (c) we use the definition of $\hat{x}_t$, i.e., $\hat{x}_t := q_t x_t + (1-q_t)x_{t-1}$. It is worth mentioning that the bounds (90) and (91) holds without the expectation $\mathbb{E}[\cdot]$.

Thus, using (94), (91) we obtain

$$\mathbb{E}\left[\left\|\sum_{j=0}^{t-1}(1-\alpha)^{t-j-1}\hat{R}_{j+1}\right\|\right] \overset{(93)}{\leq} 4\gamma\min\{L_1,\gamma L_2\}\left(\sum_{j=0}^{t-1}(1-\alpha)^{2(t-j-1)}\right)^{\frac{1}{2}} \leq 4\gamma\min\{L_1,\gamma L_2\}\,\alpha^{-\frac{1}{2}}. \tag{92}$$

Finally, for the last term, we can write

$$\mathbb{E}\left[\left\|\sum_{j=0}^{t-1}(1-\alpha)^{t-j-1}\hat{S}_{j+1}\right\|\right] \overset{\text{Lem. H.9}}{\leq} \left(\mathbb{E}\left[\left\|\sum_{j=0}^{t-1}(1-\alpha)^{t-j-1}\hat{S}_{j+1}\right\|^q\right]\right)^{\frac{1}{q}} \overset{\text{Lem. G.1}}{\leq} \left(2\sum_{j=0}^{t-1}(1-\alpha)^{q(t-j-1)}\mathbb{E}\left[\left\|\hat{S}_{j+1}\right\|^q\right]\right)^{\frac{1}{q}}, \tag{93}$$

and, using Assumption 2.5 we obtain

$$\begin{aligned}\mathbb{E}\left[\left\|\hat{S}_{j+1}\right\|^q\right] &\overset{(86)}{=} \mathbb{E}\left[\left\|\nabla^2 F(\hat{x}_t)(x_{t-1}-x_t)-\nabla^2 f(\hat{x}_t,\hat{\xi}_t)(x_{t-1}-x_t)\right\|^q\right]\\ &\leq \mathbb{E}\left[\left\|\nabla^2 F(\hat{x}_t)-\nabla^2 f(\hat{x}_t,\hat{\xi}_t)\right\|_{\text{op}}^q\cdot\|x_t-x_{t-1}\|^q\right]\\ &\overset{\text{Ass. 2.5}}{\leq} \gamma^q\sigma_2^q,\end{aligned} \tag{94}$$

thus, using (94) and the fact that $2^{\frac{1}{q}}\leq 2$ we obtain

$$\mathbb{E}\left[\left\|\sum_{j=0}^{t-1}(1-\alpha)^{t-j-1}\hat{S}_{j+1}\right\|\right] \overset{(93)}{\leq} 2\gamma\sigma_2\left(\sum_{j=0}^{t-1}(1-\alpha)^{q(t-j-1)}\right)^{\frac{1}{q}} \leq 2\gamma\sigma_2\alpha^{-\frac{1}{q}}. \tag{95}$$

Then, combining the bounds (88), (92) and (95) we have

$$\mathbb{E}\left[\|\hat{e}_t\|\right] \leq (1-\alpha)^t\mathbb{E}\left[\|\hat{e}_0\|\right] + 2\sigma_1\alpha^{\frac{p-1}{p}} + 2\gamma\sigma_2\alpha^{-\frac{1}{q}} + 4\gamma\min\{L_1,\gamma L_2\}\,\alpha^{-\frac{1}{2}},$$

which achieves the proof of the lemma. □

E.2 Proof of Theorem 5.1

Thanks to Lemmas D.4 and E.1 we can now establish the convergence analysis (in expectation) of Algorithm 2.

Theorem 5.1. *Under Assumptions 2.1 to 2.5, let the initial gradient estimate g_0 be given by*

$$g_0 = \frac{1}{B_{\text{init}}}\sum_{j=0}^{B_{\text{init}}-1}\nabla f\left(x_0,\xi_{0,j}\right),$$

where $B_{\text{init}} = \max\left\{1,\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}\right\}$, let the stepsize $\gamma = \min\left\{\sqrt{\frac{\Delta}{L_1 T}},\sqrt{\frac{\Delta\alpha^{1/q}}{\sigma_2 T}},\sqrt[3]{\frac{\Delta\alpha^{1/2}}{L_2 T}}\right\}$, the momentum parameter $\alpha = \min\{1,\alpha_{\text{eff}}\}$ where

$$\alpha_{\text{eff}} = \max\left\{\left(\frac{\varepsilon}{\sigma_1 T}\right)^{\frac{p}{2p-1}},\left(\frac{\Delta\sigma_2}{\sigma_1^2 T}\right)^{\frac{pq}{p(2q+1)-2q}},\left(\frac{L_2^{1/2}\Delta}{\sigma_1^{3/2}T}\right)^{\frac{4p}{7p-6}}\right\}. \tag{96}$$

Then, Algorithm 2 guarantees to find an ε–stationary point with the total sample complexity

$$\mathcal{O}\left(\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}} + \frac{\Delta}{\varepsilon}\left(\frac{L_1+\sigma_2}{\varepsilon}+\sqrt{\frac{L_2}{\varepsilon}}\right) + \frac{\Delta\sigma_2}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{q(p-1)}} + \frac{L_2^{1/2}\Delta\sigma_1^{1/4}}{\varepsilon^{7/4}}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{4(p-1)}}\right).$$

Proof. According to Lemma D.4 we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F(x_t)\| \leq \frac{\Delta}{\gamma T} + \frac{2}{T}\sum_{t=0}^{T-1}\|\hat{e}_t\| + \frac{\gamma L_1}{2},$$

and using Lemma E.1 this yields

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\hat{e}_t\|\right] &\leq \frac{1}{T}\sum_{t=0}^{T-1}(1-\alpha)^t\mathbb{E}\left[\|\hat{e}_0\|\right] + 2\sigma_1\alpha^{\frac{p-1}{p}} + 2\gamma\sigma_2\alpha^{-\frac{1}{q}} + 4\gamma^2 L_2\alpha^{-\frac{1}{2}} \\ &\leq \frac{\mathbb{E}\left[\|\hat{e}_0\|\right]}{\alpha T} + 2\sigma_1\alpha^{\frac{p-1}{p}} + 2\gamma\sigma_2\alpha^{-\frac{1}{q}} + 4\gamma^2 L_2\alpha^{-\frac{1}{2}},\end{aligned}$$

where in the last term we drop the $\min\{\cdots\}$ to only keep the $\gamma^2 L_2$ term. Hence we obtain the bound

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla F(x_t)\|\right] \leq \frac{\Delta}{\gamma T} + \frac{2\mathbb{E}\left[\|\hat{e}_0\|\right]}{\alpha T} + 4\sigma_1\alpha^{\frac{p-1}{p}} + 4\gamma\sigma_2\alpha^{-\frac{1}{q}} + 8\gamma^2 L_2\alpha^{-\frac{1}{2}} + \frac{\gamma L_1}{2}.$$

Now, using $\gamma = \min\left\{\sqrt{\frac{\Delta}{L_1 T}}, \sqrt{\frac{\Delta\alpha^{1/q}}{\sigma_2 T}}, \sqrt[3]{\frac{\Delta\alpha^{1/2}}{L_2 T}}\right\}$ we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla F(x_t)\|\right] = \mathcal{O}\left(\sqrt{\frac{L_1\Delta}{T}} + \alpha^{-\frac{1}{2q}}\sqrt{\frac{\Delta\sigma_2}{T}} + \alpha^{-\frac{1}{6}}\left(\frac{L_2^{1/2}\Delta}{T}\right)^{\frac{2}{3}} + \frac{\mathbb{E}\left[\|\hat{e}_0\|\right]}{\alpha T} + \sigma_1\alpha^{\frac{p-1}{p}}\right). \tag{97}$$

Now, by our choice of g_0 we have, as in we did in the previous section

$$\mathbb{E}\left[\|\hat{e}_0\|\right] \overset{(77)}{\leq} \frac{2\sigma_1}{B_{\text{init}}^{\frac{p-1}{p}}}, \tag{98}$$

and since $B_{\text{init}} = \max\left\{1, \left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}}\right\}$ we have $\mathbb{E}\left[\|\hat{e}_0\|\right] \leq \sigma_1 \times \left(\frac{\varepsilon}{\sigma_1}\right) = \varepsilon$. This gives

$$\begin{aligned}&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla F(x_t)\|\right] \\ &\overset{(97)}{=} \mathcal{O}\left(\sqrt{\frac{L_1\Delta}{T}} + \alpha^{-\frac{1}{2q}}\sqrt{\frac{\Delta\sigma_2}{T}} + \alpha^{-\frac{1}{6}}\left(\frac{L_2^{1/2}\Delta}{T}\right)^{\frac{2}{3}} + \frac{\mathbb{E}\left[\|\hat{e}_0\|\right]}{\alpha T} + \sigma_1\alpha^{\frac{p-1}{p}}\right) \\ &\overset{(98)}{=} \mathcal{O}\left(\sqrt{\frac{L_1\Delta}{T}} + \alpha^{-\frac{1}{2q}}\sqrt{\frac{\Delta\sigma_2}{T}} + \alpha^{-\frac{1}{6}}\left(\frac{L_2^{1/2}\Delta}{T}\right)^{\frac{2}{3}} + \frac{\varepsilon}{\alpha T} + \sigma_1\alpha^{\frac{p-1}{p}}\right) \\ &= \mathcal{O}\left(\sqrt{\frac{L_1\Delta}{T}} + \left[\sqrt{\frac{\Delta\sigma_2}{T}} + \alpha_{\text{eff}}^{-\frac{1}{2q}}\sqrt{\frac{\Delta\sigma_2}{T}}\right] + \frac{\varepsilon}{T}\alpha_{\text{eff}}^{-1} + \sigma_1\alpha_{\text{eff}}^{\frac{p-1}{p}} + \left[\left(\frac{L_2^{1/2}\Delta}{T}\right)^{\frac{2}{3}} + \alpha_{\text{eff}}^{-\frac{1}{6}}\left(\frac{L_2^{1/2}\Delta}{T}\right)^{\frac{2}{3}}\right]\right) \\ &\overset{(a)}{=} \mathcal{O}\left(\sqrt{\frac{L_1\Delta}{T}} + \sqrt{\frac{\Delta\sigma_2}{T}} + \left(\frac{L_2^{1/2}\Delta}{T}\right)^{\frac{2}{3}} + \sigma_1\left(\frac{\Delta\sigma_2}{\sigma_1^2 T}\right)^{\frac{q(p-1)}{p(2q+1)-2q}} + \sigma_1\left(\frac{\varepsilon}{\sigma_1 T}\right)^{\frac{p-1}{2p-1}} + \sigma_1\left(\frac{L_2^{1/2}\Delta}{\sigma_1^{3/2}T}\right)^{\frac{4(p-1)}{7p-6}}\right),\end{aligned} \tag{99}$$

where in (a) we use the choice of α_{eff} from (96), especially, we have

$$\sigma_1\alpha_{\text{eff}}^{\frac{p-1}{p}} \leq \sigma_1\left(\frac{\Delta\sigma_2}{\sigma_1^2 T}\right)^{\frac{q(p-1)}{p(2q+1)-2q}} + \sigma_1\left(\frac{\varepsilon}{\sigma_1 T}\right)^{\frac{p-1}{2p-1}} + \sigma_1\left(\frac{L_2^{1/2}\Delta}{\sigma_1^{3/2}T}\right)^{\frac{4(p-1)}{7p-6}}.$$

Finally, from the bound (99) we deduce that the sample complexity of Algorithm 2 is exactly

$$\mathcal{O}\left(\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{p-1}} + \frac{\Delta}{\varepsilon}\left(\frac{L_1+\sigma_2}{\varepsilon} + \sqrt{\frac{L_2}{\varepsilon}}\right) + \frac{\Delta\sigma_2}{\varepsilon^2}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{q(p-1)}} + \frac{L_2^{1/2}\Delta\sigma_1^{1/4}}{\varepsilon^{7/4}}\left(\frac{\sigma_1}{\varepsilon}\right)^{\frac{p}{4(p-1)}}\right), \tag{100}$$

as claimed. □

Remark E.1. In particular, when $\sigma_2 = 0$ we observe that the sample complexity (100) does not depend on the exponent q anymore, as in the lower bound we derived in Theorem 3.3.

F MISSING PROOFS IN SECTION 6

F.1 Preliminary Lemmas

Now we start the high-probability convergence analysis of Algorithm 3. The proof heavily follows the one of Sadiev et al. (2025).

F.1.1 Some Descent Lemma

Lemma F.1 (A Descent Lemma (for Algorithm 1)). *Under Assumptions 2.1 and 2.6, for any choice of stepsize $\gamma > 0$ and any choice of momentum parameter $\alpha \in (0, 1]$, Algorithm 3 generates iterates $\{x_t\}_{t\in\{0,\ldots,T\}}$ which satisfy almost surely (a.s.) the inequality*

$$\gamma \sum_{t=0}^{T-1} \|\nabla F(x_t)\| + \Delta_T \leq \Delta_0 + 2\gamma\alpha \sum_{t=1}^{T} \left\| \sum_{j=1}^{t} (1-\alpha)^{t-j} \theta_j \right\| + 2\gamma(1-\alpha) \sum_{t=1}^{T} \left\| \sum_{j=1}^{t} (1-\alpha)^{t-j} \omega_j \right\| + \frac{2\gamma \|\hat{e}_0\|}{\alpha} + \frac{\gamma^2 \bar{L} T}{2},$$

where for any $j \in \{1, \ldots, T-1\}$, the vectors θ_j and ω_j are defined as

$$\theta_j := \texttt{clip}\left(\nabla f(x_j, \xi_j), \lambda_2\right) - \nabla F(x_j), \tag{101}$$

$$\omega_j := \texttt{clip}\left(\nabla f(x_j, \xi_j) - \nabla f(x_{j-1}, \xi_j), \lambda_1\right) - \left(\nabla F(x_j) - \nabla F(x_{j-1})\right), \tag{102}$$

and for any $t \in \{0, \ldots, T-1\}$ we let $\Delta_t := F(x_t) - F^{\inf}$.

Proof. First of all, let us observe that Lemma D.1 still holds for Algorithm 3 (and can be proved analogously since Algorithms 1 and 3 have the same gradient update rule). Moreover, as the momentum term does not play any role in the proof of Lemma D.1, the iterates $\{x_t\}_{t\in\{0,\ldots,T\}}$ of Algorithm 3 satisfy almost surely (a.s.)

$$\gamma \sum_{t=0}^{T-1} \|\nabla F(x_t)\| + \Delta_T \leq \Delta_0 + 2\gamma \sum_{t=0}^{T-1} \|\hat{e}_t\| + \frac{\gamma^2 \bar{L} T}{2}. \tag{103}$$

Next, we bound $\|\hat{e}_t\|$ in an analogous way as we did in Lemma D.5. According to the update rule for the momentum term in line 10 of Algorithm 3, we have

$$\begin{aligned}
\hat{e}_t &= g_t - \nabla F(x_t) \\
&= (1-\alpha)\left(g_{t-1} + \texttt{clip}(\nabla f\left(x_t, \xi_t\right) - \nabla f\left(x_{t-1}, \xi_t\right), \lambda_1)\right) + \alpha\, \texttt{clip}(\nabla f\left(x_t, \xi_t\right), \lambda_2) - \nabla F(x_t) \\
&= (1-\alpha)\left(g_{t-1} - \nabla F(x_{t-1})\right) + (1-\alpha)\left(\texttt{clip}(\nabla f\left(x_t, \xi_t\right) - \nabla f\left(x_{t-1}, \xi_t\right), \lambda_1) - \left[\nabla F(x_t) - \nabla F(x_{t-1})\right]\right) \\
&\qquad + \alpha\left(\texttt{clip}(\nabla f\left(x_t, \xi_t\right), \lambda_2) - \nabla F(x_t)\right) \\
&\overset{(101)+(102)}{=} (1-\alpha)\hat{e}_{t-1} + \alpha\theta_t + (1-\alpha)\omega_t \\
&\overset{(a)}{=} (1-\alpha)^t \hat{e}_0 + \alpha \sum_{j=1}^{t} (1-\alpha)^{t-j}\theta_j + (1-\alpha)\sum_{j=1}^{t}(1-\alpha)^{t-j}\omega_j,
\end{aligned} \tag{104}$$

where in (a) we unroll the recursion. Then, taking the norm and applying the triangle inequality in this series of equalities, we obtain

$$\|\hat{e}_t\| \leq (1-\alpha)^t \|\hat{e}_0\| + \alpha \left\| \sum_{j=1}^{t} (1-\alpha)^{t-j}\theta_j \right\| + (1-\alpha)\left\| \sum_{j=1}^{t}(1-\alpha)^{t-j}\omega_j \right\|, \tag{105}$$

and plugging (105) in (103) leads to

$$\begin{aligned}\gamma\sum_{t=0}^{T-1}\|\nabla F(x_t)\|+\Delta_T &\leq \Delta_0+2\gamma\sum_{t=0}^{T-1}\|\hat{e}_t\|+\frac{\gamma^2\bar{L}T}{2}\\ &\leq \Delta_0+2\gamma\sum_{t=0}^{T-1}\left((1-\alpha)^t\|\hat{e}_0\|+\alpha\left\|\sum_{j=1}^{t}(1-\alpha)^{t-j}\theta_j\right\|+(1-\alpha)\left\|\sum_{j=1}^{t}(1-\alpha)^{t-j}\omega_j\right\|\right)\\ &\quad+\frac{\gamma^2\bar{L}T}{2}\\ &\overset{(a)}{\leq} \Delta_0+\frac{2\gamma\|\hat{e}_0\|}{\alpha}+\frac{\gamma^2\bar{L}T}{2}\\ &\quad+2\gamma\alpha\sum_{t=0}^{T-1}\left\|\sum_{j=1}^{t}(1-\alpha)^{t-j}\theta_j\right\|+2\gamma(1-\alpha)\sum_{t=0}^{T-1}\left\|\sum_{j=1}^{t}(1-\alpha)^{t-j}\omega_j\right\|,\end{aligned}$$

where in (a) we use the inequality

$$2\gamma\sum_{t=0}^{T-1}(1-\alpha)^t\|\hat{e}_0\|\leq 2\gamma\|\hat{e}_0\|\sum_{t\geq 0}(1-\alpha)^t=\frac{2\gamma\|\hat{e}_0\|}{\alpha}.$$

This proves the desired result. □

Remark F.1. In particular, if we assume $g_0=0$ in Algorithm 3 then $x_0=x_1$ and $\Delta_0=\Delta_1$. Moreover, we have

$$\|\hat{e}_0\|=\|g_0-\nabla F(x_0)\|=\|\nabla F(x_1)\|\leq\sqrt{2\bar{L}\Delta_1},$$

as by Assumption 2.6 we know that F has $\bar{L}$–Lipschitz continuous gradients.

Corollary F.1 (Another Descent Lemma)**.** *Under Assumptions 2.1 and 2.2, for any choice of stepsize $\gamma>0$ and any choice of momentum parameter $\alpha\in(0,1]$, Algorithm 3 generates iterates $\{x_t\}_{t\in\{0,\ldots,T\}}$ which satisfy almost surely (a.s.) the inequality*

$$\begin{aligned}\gamma\sum_{t=0}^{T-1}\|\nabla F(x_t)\|+\Delta_T\leq\Delta_0+2\gamma\alpha\sum_{t=1}^{T}\left\|\sum_{j=1}^{t}(1-\alpha)^{t-j}\theta_j\right\|+2\gamma(1-\alpha)\sum_{t=1}^{T}\left\|\sum_{j=1}^{t}(1-\alpha)^{t-j}\omega_j\right\|\\ +\frac{2\gamma\|\hat{e}_0\|}{\alpha}+\frac{\gamma^2L_1T}{2},\end{aligned}$$

where for any $j\in\{1,\ldots,T-1\}$, the vectors θ_j and ω_j are defined in (101) and (102) and for any $t\in\{0,\ldots,T-1\}$ we let $\Delta_t:=F(x_t)-F^{\inf}$.

Proof. The proof follows the exact same steps as in the previous lemma (Lemma F.1), with the exception that the function F has L_1–Lipschitz continuous gradients (instead of $\bar{L}$). □

Lemma F.2 (A Descent Lemma (for Algorithm 2))**.** *Under Assumptions 2.1 and 2.2, for any choice of stepsize $\gamma>0$ and any choice of momentum parameter $\alpha\in(0,1]$, Algorithm 3 generates iterates $\{x_t\}_{t\in\{0,\ldots,T\}}$ which satisfy almost surely (a.s.) the inequality*

$$\begin{aligned}\gamma\sum_{t=0}^{T-1}\|\nabla F(x_t)\|+\Delta_T\leq\Delta_0+2\gamma\alpha\sum_{t=1}^{T}\left\|\sum_{j=1}^{t}(1-\alpha)^{t-j}\theta_j\right\|+2\gamma(1-\alpha)\sum_{t=1}^{T}\left\|\sum_{j=1}^{t}(1-\alpha)^{t-j}\omega_j\right\|\\ +\frac{2\gamma\|\hat{e}_0\|}{\alpha}+\frac{\gamma^2L_1T}{2},\end{aligned}\tag{106}$$

where for any $j\in\{1,\ldots,T-1\}$, the vectors θ_j and ω_j are defined as

$$\theta_j:=\texttt{clip}\left(\nabla f(x_j,\xi_j),\lambda_2\right)-\nabla F(x_j),\tag{107}$$

$$\omega_j:=\texttt{clip}\left(\nabla^2 f(\hat{x}_j,\hat{\xi}_j)(x_j-x_{j-1}),\lambda_1\right)-\left(\nabla F(x_j)-\nabla F(x_{j-1})\right),\tag{108}$$

and for any $t\in\{0,\ldots,T-1\}$ we let $\Delta_t:=F(x_t)-F^{\inf}$.

Proof. The proof is very similar to the proof of Lemma F.1 with the exception that the here function F has L_1–Lipschitz continuous gradients (instead of $\bar{L}$) and the computations (104) varies slightly due to the use of the Hessian term $\nabla^2 f(\hat{x}_j, \hat{\xi}_j)(x_j - x_{j-1})$ instead of the difference $\nabla f(x_j, \xi_j) - \nabla f(x_{j-1}, \xi_{j-1})$. We thus have,

$$\begin{aligned}
\hat{e}_t &= g_t - \nabla F(x_t) \\
&= (1-\alpha)\left(g_{t-1} + \texttt{clip}(\nabla^2 f(\hat{x}_t, \hat{\xi}_t)(x_t - x_{t-1}), \lambda_1)\right) + \alpha\, \texttt{clip}(\nabla f(x_t, \xi_t), \lambda_2) - \nabla F(x_t) \\
&= (1-\alpha)\left(g_{t-1} - \nabla F(x_{t-1})\right) + (1-\alpha)\left(\texttt{clip}(\nabla^2 f(\hat{x}_t, \hat{\xi}_t)(x_t - x_{t-1}), \lambda_1) - [\nabla F(x_t) - \nabla F(x_{t-1})]\right) \\
&\qquad + \alpha\left(\texttt{clip}(\nabla f(x_t, \xi_t), \lambda_2) - \nabla F(x_t)\right) \\
&\overset{(101)+(102)}{=} (1-\alpha)\hat{e}_{t-1} + \alpha\theta_t + (1-\alpha)\omega_t \\
&\overset{(a)}{=} (1-\alpha)^t \hat{e}_0 + \alpha\sum_{j=1}^{t}(1-\alpha)^{t-j}\theta_j + (1-\alpha)\sum_{j=1}^{t}(1-\alpha)^{t-j}\omega_j,
\end{aligned}$$

where in (a) we unroll the recursion. The next steps are exactly as in the proof of Lemma F.1 and the claimed result (106) follows. □

F.1.2 High-Probability Analysis

From the previous Lemma F.1, we need to bound the two new terms:

$$\left\|\sum_{j=1}^{t}(1-\alpha)^{t-j}\theta_j\right\| \quad \text{and} \quad \left\|\sum_{j=1}^{t}(1-\alpha)^{t-j}\omega_j\right\|,$$

where $t \in \{0, \ldots, T-1\}$. To do so, we use the same strategy as in previous works (Gorbunov et al., 2020; Sadiev et al., 2023; Liu et al., 2023; Sadiev et al., 2025) and we introduce the biased and unbiased parts of θ_j and ω_j, that is, for any $j \in [T-1]$ we let $\theta_j = \theta_j^b + \theta_j^u$ with

$$\theta_j^b := \mathbb{E}_{\xi_t}\left[\texttt{clip}\left(\nabla f(x_t, \xi_t), \lambda_1\right)\right] - \nabla F(x_t), \tag{109}$$
$$\theta_j^u := \texttt{clip}\left(\nabla f(x_t, \xi_t), \lambda_2\right) - \mathbb{E}_{\xi_t}\left[\texttt{clip}\left(\nabla f(x_t, \xi_t), \lambda_2\right)\right], \tag{110}$$

and $\omega_j = \omega_j^b + \omega_j^u$ where

$$\omega_j^b := \mathbb{E}_{\xi_t}\left[\texttt{clip}\left(\nabla f(x_j, \xi_j) - \nabla f(x_{j-1}, \xi_j), \lambda_1\right)\right] - \left(\nabla F(x_j) - \nabla F(x_{j-1})\right), \tag{111}$$
$$\omega_j^u := \texttt{clip}\left(\nabla f(x_j, \xi_j) - \nabla f(x_{j-1}, \xi_j), \lambda_1\right) - \mathbb{E}_{\xi_t}\left[\texttt{clip}\left(\nabla f(x_j, \xi_j) - \nabla f(x_{j-1}, \xi_j), \lambda_1\right)\right] \tag{112}$$

for Lemma F.1 and Corollary F.1. For Lemma F.2 (hessian clipping) we let $\omega_j = \omega_j^r + \omega_j^b + \omega_j^u$ where

$$\omega_j^r := \nabla^2 F(\hat{x}_j)(x_j - x_{j-1}) - \left(\nabla F(x_j) - \nabla F(x_{j-1})\right), \tag{113}$$
$$\omega_j^b := \mathbb{E}_{\hat{\xi}_t}\left[\texttt{clip}\left(\nabla^2 f(\hat{x}_j, \hat{\xi}_j)(x_j - x_{j-1}), \lambda_1\right)\right] - \nabla^2 F(\hat{x}_j)(x_j - x_{j-1}), \tag{114}$$
$$\omega_j^u := \texttt{clip}\left(\nabla^2 f(\hat{x}_j, \hat{\xi}_j)(x_j - x_{j-1}), \lambda_1\right) - \mathbb{E}_{\hat{\xi}_t}\left[\texttt{clip}\left(\nabla^2 f(\hat{x}_j, \hat{\xi}_j)(x_j - x_{j-1}), \lambda_1\right)\right] \tag{115}$$

Lemma F.3. *Under Assumption 2.3, for any $\delta' \in (0, \frac{1}{2}]$ and any $t \in \{0, \ldots, T-1\}$, if the clipping threshold satisfies*

$$\lambda_2 \geq \max\left\{2\left\|\nabla F(x_j)\right\|, \sigma_1 \alpha^{-\frac{1}{p}}\right\},$$

for all $j \in [t]$ then, with probability at least $1 - 2\delta'$, we have

$$\left\|\sum_{j=1}^{t}(1-\alpha)^{t-j}\theta_j\right\| \leq 22\lambda_2 \log\frac{2}{\delta'}.$$

Proof. First of all, using (109) and (110) it follows

$$\left\|\sum_{j=1}^{t}(1-\alpha)^{t-j}\theta_j\right\| \leq \underbrace{\left\|\sum_{j=1}^{t}(1-\alpha)^{t-j}\theta_j^b\right\|}_{④} + \left\|\sum_{j=1}^{t}(1-\alpha)^{t-j}\theta_j^u\right\|,$$

and now, we need to bound both terms above. For the second term, we use Lemma G.2 with exponent 2 to obtain

$$\begin{aligned}\left\|\sum_{j=1}^{t}(1-\alpha)^{t-j}\theta_j^u\right\| &\overset{(a)}{\leq} \left|\sum_{j=1}^{t}V_j^t\right| + \sqrt{\max_{j\in[t]}\left\|(1-\alpha)^{t-j}\theta_j^u\right\|^2 + \sum_{j=1}^{t}\left\|(1-\alpha)^{t-j}\theta_j^u\right\|^2} \\ &\leq \left|\sum_{j=1}^{t}V_j^t\right| + \sqrt{2\sum_{j=1}^{t}\left\|(1-\alpha)^{t-j}\theta_j^u\right\|^2} \\ &\overset{(b)}{=} \underbrace{\left|\sum_{j=1}^{t}V_j^t\right|}_{①} + \sqrt{\underbrace{2\sum_{j=1}^{t}Y_j^t}_{②} + \underbrace{\sum_{j=1}^{t}\mathbb{E}_{\xi_j}\left[\left\|(1-\alpha)^{t-j}\theta_j^u\right\|^2\right]}_{③}},\end{aligned} \tag{116}$$

where in (a) we define the sequence $V_1^t,\ldots,V_t^t$ as in Lemma G.2, that is,

$$V_j^t := \begin{cases} 0, & \text{if } j=0; \\ \operatorname{sign}\left(\sum_{i=1}^{j-1}V_i^t\right)\frac{\left\langle\sum_{i=1}^{j-1}(1-\alpha)^{t-i}\theta_i^u,(1-\alpha)^{t-j}\theta_j^u\right\rangle}{\left\|\sum_{i=1}^{j-1}(1-\alpha)^{t-i}\theta_i^u\right\|}, & \text{if } j\neq 0 \text{ and } \sum_{i=1}^{j-1}(1-\alpha)^{t-i}\theta_i^u\neq 0; \\ 0, & \text{if } j\neq 0 \text{ and } \sum_{i=1}^{j-1}(1-\alpha)^{t-i}\theta_i^u = 0. \end{cases}$$

while in (b), for any $j\in[t]$ we let

$$Y_j^t := \left\|(1-\alpha)^{t-j}\theta_j^u\right\|^2 - \mathbb{E}_{\xi_j}\left[\left\|(1-\alpha)^{t-j}\theta_j^u\right\|^2\right].$$

We now bound all terms ①, ②, ③ and ④.

Bounding ①. The sequence $V_1^t,\ldots,V_t^t$ is a martingale difference sequence since, by definition of θ_j^u and V_j^t, for all $j\in[t]$ we have $\mathbb{E}\left[V_j^t \mid V_{j-1}^t,\ldots,V_1^t\right]=0$ since $\mathbb{E}_{\xi_j}\left[\theta_j^u\right]=0$. Moreover, by Lemma G.2 we also have

$$\left|V_j^t\right| \leq \left\|(1-\alpha)^{t-j}\theta_j^u\right\| \leq \left\|\theta_j^u\right\| \overset{\text{Lem. G.4}}{\leq} 2\lambda_2, \tag{117}$$

and let $c_2 := 2\lambda_2$ be the upper bound on the random variables $V_1^t,\ldots,V_t^t$. Additionally, if we denote by $\sigma_j^2 := \mathbb{E}\left[\left(V_j^t\right)^2 \,\middle|\, V_{j-1}^t,\ldots,V_1^t\right]$ then by Lemma G.3 for any real number $b_2>0$ and any $G_2>0$, as $0<\delta'\leq\frac{1}{2}$ then $\log\frac{2}{\delta'}\geq 1$ and we have

$$\mathbb{P}\left(\left|\sum_{j=1}^{t}V_j^t\right| > b_2 \text{ and } \sum_{j=1}^{t}\sigma_j^2 \leq G_2\log\frac{2}{\delta'}\right) \leq 2\exp\left(-\frac{b_2^2}{2G_2\log\frac{2}{\delta'}+\frac{2b_2c_2}{3}}\right) = \delta', \tag{118}$$

where the last equality holds provided $b_2 := \left(\frac{c_2}{3} + \sqrt{\frac{c_2^2}{9} + 2G_2}\right) \log \frac{2}{\delta'} > 0$. We now need to define the constant G_2. To do so, we need to bound the sum $\sigma_1^2 + \cdots + \sigma_t^2$, this gives

$$\begin{aligned}
\sum_{j=1}^{t} \sigma_j^2 &= \sum_{j=1}^{t} \mathbb{E}\left[\left(V_j^t\right)^2 \,\middle|\, V_{j-1}^t, \ldots, V_1^t \right] \\
&\overset{(117)}{\leq} \sum_{j=1}^{n} \mathbb{E}_{\xi_j}\left[\left\| (1-\alpha)^{t-j} \theta_j^u \right\|^2 \right] \\
&= \sum_{j=1}^{t} (1-\alpha)^{2(t-j)} \, \mathbb{E}_{\xi_j}\left[\left\| \theta_j^u \right\|^2 \right] \\
&\overset{\text{Lem. G.4}}{\leq} 18\lambda_2^{2-p}\sigma_1^p \sum_{j=1}^{t} (1-\alpha)^{2(t-j)} \\
&\leq \frac{18\lambda_2^{2-p}\sigma_1^p}{1-(1-\alpha)^2} \\
&\leq \frac{18\lambda_2^{2-p}\sigma_1^p}{\alpha},
\end{aligned} \tag{119}$$

where in the application of Lemma G.4 we require $\lambda_2 \geq 2\max_{j\in[t]} \|\nabla F(x_j)\|$. Setting $G_2 := \frac{18\lambda_2^{2-p}\sigma_1^p}{\alpha} > 0$ gives the desired bound (118).

Bounding ②. As in the previous paragraph, the sequence $Y_1^t, \ldots, Y_t^t$ is also a martingale difference sequence as the definition of $(Y_j^t)_{j\in[t]}$ implies $\mathbb{E}\left[Y_j^t \,\middle|\, Y_{j-1}^t \ldots, Y_1^t \right] = \mathbb{E}_{\xi_j}\left[Y_j^t\right] = 0$ for all $j \in [t]$. Moreover, according to Lemma G.4, we also have, as we did in (117), for any $j \in [t]$

$$\left|Y_j^t\right| \leq \left\| (1-\alpha)^{t-j}\theta_j^u \right\|^2 + \mathbb{E}_{\xi_j}\left[\left\| (1-\alpha)^{t-j}\theta_j^u \right\|^2 \right] \leq 4\lambda_2^2 + 4\lambda_2^2 = 8\lambda_2^2. \tag{120}$$

Hence we define $\widetilde{c}_2 := 8\lambda_2^2$. Now, denoting the conditional variance of Y_j^t as $\widetilde{\sigma}_j^2 := \mathbb{E}\left[\left(Y_j^t\right)^2 \,\middle|\, Y_{j-1}^t, \ldots, Y_1^t \right] = \mathbb{E}_{\xi_j}\left[\left(Y_j^t\right)^2\right]$ we have the bound

$$\widetilde{\sigma}_j^2 \overset{(120)}{\leq} 8\lambda_2^2 \, \mathbb{E}_{\xi_j}\left[\left|Y_j^t\right|\right] \leq 16\lambda_2^2 \, \mathbb{E}_{\xi_j}\left[\left\| (1-\alpha)^{t-j}\theta_j^u \right\|^2 \right], \tag{121}$$

for all $j \in [t]$. Hence,

$$\sum_{j=1}^{t} \widetilde{\sigma}_j^2 \overset{(121)}{\leq} 16\lambda_2^2 \sum_{j=1}^{t} \mathbb{E}_{\xi_j}\left[\left\| (1-\alpha)^{t-j}\theta_j^u \right\|^2 \right] \overset{(119)}{\leq} 16\lambda_2^2 \cdot \frac{18\lambda_2^{2-p}\sigma_1^p}{\alpha} = \frac{288\lambda_2^{4-p}\sigma_1^p}{\alpha},$$

(where we require $\lambda_2 \geq 2\max_{j\in[t]} \|\nabla F(x_j)\|$). Next, if we define $\widetilde{G}_2 := \frac{288\lambda_2^{4-p}\sigma_1^p}{\alpha}$ then, applying Lemma G.3 we obtain for any real number $\widetilde{b}_2$

$$\mathbb{P}\left(\left| \sum_{j=1}^{t} Y_j^t \right| > \widetilde{b}_2 \text{ and } \sum_{j=1}^{t} \widetilde{\sigma}_j^2 \leq \widetilde{G}_2 \log \frac{2}{\delta'} \right) \leq 2\exp\left(-\frac{\widetilde{b}_2^2}{2\widetilde{G}_2 \log \frac{2}{\delta'} + \frac{2\widetilde{b}_2\widetilde{c}_2}{3}} \right) = \delta', \tag{122}$$

where the last identity holds if we set $\widetilde{b}_2 := \left(\frac{\widetilde{c}_2}{3} + \sqrt{\frac{\widetilde{c}_2^2}{9} + 2\widetilde{G}_2}\right) \log \frac{2}{\delta'}$. This establishes the desired bound.

Bounding ③. As we already did in the two last paragraphs, we have

$$③ := \sum_{j=1}^{t} \mathbb{E}_{\xi_j}\left[\left\| (1-\alpha)^{t-j}\theta_j^u \right\|^2 \right] \overset{(119)}{\leq} \frac{18\lambda_2^{2-p}\sigma_1^p}{\alpha},$$

as desired.

Bounding ④. If we assume that $\lambda_2 \geq 2\left\|\nabla F(x_j)\right\|$ for all $j \in [t]$ then, with probability one we have

$$\begin{aligned}
④ &:= \left\|\sum_{j=1}^{t}(1-\alpha)^{t-j}\theta_j^b\right\| \\
&\leq \sum_{j=1}^{t}(1-\alpha)^{t-j}\left\|\theta_j^b\right\| \\
&\overset{\text{(a)}}{\leq} 4\lambda_2^{1-p}\sigma_1^p\sum_{j=1}^{t}(1-\alpha)^{t-j} \\
&\leq \frac{4\lambda_2^{1-p}\sigma_1^p}{\alpha},
\end{aligned}$$

where in (a) we use Lemma G.4, more precisely, for any $j \in [t]$,

$$\left\|\theta_j^b\right\| \leq 2^p\lambda_2^{1-p}\sigma_1^p \leq 4\lambda_2^{1-p}\sigma_1^p.$$

Summing up all bounds ①, ②, ③ and ④. To sum up, we introduce the event $E_{①,t}$ as follows

$$E_{①,t} := \left\{\left|\sum_{j=1}^{t}V_j^t\right| \leq b_2 \ \text{ or } \ \sum_{j=1}^{t}\sigma_j^2 > G_2\log\frac{2}{\delta'}\right\},$$

where we defined $c_2 := 2\lambda_2$, $G_2 := \frac{18\lambda_2^{2-p}\sigma_1^p}{\alpha}$ and $b_2 := \left(\frac{c_2}{3}+\sqrt{\frac{c_2^2}{9}+2G_2}\right)\log\frac{2}{\delta'}$ and we can bound b_2 as

$$\begin{aligned}
b_2 := \left(\frac{c_2}{3}+\sqrt{\frac{c_2^2}{9}+2G_2}\right)\log\frac{2}{\delta'} &\leq \left(\frac{2c_2}{3}+\sqrt{2G_2}\right)\log\frac{2}{\delta'} \\
&= \left(\frac{4\lambda_2}{3}+\sqrt{\frac{36\lambda_2^{2-p}\sigma_1^p}{\alpha}}\right)\log\frac{2}{\delta'} \\
&= \lambda_2\left(\frac{4}{3}+6\sqrt{\frac{1}{\alpha}\left(\frac{\sigma_1}{\lambda_2}\right)^p}\right)\log\frac{2}{\delta'} \\
&\overset{\text{(a)}}{\leq} \frac{22\lambda_2}{3}\log\frac{2}{\delta'},
\end{aligned} \tag{123}$$

where (a) holds provided $\lambda_2 \geq \sigma_1\alpha^{-\frac{1}{p}}$. On the other hand, we also define the event $E_{②,t}$ as follows

$$E_{②,t} := \left\{\left|\sum_{j=1}^{t}Y_j^t\right| \leq \widetilde{b}_2 \ \text{ or } \ \sum_{j=1}^{t}\widetilde{\sigma}_j^2 > \widetilde{G}_2\log\frac{2}{\delta'}\right\},$$

where we defined $\widetilde{c}_2 := 8\lambda_2^2$, $\widetilde{G}_2 := \frac{288\lambda_2^{4-p}\sigma_1^p}{\alpha}$ and $\widetilde{b}_2 := \left(\frac{\widetilde{c}_2}{3}+\sqrt{\frac{\widetilde{c}_2^2}{9}+2\widetilde{G}_2}\right)\log\frac{2}{\delta'}$ and we can bound $\widetilde{b}_2$ as

$$\begin{aligned}
\widetilde{b}_2 := \left(\frac{\widetilde{c}_2}{3}+\sqrt{\frac{\widetilde{c}_2^2}{9}+2\widetilde{G}_2}\right)\log\frac{2}{\delta'} &\leq \left(\frac{2\widetilde{c}_2}{3}+\sqrt{2\widetilde{G}_2}\right)\log\frac{2}{\delta'} \\
&= \left(\frac{16\lambda_2^2}{3}+\sqrt{\frac{576\lambda_2^{4-p}\sigma_1^p}{\alpha}}\right)\log\frac{2}{\delta'} \\
&= \lambda_2^2\left(\frac{16}{3}+24\sqrt{\frac{1}{\alpha}\left(\frac{\sigma_1}{\lambda_2}\right)^p}\right)\log\frac{2}{\delta'} \\
&\leq \frac{88\lambda_2^2}{3}\log\frac{2}{\delta'},
\end{aligned} \tag{124}$$

where the last inequality holds given $\lambda_2 \geq \sigma_1 \alpha^{-\frac{1}{p}}$.

Next, given $\lambda_2 \geq \max\left\{2\left\|\nabla F(x_j)\right\|, \sigma_1\alpha^{-\frac{1}{p}}\right\}$ for all $j \in [t]$ we proved in (118) and (122) that $\mathbb{P}\left(E_{①,t}\right) \geq 1-\delta'$ and $\mathbb{P}\left(E_{②,t}\right) \geq 1-\delta'$ hence, by the union bound inequality we have

$$\mathbb{P}\left(E_{①,t} \cap E_{②,t}\right) \geq 1 - 2\delta',$$

and on the event $E_{①,t} \cap E_{②,t}$ we obtain the inequality

$$\begin{aligned}
\left\|\sum_{j=1}^{t}(1-\alpha)^{t-j}\theta_j\right\| &\leq \left\|\sum_{j=1}^{t}(1-\alpha)^{t-j}\theta_j^b\right\| + \left\|\sum_{j=1}^{t}(1-\alpha)^{t-j}\theta_j^u\right\| \\
&\overset{(116)}{\leq} ① + \sqrt{2\cdot② + 2\cdot③} + ④ \\
&\overset{(123)+(124)}{\leq} \frac{22\lambda_2}{3}\log\frac{2}{\delta'} + \sqrt{2\cdot\frac{88\lambda_2^2}{3}\log\frac{2}{\delta'} + 2\cdot\frac{18\lambda_2^{2-p}\sigma_1^p}{\alpha}} + \frac{4\lambda_2^{1-p}\sigma_1^p}{\alpha} \\
&\overset{(a)}{\leq} \lambda_2\left(\frac{22}{3} + \sqrt{\frac{176}{3} + \frac{36}{\alpha}\left(\frac{\sigma_1}{\lambda_2}\right)^p} + \frac{4}{\alpha}\left(\frac{\sigma_1}{\lambda_2}\right)^p\right)\log\frac{2}{\delta'} \\
&\overset{(b)}{\leq} \lambda_2\left(\frac{22}{3} + \sqrt{\frac{176}{3}+36} + 4\right)\log\frac{2}{\delta'} \\
&\leq 22\lambda_2\log\frac{2}{\delta'},
\end{aligned}$$

where in (a) we use the fact that $\log\frac{2}{\delta'} \geq 1$ and in (b) we use $\lambda_2 \geq \sigma_1\alpha^{-\frac{1}{p}}$. This proves the lemma. □

Lemma F.4. *Under Assumption 2.6, for any $\delta'' \in (0, \frac{1}{2}]$ and any $t \in \{0, \ldots, T-1\}$, if the clipping threshold satisfies*

$$\lambda_1 \geq \max\left\{2\gamma\bar{L}, \gamma\bar{L}\alpha^{-\frac{1}{q}}\right\},$$

for all $j \in [t]$ then, with probability at least $1-2\delta''$, we have

$$\left\|\sum_{j=1}^{t}(1-\alpha)^{t-j}\omega_j\right\| \leq 46\lambda_1\log\frac{2}{\delta''}.$$

Proof. First, using (111) and (112) we have

$$\left\|\sum_{j=1}^{t}(1-\alpha)^{t-j}\omega_j\right\| \leq \underbrace{\left\|\sum_{j=1}^{t}(1-\alpha)^{t-j}\omega_j^b\right\|}_{⑧} + \left\|\sum_{j=1}^{t}(1-\alpha)^{t-j}\omega_j^u\right\|, \tag{125}$$

and, as before, we need to bound both terms above. For the second term, we use Lemma G.2 with exponent 2 to obtain

$$\begin{aligned}
\left\|\sum_{j=1}^{t}(1-\alpha)^{t-j}\omega_j^u\right\| &\overset{(a)}{\leq} \left|\sum_{j=1}^{t}W_j^t\right| + \sqrt{\max_{j\in[t]}\left\|(1-\alpha)^{t-j}\omega_j^u\right\|^2 + \sum_{j=1}^{t}\left\|(1-\alpha)^{t-j}\omega_j^u\right\|^2} \\
&\leq \left|\sum_{j=1}^{t}W_j^t\right| + \sqrt{2\sum_{j=1}^{t}\left\|(1-\alpha)^{t-j}\omega_j^u\right\|^2} \\
&\overset{(b)}{=} \underbrace{\left|\sum_{j=1}^{t}W_j^t\right|}_{⑤} + \sqrt{\underbrace{2\sum_{j=1}^{t}Z_j^t}_{⑥} + \underbrace{\sum_{j=1}^{t}\mathbb{E}_{\xi_j}\left[\left\|(1-\alpha)^{t-j}\omega_j^u\right\|^2\right]}_{⑦}},
\end{aligned} \tag{126}$$

where in (a) we define the sequence $V_1^t, \dots, V_t^t$ as in Lemma G.2, that is,

$$W_j^t := \begin{cases} 0, & \text{if } j = 0; \\ \operatorname{sign}\left(\sum_{i=1}^{j-1} W_i^t\right) \frac{\left\langle \sum_{i=1}^{j-1} (1-\alpha)^{t-i}\omega_i^u, (1-\alpha)^{t-j}\omega_j^u \right\rangle}{\left\| \sum_{i=1}^{j-1} (1-\alpha)^{t-i}\omega_i^u \right\|}, & \text{if } j \neq 0 \text{ and } \sum_{i=1}^{j-1} (1-\alpha)^{t-i}\omega_i^u \neq 0; \\ 0, & \text{if } j \neq 0 \text{ and } \sum_{i=1}^{j-1} (1-\alpha)^{t-i}\omega_i^u = 0. \end{cases}$$

while in (b), for any $j \in [t]$ we let

$$Z_j^t := \left\|(1-\alpha)^{t-j}\omega_j^u\right\|^2 - \mathbb{E}_{\xi_j}\left[\left\|(1-\alpha)^{t-j}\omega_j^u\right\|^2\right].$$

We now bound all terms ⑤, ⑥, ⑦ and ⑧.

Bounding ⑤. The sequence $W_1^t, \dots, W_t^t$ is a martingale difference sequence since, by definition of ω_j^u and W_j^t, for all $j \in [t]$ we have $\mathbb{E}\left[W_j^t \,\middle|\, W_{j-1}^t, \dots, W_1^t\right] = 0$ since $\mathbb{E}_{\xi_j}\left[\omega_j^u\right] = 0$. Moreover, by Lemma G.2 we also have

$$\left|W_j^t\right| \leq \left\|(1-\alpha)^{t-j}\omega_j^u\right\| \leq \left\|\omega_j^u\right\| \overset{\text{Lem. G.4}}{\leq} 2\lambda_1, \tag{127}$$

and let $c_1 := 2\lambda_1$ be the upper bound on the random variables $W_1^t, \dots, W_t^t$. Additionally, if we denote by $\sigma_j^2 := \mathbb{E}\left[\left(W_j^t\right)^2 \,\middle|\, W_{j-1}^t, \dots, W_1^t\right]$ then by Lemma G.3 for any real number $b_1 > 0$ and any $G_1 > 0$, as $0 < \delta'' \leq \frac{1}{2}$ then $\log\frac{2}{\delta''} \geq 1$ and we have

$$\mathbb{P}\left(\left|\sum_{j=1}^t W_j^t\right| > b_1 \text{ and } \sum_{j=1}^t \sigma_j^2 \leq G_1 \log\frac{2}{\delta''}\right) \leq 2\exp\left(-\frac{b_1^2}{2G_1\log\frac{2}{\delta''} + \frac{2b_1c_1}{3}}\right) = \delta'', \tag{128}$$

where the last equality holds provided $b_1 := \left(\frac{c_1}{3} + \sqrt{\frac{c_1^2}{9} + 2G_1}\right)\log\frac{2}{\delta''} > 0$. We now need to define the constant G_1. To do so, we need to bound the sum $\sigma_1^2 + \dots + \sigma_t^2$, this gives

$$\begin{aligned} \sum_{j=1}^t \sigma_j^2 &= \sum_{j=1}^t \mathbb{E}\left[\left(W_j^t\right)^2 \,\middle|\, W_{j-1}^t, \dots, W_1^t\right] \\ &\overset{(127)}{\leq} \sum_{j=1}^n \mathbb{E}_{\xi_j}\left[\left\|(1-\alpha)^{t-j}\omega_j^u\right\|^2\right] \\ &= \sum_{j=1}^t (1-\alpha)^{2(t-j)}\,\mathbb{E}_{\xi_j}\left[\left\|\omega_j^u\right\|^2\right] \\ &\overset{\text{Lem. G.4}}{\leq} 72\lambda_1^{2-q}\gamma^q\bar{L}^q \sum_{j=1}^t (1-\alpha)^{2(t-j)} \\ &\leq \frac{72\lambda_1^{2-q}\gamma^q\bar{L}^q}{1-(1-\alpha)^2} \\ &\leq \frac{72\lambda_1^{2-q}\gamma^q\bar{L}^q}{\alpha}, \end{aligned} \tag{129}$$

where in the application of Lemma G.4 we use the bound we proved earlier in Lemma D.5 (see more precisely at (71)), that is,

$$\mathbb{E}_{\xi_j}\left[\left\|\left[\nabla F(x_j) - \nabla F(x_{j-1})\right] - \left[\nabla f(x_j,\xi_j) - \nabla f(x_{j-1},\xi_j)\right]\right\|^q\right] \leq 2^q\gamma^q\bar{L}^q.$$

Moreover, applying Lemma G.4 requires to take $\lambda_1 \geq 2\max_{j\in[t]}\left\|\nabla F(x_j) - \nabla F(x_{j-1})\right\|$ and, since

$$\left\|\nabla F(x_j) - \nabla F(x_{j-1})\right\| \overset{\text{Ass. 2.6}}{\leq} \bar{L}\left\|x_j - x_{j-1}\right\| = \gamma\bar{L},$$

then it is enough to have $\lambda_1 \geq 2\gamma\bar{L}$. Setting $G_1 := \frac{72\lambda_1^{2-q}\gamma^q\bar{L}^q}{\alpha} > 0$ gives the desired bound (128).

Bounding ⑥. As in the previous paragraph, the sequence $Z_1^t, \ldots, Z_t^t$ is also a martingale difference sequence as the definition of $(Z_j^t)_{j\in[t]}$ implies $\mathbb{E}\left[Z_j^t \,\middle|\, Z_{j-1}^t \ldots, Z_1^t\right] = \mathbb{E}_{\xi_j}\left[Z_j^t\right] = 0$ for all $j \in [t]$. Moreover, according to Lemma G.4, we also have, as we did in (127), for any $j \in [t]$

$$\left|Z_j^t\right| \leq \left\|(1-\alpha)^{t-j}\omega_j^u\right\|^2 + \mathbb{E}_{\xi_j}\left[\left\|(1-\alpha)^{t-j}\omega_j^u\right\|^2\right] \leq 4\lambda_1^2 + 4\lambda_1^2 = 8\lambda_1^2. \tag{130}$$

Hence we define $\widetilde{c}_1 := 8\lambda_1^2$. Now, denoting the conditional variance of Z_j^t as $\widetilde{\sigma}_j^2 := \mathbb{E}\left[\left(Z_j^t\right)^2 \,\middle|\, Z_{j-1}^t, \ldots, Z_1^t\right] = \mathbb{E}_{\xi_j}\left[\left(Z_j^t\right)^2\right]$ we have the bound

$$\widetilde{\sigma}_j^2 \overset{(130)}{\leq} 8\lambda_1^2\,\mathbb{E}_{\xi_j}\left[\left|Z_j^t\right|\right] \leq 16\lambda_1^2\,\mathbb{E}_{\xi_j}\left[\left\|(1-\alpha)^{t-j}\omega_j^u\right\|^2\right], \tag{131}$$

for all $j \in [t]$. Hence,

$$\sum_{j=1}^t \widetilde{\sigma}_j^2 \overset{(131)}{\leq} 16\lambda_1^2 \sum_{j=1}^t \mathbb{E}_{\xi_j}\left[\left\|(1-\alpha)^{t-j}\omega_j^u\right\|^2\right] \overset{(129)}{\leq} 16\lambda_1^2 \cdot \frac{72\lambda_1^{2-q}\gamma^q\bar{L}^q}{\alpha} = \frac{1152\lambda_1^{4-q}\gamma^q\bar{L}^q}{\alpha},$$

(where we require $\lambda_1 \geq 2\gamma\bar{L}$). Next, if we define $\widetilde{G}_1 := \frac{1152\lambda_1^{4-q}\gamma^q\bar{L}^q}{\alpha}$ then, applying Lemma G.3 we obtain for any real number $\widetilde{b}_2$

$$\mathbb{P}\left(\left|\sum_{j=1}^t Z_j^t\right| > \widetilde{b}_1 \text{ and } \sum_{j=1}^t \widetilde{\sigma}_j^2 \leq \widetilde{G}_1 \log\frac{2}{\delta''}\right) \leq 2\exp\left(-\frac{\widetilde{b}_1^2}{2\widetilde{G}_1\log\frac{2}{\delta''} + \frac{2\widetilde{b}_1\widetilde{c}_1}{3}}\right) = \delta'', \tag{132}$$

where the last identity holds if we set $\widetilde{b}_1 := \left(\frac{\widetilde{c}_1}{3} + \sqrt{\frac{\widetilde{c}_1^2}{9} + 2\widetilde{G}_1}\right)\log\frac{2}{\delta''}$. This establishes the desired bound.

Bounding ⑦. As we already did in the two last paragraphs, we have

$$⑦ := \sum_{j=1}^t \mathbb{E}_{\xi_j}\left[\left\|(1-\alpha)^{t-j}\omega_j^u\right\|^2\right] \overset{(129)}{\leq} \frac{72\lambda_1^{2-q}\gamma^q\bar{L}^q}{\alpha},$$

as desired.

Bounding ⑧. If we assume that $\lambda_1 \geq 2\gamma\bar{L}$ then, with probability one we have

$$\begin{aligned}
⑧ &:= \left\|\sum_{j=1}^t (1-\alpha)^{t-j}\omega_j^b\right\| \\
&\leq \sum_{j=1}^t (1-\alpha)^{t-j}\left\|\omega_j^b\right\| \\
&\overset{(a)}{\leq} 16\lambda_1^{1-q}\gamma^q\bar{L}^q \sum_{j=1}^t (1-\alpha)^{t-j} \\
&\leq \frac{16\lambda_1^{1-q}\gamma^q\bar{L}^q}{\alpha},
\end{aligned}$$

where in (a) we use Lemma G.4, more precisely, for any $j \in [t]$,

$$\left\|\omega_j^b\right\| \leq 4^q\lambda_1^{1-q}\gamma^q\bar{L}^q \leq 16\lambda_1^{1-q}\gamma^q\bar{L}^q,$$

since $\mathbb{E}\left[\left\|\left[\nabla F(x_t) - \nabla F(x_{t-1})\right] - \left[\nabla f(x_t,\xi_t) - \nabla f(x_{t-1},\xi_t)\right]\right\|^q\right] \leq 2^q\gamma^q\bar{L}^q$ by Lemma D.5 (and more precisely by (71)).

Summing up all bounds ⑤, ⑥, ⑦ and ⑧. To sum up, we introduce the event $E_{⑤,t}$ as follows

$$E_{⑤,t} := \left\{ \left| \sum_{j=1}^{t} W_j^t \right| \leq b_1 \ \text{ or } \ \sum_{j=1}^{t} \sigma_j^2 > G_1 \log \frac{2}{\delta''} \right\}, \tag{133}$$

where we defined $c_1 := 2\lambda_1$, $G_1 := \frac{72\lambda_1^{2-q}\gamma^q \bar{L}^q}{\alpha}$ and $b_1 := \left(\frac{c_1}{3} + \sqrt{\frac{c_1^2}{9} + 2G_1} \right) \log \frac{2}{\delta''}$ and we can bound b_1 as

$$\begin{aligned} b_1 := \left(\frac{c_1}{3} + \sqrt{\frac{c_1^2}{9} + 2G_1} \right) \log \frac{2}{\delta''} &\leq \left(\frac{2c_1}{3} + \sqrt{2G_1} \right) \log \frac{2}{\delta''} \\ &= \left(\frac{4\lambda_1}{3} + \sqrt{\frac{144\lambda_1^{2-q}\gamma^q \bar{L}^q}{\alpha}} \right) \log \frac{2}{\delta''} \\ &= \lambda_1 \left(\frac{4}{3} + 12\sqrt{\frac{1}{\alpha}\left(\frac{\gamma \bar{L}}{\lambda_1}\right)^q} \right) \log \frac{2}{\delta''} \\ &\overset{\text{(a)}}{\leq} \frac{40\lambda_1}{3} \log \frac{2}{\delta''}, \end{aligned} \tag{134}$$

where (a) holds provided $\lambda_1 \geq \gamma \bar{L} \alpha^{-\frac{1}{q}}$. On the other hand, we also define the event $E_{⑥,t}$ as follows

$$E_{⑥,t} := \left\{ \left| \sum_{j=1}^{t} Z_j^t \right| \leq \widetilde{b}_1 \ \text{ or } \ \sum_{j=1}^{t} \widetilde{\sigma}_j^2 > \widetilde{G}_1 \log \frac{2}{\delta''} \right\}, \tag{135}$$

where we defined $\widetilde{c}_1 := 8\lambda_1^2$, $\widetilde{G}_1 := \frac{1152\lambda_1^{4-q}\gamma^q \bar{L}^q}{\alpha}$ and $\widetilde{b}_1 := \left(\frac{\widetilde{c}_1}{3} + \sqrt{\frac{\widetilde{c}_1^2}{9} + 2\widetilde{G}_1} \right) \log \frac{2}{\delta''}$ and we can bound $\widetilde{b}_1$ as

$$\begin{aligned} \widetilde{b}_1 := \left(\frac{\widetilde{c}_1}{3} + \sqrt{\frac{\widetilde{c}_1^2}{9} + 2\widetilde{G}_1} \right) \log \frac{2}{\delta''} &\leq \left(\frac{2\widetilde{c}_1}{3} + \sqrt{2\widetilde{G}_1} \right) \log \frac{2}{\delta''} \\ &= \left(\frac{16\lambda_1^2}{3} + \sqrt{\frac{2304\lambda_1^{4-q}\gamma^q \bar{L}^q}{\alpha}} \right) \log \frac{2}{\delta''} \\ &= \lambda_1^2 \left(\frac{16}{3} + 48\sqrt{\frac{1}{\alpha}\left(\frac{\gamma \bar{L}}{\lambda_1}\right)^q} \right) \log \frac{2}{\delta''} \\ &\leq \frac{160\lambda_1^2}{3} \log \frac{2}{\delta''}, \end{aligned} \tag{136}$$

where the last inequality holds given $\lambda_1 \geq \gamma \bar{L} \alpha^{-\frac{1}{q}}$.

Next, given $\lambda_1 \geq \max\left\{ 2\gamma \bar{L}, \gamma \bar{L} \alpha^{-\frac{1}{q}} \right\}$, we proved in (128) and (132) that $\mathbb{P}\left(E_{⑤,t}\right) \geq 1 - \delta''$ and $\mathbb{P}\left(E_{⑥,t}\right) \geq 1 - \delta''$ hence, by the union bound inequality we have

$$\mathbb{P}\left(E_{⑤,t} \cap E_{⑥,t}\right) \geq 1 - 2\delta'',$$

and on the event $E_{⑤,t} \cap E_{⑥,t}$ we obtain the inequality

$$\begin{aligned}
\left\|\sum_{j=1}^{t}(1-\alpha)^{t-j}\omega_j\right\| &\leq \left\|\sum_{j=1}^{t}(1-\alpha)^{t-j}\omega_j^b\right\| + \left\|\sum_{j=1}^{t}(1-\alpha)^{t-j}\omega_j^u\right\| \\
&\overset{(126)}{\leq} ⑤ + \sqrt{2\cdot ⑥ + 2\cdot ⑦} + ⑧ \\
&\overset{(134)+(136)}{\leq} \frac{40\lambda_1}{3}\log\frac{2}{\delta''} + \sqrt{2\cdot\frac{160\lambda_1^2}{3}\log\frac{2}{\delta''} + 2\cdot\frac{72\lambda_1^{2-q}\gamma^q\bar{L}^q}{\alpha}} + \frac{16\lambda_1^{1-q}\gamma^q\bar{L}^q}{\alpha} \\
&\overset{(a)}{\leq} \lambda_1\left(\frac{40}{3} + \sqrt{\frac{320}{3} + \frac{144}{\alpha}\left(\frac{\gamma\bar{L}}{\lambda_1}\right)^q} + \frac{16}{\alpha}\left(\frac{\gamma\bar{L}}{\lambda_1}\right)^q\right)\log\frac{2}{\delta''} \\
&\overset{(b)}{\leq} \lambda_1\left(\frac{40}{3} + \sqrt{\frac{320}{3} + 144} + 16\right)\log\frac{2}{\delta''} \\
&\leq 46\lambda_1\log\frac{2}{\delta''},
\end{aligned}$$

where in (a) we use the fact that $\log\frac{2}{\delta''} \geq 1$ and in (b) we use $\lambda_1 \geq \gamma\bar{L}\alpha^{-\frac{1}{q}}$. This proves the lemma. □

Lemma F.5. *Under Assumptions 2.2 and 2.7, for any $\delta'' \in (0, \frac{1}{2}]$ and any $t \in \{0, \ldots, T-1\}$, if the clipping threshold satisfies*

$$\lambda_1 \geq \max\left\{2\gamma L_1, \gamma\bar{\delta}\alpha^{-\frac{1}{q}}\right\},$$

for all $j \in [t]$ then, with probability at least $1-2\delta''$, we have

$$\left\|\sum_{j=1}^{t}(1-\alpha)^{t-j}\omega_j\right\| \leq 22\lambda_1\log\frac{2}{\delta''}.$$

Proof. First, note that the computations (125) and (126) are identical and we now need to bound all terms ⑤, ⑥, ⑦ and ⑧ where

$$⑤ := \left|\sum_{j=1}^{t} W_j^t\right|, \quad ⑥ := \sum_{j=1}^{t} Z_j^t, \quad ⑦ := \sum_{j=1}^{t}\mathbb{E}_{\xi_j}\left[\left\|(1-\alpha)^{t-j}\omega_j^u\right\|^2\right] \text{ and } ⑧ := \left\|\sum_{j=1}^{t}(1-\alpha)^{t-j}\omega_j^b\right\|.$$

Bounding ⑤. As in the previous lemma, we choose $c_1 := 2\lambda_1$ and $b_1 := \left(\frac{c_1}{3} + \sqrt{\frac{c_1^2}{9} + 2G_1}\right)\log\frac{2}{\delta''} > 0$. Then, to select $G_1 > 0$ we need to bound the sum $\sigma_1^2 + \cdots + \sigma_t^2$, i.e.,

$$\begin{aligned}
\sum_{j=1}^{t}\sigma_j^2 &\overset{(129)}{\leq} \sum_{j=1}^{t}(1-\alpha)^{2(t-j)}\,\mathbb{E}_{\xi_j}\left[\left\|\omega_j^u\right\|^2\right] \\
&\overset{\text{Lem. G.4}}{\leq} 18\lambda_1^{2-q}\gamma^q\bar{\delta}^q\sum_{j=1}^{t}(1-\alpha)^{2(t-j)} \\
&\leq \frac{18\lambda_1^{2-q}\gamma^q\bar{\delta}^q}{\alpha},
\end{aligned} \tag{137}$$

where $\bar{\delta} \geq 0$ is the parameter in Assumption 2.7. In the application of Lemma G.4 we use Assumption 2.7

$$\mathbb{E}_{\xi_j}\left[\left\|[\nabla F(x_j) - \nabla F(x_{j-1})] - [\nabla f(x_j,\xi_j) - \nabla f(x_{j-1},\xi_j)]\right\|^q\right] \leq \bar{\delta}^q\left\|x_j - x_{j-1}\right\|^q \leq \gamma^q\bar{\delta}^q.$$

Moreover, applying Lemma G.4 requires to take $\lambda_1 \geq 2\max_{j\in[t]}\|\nabla F(x_j) - \nabla F(x_{j-1})\|$ and, since

$$\|\nabla F(x_j) - \nabla F(x_{j-1})\| \overset{\text{Ass. 2.2}}{\leq} L_1\|x_j - x_{j-1}\| = \gamma L_1,$$

then it is enough to have $\lambda_1 \geq 2\gamma L_1$. Setting $G_1 := \frac{18\lambda_1^{2-q}\gamma^q\bar{\delta}^q}{\alpha} > 0$ gives the desired bound.

Bounding ⑥. Similarly to the previous lemma, we set $\widetilde{c}_1 := 8\lambda_1^2$ and $\widetilde{b}_1 := \left(\frac{\widetilde{c}_1}{3} + \sqrt{\frac{\widetilde{c}_1^2}{9} + 2\widetilde{G}_1}\right)\log\frac{2}{\delta''}$. For the choice $\widetilde{G}_1$, we have the bound

$$\sum_{j=1}^{t}\widetilde{\sigma}_j^2 \overset{(131)}{\leq} 16\lambda_1^2\sum_{j=1}^{t}\mathbb{E}_{\xi_j}\left[\left\|(1-\alpha)^{t-j}\omega_j^u\right\|^2\right] \overset{(137)}{\leq} 16\lambda_1^2\cdot\frac{18\lambda_1^{2-q}\gamma^q\bar{\delta}^q}{\alpha} = \frac{288\lambda_1^{4-q}\gamma^q\bar{\delta}^q}{\alpha},$$

(where we require $\lambda_1 \geq 2\gamma L_1$). Hence, if we let $\widetilde{G}_1 := \frac{288\lambda_1^{4-q}\gamma^q\bar{\delta}^q}{\alpha}$ this establishes the desired bound.

Bounding ⑦. As we already did in the two last paragraphs, we have

$$⑦ := \sum_{j=1}^{t}\mathbb{E}_{\xi_j}\left[\left\|(1-\alpha)^{t-j}\omega_j^u\right\|^2\right] \overset{(137)}{\leq} \frac{18\lambda_1^{2-q}\gamma^q\bar{\delta}^q}{\alpha},$$

as desired.

Bounding ⑧. For the last bound, if we assume that $\lambda_1 \geq 2\gamma L_1$ then, with probability one we have

$$\begin{aligned}
⑧ &:= \left\|\sum_{j=1}^{t}(1-\alpha)^{t-j}\omega_j^b\right\| \\
&\leq \sum_{j=1}^{t}(1-\alpha)^{t-j}\left\|\omega_j^b\right\| \\
&\overset{(a)}{\leq} 4\lambda_1^{1-q}\gamma^q\bar{\delta}^q\sum_{j=1}^{t}(1-\alpha)^{t-j} \\
&\leq \frac{4\lambda_1^{1-q}\gamma^q\bar{\delta}^q}{\alpha},
\end{aligned}$$

where in (a) we use Lemma G.4, more precisely, for any $j \in [t]$,

$$\left\|\omega_j^b\right\| \leq 2^q\lambda_1^{1-q}\gamma^q\bar{\delta}^q \leq 4\lambda_1^{1-q}\gamma^q\bar{\delta}^q,$$

since $\mathbb{E}\left[\|[\nabla F(x_t) - \nabla F(x_{t-1})] - [\nabla f(x_t,\xi_t) - \nabla f(x_{t-1},\xi_t)]\|^q\right] \leq \gamma^q\bar{\delta}^q$ by Assumption 2.7.

Summing up all bounds ⑤, ⑥, ⑦ and ⑧. We introduce the same events $E_{⑤,t}$ and $E_{⑥,t}$ as in (133) and (135) respectively. Then, by our choice of c_1, b_1 and G_1 we have the bound

$$\begin{aligned}
b_1 := \left(\frac{c_1}{3} + \sqrt{\frac{c_1^2}{9} + 2G_1}\right)\log\frac{2}{\delta''} &\leq \left(\frac{2c_1}{3} + \sqrt{2G_1}\right)\log\frac{2}{\delta''} \\
&= \left(\frac{4\lambda_1}{3} + \sqrt{\frac{36\lambda_1^{2-q}\gamma^q\bar{\delta}^q}{\alpha}}\right)\log\frac{2}{\delta''} \\
&= \lambda_1\left(\frac{4}{3} + 6\sqrt{\frac{1}{\alpha}\left(\frac{\gamma\bar{\delta}}{\lambda_1}\right)^q}\right)\log\frac{2}{\delta''} \\
&\overset{(a)}{\leq} \frac{22\lambda_1}{3}\log\frac{2}{\delta''},
\end{aligned} \tag{138}$$

where (a) holds provided $\lambda_1 \geq \gamma\bar{\delta}\alpha^{-\frac{1}{q}}$. Moreover, by our choice of $\widetilde{c}_1$, $\widetilde{b}_1$ and $\widetilde{G}_1$ we also have the bound

$$\begin{aligned}\widetilde{b}_1 := \left(\frac{\widetilde{c}_1}{3} + \sqrt{\frac{\widetilde{c}_1^2}{9} + 2\widetilde{G}_1}\right)\log\frac{2}{\delta''} &\leq \left(\frac{2\widetilde{c}_1}{3} + \sqrt{2\widetilde{G}_1}\right)\log\frac{2}{\delta''} \\ &= \left(\frac{16\lambda_1^2}{3} + \sqrt{\frac{576\lambda_1^{4-q}\gamma^q\bar{\delta}^q}{\alpha}}\right)\log\frac{2}{\delta''} \\ &= \lambda_1^2\left(\frac{16}{3} + 24\sqrt{\frac{1}{\alpha}\left(\frac{\gamma\bar{\delta}}{\lambda_1}\right)^q}\right)\log\frac{2}{\delta''} \\ &\leq \frac{88\lambda_1^2}{3}\log\frac{2}{\delta''}, \end{aligned} \tag{139}$$

where the last inequality holds given $\lambda_1 \geq \gamma\bar{\delta}\alpha^{-\frac{1}{q}}$.

Next, given $\lambda_1 \geq \max\left\{2\gamma L_1, \gamma\bar{\delta}\alpha^{-\frac{1}{q}}\right\}$, as in the previous lemma, we have $\mathbb{P}\left(E_{⑤,t}\right) \geq 1-\delta''$ and $\mathbb{P}\left(E_{⑥,t}\right) \geq 1-\delta''$ hence, by the union bound inequality we have

$$\mathbb{P}\left(E_{⑤,t} \cap E_{⑥,t}\right) \geq 1 - 2\delta'',$$

and on the event $E_{⑤,t} \cap E_{⑥,t}$ we obtain the inequality

$$\begin{aligned}\left\|\sum_{j=1}^t (1-\alpha)^{t-j}\omega_j\right\| &\leq \left\|\sum_{j=1}^t (1-\alpha)^{t-j}\omega_j^b\right\| + \left\|\sum_{j=1}^t (1-\alpha)^{t-j}\omega_j^u\right\| \\ &\overset{(126)}{\leq} ⑤ + \sqrt{2\cdot⑥ + 2\cdot⑦} + ⑧ \\ &\overset{(138)+(139)}{\leq} \frac{22\lambda_1}{3}\log\frac{2}{\delta''} + \sqrt{2\cdot\frac{176\lambda_1^2}{3}\log\frac{2}{\delta''} + 2\cdot\frac{18\lambda_1^{2-q}\gamma^q\bar{\delta}^q}{\alpha}} + \frac{4\lambda_1^{1-q}\gamma^q\bar{\delta}^q}{\alpha} \\ &\overset{(a)}{\leq} \lambda_1\left(\frac{22}{3} + \sqrt{\frac{176}{3} + \frac{36}{\alpha}\left(\frac{\gamma\bar{\delta}}{\lambda_1}\right)^q} + \frac{4}{\alpha}\left(\frac{\gamma\bar{\delta}}{\lambda_1}\right)^q\right)\log\frac{2}{\delta''} \\ &\overset{(b)}{\leq} \lambda_1\left(\frac{22}{3} + \sqrt{\frac{176}{3} + 36} + 4\right)\log\frac{2}{\delta''} \\ &\leq 22\lambda_1\log\frac{2}{\delta''}, \end{aligned}$$

where in (a) we use the fact that $\log\frac{2}{\delta''} \geq 1$ and in (b) we use $\lambda_1 \geq \gamma\bar{\delta}\alpha^{-\frac{1}{q}}$. This proves the lemma. □

Lemma F.6. *Under Assumptions 2.2, 2.4 and 2.5, for any $\delta'' \in (0, \frac{1}{2}]$ and any $t \in \{0, \ldots, T-1\}$, if the clipping threshold satisfies*

$$\lambda_1 \geq \max\left\{2\gamma L_1, \gamma\sigma_2\alpha^{-\frac{1}{q}}\right\},$$

for all $j \in [t]$ then, with probability at least $1-2\delta''$, we have

$$\left\|\sum_{j=1}^t (1-\alpha)^{t-j}\omega_j\right\| \leq \frac{2\min\left\{\gamma L_1, \gamma^2 L_2\right\}}{\alpha} + 22\lambda_1\log\frac{2}{\delta''}.$$

Proof. First, using (113), (114) and (115) we have

$$\left\|\sum_{j=1}^t (1-\alpha)^{t-j}\omega_j\right\| \leq \underbrace{\left\|\sum_{j=1}^t (1-\alpha)^{t-j}\omega_j^r\right\|}_{⑨} + \underbrace{\left\|\sum_{j=1}^t (1-\alpha)^{t-j}\omega_j^b\right\|}_{⑧} + \left\|\sum_{j=1}^t (1-\alpha)^{t-j}\omega_j^u\right\|, \tag{140}$$

As in the previous lemmas, the computation (126) is identical and we now need to bound all terms ⑤, ⑥, ⑦, ⑧ and ⑨ where

$$⑤ := \left|\sum_{j=1}^{t} W_j^t\right|, \quad ⑥ := \sum_{j=1}^{t} Z_j^t \text{ and } ⑦ := \sum_{j=1}^{t} \mathbb{E}_{\hat{\xi}_j}\left[\left\|(1-\alpha)^{t-j}\omega_j^u\right\|^2\right],$$

and $W_1^t, \ldots, W_t^t$ is a martingale difference sequence since, by definition of ω_j^u and W_j^t, for all $j \in [t]$ we have $\mathbb{E}\left[W_j^t \,\middle|\, W_{j-1}^t, \ldots, W_1^t\right] = 0$ since

$$\mathbb{E}_{q_j, \hat{\xi}_j}\left[\omega_j^u\right] = \mathbb{E}_{q_j}\left[\mathbb{E}_{\hat{\xi}_j}\left[\omega_j^u \mid q_j\right]\right] = 0.$$

The same argument applies to the sequence $Z_1^t, \ldots, Z_t^t$.

Bounding ⑤. As already done before, we choose $c_1 := 2\lambda_1$ and $b_1 := \left(\frac{c_1}{3} + \sqrt{\frac{c_1^2}{9} + 2G_1}\right)\log\frac{2}{\delta''} > 0$. Then, to select $G_1 > 0$ we need to bound the sum $\sigma_1^2 + \cdots + \sigma_t^2$, i.e.,

$$\begin{aligned}\sum_{j=1}^{t}\sigma_j^2 &\overset{(129)}{\leq} \sum_{j=1}^{t}(1-\alpha)^{2(t-j)}\,\mathbb{E}_{\hat{\xi}_j}\left[\left\|\omega_j^u\right\|^2\right] \\ &\overset{\text{Lem. G.4}}{\leq} 18\lambda_1^{2-q}\gamma^q\sigma_2^q\sum_{j=1}^{t}(1-\alpha)^{2(t-j)} \\ &\leq \frac{18\lambda_1^{2-q}\gamma^q\sigma_2^q}{\alpha}. \end{aligned} \tag{141}$$

In the application of Lemma G.4 we use Assumption 2.5

$$\mathbb{E}_{\hat{\xi}_j}\left[\left\|\nabla^2 f(\hat{x}_j, \hat{\xi}_j)(x_j - x_{j-1}) - \nabla^2 F(\hat{x}_j)(x_j - x_{j-1})\right\|^q\right] \leq \mathbb{E}_{\hat{\xi}_j}\left[\left\|\nabla^2 f(\hat{x}_j, \hat{\xi}_j) - \nabla^2 F(\hat{x}_j)\right\|_{\text{op}}^q \cdot \|x_j - x_{j-1}\|^q\right] \leq \gamma^q\sigma_2^q,$$

as

$$\mathbb{E}_{\hat{\xi}_j}\left[\nabla^2 f(\hat{x}_j, \hat{\xi}_j)(x_j - x_{j-1})\right] = \nabla^2 F(\hat{x}_j)(x_j - x_{j-1}).$$

Moreover, applying Lemma G.4 requires to take $\lambda_1 \geq 2\max_{j\in[t]}\left\|\nabla^2 F(\hat{x}_j)(x_j - x_{j-1})\right\|$ and, since

$$\left\|\nabla^2 F(\hat{x}_j)(x_j - x_{j-1})\right\| \leq \left\|\nabla^2 F(\hat{x}_j)\right\|_{\text{op}} \cdot \|x_j - x_{j-1}\| \overset{\text{Ass. 2.2}}{\leq} L_1\|x_j - x_{j-1}\| = \gamma L_1,$$

then it is enough to have $\lambda_1 \geq 2\gamma L_1$. Setting $G_1 := \frac{18\lambda_1^{2-q}\gamma^q\sigma_2^q}{\alpha} > 0$ gives the desired bound.

Bounding ⑥. Similarly to the previous lemma, we set $\widetilde{c}_1 := 8\lambda_1^2$ and $\widetilde{b}_1 := \left(\frac{\widetilde{c}_1}{3} + \sqrt{\frac{\widetilde{c}_1^2}{9} + 2\widetilde{G}_1}\right)\log\frac{2}{\delta''}$. For the choice $\widetilde{G}_1$, we have the bound

$$\sum_{j=1}^{t}\widetilde{\sigma}_j^2 \overset{(131)}{\leq} 16\lambda_1^2\sum_{j=1}^{t}\mathbb{E}_{q_j,\hat{\xi}_j}\left[\left\|(1-\alpha)^{t-j}\omega_j^u\right\|^2\right] \overset{(141)}{\leq} 16\lambda_1^2\cdot\frac{18\lambda_1^{2-q}\gamma^q\sigma_2^q}{\alpha} = \frac{288\lambda_1^{4-q}\gamma^q\sigma_2^q}{\alpha},$$

(where we require $\lambda_1 \geq 2\gamma L_1$). Hence, if we let $\widetilde{G}_1 := \frac{288\lambda_1^{4-q}\gamma^q\sigma_2^q}{\alpha}$ this establishes the desired bound.

Bounding ⑦. As we already did in the two last paragraphs, we have

$$⑦ := \sum_{j=1}^{t}\mathbb{E}_{\hat{\xi}_j}\left[\left\|(1-\alpha)^{t-j}\omega_j^u\right\|^2\right] \overset{(141)}{\leq} \frac{18\lambda_1^{2-q}\gamma^q\sigma_2^q}{\alpha},$$

as desired.

Bounding ⑧. For this bound, if we assume that $\lambda_1 \geq 2\gamma L_1$ then, with probability one we have

$$\begin{aligned}
⑧ &:= \left\| \sum_{j=1}^{t} (1-\alpha)^{t-j} \omega_j^b \right\| \\
&\leq \sum_{j=1}^{t} (1-\alpha)^{t-j} \left\| \omega_j^b \right\| \\
&\overset{(a)}{\leq} 4\lambda_1^{1-q} \gamma^q \sigma_2^q \sum_{j=1}^{t} (1-\alpha)^{t-j} \\
&\leq \frac{4\lambda_1^{1-q} \gamma^q \sigma_2^q}{\alpha},
\end{aligned}$$

where in (a) we use Lemma G.4 (with $\mathbb{E}_{\hat{\xi}_j}[\cdot \mid q_j]$), more precisely, for any $j \in [t]$,

$$\left\| \omega_j^b \right\| \leq 2^q \lambda_1^{1-q} \gamma^q \sigma_2^q \leq 4\lambda_1^{1-q} \gamma^q \sigma_2^q,$$

since

$$\begin{aligned}
\mathbb{E}\left[\left\| \nabla^2 f(\hat{x}_j, \hat{\xi}_j)(x_j - x_{j-1}) - \nabla^2 F(\hat{x}_j)(x_j - x_{j-1}) \right\|^q \right] &\leq \mathbb{E}\left[\left\| \nabla^2 f(\hat{x}_j, \hat{\xi}_j) - \nabla^2 F(\hat{x}_j) \right\|_{\mathrm{op}}^q \cdot \|x_j - x_{j-1}\|^q \right] \\
&\overset{\text{Lem. H.2}}{=} \gamma^q \, \mathbb{E}\left[\mathbb{E}\left[\left\| \nabla^2 f(\hat{x}_j, \hat{\xi}_j) - \nabla^2 F(\hat{x}_j) \right\|_{\mathrm{op}}^q \,\middle|\, \hat{x}_j \right] \right] \\
&\overset{\text{Ass. 2.5}}{\leq} \gamma^q \sigma_2^q.
\end{aligned}$$

Bounding ⑨. For the last bound, we use the triangle inequality, this gives

$$\begin{aligned}
⑨ &:= \left\| \sum_{j=1} 4(1-\alpha)^{tj} \omega_j^r \right\| \\
&\leq \sum_{j=1}^{t} (1-\alpha)^{t-j} \left\| \omega_j^r \right\| \\
&= \sum_{j=1}^{t} (1-\alpha)^{t-j} \left\| \nabla^2 F(\hat{x}_j)(x_j - x_{j-1}) - [\nabla F(x_j) - \nabla F(x_{j-1})] \right\| \\
&\overset{(a)}{\leq} \min\left\{ 2\gamma L_1, \frac{\gamma^2 L_2}{2} \right\} \sum_{j=1}^{t} (1-\alpha)^{t-j} \\
&\leq \frac{2\min\left\{ \gamma L_1, \gamma^2 L_2 \right\}}{\alpha},
\end{aligned} \tag{142}$$

where in (a) we use the bounds (90) and (91), which holds with probability one.

Summing up all bounds ⑤, ⑥, ⑦, ⑧ and ⑨. We introduce the same events $E_{⑤,t}$ and $E_{⑥,t}$ as in (133) and (135) respectively. Then, by our choice of c_1, b_1 and G_1 we have the bound

$$\begin{aligned} b_1 := \left(\frac{c_1}{3} + \sqrt{\frac{c_1^2}{9} + 2G_1}\right)\log\frac{2}{\delta''} &\leq \left(\frac{2c_1}{3} + \sqrt{2G_1}\right)\log\frac{2}{\delta''} \\ &= \left(\frac{4\lambda_1}{3} + \sqrt{\frac{36\lambda_1^{2-q}\gamma^q\sigma_2^q}{\alpha}}\right)\log\frac{2}{\delta''} \\ &= \lambda_1\left(\frac{4}{3} + 6\sqrt{\frac{1}{\alpha}\left(\frac{\gamma\sigma_2}{\lambda_1}\right)^q}\right)\log\frac{2}{\delta''} \\ &\overset{(a)}{\leq} \frac{22\lambda_1}{3}\log\frac{2}{\delta''}, \end{aligned} \tag{143}$$

where (a) holds provided $\lambda_1 \geq \gamma\bar{\delta}\alpha^{-\frac{1}{q}}$. Moreover, by our choice of $\widetilde{c}_1$, $\widetilde{b}_1$ and $\widetilde{G}_1$ we also have the bound

$$\begin{aligned} \widetilde{b}_1 := \left(\frac{\widetilde{c}_1}{3} + \sqrt{\frac{\widetilde{c}_1^2}{9} + 2\widetilde{G}_1}\right)\log\frac{2}{\delta''} &\leq \left(\frac{2\widetilde{c}_1}{3} + \sqrt{2\widetilde{G}_1}\right)\log\frac{2}{\delta''} \\ &= \left(\frac{16\lambda_1^2}{3} + \sqrt{\frac{576\lambda_1^{4-q}\gamma^q\sigma_2^q}{\alpha}}\right)\log\frac{2}{\delta''} \\ &= \lambda_1^2\left(\frac{16}{3} + 24\sqrt{\frac{1}{\alpha}\left(\frac{\gamma\sigma_2}{\lambda_1}\right)^q}\right)\log\frac{2}{\delta''} \\ &\leq \frac{88\lambda_1^2}{3}\log\frac{2}{\delta''}, \end{aligned} \tag{144}$$

where the last inequality holds given $\lambda_1 \geq \gamma\sigma_2\alpha^{-\frac{1}{q}}$.

Next, given $\lambda_1 \geq \max\left\{2\gamma L_1, \gamma\sigma_2\alpha^{-\frac{1}{q}}\right\}$, as in the previous lemma, we have $\mathbb{P}\left(E_{⑤,t}\right) \geq 1-\delta''$ and $\mathbb{P}\left(E_{⑥,t}\right) \geq 1-\delta''$ hence, by the union bound inequality we have

$$\mathbb{P}\left(E_{⑤,t} \cap E_{⑥,t}\right) \geq 1 - 2\delta'',$$

and on the event $E_{⑤,t} \cap E_{⑥,t}$ we obtain the inequality

$$\begin{aligned} \left\|\sum_{j=1}^t (1-\alpha)^{t-j}\omega_j\right\| &\leq \left\|\sum_{j=1}^t (1-\alpha)^{t-j}\omega_j^r\right\| + \left\|\sum_{j=1}^t (1-\alpha)^{t-j}\omega_j^b\right\| + \left\|\sum_{j=1}^t (1-\alpha)^{t-j}\omega_j^u\right\| \\ &\overset{(126)+(140)}{\leq} ⑨ + ⑤ + \sqrt{2\cdot⑥ + 2\cdot⑦} + ⑧ \\ &\overset{(142)+(143)+(144)}{\leq} \frac{2\min\left\{\gamma L_1, \gamma^2 L_2\right\}}{\alpha} + \frac{22\lambda_1}{3}\log\frac{2}{\delta''} + \sqrt{2\cdot\frac{176\lambda_1^2}{3}\log\frac{2}{\delta''} + 2\cdot\frac{18\lambda_1^{2-q}\gamma^q\sigma_2^q}{\alpha}} + \frac{4\lambda_1^{1-q}\gamma^q\sigma_2^q}{\alpha} \\ &\overset{(a)}{\leq} \frac{2\min\left\{\gamma L_1, \gamma^2 L_2\right\}}{\alpha} + \lambda_1\left(\frac{22}{3} + \sqrt{\frac{176}{3} + \frac{36}{\alpha}\left(\frac{\gamma\sigma_2}{\lambda_1}\right)^q} + \frac{4}{\alpha}\left(\frac{\gamma\sigma_2}{\lambda_1}\right)^q\right)\log\frac{2}{\delta''} \\ &\overset{(b)}{\leq} \frac{2\min\left\{\gamma L_1, \gamma^2 L_2\right\}}{\alpha} + \lambda_1\left(\frac{22}{3} + \sqrt{\frac{176}{3} + 36} + 4\right)\log\frac{2}{\delta''} \\ &\leq \frac{2\min\left\{\gamma L_1, \gamma^2 L_2\right\}}{\alpha} + 22\lambda_1\log\frac{2}{\delta''}, \end{aligned}$$

where in (a) we use the fact that $\log\frac{2}{\delta''} \geq 1$ and in (b) we use $\lambda_1 \geq \gamma\sigma_2\alpha^{-\frac{1}{q}}$. This proves the lemma. □

F.2 Proof of Theorem 6.1

With Lemmas F.3 and F.4 in our hands, we are now ready to prove the main result of this section, i.e., high-probability convergence guarantees for Algorithm 3, i.e., `clip-NSGD-MVR`.

Theorem 6.1. *Under Assumptions 2.1, 2.3 and 2.6, let $T \geq 1$ and $\delta \in (0,1]$ such that $\log \frac{8T}{\delta} \geq 1$ and suppose that we choose $g_0 = 0$ in Algorithm 3 and let $\Delta_1 := F(x_0) - F^{\inf}$ the initial sub-optimality. Suppose we run Algorithm 3 using momentum parameter $\alpha = \max\{T^{-\frac{p}{2p-1}}, T^{-\frac{pq}{p(2q+1)-2q}}\}$, clipping thresholds $\lambda_1 = 2\gamma\bar{L}\alpha^{-\frac{1}{q}}$ and $\lambda_2 = \max\{4\sqrt{\bar{L}\Delta_1}, \sigma_1\alpha^{-\frac{1}{p}}\}$ and with stepsize*

$$\gamma = \mathcal{O}\left(\min\left\{\sqrt{\frac{\Delta_1}{\bar{L}T}}, \alpha\sqrt{\frac{\Delta_1}{\bar{L}}}, \frac{1}{\alpha T\log\frac{T}{\delta}}\sqrt{\frac{\Delta_1}{\bar{L}}}, \frac{\Delta_1}{\sigma_1\alpha^{\frac{p-1}{p}}T\log\frac{T}{\delta}}, \sqrt{\frac{\Delta_1\alpha^{\frac{1}{q}}}{\bar{L}T\log\frac{T}{\delta}}}\right\}\right).$$

Then, with probability at least $1-\delta$, the output of Algorithm 3 satisfies

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F(x_t)\| \leq \frac{2\Delta_1}{\gamma T},$$

and, by our choice of parameters, the norm of the gradients converges at the rate

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F(x_t)\| = \mathcal{O}\left(\left(\frac{\sqrt{\bar{L}\Delta_1}+\sigma_1}{T^{\frac{p-1}{2p-1}\wedge\frac{q(p-1)}{p(2q+1)-2q}}}\right)\log\frac{T}{\delta}\right),$$

with high probability.

Proof. Let us remind from the proofs of Lemmas F.3 and F.4 that we defined

$$E_{①,t} := \left\{\left|\sum_{j=1}^{t}V_j^t\right| \leq b_2 \text{ or } \sum_{j=1}^{t}\sigma_j^2 > G_2\log\frac{2}{\delta'}\right\}, \quad E_{②,t} := \left\{\left|\sum_{j=1}^{t}Y_j^t\right| \leq \widetilde{b}_2 \text{ or } \sum_{j=1}^{t}\widetilde{\sigma}_j^2 > \widetilde{G}_2\log\frac{2}{\delta'}\right\},$$
$$E_{⑤,t} := \left\{\left|\sum_{j=1}^{t}W_j^t\right| \leq b_1 \text{ or } \sum_{j=1}^{t}\sigma_j^2 > G_1\log\frac{2}{\delta''}\right\}, \quad E_{⑥,t} := \left\{\left|\sum_{j=1}^{t}Z_j^t\right| \leq \widetilde{b}_1 \text{ or } \sum_{j=1}^{t}\widetilde{\sigma}_j^2 > \widetilde{G}_1\log\frac{2}{\delta''}\right\},$$

where

$$c_1 := 2\lambda_1, \quad \widetilde{c}_1 := 8\lambda_1^2, \quad c_2 := 2\lambda_2, \quad \widetilde{c}_2 := 8\lambda_2^2,$$
$$G_1 := \frac{72\lambda_1^{2-q}\gamma^q\bar{L}^q}{\alpha}, \quad \widetilde{G}_1 := \frac{1152\lambda_1^{4-q}\gamma^q\bar{L}^q}{\alpha}, \quad G_2 := \frac{18\lambda_2^{2-p}\sigma_1^p}{\alpha}, \quad \widetilde{G}_2 := \frac{288\lambda_2^{4-p}\sigma_1^p}{\alpha},$$
$$b_1 \leq \frac{40\lambda_1}{3}\log\frac{2}{\delta''}, \quad \widetilde{b}_1 \leq \frac{160\lambda_1^2}{3}\log\frac{2}{\delta''}, \quad b_2 \leq \frac{22\lambda_2}{3}\log\frac{2}{\delta'}, \quad \widetilde{b}_2 \leq \frac{88\lambda_2^2}{3}\log\frac{2}{\delta'},$$

and we have shown the following bounds:

$$\mathbb{P}\left(E_{①,t}\right) \geq 1-\delta', \quad \mathbb{P}\left(E_{②,t}\right) \geq 1-\delta', \quad \mathbb{P}\left(E_{⑤,t}\right) \geq 1-\delta'', \quad \mathbb{P}\left(E_{⑥,t}\right) \geq 1-\delta'', \tag{145}$$

valid for all $t \in \{0,\ldots,T-1\}$.

The idea of the proof is based on a technique from previous works in the literature (Gorbunov et al., 2020; Sadiev et al., 2023; Liu et al., 2023; Sadiev et al., 2025) and proceeds by a mathematical induction. Our goal is to prove that for any $\tau \in \{0,\ldots,T-1\}$, the event

$$G_\tau := E_\tau \cap E_{1,\tau} \cap E_{2,\tau} \cap E_{5,\tau} \cap E_{6,\tau},$$

holds with probability at least $1-\frac{\tau\delta}{T}$, where the event E_τ is defined as

$$E_\tau := \bigcap_{t=0}^{\tau}\left\{\gamma\sum_{j=0}^{t}\|\nabla F(x_j)\| + \Delta_{t+1} \leq 2\Delta_1\right\},$$

and

$$E_{1,\tau} := \bigcap_{t=1}^{\tau}E_{①,t}, \quad E_{2,\tau} := \bigcap_{t=1}^{\tau}E_{②,t}, \quad E_{5,\tau} := \bigcap_{t=1}^{\tau}E_{⑤,t}, \quad E_{6,\tau} := \bigcap_{t=1}^{\tau}E_{⑥,t}.$$

Base case. For the base case $\tau = 0$ we have $G_0 = E_0 = \{\Delta_1 \leq 2\Delta_1\}$ which holds with probability $1 = 1 - \frac{\tau\delta}{T}$ since $\tau = 0$.

Inductive case. Now, let us assume that the induction hypothesis holds for $\tau - 1$, i.e., $\mathbb{P}\left(G_{\tau-1}\right) \geq 1 - \frac{(\tau-1)\delta}{T}$ and we need to prove that is holds for τ also, that is, $\mathbb{P}\left(G_\tau\right) \geq 1 - \frac{\tau\delta}{T}$. First, let us observe that

$$E_{\tau-1} \cap E_{1,\tau} \cap E_{2,\tau} \cap E_{5,\tau} \cap E_{6,\tau} = G_{\tau-1} \cap E_{①,\tau} \cap E_{②,\tau} \cap E_{⑤,\tau} \cap E_{⑥,\tau}. \tag{146}$$

Then, we have

$$\begin{aligned}
\mathbb{P}\left(E_{\tau-1} \cap E_{1,\tau} \cap E_{2,\tau} \cap E_{5,\tau} \cap E_{6,\tau}\right) &\overset{(146)}{=} \mathbb{P}\left(G_{\tau-1} \cap E_{①,\tau} \cap E_{②,\tau} \cap E_{⑤,\tau} \cap E_{⑥,\tau}\right) \\
&= 1 - \mathbb{P}\left(\overline{G_{\tau-1} \cap E_{①,\tau} \cap E_{②,\tau} \cap E_{⑤,\tau} \cap E_{⑥,\tau}}\right) \\
&\geq 1 - \mathbb{P}\left(\overline{G_{\tau-1}}\right) - \mathbb{P}\left(\overline{E_{①,\tau}}\right) - \mathbb{P}\left(\overline{E_{②,\tau}}\right) - \mathbb{P}\left(\overline{E_{⑤,\tau}}\right) - \mathbb{P}\left(\overline{E_{⑥,\tau}}\right) \\
&\overset{(145)}{\geq} 1 - \frac{(\tau-1)\delta}{T} - 2\delta' - 2\delta'' \\
&= 1 - \frac{\tau\delta}{T} + \left(\frac{\delta}{T} - 2\delta' - 2\delta''\right) \\
&= 1 - \frac{\tau\delta}{T},
\end{aligned} \tag{147}$$

provided $\delta' = \delta'' = \frac{\delta}{4T}$. Now, given $\tau \in \{0, \ldots, T-1\}$, on the event $E_{\tau-1}$ we have

$$\Delta_t \leq \gamma \sum_{j=0}^{t-1} \|\nabla F(x_j)\| + \Delta_t \leq 2\Delta_1$$

for all integer $1 \leq t \leq \tau$ and therefore, for all $j \in \{0, \ldots, \tau-1\}$, we have

$$\|\nabla F(x_j)\| \leq \sqrt{2\bar{L}\Delta_j} \leq 2\sqrt{\bar{L}\Delta_1},$$

hence, by our choice of clipping threshold $\lambda_2 = \max\left\{4\sqrt{\bar{L}\Delta_1}, \sigma_1\alpha^{-\frac{1}{p}}\right\}$ we have $\lambda_2 \geq 2\max_{j\in[t]}\|\nabla F(x_j)\|$. Additionally, if we assume the clipping level $\lambda_1 = 2\gamma\bar{L}\alpha^{-\frac{1}{q}}$ then, on the event $E_{\tau-1} \cap E_{1,\tau} \cap E_{2,\tau} \cap E_{5,\tau} \cap E_{6,\tau}$, we have

$$\begin{aligned}
\gamma \sum_{t=0}^{\tau} \|\nabla F(x_t)\| + \Delta_{\tau+1} &\overset{\text{Lem. F.1}}{\leq} \Delta_1 + \frac{2\gamma\sqrt{2\bar{L}\Delta_1}}{\alpha} + \frac{\gamma^2\bar{L}(\tau+1)}{2} \\
&\quad + 2\gamma\alpha \sum_{t=1}^{\tau} \left\| \sum_{j=1}^{t} (1-\alpha)^{t-j}\theta_j \right\| + 2\gamma(1-\alpha) \sum_{t=1}^{\tau} \left\| \sum_{j=1}^{t} (1-\alpha)^{t-j}\omega_j \right\| \\
&\overset{\text{Lem. F.3+F.4}}{\leq} \Delta_1 + \frac{2\gamma\sqrt{2\bar{L}\Delta_1}}{\alpha} + \frac{\gamma^2\bar{L}(\tau+1)}{2} \\
&\quad + 44\gamma\alpha\tau\lambda_2 \log\frac{8T}{\delta} + 92\gamma(1-\alpha)\tau\lambda_1 \log\frac{8T}{\delta},
\end{aligned}$$

and, as $0 \leq \tau \leq T-1$ then, by our choice of stepsize

$$\gamma = \min\left\{\sqrt{\frac{\Delta_1}{2\bar{L}T}}, \frac{\alpha}{8}\sqrt{\frac{\Delta_1}{2\bar{L}}}, \frac{1}{704\alpha T\log\frac{8T}{\delta}}\sqrt{\frac{\Delta_1}{\bar{L}}}, \frac{\Delta_1}{176\sigma_1\alpha^{\frac{p-1}{p}}T\log\frac{8T}{\delta}}, \sqrt{\frac{\Delta_1\alpha^{\frac{1}{q}}}{736\bar{L}T\log\frac{8T}{\delta}}}\right\}, \tag{148}$$

we have

$$\gamma \sum_{t=0}^{\tau} \|\nabla F(x_t)\| + \Delta_{\tau+1} \overset{(148)}{\leq} \Delta_1 + \frac{\Delta_1}{4} + \frac{\Delta_1}{4} + \frac{\Delta_1}{4} + \frac{\Delta_1}{4} = 2\Delta_1,$$

therefore we have $E_\tau \cap E_{1,\tau} \cap E_{2,\tau} \cap E_{5,\tau} \cap E_{6,\tau} = E_{\tau-1} \cap E_{1,\tau} \cap E_{2,\tau} \cap E_{5,\tau} \cap E_{6,\tau}$ which leads to

$$\begin{aligned}\mathbb{P}(G_\tau) &= \mathbb{P}(E_\tau \cap E_{1,\tau} \cap E_{2,\tau} \cap E_{5,\tau} \cap E_{6,\tau}) \\ &= \mathbb{P}(E_{\tau-1} \cap E_{1,\tau} \cap E_{2,\tau} \cap E_{5,\tau} \cap E_{6,\tau}) \\ &\overset{(147)}{\geq} 1 - \frac{\tau\delta}{T},\end{aligned} \tag{149}$$

as claimed. This achieves the proof of the induction.

Bounding $\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F(x_t)\|$ in high-probability. From the previous paragraph, we have for $\tau = T$

$$\mathbb{P}(E_T) \geq \mathbb{P}(G_T) \geq 1 - \delta,$$

hence, with probability at least $1-\delta$ we have

$$\gamma \sum_{t=0}^{T-1} \|\nabla F(x_t)\| + \Delta_T \leq 2\Delta_1,$$

which implies the bound

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla F(x_t)\| \leq \frac{2\Delta_1}{\gamma T},$$

with probability at least $1-\delta$. By our choice of stepsize (148) we obtain

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla F(x_t)\| \leq \frac{2\Delta_1}{\gamma T} = \mathcal{O}\left(\max\left\{\sqrt{\frac{\bar{L}\Delta_1}{T}}, \frac{\sqrt{\bar{L}\Delta_1}}{\alpha T}, \alpha\sqrt{\bar{L}\Delta_1}\log\frac{T}{\delta}, \sigma_1\alpha^{\frac{p-1}{p}}\log\frac{T}{\delta}, \sqrt{\frac{\bar{L}\Delta_1\log\frac{T}{\delta}}{T\alpha^{\frac{1}{q}}}}\right\}\right),$$

and choosing $\alpha = \max\left\{T^{-\frac{p}{2p-1}}, T^{-\frac{pq}{p(2q+1)-2q}}\right\}$ we have

$$\alpha^{\frac{p-1}{p}} = \max\left\{T^{-\frac{p-1}{2p-1}}, T^{-\frac{q(p-1)}{p(2q+1)-2q}}\right\}, \quad \text{and} \quad T^{\frac{1}{2}}\alpha^{\frac{1}{2q}} \geq T^{\frac{1}{2}\left(1-\frac{p}{p(2q+1)-2q}\right)} = T^{\frac{q(p-1)}{p(2q+1)-2q}},$$

and

$$\alpha T \geq T^{1-\frac{p}{2p-1}} = T^{\frac{p-1}{2p-1}},$$

thus, assuming $\log\frac{8T}{\delta} \geq 1$ we get

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla F(x_t)\| &= \mathcal{O}\left(\max\left\{\sqrt{\frac{\bar{L}\Delta_1}{T}}, \frac{\sqrt{\bar{L}\Delta_1}}{T^{\frac{p-1}{2p-1}}}\log\frac{T}{\delta}, \frac{\sqrt{\bar{L}\Delta_1}}{T^{\frac{q(p-1)}{p(2q+1)-2q}}}\log\frac{T}{\delta}, \frac{\sigma_1}{T^{\frac{p-1}{2p-1}}}\log\frac{T}{\delta}, \frac{\sigma_1}{T^{\frac{q(p-1)}{p(2q+1)-2q}}}\log\frac{T}{\delta}, \frac{\sqrt{\bar{L}\Delta_1\log\frac{T}{\delta}}}{T^{\frac{q(p-1)}{p(2q+1)-2q}}}\right\}\right) \\ &\overset{(a)}{=} \mathcal{O}\left(\left(\frac{\sqrt{\bar{L}\Delta_1}+\sigma_1}{T^{\frac{p-1}{2p-1}\wedge\frac{q(p-1)}{p(2q+1)-2q}}}\right)\log\frac{T}{\delta}\right)\end{aligned}$$

since $\frac{p-1}{2p-1} \leq \frac{1}{2}$. In (a) we use $\wedge$ to denote the minimum between the two exponents $\frac{p-1}{2p-1}$ and $\frac{q(p-1)}{p(2q+1)-2q}$.

This concludes the proof of the theorem. □

F.3 Proof of Theorem F.1

Theorem F.1. *Under Assumptions 2.1 to 2.3 and 2.7, let $T \geq 1$ and $\delta \in (0,1]$ such that $\log\frac{8T}{\delta} \geq 1$ and suppose that we choose $g_0 = 0$ in Algorithm 3 and let $\Delta_1 := F(x_0) - F^{\inf}$ the initial sub-optimality. Suppose we run Algorithm 3 using momentum parameter $\alpha = \max\{T^{-\frac{p}{2p-1}}, T^{-\frac{pq}{p(2q+1)-2q}}\}$, clipping thresholds $\lambda_1 = \max\{2\gamma L_1, \gamma\bar{\delta}\alpha^{-\frac{1}{q}}\}$ and $\lambda_2 = \max\{4\sqrt{L_1\Delta_1}, \sigma_1\alpha^{-\frac{1}{p}}\}$ and with stepsize*

$$\gamma = \mathcal{O}\left(\min\left\{\alpha\sqrt{\frac{\Delta_1}{L_1}}, \frac{1}{\alpha T\log\frac{T}{\delta}}\sqrt{\frac{\Delta_1}{L_1}}, \frac{\Delta_1}{\sigma_1\alpha^{\frac{p-1}{p}}T\log\frac{T}{\delta}}, \sqrt{\frac{\Delta_1\alpha^{\frac{1}{q}}}{\bar{\delta}T\log\frac{T}{\delta}}}, \sqrt{\frac{\Delta_1}{L_1T\log\frac{T}{\delta}}}\right\}\right).$$

Then, with probability at least $1-\delta$, the output of Algorithm 3 satisfies

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F(x_t)\| \leq \frac{2\Delta_1}{\gamma T},$$

and, by our choice of parameters, the norm of the gradients converges at the rate

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F(x_t)\| = \mathcal{O}\left(\left(\frac{\sqrt{L_1\Delta_1}+\sigma_1}{T^{\frac{p-1}{2p-1}\wedge\frac{q(p-1)}{p(2q+1)-2q}}}\right)\log\frac{T}{\delta} + \frac{\sqrt{\bar{\delta}\Delta_1}}{T^{\frac{q(p-1)}{p(2q+1)-2q}}}\sqrt{\log\frac{T}{\delta}}\right),$$

with high probability.

Proof. From the proofs of Lemmas F.3 and F.5 we have

$$E_{①,t} := \left\{\left|\sum_{j=1}^t V_j^t\right| \leq b_2 \text{ or } \sum_{j=1}^t \sigma_j^2 > G_2\log\frac{2}{\delta'}\right\}, \quad E_{②,t} := \left\{\left|\sum_{j=1}^t Y_j^t\right| \leq \widetilde{b}_2 \text{ or } \sum_{j=1}^t \widetilde{\sigma}_j^2 > \widetilde{G}_2\log\frac{2}{\delta'}\right\},$$
$$E_{⑤,t} := \left\{\left|\sum_{j=1}^t W_j^t\right| \leq b_1 \text{ or } \sum_{j=1}^t \sigma_j^2 > G_1\log\frac{2}{\delta''}\right\}, \quad E_{⑥,t} := \left\{\left|\sum_{j=1}^t Z_j^t\right| \leq \widetilde{b}_1 \text{ or } \sum_{j=1}^t \widetilde{\sigma}_j^2 > \widetilde{G}_1\log\frac{2}{\delta''}\right\},$$

where

$$c_1 := 2\lambda_1, \quad \widetilde{c}_1 := 8\lambda_1^2, \quad c_2 := 2\lambda_2, \quad \widetilde{c}_2 := 8\lambda_2^2,$$
$$G_1 := \frac{18\lambda_1^{2-q}\gamma^q\bar{\delta}^q}{\alpha}, \quad \widetilde{G}_1 := \frac{288\lambda_1^{4-q}\gamma^q\bar{\delta}^q}{\alpha}, \quad G_2 := \frac{18\lambda_2^{2-p}\sigma_1^p}{\alpha}, \quad \widetilde{G}_2 := \frac{288\lambda_2^{4-p}\sigma_1^p}{\alpha},$$
$$b_1 \leq \frac{22\lambda_1}{3}\log\frac{2}{\delta''}, \quad \widetilde{b}_1 \leq \frac{88\lambda_1^2}{3}\log\frac{2}{\delta''}, \quad b_2 \leq \frac{22\lambda_2}{3}\log\frac{2}{\delta'}, \quad \widetilde{b}_2 \leq \frac{88\lambda_2^2}{3}\log\frac{2}{\delta'},$$

and we have shown the following bounds:

$$\mathbb{P}\left(E_{①,t}\right) \geq 1-\delta', \quad \mathbb{P}\left(E_{②,t}\right) \geq 1-\delta', \quad \mathbb{P}\left(E_{⑤,t}\right) \geq 1-\delta'', \quad \mathbb{P}\left(E_{⑥,t}\right) \geq 1-\delta'', \tag{150}$$

valid for all $t \in \{0,\ldots,T-1\}$.

Now, we follows the exact same steps as in the proof of the previous theorem, up to (147). Then, given $\tau \in \{0,\ldots,T-1\}$, on the event $E_{\tau-1}$ we have

$$\Delta_t \leq \gamma\sum_{j=0}^{t-1}\|\nabla F(x_j)\| + \Delta_t \leq 2\Delta_1$$

for all integer $1 \leq t \leq \tau$ and therefore, for all $j \in \{0,\ldots,\tau-1\}$, we have

$$\|\nabla F(x_j)\| \leq \sqrt{2\bar{L}\Delta_j} \leq 2\sqrt{L_1\Delta_1},$$

hence, by our choice of clipping threshold $\lambda_2 = \max\left\{4\sqrt{L_1\Delta_1}, \sigma_1\alpha^{-\frac{1}{p}}\right\}$ we have $\lambda_2 \geq 2\max_{j\in[t]}\|\nabla F(x_j)\|$. Additionally, if we assume the clipping level $\lambda_1 = \max\left\{2\gamma L_1, \gamma\bar{\delta}\alpha^{-\frac{1}{q}}\right\}$ then, on the event $E_{\tau-1}\cap E_{1,\tau}\cap E_{2,\tau}\cap E_{5,\tau}\cap E_{6,\tau}$, we have

$$\begin{aligned}
\gamma\sum_{t=0}^{\tau}\|\nabla F(x_t)\| + \Delta_{\tau+1} &\overset{\text{Lem. F.1}}{\leq} \Delta_1 + \frac{2\gamma\sqrt{2L_1\Delta_1}}{\alpha} + \frac{\gamma^2 L_1(\tau+1)}{2} \\
&\quad + 2\gamma\alpha\sum_{t=1}^{\tau}\left\|\sum_{j=1}^t(1-\alpha)^{t-j}\theta_j\right\| + 2\gamma(1-\alpha)\sum_{t=1}^{\tau}\left\|\sum_{j=1}^t(1-\alpha)^{t-j}\omega_j\right\| \\
&\overset{\text{Lem. F.3+F.5}}{\leq} \Delta_1 + \frac{2\gamma\sqrt{2L_1\Delta_1}}{\alpha} + \frac{\gamma^2 L_1(\tau+1)}{2} \\
&\quad + 44\gamma\alpha\tau\lambda_2\log\frac{8T}{\delta} + 44\gamma(1-\alpha)\tau\lambda_1\log\frac{8T}{\delta},
\end{aligned} \tag{151}$$

and, as $0 \leq \tau \leq T-1$ then, by our choice of stepsize

$$\gamma = \min\left\{\sqrt{\frac{\Delta_1}{2L_1T}}, \frac{\alpha}{8}\sqrt{\frac{\Delta_1}{2L_1}}, \frac{1}{704\alpha T\log\frac{8T}{\delta}}\sqrt{\frac{\Delta_1}{L_1}}, \frac{\Delta_1}{176\sigma_1\alpha^{\frac{p-1}{p}}T\log\frac{8T}{\delta}}, \sqrt{\frac{\Delta_1}{352L_1T\log\frac{8T}{\delta}}}, \sqrt{\frac{\Delta_1\alpha^{\frac{1}{q}}}{176\bar{\delta}T\log\frac{8T}{\delta}}}\right\}, \tag{152}$$

we have

$$\gamma\sum_{t=0}^{\tau}\|\nabla F(x_t)\| + \Delta_{\tau+1} \overset{(152)}{\leq} \Delta_1 + \frac{\Delta_1}{4} + \frac{\Delta_1}{4} + \frac{\Delta_1}{4} + \frac{\Delta_1}{4} = 2\Delta_1,$$

and we can achieve the proof by induction as in (149).

Then, as before, with probability as least $1-\delta$, we have

$$\gamma\sum_{t=0}^{T-1}\|\nabla F(x_t)\| + \Delta_T \leq 2\Delta_1,$$

which implies the bound

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F(x_t)\| \leq \frac{2\Delta_1}{\gamma T},$$

with probability at least $1-\delta$. By our choice of stepsize (152) we obtain

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F(x_t)\| \leq \frac{2\Delta_1}{\gamma T} &= \mathcal{O}\left(\max\left\{\sqrt{\frac{L_1\Delta_1}{T}}, \frac{\sqrt{L_1\Delta_1}}{\alpha T}, \alpha\sqrt{L_1\Delta_1}\log\frac{T}{\delta}, \sigma_1\alpha^{\frac{p-1}{p}}\log\frac{T}{\delta}, \sqrt{\frac{L_1\Delta_1\log\frac{T}{\delta}}{T}}, \sqrt{\frac{\bar{\delta}\Delta_1\log\frac{T}{\delta}}{T\alpha^{\frac{1}{q}}}}\right\}\right)\\ &= \mathcal{O}\left(\max\left\{\frac{\sqrt{L_1\Delta_1}}{\alpha T}, \alpha\sqrt{L_1\Delta_1}\log\frac{T}{\delta}, \sigma_1\alpha^{\frac{p-1}{p}}\log\frac{T}{\delta}, \sqrt{\frac{L_1\Delta_1\log\frac{T}{\delta}}{T}}, \sqrt{\frac{\bar{\delta}\Delta_1\log\frac{T}{\delta}}{T\alpha^{\frac{1}{q}}}}\right\}\right),\end{aligned}$$

as we assumed $\log\frac{8T}{\delta} \geq 1$. Now, choosing $\alpha = \max\left\{T^{-\frac{p}{2p-1}}, T^{-\frac{pq}{p(2q+1)-2q}}\right\}$ we have

$$\alpha^{\frac{p-1}{p}} = \max\left\{T^{-\frac{p-1}{2p-1}}, T^{-\frac{q(p-1)}{p(2q+1)-2q}}\right\}, \quad \text{and} \quad T^{\frac{1}{2}}\alpha^{\frac{1}{2q}} \geq T^{\frac{1}{2}\left(1-\frac{p}{p(2q+1)-2q}\right)} = T^{\frac{q(p-1)}{p(2q+1)-2q}},$$

and

$$\alpha T \geq T^{1-\frac{p}{2p-1}} = T^{\frac{p-1}{2p-1}},$$

thus, given $\log\frac{8T}{\delta} \geq 1$ we get

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F(x_t)\| &= \mathcal{O}\left(\max\left\{\frac{\sqrt{L_1\Delta_1}}{T^{\frac{p-1}{2p-1}}}\log\frac{T}{\delta}, \frac{\sqrt{L_1\Delta_1}}{T^{\frac{q(p-1)}{p(2q+1)-2q}}}\log\frac{T}{\delta}, \frac{\sigma_1}{T^{\frac{p-1}{2p-1}}}\log\frac{T}{\delta}, \frac{\sigma_1}{T^{\frac{q(p-1)}{p(2q+1)-2q}}}\log\frac{T}{\delta}, \frac{\sqrt{\bar{\delta}\Delta_1\log\frac{T}{\delta}}}{T^{\frac{q(p-1)}{p(2q+1)-2q}}}\right\}\right)\\ &\overset{(a)}{=} \mathcal{O}\left(\left(\frac{\sqrt{L_1\Delta_1}+\sigma_1}{T^{\frac{p-1}{2p-1}\wedge\frac{q(p-1)}{p(2q+1)-2q}}}\right)\log\frac{T}{\delta} + \frac{\sqrt{\bar{\delta}\Delta_1}}{T^{\frac{q(p-1)}{p(2q+1)-2q}}}\sqrt{\log\frac{T}{\delta}}\right)\end{aligned}$$

since $\frac{p-1}{2p-1} \leq \frac{1}{2}$ and $\sqrt{\log\frac{T}{\delta}} \lesssim \log\frac{T}{\delta}$. In (a) we use $\wedge$ to denote the minimum between the two exponents $\frac{p-1}{2p-1}$ and $\frac{q(p-1)}{p(2q+1)-2q}$.

This concludes the proof of the theorem. □

F.4 Proof of Theorem F.2

Theorem F.2. *Under Assumptions 2.1 to 2.5, let $T \geq 1$ and $\beta \in (0,1]$ such that $\log\frac{8T}{\beta} \geq 1$ and suppose that we choose $g_0 = 0$ in Algorithm 3 and let $\Delta_1 := F(x_0) - F^{\inf}$ the initial sub-optimality. Suppose we run Algorithm 3*

Algorithm 4: Clipped NSGD-Hess (Clipped Normalized SGD with Hessian-corrected Momentum)

Initialization:
- $x_0 \in \mathbb{R}^d$, the starting point
- $T > 0$, the number of iterations
- $g_0 \in \mathbb{R}^d$, an initial vector
- $\gamma > 0$, the stepsize
- $\alpha \in (0, 1]$, the momentum parameter for Hess
- $\lambda_1, \lambda_2 > 0$, the clipping thresholds

$x_1 \leftarrow x_0 - \gamma \frac{g_0}{\|g_0\|}$

For $t = 1, 2, \ldots, T-1$ **do**

Sample $q_t \sim \mathcal{U}([0, 1])$

$\hat{x}_t \leftarrow q_t x_t + (1 - q_t) x_{t-1}$

// Apply Hess, here $\xi_t, \hat{\xi}_t \sim \mathcal{D}$ are independent.

$g_t \leftarrow (1-\alpha)\left(g_{t-1} + \texttt{clip}\left(\nabla^2 f(\hat{x}_t, \hat{\xi}_t)(x_t - x_{t-1}), \lambda_1\right)\right) + \alpha \texttt{clip}\left(\nabla f\left(x_t, \xi_t\right), \lambda_2\right)$

// Do one descent step.

$x_{t+1} \leftarrow x_t - \gamma \frac{g_t}{\|g_t\|}$

Output: x_T

using momentum parameter $\alpha = \max\{T^{-\frac{p}{2p-1}}, T^{-\frac{pq}{p(2q+1)-2q}}\}$, *clipping thresholds* $\lambda_1 = \max\{2\gamma L_1, \gamma\sigma_2\alpha^{-\frac{1}{q}}\}$ *and* $\lambda_2 = \max\{4\sqrt{L_1\Delta_1}, \sigma_1\alpha^{-\frac{1}{p}}\}$ *and with stepsize*

$$\gamma = \mathcal{O}\left(\min\left\{\alpha\sqrt{\frac{\Delta_1}{L_1}}, \sqrt[3]{\frac{\Delta_1\alpha}{L_2 T}}, \frac{1}{\alpha T\log\frac{T}{\beta}}\sqrt{\frac{\Delta_1}{L_1}}, \frac{\Delta_1}{\sigma_1\alpha^{\frac{p-1}{p}}T\log\frac{T}{\beta}}, \sqrt{\frac{\Delta_1\alpha^{\frac{1}{q}}}{\sigma_2 T\log\frac{T}{\beta}}}, \sqrt{\frac{\Delta_1}{L_1 T\log\frac{T}{\beta}}}\right\}\right).$$

Then, with probability at least $1-\beta$, *the output of Algorithm 3 satisfies*

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F(x_t)\| \leq \frac{2\Delta_1}{\gamma T},$$

and, by our choice of parameters, the norm of the gradients converges at the rate

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F(x_t)\| = \mathcal{O}\left(\frac{\left(L_2^{1/2}\Delta_1\right)^{\frac{2}{3}}}{T^{\frac{2(p-1)}{4p-3}}} + \left(\frac{\sqrt{L_1\Delta_1}+\sigma_1}{T^{\frac{p-1}{2p-1}\wedge\frac{q(p-1)}{p(2q+1)-2q}}}\right)\log\frac{T}{\beta} + \frac{\sqrt{\sigma_2\Delta_1}}{T^{\frac{q(p-1)}{p(2q+1)-2q}}}\sqrt{\log\frac{T}{\beta}}\right),$$

with high probability.

Proof. The proof readily follows from the proof of Theorem F.1 where $\bar{\delta}$ is replaced by σ_2. Additionally, by Lemma F.6 we need to bound an extra term in (151), i.e., we have

$$\begin{aligned}
\gamma\sum_{t=0}^{\tau}\|\nabla F(x_t)\| + \Delta_{\tau+1} &\overset{\text{Lem. F.1}}{\leq} \Delta_1 + \frac{2\gamma\sqrt{2L_1\Delta_1}}{\alpha} + \frac{\gamma^2 L_1(\tau+1)}{2} \\
&\quad + 2\gamma\alpha\sum_{t=1}^{\tau}\left\|\sum_{j=1}^{t}(1-\alpha)^{t-j}\theta_j\right\| + 2\gamma(1-\alpha)\sum_{t=1}^{\tau}\left\|\sum_{j=1}^{t}(1-\alpha)^{t-j}\omega_j\right\| \\
&\overset{\text{Lem. F.6+F.6}}{\leq} \Delta_1 + \frac{2\gamma\sqrt{2L_1\Delta_1}}{\alpha} + \frac{\gamma^2 L_1(\tau+1)}{2} + \frac{4(1-\alpha)\gamma^3 L_2(\tau+1)}{\alpha} \\
&\quad + 44\gamma\alpha\tau\lambda_2\log\frac{8T}{\delta} + 44\gamma(1-\alpha)\tau\lambda_1\log\frac{8T}{\delta},
\end{aligned}$$

where in the last inequality, we drop the $\min\{\cdots\}$ term and simply keep the $\gamma^3 L_2$ term. Additionally, as we have $0 \leq \tau \leq T-1$ then choosing

$$\gamma = \min\left\{\sqrt{\frac{2\Delta_1}{5L_1T}}, \frac{\alpha}{10}\sqrt{\frac{\Delta_1}{2L_1}}, \sqrt[3]{\frac{\Delta_1\alpha}{20L_2T}}, \frac{1}{880\alpha T\log\frac{8T}{\delta}}\sqrt{\frac{\Delta_1}{L_1}}, \frac{\Delta_1}{220\sigma_1\alpha^{\frac{p-1}{p}}T\log\frac{8T}{\delta}}, \sqrt{\frac{\Delta_1}{440L_1T\log\frac{8T}{\delta}}}, \sqrt{\frac{\Delta_1\alpha^{\frac{1}{q}}}{220\sigma_2T\log\frac{8T}{\delta}}}\right\}, \tag{153}$$

gives

$$\gamma\sum_{t=0}^{\tau}\|\nabla F(x_t)\| + \Delta_{\tau+1} \overset{(153)}{\leq} \Delta_1 + \frac{\Delta_1}{5} + \frac{\Delta_1}{5} + \frac{\Delta_1}{5} + \frac{\Delta_1}{5} + \frac{\Delta_1}{5} = 2\Delta_1.$$

Then, as before, with probability as least $1-\delta$, we have

$$\gamma\sum_{t=0}^{T-1}\|\nabla F(x_t)\| + \Delta_T \leq 2\Delta_1,$$

which implies the bound

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F(x_t)\| \leq \frac{2\Delta_1}{\gamma T},$$

with probability at least $1-\delta$. By our choice of stepsize (153) we obtain

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F(x_t)\| \leq \frac{2\Delta_1}{\gamma T} &= \mathcal{O}\left(\max\left\{\sqrt{\frac{L_1\Delta_1}{T}}, \frac{\sqrt{L_1\Delta_1}}{\alpha T}, \alpha^{-\frac{1}{3}}\left(\frac{L_2^{1/2}\Delta_1}{T}\right)^{\frac{2}{3}}, \alpha\sqrt{L_1\Delta_1}\log\frac{T}{\delta}, \sigma_1\alpha^{\frac{p-1}{p}}\log\frac{T}{\delta},\right.\right.\\ &\qquad\left.\left.\sqrt{\frac{L_1\Delta_1\log\frac{T}{\delta}}{T}}, \sqrt{\frac{\sigma_2\Delta_1\log\frac{T}{\delta}}{T\alpha^{\frac{1}{q}}}}\right\}\right)\\ &= \mathcal{O}\left(\max\left\{\frac{\sqrt{L_1\Delta_1}}{\alpha T}, \alpha^{-\frac{1}{3}}\left(\frac{L_2^{1/2}\Delta_1}{T}\right)^{\frac{2}{3}}, \alpha\sqrt{L_1\Delta_1}\log\frac{T}{\delta}, \sigma_1\alpha^{\frac{p-1}{p}}\log\frac{T}{\delta},\right.\right.\\ &\qquad\left.\left.\sqrt{\frac{L_1\Delta_1\log\frac{T}{\delta}}{T}}, \sqrt{\frac{\sigma_2\Delta_1\log\frac{T}{\delta}}{T\alpha^{\frac{1}{q}}}}\right\}\right),\end{aligned}$$

as we assumed $\log\frac{8T}{\delta} \geq 1$. Now, choosing

$$\alpha = \max\left\{T^{-\frac{p}{2p-1}}, T^{-\frac{pq}{p(2q+1)-2q}}, T^{-\frac{2p}{4p-3}}\right\} = \max\left\{T^{-\frac{p}{2p-1}}, T^{-\frac{pq}{p(2q+1)-2q}}\right\},$$

since $\frac{p}{2p-1} \leq \frac{2p}{4p-3}$, we have

$$\alpha^{\frac{p-1}{p}} = \max\left\{T^{-\frac{p-1}{2p-1}}, T^{-\frac{q(p-1)}{p(2q+1)-2q}}\right\}, \quad T^{\frac{1}{2}}\alpha^{\frac{1}{2q}} \geq T^{\frac{1}{2}\left(1-\frac{p}{p(2q+1)-2q}\right)} = T^{\frac{q(p-1)}{p(2q+1)-2q}},$$

and

$$\alpha T \geq T^{1-\frac{p}{2p-1}} = T^{\frac{p-1}{2p-1}}, \quad T^{\frac{2}{3}}\alpha^{\frac{1}{3}} \geq T^{\frac{2}{3}-\frac{2p}{3(4p-3)}} = T^{-\frac{2(p-1)}{4p-3}}$$

thus, given $\log\frac{8T}{\delta} \geq 1$ we get

$$\begin{aligned}&\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F(x_t)\|\\ &\quad= \mathcal{O}\left(\max\left\{\frac{\left(L_2^{1/2}\Delta_1\right)^{\frac{2}{3}}}{T^{\frac{2(p-1)}{4p-3}}}, \frac{\sqrt{L_1\Delta_1}}{T^{\frac{p-1}{2p-1}}}\log\frac{T}{\delta}, \frac{\sqrt{L_1\Delta_1}}{T^{\frac{q(p-1)}{p(2q+1)-2q}}}\log\frac{T}{\delta}, \frac{\sigma_1}{T^{\frac{p-1}{2p-1}}}\log\frac{T}{\delta}, \frac{\sigma_1}{T^{\frac{q(p-1)}{p(2q+1)-2q}}}\log\frac{T}{\delta}, \frac{\sqrt{\sigma_2\Delta_1\log\frac{T}{\delta}}}{T^{\frac{q(p-1)}{p(2q+1)-2q}}}\right\}\right)\\ &\quad\overset{(a)}{=} \mathcal{O}\left(\frac{\left(L_2^{1/2}\Delta_1\right)^{\frac{2}{3}}}{T^{\frac{2(p-1)}{4p-3}}} + \left(\frac{\sqrt{L_1\Delta_1}+\sigma_1}{T^{\frac{p-1}{2p-1}\wedge\frac{q(p-1)}{p(2q+1)-2q}}}\right)\log\frac{T}{\delta} + \frac{\sqrt{\sigma_2\Delta_1}}{T^{\frac{q(p-1)}{p(2q+1)-2q}}}\sqrt{\log\frac{T}{\delta}}\right),\end{aligned}$$

since $\frac{p-1}{2p-1} \leq \frac{1}{2}$ and $\sqrt{\log \frac{T}{\delta}} \lesssim \log \frac{T}{\delta}$. In (a) we use $\wedge$ to denote the minimum between the two exponents $\frac{p-1}{2p-1}$ and $\frac{q(p-1)}{p(2q+1)-2q}$.

This concludes the proof of the theorem. $\square$

G TECHNICAL LEMMAS

We list below the technical lemmas used throughout this paper.

Lemma G.1 (Hübler et al. (2025, Lemma 10)). *Let $p \in [1,2]$, and $X_1, \ldots, X_n \in \mathbb{R}^d$ be a martingale difference sequence (MDS), i.e., $\mathbb{E}\left[X_j \mid X_{j-1}, \ldots, X_1\right] = 0$ a.s. for all $j = 1, \ldots, n$ satisfying*

$$\mathbb{E}\left[\|X_j\|^p\right] < \infty, \qquad \text{for all } j = 1, \ldots, n.$$

Define $S_n := \sum_{j=1}^n X_j$, then

$$\mathbb{E}\left[\|S_n\|^p\right] \leq 2 \sum_{j=1}^n \mathbb{E}\left[\|X_j\|^p\right].$$

The following lemma will be useful to derive high-probability convergence bounds. The constant 2^{2-p} appearing in (154) is a consequence of He et al. (2025, Lemma 1) which we recall in Lemma H.7.

Lemma G.2 ((Cutkosky and Mehta, 2021, Lemma 10)). *Let $X_1, \ldots, X_n \in \mathbb{R}^d$ be a sequence of random vectors. Let us define the following sequence $w_1, \ldots, w_n$ recursively by*

1. *$w_0 = 0$,*
2. *if $\sum_{i=1}^{j-1} X_i \neq 0$, then we set*

$$w_j = \operatorname{sign}\left(\sum_{i=1}^{j-1} w_i\right) \frac{\left\langle \sum_{i=1}^{j-1} X_i, X_j \right\rangle}{\left\| \sum_{i=1}^{j-1} X_i \right\|},$$

3. *if $\sum_{i=1}^{j-1} X_i = 0$, we set $w_j = 0$.*

Then, $|w_j| \leq \|X_j\|$ for all $j \in \{1, \ldots, n\}$ and

$$\left\| \sum_{i=1}^n X_i \right\| \leq \left| \sum_{i=1}^n w_i \right| + \left(\max_{i \in [n]} \|X_i\|^p + 2^{2-p} \sum_{i=1}^n \|X_i\|^p \right)^{\frac{1}{p}}. \tag{154}$$

Lemma G.3 (Freedman's Inequality). *Let $X_1, \ldots, X_n \in \mathbb{R}^d$ be a martingale difference sequence (MDS), i.e., $\mathbb{E}\left[X_j \mid X_{j-1}, \ldots, X_1\right] = 0$ a.s. for $j = 1, \ldots, n$. Assume there exists a constant $c > 0$ such that $|X_j| \leq c$ a.s. for $j = 1, \ldots, n$ and define $\sigma_j^2 := \mathbb{E}\left[X_j^2 \mid X_{j-1}, \ldots, X_1\right]$. Then, for all $b > 0$ and all $G > 0$ we have*

$$\mathbb{P}\left(\left| \sum_{j=1}^n X_j \right| > b \text{ and } \sum_{j=1}^n \sigma_j^2 \leq G \right) \leq 2 \exp\left(-\frac{b^2}{2G + \frac{2bc}{3}} \right).$$

The next lemma is a generalization of (Sadiev et al., 2023, Lemma 5.1).

Lemma G.4. *Let $\lambda > 0$ be a scalar, $X \in \mathbb{R}^d$ be a random vector and let $\widetilde{X} = \texttt{clip}(X, \lambda)$ then*

$$\left\| \widetilde{X} - \mathbb{E}\left[\widetilde{X}\right] \right\| \leq 2\lambda. \tag{155}$$

Moreover, if for some $\sigma_1 \geq 0$ and $p \in (1,2]$ we have $\mathbb{E}[X] = x \in \mathbb{R}^d$, $\mathbb{E}\left[\|X - x\|^p\right] \leq \sigma_1^p$, and $\|x\| \leq \frac{\lambda}{2}$ then, for any $q \in [p, 2]$ we have

$$\left\| \mathbb{E}\left[\widetilde{X}\right] - x \right\| \leq \frac{2^p \sigma_1^p}{\lambda^{p-1}}, \tag{156}$$

$$\mathbb{E}\left[\left\| \widetilde{X} - x \right\|^q \right] \leq 2 \cdot 3^q \lambda^{q-p} \sigma_1^p, \tag{157}$$

$$\mathbb{E}\left[\left\| \widetilde{X} - \mathbb{E}\left[\widetilde{X}\right] \right\|^q \right] \leq 2^{3-q} \cdot 3^q \lambda^{q-p} \sigma_1^p. \tag{158}$$

Proof. The proof follows the same strategy as in Sadiev et al. (2023); inequalities (155) and (156) have already been established in Sadiev et al. (2023).

As in Sadiev et al. (2023), let us define the random variables

$$\chi = \mathbb{I}\left\{\|X\| > \lambda\right\} = \begin{cases} 1, & \text{if } \|X\| > \lambda; \\ 0, & \text{otherwise;} \end{cases} \quad \text{and} \quad \eta = \mathbb{I}\left\{\|X - x\| > \frac{\lambda}{2}\right\} = \begin{cases} 1, & \text{if } \|X - x\| > \frac{\lambda}{2}; \\ 0, & \text{otherwise;} \end{cases}$$

then, observe that

$$\texttt{clip}(X, \lambda) = \min\left\{1, \frac{\lambda}{\|X\|}X\right\} = \frac{\lambda X}{\|X\|}\chi + X(1-\chi) = X + \chi\left(\frac{\lambda}{\|X\|} - 1\right)X,$$

and $\chi \leq \eta$ since $\|X\| > \lambda$ implies

$$\|X - x\| \geq \|X\| - \|x\| > \lambda - \|x\| \geq \frac{\lambda}{2}.$$

Moreover, thanks to Markov's inequality, we have

$$\begin{aligned} \mathbb{E}\left[\eta\right] &= \mathbb{P}\left(\|X - \lambda\| > \frac{\lambda}{2}\right) \\ &= \mathbb{P}\left(\|X - x\|^p > \left(\frac{\lambda}{2}\right)^p\right) \\ &\overset{\text{Lem. H.11}}{\leq} \left(\frac{2}{\lambda}\right)^p \mathbb{E}\left[\|X - x\|^p\right] \\ &\leq \left(\frac{2}{\lambda}\right)^p \sigma_1^p. \end{aligned} \tag{159}$$

Proof of (157): using $\|x\| \leq \frac{\lambda}{2}$ and $\left\|\widetilde{X} - x\right\| \leq \left\|\widetilde{X}\right\| + \|x\| \leq \lambda + \frac{\lambda}{2} = \frac{3\lambda}{2}$ we have for any $q \in [p, 2]$

$$\begin{aligned} \mathbb{E}\left[\left\|\widetilde{X} - x\right\|^q\right] &= \mathbb{E}\left[\left\|\widetilde{X} - x\right\|^p \cdot \left\|\widetilde{X} - x\right\|^{q-p}\right] \\ &\leq \left(\frac{3\lambda}{2}\right)^{q-p} \mathbb{E}\left[\left\|\widetilde{X} - x\right\|^p \chi + \left\|\widetilde{X} - x\right\|^p (1-\chi)\right] \\ &\overset{\text{(a)}}{=} \left(\frac{3\lambda}{2}\right)^{q-p} \mathbb{E}\left[\left\|\frac{\lambda X}{\|X\|} - x\right\|^p \chi + \|X - x\|^p (1-\chi)\right] \\ &\leq \left(\frac{3\lambda}{2}\right)^{q-p} \mathbb{E}\left[\left(\left\|\frac{\lambda X}{\|X\|}\right\| + \|x\|\right)^p \chi + \|X - x\|^p (1-\chi)\right] \\ &\leq \left(\frac{3\lambda}{2}\right)^{q-p} \mathbb{E}\left[\left(\frac{3\lambda}{2}\right)^p \chi + \|X - x\|^p (1-\chi)\right] \\ &\overset{\text{(b)}}{\leq} \left(\frac{3\lambda}{2}\right)^q \mathbb{E}\left[\chi\right] + \left(\frac{3\lambda}{2}\right)^{q-p} \sigma_1^p \\ &\overset{(159)}{\leq} \left[\left(\frac{3\lambda}{2}\right)^q \left(\frac{2}{\lambda}\right)^p + \left(\frac{3\lambda}{2}\right)^{q-p}\right] \sigma_1^p \\ &\leq 2 \cdot 3^q \lambda^{q-p} \sigma_1^p, \end{aligned}$$

where in (a) we use the definition of $\texttt{clip}$ and χ. In (b) we use $1 - \chi \leq 1$ and $\mathbb{E}\left[\|X - x\|^p\right] \leq \sigma_1^p$.

Proof of (158): according to Lemma H.8 with $c = x$, we have

$$\begin{aligned} \mathbb{E}\left[\left\|\widetilde{X} - \mathbb{E}\left[\widehat{X}\right]\right\|^q\right] &\leq 2^{2-q}\mathbb{E}\left[\left\|\widetilde{X} - x\right\|^q\right] \\ &\overset{(157)}{\leq} 2^{3-q} \cdot 3^q \lambda^{q-p} \sigma_1^p, \end{aligned}$$

as claimed. □

H USEFUL IDENTITIES AND INEQUALITIES

For any vectors $x, y \in \mathbb{R}^d$, we have

$$2\langle x, y\rangle = \|x\|^2 + \|y\|^2 - \|x - y\|^2 . \tag{160}$$

Lemma H.1 (Variance Decomposition). *For any random vector $X \in \mathbb{R}^d$ and any non-random vector $c \in \mathbb{R}^d$ we have*

$$\mathbb{E}\left[\|X - c\|^2\right] = \mathbb{E}\left[\|X - \mathbb{E}[X]\|^2\right] + \|\mathbb{E}[X] - c\|^2 .$$

Remark H.1. In particular, this implies that for any random vector $X \in \mathbb{R}^d$ and any non-random vector $c \in \mathbb{R}^d$ we have

$$\mathbb{E}\left[\|X - \mathbb{E}[X]\|^2\right] \leq \mathbb{E}\left[\|X - c\|^2\right] .$$

In Lemma H.8 we extend this inequality to any exponent $\alpha > 1$ instead of just 2.

Lemma H.2 (Tower Property of the Expectation). *For any random variables $X \in \mathbb{R}^d$ and $Y_1, \ldots, Y_n$ we have*

$$\mathbb{E}[\mathbb{E}[X \mid Y_1, \ldots, Y_n]] = \mathbb{E}[X] .$$

Lemma H.3 (Cauchy Schwarz's Inequality). *For any vectors $a, b \in \mathbb{R}^d$ we have*

$$\langle a, b\rangle \leq |\langle a, b\rangle| \leq \|a\| \, \|b\| .$$

Lemma H.4 (Young's Inequality (Norm Form)). *For any vectors $a, b \in \mathbb{R}^d$ and any scalar $\alpha > 0$ we have*

$$\|a + b\|^2 \leq (1 + \alpha) \|x\|^2 + \left(1 + \frac{1}{\alpha}\right) \|y\|^2 .$$

Lemma H.5 (Young's Inequality (Inner Product Form)). *For any vectors $a, b \in \mathbb{R}^d$ and any scalar $\alpha > 0$ we have*

$$2\langle a, b\rangle \leq 2|\langle a, b\rangle| \leq \alpha \|x\|^2 + \frac{1}{\alpha} \|y\|^2 . \tag{161}$$

Proof. It's enough to prove inequality (161) when $d = 1$. Hence, consider $a, b \in \mathbb{R}$, we have given $\alpha > 0$

$$2ab \leq 2|ab| = 2|a| \cdot |b| = 2\left|\sqrt{\alpha}\, a\right| \cdot \left|\frac{b}{\sqrt{\alpha}}\right| \overset{(a)}{\leq} \alpha |a|^2 + \frac{1}{\alpha} |b|^2 = \alpha a^2 + \frac{b^2}{\alpha},$$

where in (a) we use the arithmetic-geometric inequality in $n = 2$ variables $\left(\sqrt{\alpha}\,|a|, \frac{1}{\sqrt{\alpha}}|b|\right)$. □

Lemma H.6. *For any vectors $a, b \in \mathbb{R}^d$ and any scalar $\alpha \geq 1$ we have*

$$\|a + b\|^\alpha \leq 2^{\alpha - 1} \left(\|a\|^\alpha + \|b\|^\alpha\right) . \tag{162}$$

Proof. Note that the function $x \mapsto \|x\|^\alpha$ is convex over $\mathbb{R}^d$ since $\|\cdot\|$ is convex (as a norm) and $t \mapsto t^\alpha$ is convex since $\alpha \geq 1$. Hence, by the Jensen's inequality (Lemma H.9), we have

$$\left\|\frac{a + b}{2}\right\|^\alpha \leq \frac{1}{2} \left(\|a\|^\alpha + \|b\|^\alpha\right) ,$$

i.e.,

$$\|a + b\|^\alpha \leq 2^{\alpha - 1} \left(\|a\|^\alpha + \|b\|^\alpha\right) ,$$

as claimed. □

Remark H.2. Similar inequalities to (162) appear in He et al. (2025). In Lemma H.7 we state one of them that we use in our proofs.

Lemma H.7 ((He et al., 2025, Lemma 1)). *For any vectors $a, b \in \mathbb{R}^d$, $a \neq \mathbf{0}$, and any scalar $\alpha \in (1, 2]$ we have*[3]

$$\|a+b\|^\alpha \leq \|a\|^\alpha + \alpha \|a\|^{\alpha-2} \langle a, b\rangle + 2^{2-\alpha} \|b\|^\alpha .$$

Lemma H.8. *For any scalar $\alpha \in (1, 2]$, any random vector $X \in \mathbb{R}^d$ and any non-random vector $c \in \mathbb{R}^d$ we have*

$$\mathbb{E}\left[\|X - \mathbb{E}[X]\|^\alpha\right] \leq 2^{2-\alpha} \mathbb{E}\left[\|X - c\|^\alpha\right].$$

Proof. By Lemma H.7 we have

$$\begin{aligned}
\mathbb{E}\left[\|X - \mathbb{E}[X]\|^\alpha\right] &= \mathbb{E}\left[\|(X-c) + (c - \mathbb{E}[X])\|^\alpha\right] \\
&\overset{\text{Lem. H.7}}{\leq} \|c - \mathbb{E}[X]\|^\alpha + \alpha \|c - \mathbb{E}[X]\|^{\alpha-2} \mathbb{E}\left[\langle c - \mathbb{E}[X], X - c\rangle\right] + 2^{2-\alpha}\mathbb{E}\left[\|X-c\|^\alpha\right] \\
&= \|c - \mathbb{E}[X]\|^\alpha - \alpha \|c - \mathbb{E}[X]\|^\alpha + 2^{2-\alpha}\mathbb{E}\left[\|X-c\|^\alpha\right] \\
&\overset{(a)}{\leq} 2^{2-\alpha}\mathbb{E}\left[\|X-c\|^\alpha\right],
\end{aligned}$$

where in (a) we use $\alpha > 1$. □

Lemma H.9 (Jensen's Inequality). *Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be a convex function then*

1. *(Probabilistic Form) for any random vector $X \in \mathbb{R}^d$ we have*

$$\mathbb{E}\left[f(X)\right] \geq f\left(\mathbb{E}[X]\right).$$

2. *(Deterministic Form) for any vectors $v_1, \ldots, v_n \in \mathbb{R}^d$ and scalars $\lambda_1, \ldots, \lambda_n \in \mathbb{R}_+$ we have*

$$\sum_{i=1}^n \lambda_i f(v_i) \geq f\left(\sum_{i=1}^n \lambda_i v_i\right),$$

provided $\lambda_i \geq 0$ for all $i \in [n]$ and $\sum_{i=1}^n \lambda_i = 1$.

Lemma H.10. *For any vectors $v_1, \ldots, v_n \in \mathbb{R}^d$ we have*

$$\left\|\sum_{i=1}^n v_i\right\|^2 \leq n \sum_{i=1}^n \|v_i\|^2 .$$

Proof. The function $\|\cdot\|^2 : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex with $\mu = 2$ so is convex thus applying Jensen's inequality H.9 with $\lambda_1 = \cdots = \lambda_n = \frac{1}{n}$ gives

$$\left\|\sum_{i=1}^n \frac{v_i}{n}\right\|^2 \leq \frac{1}{n}\sum_{i=1}^n \|v_i\|^2 ,$$

and multiplying both sides by n^2 gives the desired inequality. □

Lemma H.11 (Markov's Inequality). *For any non-negative random variable X and any scalar $a > 0$, we have*

$$\mathbb{P}\left(X \geq a\right) \leq \frac{\mathbb{E}[X]}{a}.$$

Lemma H.12 (Bernoulli's Inequality). *For any real number $x \geq -1$ and $r \in \{0\} \cup [1, +\infty)$ we have*

$$(1+x)^r \geq 1 + rx.$$

[3] Note that, when $a = 0$ we have $\|a\|^{\alpha-2} a = 0$ since $\alpha > 1$ thus $\alpha - 2 > -1$.

Lemma H.13 (Absolute p^{th} Central Moment of Normal Distribution (Andreas, 2014, eq. (18))). *For any real number $p > -1$, $\sigma_1 > 0$ and $\mu \in \mathbb{R}$, if $X \sim \mathcal{N}(\mu, \sigma_1^2)$ then*

$$\mathbb{E}\left[|X-\mu|^p\right] = 2^{\frac{p}{2}}\sigma_1^p \cdot \frac{\Gamma\left(\frac{p+1}{2}\right)}{\sqrt{\pi}},$$

where Γ is the Euler's gamma function (see DLMF (2025, (5.2.1)) for formal definition).

Definition H.1 (Operator Norm). Given a $d \times d$ matrix $A \in \mathbb{R}^{d\times d}$, the operator norm of A is the norm

$$\|A\|_{\text{op}} := \sup_{y\in\mathbb{R}^d, \|y\|=1} \|Ay\|,$$

where $\|\cdot\|$ is the standard euclidean norm in $\mathbb{R}^d$.

Lemma H.14 (Lipchitz Gradients Implies Bounded Hessian). *Given $F\colon \mathbb{R}^d \to \mathbb{R}$ a twice continuously differentiable function over $\mathbb{R}^d$ such that its gradients are L_1–Lipschitz continuous for some constant $L_1 \geq 0$ then, for all $x \in \mathbb{R}^d$ we have*

$$\left\|\nabla^2 F(x)\right\|_{\text{op}} \leq L_1.$$

The above lemma can be extended to p^{th}-order continuously differentiable function, $p \geq 2$ for which the $(p-1)^{\text{th}}$-order derivative of F is Lipschitz continuous in the operator norm induced by ℓ^2–norm.

Lemma H.15. *Given $F\colon \mathbb{R}^d \to \mathbb{R}$ a continuously differentiable and lower bounded function over $\mathbb{R}^d$ such that its gradients are L_1–Lipschitz continuous for some constant $L_1 \geq 0$ then, for all $x \in \mathbb{R}^d$ we have*

$$\|\nabla F(x)\|^2 \leq 2L_1\left(F(x) - F^{\text{inf}}\right),$$

for any $F^{\text{inf}} \leq \inf_{x\in\mathbb{R}^d} F(x)$.

Lemma H.16. *Given $F\colon \mathbb{R}^d \to \mathbb{R}$ a twice continuously differentiable and lower bounded function over $\mathbb{R}^d$ such that its Hessians are L_2–Lipschitz continuous for some constant $L_2 \geq 0$ then, for all $x, y \in \mathbb{R}^d$ we have*

$$\left\|\nabla F(x) - \nabla F(y) - \nabla^2 F(y)(x-y)\right\| \leq \frac{L_2}{2}\|x-y\|^2.$$

Lemma H.17 (Landau-Kolmogorov's Inequalities on a Finite Interval (Fabry, 1987, Theorem 2)). *Let $p \geq 1$ be an integer, $a \leq b$ be real numbers and $F \in \mathcal{C}^p([a,b],\mathbb{R})$. For any $k \in \{0,\ldots,p\}$ let $M_k := \sup_{x\in[a,b]}\left|F^{(k)}(x)\right| \in [0,+\infty]$ then, if M_0 and M_p are finite, we have for all $k \in [p]$,*

1. *M_k is finite,*
2. *the inequality*

$$M_k \leq c_{k,p} M_0^{\frac{p-k}{p}} \max\left\{\frac{M_p}{2^{p-1}p!}, 4^p(b-a)^{-p}M_0\right\}^{\frac{k}{p}}, \tag{163}$$

holds where $c_{k,p} > 0$ is the universal constant

$$c_{k,p} := \frac{p2^k k!}{p+k}\binom{p+k}{p-k}.$$

Remark H.3. Notably, the constant $c_{k,p}$ appearing in the inequality (163) does not depend on the choice of the function F while the $\{M_k\}_{k\in\{0,\ldots,p\}}$ do.

Remark H.4. Other related inequalities can be found in Mitrinović et al. (1991); Chen (1993).

Lemma H.18. *Let $F\colon \mathbb{R}^d \to \mathbb{R}$ be a continuously differentiable function over $\mathbb{R}^d$ with L–Lipschitz gradients for some $L > 0$ and $x_0 \in \mathbb{R}^d$. Assume that $\|\nabla F(x_0)\| > 0$ then for all $t \in [0,2]$ we have*

$$F(x_0 + tv) \leq F(x_0),$$

where $v = -\frac{1}{L}\nabla F(x_0)$.

Proof. Using the fact that the function F has L–Lipschitz gradients then we know F is L–smooth (Nesterov (2018)) hence, for all $t \in [0,2]$ we have

$$D_F(x_0+tv,x_0) := F(x_0+tv) - F(x_0) - \langle \nabla F(x_0), (x_0+tv)-x_0\rangle \leq \frac{L}{2}\left\|(x_0+tv)-x_0\right\|^2 = \frac{Lt^2}{2}\left\|v\right\|^2,$$

where $D_F(x,y)$ denotes the Bregman divergence of F at $x,y \in \mathbb{R}^d$. Rewriting the above inequality using the choice of v gives

$$\begin{aligned}
F(x_0+tv) &\leq F(x_0) + t\left\langle \nabla F(x_0), v\right\rangle + \frac{Lt^2}{2}\left\|v\right\|^2 \\
&= F(x_0) - \frac{t}{L}\left\|\nabla F(x_0)\right\|^2 + \frac{t^2}{2L}\left\|\nabla F(x_0)\right\|^2 \\
&= F(x_0) - \frac{t}{L}\left\|\nabla F(x_0)\right\|^2\left(1-\frac{t}{2}\right) \\
&\leq F(x_0),
\end{aligned}$$

since $0 \leq t \leq 2$ and $L > 0$. This achieves the proof of the lemma. □

Remark H.5. The above lemme still holds when $\|\nabla F(x_0)\| = 0$, i.e., $v = 0$ but it is not of interest.

Lemma H.19. *Let $F\colon \mathbb{R}^d \to \mathbb{R}$ be $p+1$ times continuously differentiable ($p \geq 1$) over $\mathbb{R}^d$. Assume that*

1. *for all $k \in [p]$, the function $\nabla^k F$ is L_k–Lipschitz continuous for some $L_k > 0$,*
2. *the function F is lower bounded over $\mathbb{R}^d$ and we denote $F^{\inf} := \inf_{x\in\mathbb{R}^d} F(x)$,*

then, for all $k \in [p]$ there exists a constant $c_k := (k+1)^2 > 0$, depending only on k such that for all $x_0 \in \mathbb{R}^d$

$$\left\|\nabla F(x_0)\right\| \leq c_k\left(F(x_0)-F^{\inf}\right)^{k/k+1}\max\left\{\frac{L_k^{\frac{1}{k+1}}}{2^{\frac{k}{k+1}}[(k+1)!]^{\frac{1}{k+1}}}, \frac{2L_1\left(F(x_0)-F^{\inf}\right)^{\frac{1}{k+1}}}{\left\|\nabla F(x_0)\right\|}\right\}. \tag{164}$$

In particular, if the gradient of F at x_0 is large (say, $\|\nabla F(x_0)\| = \Omega(\sqrt{L_1\left(F(x_0)-F^{\inf}\right)})$, see Lemma H.15) and the Lipschitz constant L_k dominates, then (164) simplifies to

$$\left\|\nabla F(x_0)\right\| \leq \widetilde{c}_k L_k^{1/k+1}\left(F(x_0)-F^{\inf}\right)^{k/k+1}, \tag{165}$$

for some universal constant $\widetilde{c}_k$ which depends only on k.

Remark H.6. The bound (164) is (almost) a generalization to high-order Lipschitz constant of the well-known inequality

$$\left\|\nabla F(x_0)\right\| \leq \sqrt{2L_1\left(F(x_0)-F^{\inf}\right)},$$

which we recall in Lemma H.15 and which corresponds to the case $k = 1$.

While it is hopeless[4] to obtain the inequality (165) in full generality, i.e., without extra assumptions, in practical settings the Lipschitz constants are very large and the gradient of the objective at the initial point is also large.

Proof. Let $F\colon \mathbb{R}^d \to \mathbb{R}$ be defined as above. Then, if $\|\nabla F(x_0)\| = 0$ the inequality (164) holds since $c_k > 0$ and $F(x_0) - F^{\inf} \geq 0$. Now, assume $\|\nabla F(x_0)\| > 0$, we define the scalar function $\varphi\colon [0,2] \to \mathbb{R}$ as

$$\varphi\colon t \mapsto F(x_0+tv) - F^{\inf},$$

[4]Inequality (165) already fails for the case $k = 2$ with quadratic functions. For instance, consider $F\colon x \mapsto \frac{1}{2}\|x\|^2$ over $\mathbb{R}^d$ and let $x_0 = (1,0,\ldots,0)$ then $\|\nabla F(x_0)\| = \|x_0\| = 1$, but $\nabla^2 F(x) = \mathrm{Id}$ for all $x \in \mathbb{R}^d$ hence we can take $L_2 = 0$ since the Hessian of F is constant, but then

$$L_2^{1/3}\left(F(x_0)-F^{\inf}\right)^{2/3} = 0,$$

and there do not exists universal constant c_2 for which (165) can hold.

where $v = -\frac{1}{L_1}\nabla F(x_0) \neq 0$ and L_1 is the Lipschitz constant of ∇F. Then, the above function φ is $p+1$ time continuously differentiable over $[0,2]$, and for all $t \in [0,2]$ and all $k \in \{2,\dots,p+1\}$ we have

$$\varphi^{(k)}(t) = \nabla^k F(x_0+tv)[v,\dots,v] = \frac{(-1)^k}{L_1^k}\nabla^k F(x_0+tv)[g,\dots,g], \tag{166}$$

where we let $g := \nabla F(x_0)$ and $\nabla^k F(x_0+tv)[\cdot,\dots,\cdot]$ is the k–linear form induced by the k^{th}–derivative of F. From equality (166) we obtain

$$\left|\varphi^{(k)}(t)\right| = \frac{1}{L_1^k}\left|\nabla^k F(x_0+tv)[g,\dots,g]\right| \overset{(a)}{\leq} \frac{1}{L_1^k}\left\|\nabla^k F(x_0+tv)\right\|_{\text{op}} \cdot \|g\|^k \overset{(b)}{\leq} \left(\frac{\|\nabla F(x_0)\|}{L_1}\right)^k L_{k-1}, \tag{167}$$

where in (a) we use the Cauchy-Schwarz's inequality (Lemma H.3) while in (b) we use (166) and the fact that $\nabla^r F$ is L_r–Lipschitz continuous for any $r \in [p]$ which implies that $\nabla^k F$ is bounded in operator norm as long as $k \geq 2$ (see Lemma H.14 for the case $k=2$). Hence, if we define $M_k := \sup_{t\in[0,2]}\left|\varphi^{(k)}(t)\right|$ we have, by (167),

$$M_k \leq \left(\frac{\|\nabla F(x_0)\|}{L_1}\right)^k L_{k-1} < +\infty, \tag{168}$$

since $L_k > 0$ for all $k \in [p]$ and

$$M_1 = \sup_{t\in[0,2]}|\varphi'(t)| = \sup_{t\in[0,2]}|\langle \nabla F(x_0+tv), v\rangle| = \frac{1}{L_1}\sup_{t\in[0,2]}|\langle \nabla F(x_0+tv), \nabla F(x_0)\rangle| \geq \frac{1}{L_1}\|\nabla F(x_0)\|^2. \tag{169}$$

Additionally,

$$M_0 := \sup_{t\in[0,2]}|\varphi(t)| \overset{(a)}{=} \sup_{t\in[0,2]}\left(F(x_0+tv) - F^{\inf}\right) \overset{(b)}{=} F(x_0) - F^{\inf} < +\infty, \tag{170}$$

where (a) follows from $F \geq F^{\inf}$ while (b) follows from Lemma H.18 since F is continuously differentiable and has L_1–Lipschitz gradients.

Then for any $k \in [p]$, since the function φ is $k+1 \geq 2$ times continuously differentiable over $\mathbb{R}$ and because the $\{M_\ell\}_{\ell\in\{0,\dots,p\}}$ are all finite from (168) and (170) then, by the Landau-Kolmogorov's inequalities (Lemma H.17),

$$M_1 \leq c_k\, M_0^{\frac{k}{k+1}} \max\left\{\frac{M_{k+1}}{2^k(k+1)!}, 2^{k+1}M_0\right\}^{\frac{1}{k+1}}, \tag{171}$$

where $c_k := c_{1,k+1} = (k+1)^2$ is an universal constant depending only on k (by Lemma H.17). From the inequalities (168) and (169), (170) along with (171) we obtain

$$\begin{aligned}\frac{1}{L_1}\|\nabla F(x_0)\|^2 &\leq c_k\left(F(x_0)-F^{\inf}\right)^{\frac{k}{k+1}} \max\left\{\left(\left(\frac{\|\nabla F(x_0)\|}{L_1}\right)^{k+1}\frac{L_k}{2^k(k+1)!}\right)^{\frac{1}{k+1}}, 2M_0^{\frac{1}{k+1}}\right\}\\ &= \frac{c_k}{L_1}\left(F(x_0)-F^{\inf}\right)^{k/k+1}\|\nabla F(x_0)\| \max\left\{\frac{L_k^{\frac{1}{k+1}}}{2^{\frac{k}{k+1}}[(k+1)!]^{\frac{1}{k+1}}}, \frac{2L_1\left(F(x_0)-F^{\inf}\right)^{\frac{1}{k+1}}}{\|\nabla F(x_0)\|}\right\},\end{aligned}$$

which simplifies to

$$\|\nabla F(x_0)\| \leq c_k\left(F(x_0)-F^{\inf}\right)^{k/k+1} \max\left\{\frac{L_k^{\frac{1}{k+1}}}{2^{\frac{k}{k+1}}[(k+1)!]^{\frac{1}{k+1}}}, \frac{2L_1\left(F(x_0)-F^{\inf}\right)^{\frac{1}{k+1}}}{\|\nabla F(x_0)\|}\right\}.$$

and leads to the desired inequality. □

Algorithm 5: SGD-MVR (SGD with MVR)

Initialization:
- $x_0 \in \mathbb{R}^d$, the starting point
- $T > 0$, the number of iterations
- $g_0 \in \mathbb{R}^d$, an initial vector
- $\gamma > 0$, the stepsize
- $\alpha \in (0, 1]$, the momentum parameter for MVR

$x_1 \leftarrow x_0 - \gamma g_0$
For $t = 1, 2, \ldots, T-1$ **do**
  // Apply MVR.
  $g_t \leftarrow (1-\alpha)\left(g_{t-1} + \nabla f\left(x_t, \xi_t\right) - \nabla f\left(x_{t-1}, \xi_t\right)\right) + \alpha \nabla f\left(x_t, \xi_t\right)$
  // Do one descent step.
  $x_{t+1} \leftarrow x_t - \gamma g_t$
Output: x_T

I ADDITIONAL ANALYSIS IN THE CASE $p = q = 2$

In this section, which is of independent interest, we study regular, not normalized, SGD with Momentum Variance Reduction (SGD-MVR). This method is the original one by Cutkosky and Orabona (2019), where it is called STORM. The method is shown as Algorithm 5. We show that in the case $p = q = 2$, it achieves the same optimal complexity as NSGD-MVR. That is, normalization is not required in this particular case.

Theorem I.1. *Under Assumptions 2.1 to 2.3 and 2.7, with $p = q = 2$, let the initial gradient estimate g_0 be given by*

$$g_0 = \frac{1}{B_{\text{init}}} \sum_{j=1}^{B_{\text{init}}-1} \nabla f\left(x_0, \xi_{0,j}\right),$$

with $B_{\text{init}} = \max\left\{1, \frac{\sigma_1^2}{\varepsilon^2}\right\}$, momentum parameter $\alpha = \min\left\{1, \frac{\varepsilon^2}{\sigma_1^2}\right\}$, stepsize $\gamma = \frac{1}{L_1 + \frac{\delta\sqrt{2}(1-\alpha)}{\sqrt{\alpha}}}$. Then Algorithm 5 is guaranteed to find an ε-stationary point with total sample complexity

$$\mathcal{O}\left(\frac{\sigma_1^2}{\varepsilon^2} + \frac{(L_1+\delta)\Delta}{\varepsilon^2} + \frac{\delta\Delta\sigma_1}{\varepsilon^3}\right).$$

Proof. Let $t \geq 1$. We have the descent lemma

$$\begin{aligned}
F(x^t) &\leq F(x^{t-1}) + \langle x^t - x^{t-1}, \nabla F(x^{t-1})\rangle + \frac{L}{2}\left\|x^t - x^{t-1}\right\|^2 \\
&= F(x^{t-1}) - \gamma\langle g_{t-1}, \nabla F(x^{t-1})\rangle + \frac{L}{2}\left\|x^t - x^{t-1}\right\|^2 \\
&= F(x^{t-1}) + \frac{\gamma}{2}\left\|g_{t-1} - \nabla F(x^{t-1})\right\|^2 - \frac{\gamma}{2}\left\|\nabla F(x^{t-1})\right\|^2 - \frac{\gamma}{2}\left\|g_{t-1}\right\|^2 + \frac{L}{2}\left\|x^t - x^{t-1}\right\|^2 \\
&= F(x^{t-1}) + \frac{\gamma}{2}\left\|g_{t-1} - \nabla F(x^{t-1})\right\|^2 - \frac{\gamma}{2}\left\|\nabla F(x^{t-1})\right\|^2 + \left(\frac{L}{2} - \frac{1}{2\gamma}\right)\left\|x^t - x^{t-1}\right\|^2.
\end{aligned}$$

Therefore, denoting by $\mathcal{F}_{t-1}$ the filtration $\sigma(g_0, \xi_1, \ldots, \xi_{t-1})$, we have

$$\begin{aligned}
\mathbb{E}\left[F(x^t) - F^{\inf} \mid \mathcal{F}_{t-1}\right] \leq F(x^{t-1}) - F^{\inf} - \frac{\gamma}{2}\left\|\nabla F(x^{t-1})\right\|^2 + \frac{\gamma}{2}\left\|g_{t-1} - \nabla F(x^{t-1})\right\|^2 \\
+ \left(\frac{L}{2} - \frac{1}{2\gamma}\right)\left\|x^t - x^{t-1}\right\|^2.
\end{aligned}$$

We need to control the deviation $\left\|g_{t-1} - \nabla F(x^{t-1})\right\|^2$, that will appear in the Lyapunov function. So, let us study $\mathbb{E}\left[\left\|g_t - \nabla F(x^t)\right\|^2 \,\middle|\, \mathcal{F}_{t-1}\right]$. From a bias–variance decomposition, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|g_t - \nabla F(x^t)\right\|^2 \,\middle|\, \mathcal{F}_{t-1}\right] \\
&= \mathbb{E}\left[\left\|\nabla f(x^t,\xi^t) - \nabla F(x^t) + (1-\alpha)\big(g_{t-1} - \nabla f(x^{t-1},\xi^t)\big)\right\|^2 \,\middle|\, \mathcal{F}_{t-1}\right] \\
&= (1-\alpha)^2\left\|g_{t-1} - \nabla F(x^{t-1})\right\|^2 \\
&\quad + \mathbb{E}\left[\left\|\nabla f(x^t,\xi^t) - \nabla F(x^t) + (1-\alpha)\big(\nabla F(x^{t-1}) - \nabla f(x^{t-1},\xi^t)\big)\right\|^2 \,\middle|\, \mathcal{F}_{t-1}\right] \\
&= (1-\alpha)^2\left\|g_{t-1} - \nabla F(x^{t-1})\right\|^2 \\
&\quad + \mathbb{E}\left[\left\|\alpha\big(\nabla f(x^t,\xi^t) - \nabla F(x^t)\big) + (1-\alpha)\big(\nabla f(x^t,\xi^t) - \nabla F(x^t) - \nabla f(x^{t-1},\xi^t) + \nabla F(x^{t-1})\big)\right\|^2 \,\middle|\, \mathcal{F}_{t-1}\right] \\
&\le (1-\alpha)^2\left\|g_{t-1} - \nabla F(x^{t-1})\right\|^2 + 2\alpha^2\mathbb{E}\left[\left\|\nabla f(x^t,\xi^t) - \nabla F(x^t)\right\|^2 \,\middle|\, \mathcal{F}_{t-1}\right] \\
&\quad + 2(1-\alpha)^2\mathbb{E}\left[\left\|\nabla f(x^t,\xi^t) - \nabla F(x^t) - \nabla f(x^{t-1},\xi^t) + \nabla F(x^{t-1})\right\|^2 \,\middle|\, \mathcal{F}_{t-1}\right] \\
&\le (1-\alpha)^2\left\|g_{t-1} - \nabla F(x^{t-1})\right\|^2 + 2\alpha^2\sigma_1^2 + 2(1-\alpha)^2\delta^2\left\|x^t - x^{t-1}\right\|^2.
\end{aligned}$$

We introduce the Lyapunov function

$$H^t := F(x^t) - F^{\inf} + \frac{\gamma}{2\alpha}\left\|g_t - \nabla F(x^t)\right\|^2.$$

We have

$$\begin{aligned}
\mathbb{E}\left[H^t \,\middle|\, \mathcal{F}_{t-1}\right] &\le F(x^{t-1}) - F^{\inf} - \frac{\gamma}{2}\left\|\nabla F(x^{t-1})\right\|^2 + \frac{\gamma}{2}\left\|g_{t-1} - \nabla F(x^{t-1})\right\|^2 \\
&\quad + \left(\frac{L}{2} - \frac{1}{2\gamma}\right)\mathbb{E}\left[\left\|x^t - x^{t-1}\right\|^2 \,\middle|\, \mathcal{F}_{t-1}\right] \\
&\quad + (1-\alpha)^2\frac{\gamma}{2\alpha}\left\|g_{t-1} - \nabla F(x^{t-1})\right\|^2 + \gamma\alpha\sigma_1^2 \\
&\quad + \frac{(1-\alpha)^2\gamma\delta^2}{\alpha}\mathbb{E}\left[\left\|x^t - x^{t-1}\right\|^2 \,\middle|\, \mathcal{F}_{t-1}\right] \\
&\le F(x^{t-1}) - F^{\inf} - \frac{\gamma}{2}\left\|\nabla F(x^{t-1})\right\|^2 + \gamma\alpha\sigma_1^2 \\
&\quad + \big(\alpha + (1-\alpha)^2\big)\frac{\gamma}{2\alpha}\left\|g_{t-1} - \nabla F(x^{t-1})\right\|^2 \\
&\quad + \left(\frac{L}{2} - \frac{1}{2\gamma} + \frac{(1-\alpha)^2\gamma\delta^2}{\alpha}\right)\mathbb{E}\left[\left\|x^t - x^{t-1}\right\|^2 \,\middle|\, \mathcal{F}_{t-1}\right].
\end{aligned}$$

We have $\alpha + (1-\alpha)^2 \le \alpha + (1-\alpha) = 1$. A sufficient condition for $\frac{L}{2} - \frac{1}{2\gamma} + \frac{(1-\alpha)^2\gamma\delta^2}{\alpha} \le 0$ is

$$\gamma \le \frac{1}{L_1 + \delta\frac{\sqrt{2}(1-\alpha)}{\sqrt{\alpha}}}.$$

Assuming now that this condition holds, we have

$$\mathbb{E}\left[H^t \,\middle|\, \mathcal{F}_{t-1}\right] \le H^{t-1} - \frac{\gamma}{2}\left\|\nabla F(x^{t-1})\right\|^2 + \gamma\alpha\sigma_1^2. \tag{172}$$

Let $T \ge 1$. Unrolling the recursion, we have

$$\sum_{t=0}^{T-1}\left(\frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla F(x^t)\right\|^2\right] - \gamma\alpha\sigma^2\right) \le \Delta + \frac{\gamma}{2\alpha}\mathbb{E}\left[\left\|g^0 - \nabla F(x^0)\right\|^2\right], \tag{173}$$

so that

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(x^t)\right\|^2\right] \le \frac{2\Delta}{\gamma T} + 2a\sigma^2 + \frac{1}{\alpha T}\mathbb{E}\left[\left\|g^0 - \nabla F(x^0)\right\|^2\right]. \tag{174}$$

We form g^0 as an unbiased estimate of $\nabla F(x^0)$ using a minibatch of size B_{init}. Hence $\mathbb{E}\left[\left\|g^0 - \nabla F(x^0)\right\|^2\right] \leq \frac{\sigma_1^2}{B_{\text{init}}}$ and

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(x^t)\right\|^2\right] \leq \frac{2\Delta}{\gamma T} + \left(2a + \frac{1}{\alpha B_{\text{init}} T}\right)\sigma_1^2. \tag{175}$$

We interpret the left-hand side as $\mathbb{E}\left[\left\|\nabla F(\hat{x}^T)\right\|^2\right]$ for $\hat{x}^T$ chosen uniformly at random in $x^0, \ldots, x^{T-1}$. Hence, given $\varepsilon > 0$, with $\alpha = \min\left\{1, \frac{\varepsilon^2}{\sigma_1^2}\right\}$, $\gamma = \frac{1}{L_1 + \frac{\delta\sqrt{2}(1-\alpha)}{\sqrt{\alpha}}}$, $B_{\text{init}} = \alpha^{-1} = \max\left\{1, \frac{\sigma_1^2}{\varepsilon^2}\right\}$, we have

$$\mathbb{E}\left[\left\|\nabla F(\hat{x}^T)\right\|^2\right] \leq \frac{2\Delta}{T}\left(L_1 + \delta\sigma_1\varepsilon^{-1}\right) + 2\varepsilon^2 + \frac{\sigma_1^2}{T}. \tag{176}$$

With $T = \max\left(\frac{\sigma_1^2}{\varepsilon^2}, 2\Delta\left(L_1\varepsilon^{-2} + \delta\sigma_1\varepsilon^{-3}\right)\right)$, we have $\frac{\sigma_1^2}{T} \leq \varepsilon^2$ and $\frac{2\Delta}{T}\left(L_1 + \delta\sigma\varepsilon^{-1}\right) \leq \epsilon^2$. Hence,

$$\mathbb{E}\left[\left\|\nabla F(\hat{x}^T)\right\|^2\right] \leq 4\varepsilon^2, \tag{177}$$

with $B_{\text{init}} + 2(T-1) = \mathcal{O}\left(\frac{\sigma_1^2}{\varepsilon^2} + \frac{\Delta(L_1+\delta)}{\varepsilon^2} + \frac{\Delta\delta\sigma_1}{\epsilon^3}\right)$ stochastic gradient evaluations. □

J EMPIRICAL EVALUATION

To validate our theoretical findings and provide valuable intuition regarding the behavior and stability of the proposed algorithms, we conduct a series of synthetic experiments. These experiments evaluate convergence trajectories, the necessity of our double-clipping mechanism, and how algorithmic complexity scales with the heavy-tail index p.

J.1 Experimental Setup

We consider the minimization of a highly ill-conditioned quadratic objective $F(x) = \frac{1}{2}x^\top Ax$, where $A \in \mathbb{R}^{d\times d}$ is a diagonal matrix with $A_{1,1} = 0.01$ and $A_{i,i} = 1$ for $i > 1$, and dimension $d = 10$. To simulate heavy-tailed noise under the p-BCM assumption (Assumption 2.3), we inject synthetic noise into the exact gradients and Hessian-vector products. Specifically, the noise vectors are drawn from a symmetric Pareto-like distribution generated via $s \cdot (u^{-1/p} - 1)$, where $u \sim \mathcal{U}(0, 1)$ and $s \sim \{-1, 1\}$ uniformly. This ensures that the noise has an infinite variance when $p < 2$, strictly adhering to our theoretical noise model.

J.2 The Necessity of Double-Clipping and Algorithmic Stability

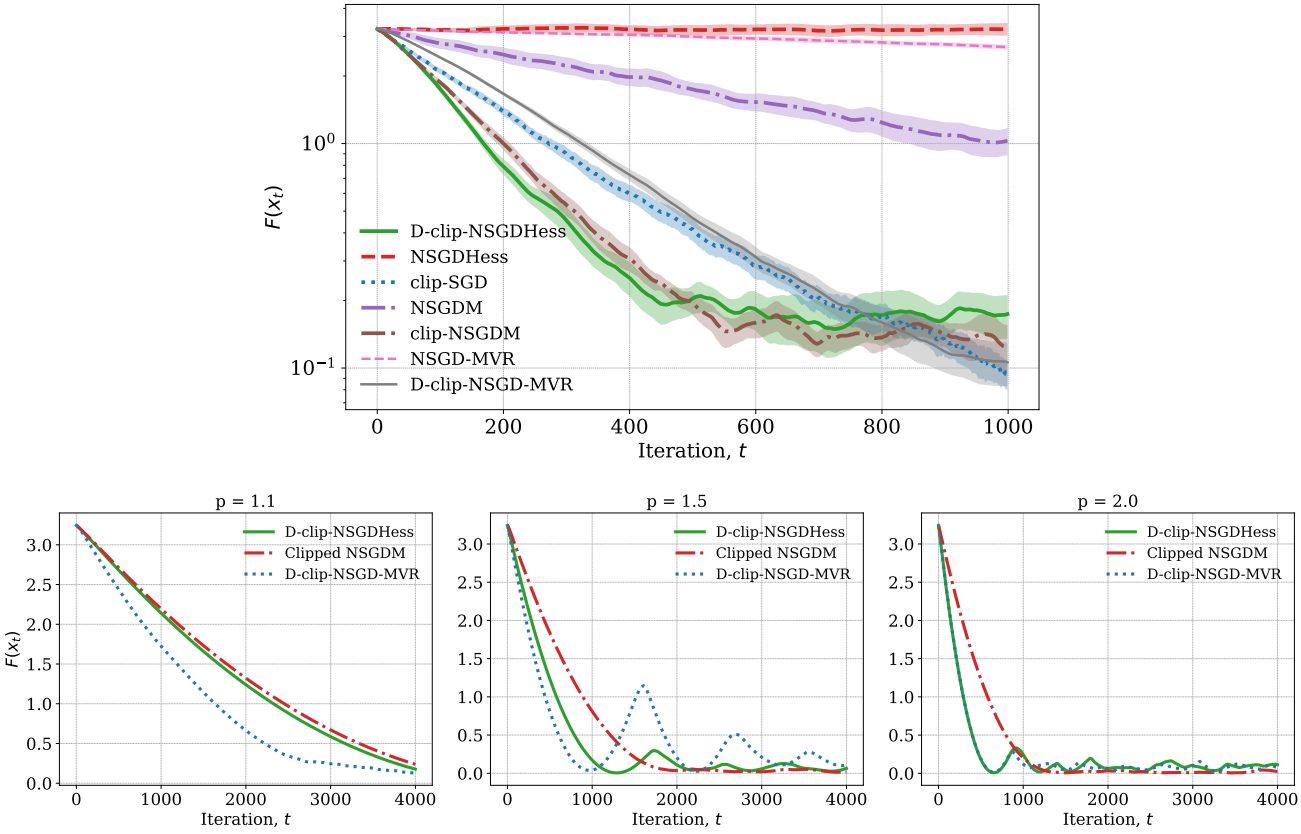

Figure 1: **Top:** Convergence trajectories of $F(x_t)$ over $T = 1000$ iterations for $p = 1.1$. Solid/dashed lines represent the mean over 20 independent runs, and shaded regions denote the standard deviation. **Bottom:** Algorithm performance across different tail indices $p \in \{1.1, 1.5, 2.0\}$. Hyperparameters are scaled strictly according to their respective theoretical optimal rates.

In our first experiment (Figure 1, Left), we fix the tail index to $p = 1.1$ (an extreme heavy-tailed regime) and compare our proposed double-clipped methods (`D-clip-NSGD-MVR` and `D-clip-NSGDHess`) against standard clipped SGD Zhang et al. (2020), clipped NSGD with momentum (`clip-NSGDM` (Cutkosky and Mehta, 2021; Liu et al., 2023)), and their unclipped counterparts. We run $T = 1000$ iterations and average the results over 20 independent trials.

The results clearly demonstrate the necessity of clipping in this regime: unclipped methods (NSGDM (Cutkosky and Mehta, 2020; Hübler et al., 2025) and NSGD-MVR) fail to converge and exhibit severe instability, eventually diverging. In contrast, all clipped variants successfully decrease the objective. Notably, our proposed D-clip-NSGD-MVR and D-clip-NSGDHess exhibit the fastest and most stable convergence, validating the theoretical benefit of combining variance reduction or Hessian-correction with our proposed double-clipping mechanism.

Next, we evaluate the performance of these algorithms across a spectrum of heavy-tailed noise distributions by varying the tail index $p \in \{1.1, 1.5, 2.0\}$ over $T = 4000$ iterations (Figure 1, Right). Crucially, to test our theoretical bounds, the stepsize γ, momentum parameter α, and clipping thresholds for each method are scaled *strictly* according to the theoretically optimal T-dependent exponents derived in our main theorems. The results confirm our theoretical characterizations: in the most challenging regime ($p = 1.1$), D-clip-NSGD-MVR significantly outperforms the standard clip-NSGDM (Cutkosky and Mehta, 2021; Liu et al., 2023) baseline. As $p \to 2.0$ (approaching the standard bounded variance setting), the gap between the methods narrows, and all algorithms exhibit rapid, stable convergence.

J.3 Theoretical vs. Empirical Complexity Scaling

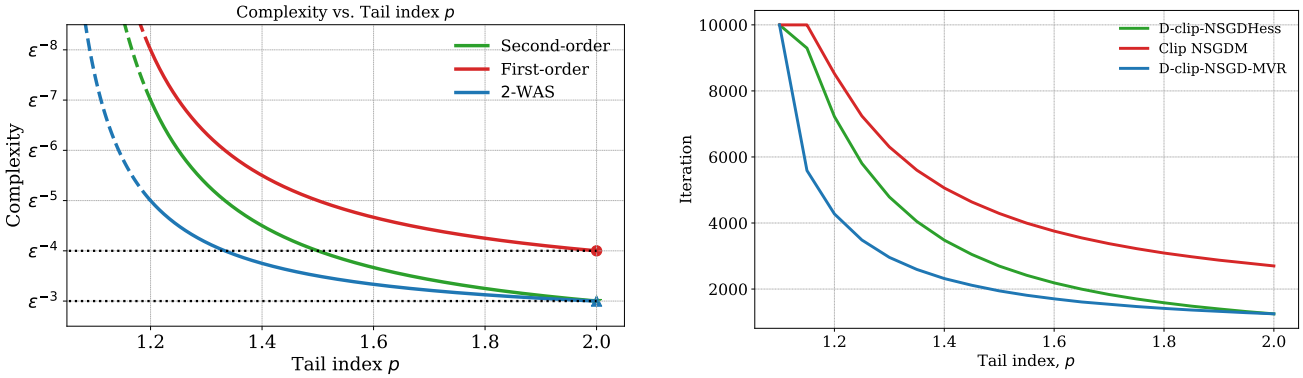

Figure 2: **Left:** Theoretical complexity $\mathcal{O}(\varepsilon^{-c})$ vs. tail index p. The y-axis represents the exponent c of ε. **Right:** Empirical iterations required to reach a target suboptimality $F(x_t) < 1.2$ across different tail indices p. The empirical performance perfectly mirrors the theoretical scaling, demonstrating that D-clip-NSGD-MVR (2-WAS) provides the most robust acceleration in severe heavy-tailed regimes ($p < 2$).

To further illustrate the robustness of our analysis, we investigate how the heavy-tail index p explicitly impacts sample complexity. We compare three distinct regimes: standard first-order optimization (clipNSGDM, (Cutkosky and Mehta, 2021; Liu et al., 2023)), second-order optimization (D-clip-NSGDHess), and first-order optimization with variance reduction under the 2-WAS assumption (D-clip-NSGD-MVR).

Theoretical Scaling: In the left panel of Figure 2, we plot the theoretical complexity exponents c for the bound $\mathcal{O}(\varepsilon^{-c})$ as a function of the tail index $p \in (1, 2]$. As expected, in the bounded variance regime ($p = 2$), the standard first-order method requires $\mathcal{O}(\varepsilon^{-4})$, while both the second-order method and the 2-WAS variance-reduced method achieve the optimal $\mathcal{O}(\varepsilon^{-3})$ complexity. However, as the noise becomes increasingly heavy-tailed ($p \to 1$), the required complexity approaches infinity across all methods. Notably, the theoretical curves reveal that under the 2-WAS assumption, our variance-reduced method (D-clip-NSGD-MVR) achieves a strictly better polynomial dependence on ε than even the second-order method for all $p < 2$.

Empirical Validation: To validate these theoretical scaling laws empirically, we track the exact number of iterations required for each algorithm to reach a fixed suboptimality threshold ($F(x_t) < 1.2$) as we vary p from 1.1 to 2.0 (Figure 2, Right). The empirical results perfectly mirror the theoretical predictions. clip-NSGDM consistently requires the most iterations to converge. Furthermore, the empirical iteration count confirms the exact crossover phenomenon predicted by our lower bounds: D-clip-NSGD-MVR requires fewer iterations than the second-order D-clip-NSGDHess in the severe heavy-tailed regime ($p < 2$). This compellingly demonstrates that variance reduction under the 2-WAS assumption is a highly effective mechanism for accelerating convergence under extreme noise, completely aligning with our established theory.