
Second-order Optimization under Heavy-Tailed Noise: Hessian Clipping and Sample Complexity Limits

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 Heavy-tailed noise is pervasive in modern machine learning applications, arising
2 from data heterogeneity, outliers, and stochastic non-stationary environments.
3 While second-order methods can significantly accelerate convergence in light-tailed
4 or bounded-noise settings, such algorithms are often brittle and lack guarantees
5 under heavy-tailed noise—precisely the regimes where robustness is most critical.
6 In this work, we take a first step toward a theoretical understanding of second-order
7 optimization under heavy-tailed noise. We consider a setting where stochastic
8 gradients and Hessians have only bounded p -th moments for $p \in (1, 2]$, and
9 establish tight lower bounds on the sample complexity of any second-order method.
10 We then develop a variant of normalized stochastic gradient descent that leverages
11 second-order information and provably matches these lower bounds. To address
12 the instability caused by large deviations, we introduce a novel algorithm based on
13 gradient and Hessian clipping, and prove high-probability upper bounds that nearly
14 match the fundamental limits. Our results provide the first comprehensive sample
15 complexity characterization for second-order optimization under heavy-tailed noise.
16 This positions Hessian clipping as a robust and theoretically sound strategy for
17 second-order algorithm design in heavy-tailed regimes.

18 1 Introduction

19 We consider the stochastic optimization problem

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

20 with F a smooth (but potentially nonconvex) objective and ξ a random variable drawn from an
21 unknown distribution. While first-order methods are widely used in practice due to their simplicity
22 and scalability, the application of second-order methods remains limited in stochastic settings. This is
23 in stark contrast to their strong theoretical properties: second-order methods such as Newton’s method
24 [Bennett, 1916, Kantorovich, 1948], cubic regularization [Nesterov and Polyak, 2006, Nesterov,
25 2008], quasi-Newton [Dennis and Moré, 1977], and adaptive second-order algorithms [Doikov
26 et al., 2024] can offer provably faster convergence rates. Furthermore, complexity-theoretic results
27 demonstrate that second-order methods can outperform first-order ones in oracle settings, often with
28 only moderate additional computational cost [Agarwal and Hazan, 2018, Arjevani et al., 2019b,
29 2020].

30 **Second-order methods are highly sensitive to noise.** First-order stochastic optimization (FOSO)
31 methods—such as stochastic gradient descent (SGD) and its adaptive variants—are well understood,
32 both algorithmically and from a complexity-theoretic perspective. These methods enjoy optimal
33 convergence guarantees under a broad class of noise models, including bounded variance and infinite

variance regimes [Gower et al., 2019, Khaled and Richtárik, 2020, Wang et al., 2021, Yang et al., 2023]. In contrast, the development of second-order stochastic optimization (SOSO) methods has been significantly hampered by their extreme sensitivity to noise, particularly in Hessian estimation. Existing second-order methods suffer from both practical instability and overly restrictive theoretical assumptions. For instance, stochastic extensions of cubic regularization [Ghadimi et al., 2017, Tripuraneni et al., 2018] and its momentum variants [Chayti et al., 2024] require bounded noise assumptions, which are rarely satisfied in modern large-scale learning tasks. Other frameworks—such as trust-region methods [Arjevani et al., 2020], recursive momentum schemes [Tran and Cutkosky, 2021], and extrapolation-based approaches [Antonakopoulos et al., 2022, Agafonov et al., 2023]—relax this to bounded variance, but still fall short of handling more realistic heavy-tailed noise distributions. To the best of our knowledge, no existing SOSO method can operate reliably under noise conditions with unbounded variance, highlighting a fundamental gap in the current landscape of stochastic optimization.

Heavy-tailed noise in gradients and Hessians. In recent years, the first-order optimization literature has increasingly moved beyond the bounded variance assumption, adopting more general noise models that better reflect empirical observations in modern applications. A particularly influential framework is the *bounded central moment* condition (p -BCM), which assumes

$$\mathbb{E} [\|\nabla f(x, \xi) - \nabla F(x)\|^p] \leq \sigma^p, \quad \text{for some } p \in (1, 2],$$

thereby allowing for heavy-tailed noise with unbounded variance. This model aligns well with empirical findings in deep learning, reinforcement learning (RL), and large-scale models, where heavy-tailed gradient noise is often observed [Garg et al., 2021, Zhang et al., 2020, Ahn et al., 2024, Simsekli et al., 2019, Battash et al., 2024]. Algorithms designed for robustness under p -BCM noise—such as those employing gradient clipping or normalization—have demonstrated both strong empirical performance and rigorous convergence guarantees [Zhang et al., 2020, Hübler et al., 2024]. These developments suggest that heavy-tailed noise models are not only practically relevant but also theoretically tractable, providing a principled foundation for algorithm design. Given that second-order methods are even more sensitive to noise than first-order ones, this naturally motivates the extension of such robustness principles to the second-order setting. In this work, we take this step and aim to

develop a comprehensive theory of second-order stochastic optimization under heavy-tailed noise.

Specifically, we assume access to unbiased stochastic gradients and Hessian-vector products, each satisfying a p -BCM-type condition:

$$\mathbb{E} [\|\nabla^2 f(x, \xi) - \nabla^2 F(x)\|_{\text{op}}^p] \leq \sigma_h^p, \quad \text{for some } p \in (1, 2].$$

Our goal is to characterize the fundamental performance limits and develop practical algorithms for minimizing $F(x)$ in this setting, with a focus on obtaining guarantees for finding points with small gradient norm $\|\nabla F(x)\| \leq \varepsilon$, either in expectation or with high probability.¹

From gradient to Hessian clipping. Gradient clipping has become a standard tool in modern machine learning, particularly due to its empirical success in stabilizing training under heavy-tailed noise and ill-conditioned objectives [Pascanu et al., 2013, Schulman et al., 2017]. Theoretically, it has been shown to offer robustness under relaxed smoothness and moment assumptions [Polyak and Tsytkin, 1979, Jakovetić et al., 2023], and to enable high-probability convergence guarantees with only logarithmic dependence on the failure probability. These properties are well-documented across 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., 2023] and nonconvex [Sadiev et al., 2023, Nguyen et al., 2023, Cutkosky and Mehta, 2021, Hübler et al., 2024] regimes. Crucially, high-probability results are both theoretically and practically appealing as they guarantee the performance of individual runs rather than average-case behavior.

Despite these advances, the robustness enabled by gradient clipping remains largely confined to first-order methods. In the second-order setting, where both gradients and Hessians may be corrupted by heavy-tailed noise, comparable algorithmic tools are conspicuously lacking. The core difficulty is

¹It would be interesting to extend our results to finding second-order stationary points [Nesterov and Polyak, 2006, Arjevani et al., 2019a], but we leave this investigation for future work.

not just technical but conceptual: existing SOSO algorithms do not include mechanisms to suppress the influence of extreme noise in Hessian estimates. This leads us to a fundamental question for the design of robust second-order stochastic methods:

Can we develop second-order algorithms with high-probability convergence guarantees under simultaneous heavy-tailed noise in both gradients and Hessians?

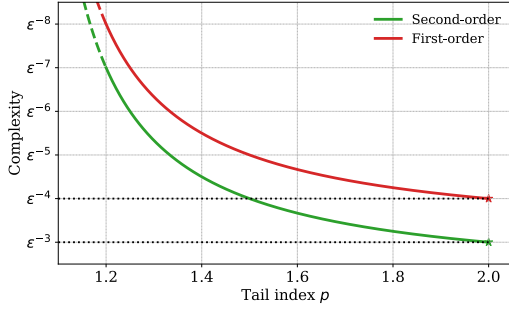


Figure 1: Sample complexity comparison for FOSO and SOSO depending on the tail index p . Each line corresponds to the leading term in the sample complexity for each class of algorithms. These leading terms match in upper and lower bounds, so this characterization is exact. We establish the characterization along the entire green line complementing the prior work [Arjevani et al., 2019a] for $p = 2$.

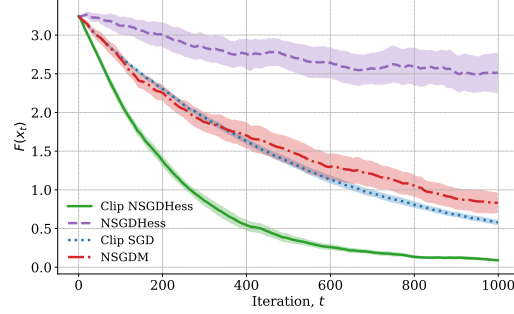


Figure 2: Performance of algorithms on a simple problem $F(x) = 0.5 \|x\|^2$, $d = 10$ with synthetic noise generated from symmetrized Pareto distribution with tail index $p = 1.1$. We observe that algorithms without clipping, NSGDM and NSGDMHess, suffer significantly from noise. This motivates our more in-depth study involving gradient and Hessian clipping for high probability convergence.

Contributions:

- **Tight lower bound.** We establish a minimax lower bound on the sample complexity of any SOSO algorithm under p -BCM noise model: $\Omega\left(\frac{\Delta\sigma_h}{\varepsilon^2}\left(\frac{\sigma}{\varepsilon}\right)^{\frac{1}{p-1}}\right)$, where Δ denotes the initial suboptimality, and σ, σ_h denote the scale of noise in gradients and Hessians, respectively. This bound improves over the best known complexity for first-order methods [Zhang et al., 2020, Hübner et al., 2024] by a factor of $1/\varepsilon$ uniformly for all $p \in (1, 2]$, demonstrating that second-order methods can yield provable advantages even in the heavy-tailed regime, see Figure 1. Moreover, our result implies that increasing the order of the algorithm (e.g., using higher-order derivatives) cannot significantly improve the rate under the same noise assumptions.
- **Optimal algorithm.** We develop a second-order stochastic algorithm that achieves the lower bound (up to constants) in the regime $L \leq \sigma_h$, where L denotes the Lipschitz constant of the gradient. The algorithm attains a sample complexity of $\mathcal{O}\left(\frac{\Delta(L+\sigma_h)}{\varepsilon^2}\left(\frac{\sigma}{\varepsilon}\right)^{\frac{1}{p-1}}\right)$, and is, to our knowledge, the first second-order method that provably works under p -BCM noise. Notably, our method does not require second-order smoothness of the objective. This result holds even in the classical setting $p = 2$, highlighting the broader implications of the approach.
- **High-probability convergence via Hessian clipping.** We propose a clipped variant of our algorithm that incorporates both gradient and *Hessian clipping*—the latter introduced here for the first time. We show that this variant achieves near-optimal sample complexity with high probability, incurring only a poly-logarithmic overhead in the failure probability. This significantly extends the robustness of SOSO methods, enabling strong guarantees not only in expectation but also with high confidence.

Our results are primarily theoretical and validated through synthetic experiments with controllable heavy-tailed noise. While one component of our algorithm (in its unclipped form) has previously been applied in reinforcement learning settings [Fatkhullin et al., 2023], its broader practical deployment

remains an open challenge. In particular, several questions must be addressed to make these methods viable in real-world applications: How can we reduce hyperparameter tuning overhead? Are there default configurations that are robust across problem classes? And is clipping truly necessary, or can it be systematically replaced by normalization techniques, as recent work suggests in the first-order context [Hübner et al., 2024]? We view our contributions as a foundational step toward a principled understanding of second-order methods under realistic, heavy-tailed noise conditions, and hope they spur further theoretical and algorithmic advances in this direction.

2 Notations and Assumptions

We use the common notation for natural numbers $\mathbb{N} = \{0, 1, \dots\}$, $[n] = \{1, 2, \dots, n\}$. Throughout this paper, $d \in \mathbb{N}_{\geq 1}$ denotes the dimension of the optimization problem (1). We use $\mathcal{O}(\cdot)$, $\Omega(\cdot)$ for sample complexity notations, which preserve the dominating term in the target accuracy $\varepsilon > 0$ along with the multiplicative constants such as initial suboptimality Δ , smoothness constants L_r , $r \geq 1$, the variance of higher-order stochastic oracles σ_r , $r \geq 1$. When we write $\bar{\mathcal{O}}(\cdot)$, we suppress the dependence on all parameters except for accuracy ε in the dominating term.

We make the following assumptions throughout the paper.

Assumption 1 (Lower Boundedness). *The objective function F is lower bounded by $F^* > -\infty$.*

Assumption 2 (L -smoothness). *The objective function F is L -smooth, i.e. F is differentiable and for all $x, y \in \mathbb{R}^d$ we have $\|\nabla F(x) - \nabla F(y)\| \leq L\|x - y\|$.*

Assumption 3 (p -BCM). *We have access to stochastic gradients $\nabla f(x, \xi)$ such that $\mathbb{E}[\nabla f(x, \xi)] = \nabla F(x)$. Moreover, there exist $p \in (1, 2]$ and $\sigma > 0$ such that*

$$\mathbb{E}[\|\nabla f(x, \xi) - \nabla F(x)\|^p] \leq \sigma^p \quad \text{for any } x \in \mathbb{R}^d.$$

Assumption 4 (p -BCM for Hessian). *The function F is twice differentiable and we additionally have access to stochastic Hessian-vector products $\nabla^2 f(x, \xi) \cdot v$ for any $v \in \mathbb{R}^d$ such that $\mathbb{E}[\nabla^2 f(x, \xi)] = \nabla^2 F(x)$. Moreover, there exist $p \in (1, 2]$ and $\sigma_h > 0$ such that, for all $x \in \mathbb{R}^d$*

$$\mathbb{E}[\|\nabla^2 f(x, \xi) - \nabla^2 F(x)\|_{\text{op}}^p] \leq \sigma_h^p.$$

3 Lower Complexity Bounds for SOSO under p -BCM

To establish lower bounds, we build on a technique developed in a series of works [Arjevani et al., 2019a, Carmon et al., 2020, Arjevani et al., 2020], which introduced a *worst-case* non-convex function exhibiting the zero-chain property: starting from the origin, each algorithmic iteration reveals information about only a single coordinate. Leveraging this framework, Zhang et al. [2020] derived the first lower bounds for FOSO under heavy-tailed noise, assuming the p -th moment of the stochastic gradient is bounded, i.e., $\mathbb{E}[\|\nabla f(x, \xi)\|^p] \leq G^p$, which differs from Assumption 3.

In our work, we assume access to a q -th order stochastic oracle, meaning that each query yields stochastic approximations of derivatives of orders 1 to q at a given point x . This oracle model was formally introduced by Arjevani et al. [2020]. Building on their proof technique, we derive new lower bounds under the assumption that each stochastic derivative has a bounded p -th central moment.

Function class. We consider the class of q -times differentiable functions ($q \geq 1$) satisfying two key properties: (i) functional boundedness, i.e., $F(0) - F^* \leq \Delta$; and (ii) Lipschitz continuity of each derivative up to order q , i.e., for all $x, y \in \mathbb{R}^d$ and $r \in [q]$,

$$\|\nabla^r F(x) - \nabla^r F(y)\|_{\text{op}} \leq L_r \|x - y\|.$$

We denote this function class by $\mathcal{F}(\Delta, L_{1:q})$, where $L_{1:q} = \{L_1, L_2, \dots, L_q\}$. Note that the constant L from Assumption 2 corresponds to L_1 in this notation.

Oracle class. We now define the oracle model. For a fixed function $F \in \mathcal{F}(\Delta, L_{1:q})$, a stochastic q -th order oracle is defined as follows: for any $x \in \mathbb{R}^d$ and $\xi \sim \mathcal{D}$,

$$\mathcal{O}_f^q(x, \xi) \stackrel{\text{def}}{=} (f(x, \xi), \nabla f(x, \xi), \nabla^2 f(x, \xi), \dots, \nabla^q f(x, \xi)),$$

where each component satisfies the unbiasedness conditions: $\mathbb{E}_{\xi \sim \mathcal{D}}[f(x, \xi)] = F(x)$, and $\mathbb{E}_{\xi \sim \mathcal{D}}[\nabla^r f(x, \xi)] = \nabla^r F(x)$, $\forall r \in [q]$. We define the oracle class $\mathcal{O}_q(F, \sigma_{1:q})$ as the set of all such stochastic oracles for which the p -th moment of the estimation error (with $p \in [1, 2)$) satisfies:

$$\mathbb{E}_{\xi \sim \mathcal{D}} \left[\|\nabla^r f(x, \xi) - \nabla^r F(x)\|_{\text{op}}^p \right] \leq \sigma_r^p, \quad \forall r \in [q].$$

For $r = 1$, this recovers $\sigma_1 = \sigma$ from Assumption 3, and for $r = 2$, we similarly have $\sigma_2 = \sigma_h$ from Assumption 4.

Algorithm class. To prove lower bounds we will work with a class of *zero-respecting* algorithms. The formal definition is provided below. The motivation why this class of algorithms is interesting for us is based on its property: on each iteration of such algorithm we can reveal only maximum one new coordinate.

Definition 1. We call a stochastic q -th order algorithm A zero-respecting if for any function F and any p -th order oracle O_F^q , the iterates $\{x^t\}_{t \in \mathbb{N}}$ generated by A via querying O_F^q satisfy the following property: $\text{supp}\{x^t\} \subset \bigcup_{i < t} \text{supp}\{O_F^q(x^i, \xi^i)\}$, $\forall t \in \mathbb{N}$ with probability one with respect the randomness of the algorithm and the realizations of $\{\xi^t\}_{t \in \mathbb{N}}$.

Theorem 1. Let $q \in \mathbb{N}$, and let $\Delta > 0$, $L_{1:q} \stackrel{\text{def}}{=} (L_1, \dots, L_q)$, $\sigma_{1:q} \stackrel{\text{def}}{=} (\sigma_1, \dots, \sigma_q)$ and $\varepsilon \leq \mathcal{O}(\sigma_1)$. Then, there exists $F \in \mathcal{F}(\Delta, L_{1:q})$ and a corresponding noisy oracle $O_F^q \in \mathcal{O}(F, \sigma_{1:q})$ such that for any q -th order zero-respecting algorithm, the number of oracle queries required to find an ε -stationary point (with constant probability) is lower bounded by

$$\Omega(1) \cdot \frac{\Delta}{\varepsilon} \left(\frac{\sigma_1}{\varepsilon} \right)^{\frac{p}{p-1}} \min \left\{ \min_{r \in \{2, \dots, q\}} \left(\frac{\sigma_r}{\sigma_1} \right)^{\frac{1}{r-1}}, \min_{r' \in [q]} \left(\frac{L_{r'}}{\varepsilon} \right)^{\frac{1}{r'}} \right\}. \quad (2)$$

Moreover, this lower bounds is realized by a construction of dimension $\Theta \left(\frac{\Delta}{\varepsilon} \min \left\{ \min_{r \in \{2, \dots, q\}} \left(\frac{\sigma_r}{\sigma_1} \right)^{\frac{1}{r-1}}, \min_{r' \in [q]} \left(\frac{L_{r'}}{\varepsilon} \right)^{\frac{1}{r'}} \right\} \right)$.

The proof of Theorem 1 is deferred to Appendix D. In Theorem 1 we provide lower bounds for any $q \in \mathbb{N}$, which corresponds to the order of an oracle used in an optimization algorithm. Since the main focus of our work is the second order information and the benefits to improving complexity compare to the first order, we discuss a complexity result from Theorem 1 for the case $q = 2$. For second-order methods $q = 2$, under Assumptions 2, 3, 4 and, additionally, (possibly) assuming the Hessian is Lipschitz continuous with constant L_h , we have

$$\Omega(1) \cdot \min \left\{ \frac{\Delta \sigma_h}{\varepsilon^2} \left(\frac{\sigma}{\varepsilon} \right)^{\frac{1}{p-1}}, \frac{\Delta L \sigma}{\varepsilon^3} \left(\frac{\sigma}{\varepsilon} \right)^{\frac{1}{p-1}}, \frac{\Delta \sqrt{L_h} \sigma}{\varepsilon^{5/2}} \left(\frac{\sigma}{\varepsilon} \right)^{\frac{1}{p-1}} \right\}, \quad (3)$$

where the parameters are that $\sigma = \sigma_1$, $\sigma_h = \sigma_2$, $L = L_1$, $L_2 = L_h$. Now we will discuss each term separately in different regimes.

First-order optimization with Lipschitz gradient. When we do not have access to the second-order information and the objective function has only Lipschitz continuous gradient, but not any higher derivative, our lower bounds (3) reduces to second term:

$$\Omega \left(\frac{\Delta L \sigma}{\varepsilon^3} \left(\frac{\sigma}{\varepsilon} \right)^{\frac{1}{p-1}} \right).$$

This bound exactly matches (up to a numerical constant) the previously known lower bounds for first-order stochastic optimization under p -BCM [Zhang et al., 2020], and the corresponding upper bound achieved by normalized SGD [Hübner et al., 2024]. Several other algorithms like Clip-SGD and NSGD with momentum and clipping also nearly match this complexity lower bound up to additional polylogarithmic terms in $1/\varepsilon$ and polynomial terms in δ , L and σ [Zhang et al., 2020, Cutkosky and Mehta, 2021, Nguyen et al., 2023].

First-order optimization under higher-order smoothness. The second case worth discussion is when we still do not have access to stochastic Hessian, but we know that an objective function has Lipschitz continuous second order derivatives. Then our lower bound in (3) becomes

$$\Omega \left(\min \left\{ \frac{\Delta L \sigma}{\varepsilon^3} \left(\frac{\sigma}{\varepsilon} \right)^{\frac{1}{p-1}}, \frac{\Delta \sqrt{L_h} \sigma}{\varepsilon^{5/2}} \left(\frac{\sigma}{\varepsilon} \right)^{\frac{1}{p-1}} \right\} \right).$$

189 The last term in the above bound nearly matches with the complexity of method called NIGT
 190 with clipping from the paper of Cutkosky and Mehta [2021] in terms of dependence on $1/\varepsilon$ up to
 191 logarithmic factors. Unfortunately, there is still a small discrepancy with the upper bound in the
 192 dependence on other parameters and with that Cutkosky and Mehta [2021] use a slightly stronger
 193 (non-central) assumption $\mathbb{E} [\|\nabla f(x, \xi)\|^p] \leq G^p$ instead of our p -BCM. We believe our bound is tight
 194 and can be achieved with a more careful analysis of NIGT type algorithm under p -BCM assumption.

195 **Second-order optimization with only Lipschitz gradient.** Finally, we wish to discuss an important
 196 setting, which is rarely discussed in the literature on second-order optimization. Higher-order
 197 smoothness can be challenging to verify in practice and even when it is possible, the estimates of L_h
 198 can be too large to be useful. Thus we pay attention to the case when we have access to stochastic
 199 Hessian, but the second derivative can be non-continuous, i.e. $L_h \rightarrow +\infty$. Our lower bound (3)
 200 provides new insights on this interesting setting, simplifying to

$$\Omega \left(\min \left\{ \frac{\Delta \sigma_h}{\varepsilon^2} \left(\frac{\sigma}{\varepsilon} \right)^{\frac{1}{p-1}}, \frac{\Delta L \sigma}{\varepsilon^3} \left(\frac{\sigma}{\varepsilon} \right)^{\frac{1}{p-1}} \right\} \right).$$

201 As we can see, the first term corresponding to the second-order optimization has better dependence on
 202 an accuracy ε , and can be potentially much smaller than the second one. Moreover, this complexity is
 203 not impacted with the constants corresponding to higher order smoothness, which potentially gives
 204 us the possibility of designing a second-order methods with this complexity that does not depend on
 205 L_h and higher order smoothness constants. However, this is merely a lower bound, which does not
 206 give us any algorithmic solution yet. The question we investigate in the next section:

207 *Can we design an algorithm handling heavy-tailed noise with complexity $\mathcal{O} \left(\frac{\Delta \sigma_h}{\varepsilon^2} \left(\frac{\sigma}{\varepsilon} \right)^{\frac{1}{p-1}} \right)$?*

208 4 Near-optimal Second-Order Method under p -BCM

209 In the last decade, the stochastic second-order optimization (SOSO) has been extensively studied
 210 under stronger noise assumptions. We will first overview the existing approaches, particularly
 211 focusing on the development in the non-convex setup. First, Tripuraneni et al. [2018] proposed and
 212 analyzed Stochastic Cubic Newton (SCN) method achieving $\tilde{\mathcal{O}}(\varepsilon^{-7/2})$ sample complexity. They
 213 require a strong bounded noise assumption, use large mini-batches and importantly this complexity is
 214 not order optimal. Later Arjevani et al. [2020] established lower bounds for SOSO to find first-order
 215 stationary points and proposed two algorithms: SGD with recursive variance reduction with stochastic
 216 Hessian-vector products (HVP-RVR) and Subsampled cubic-regularized trust-region method with
 217 HVP-RVR. Both algorithms achieve $\tilde{\mathcal{O}}(\varepsilon^{-3})$ sample complexities, which is min-max optimal when
 218 the variances of stochastic gradient and Hessian are bounded. Besides strong noise assumptions, their
 219 algorithms require large batch sizes of Hessians, since they used HVP-RVR subroutine to update the
 220 momentum term. To overcome the large batch-size requirement Tran and Cutkosky [2021] design
 221 a simpler algorithm: SGD with Hessian-corrected momentum (with optional normalization) called
 222 SGDHess. Thanks to the Hessian corrected momentum term, their methods also achieve $\tilde{\mathcal{O}}(\varepsilon^{-3})$
 223 sample complexity with any batch size ≥ 1 . Recently, Chayti et al. [2024] propose a variant of SCN
 224 with momentum without large batch requirement. However, their analysis requires bounded noise
 225 assumption and only achieves $\tilde{\mathcal{O}}(\varepsilon^{-7/2})$ sample complexity.²

226 To summarize, all above mentioned works require strong noise assumptions: bounded noise and
 227 bounded variance. Moreover, their sample complexities also depend on second-order smoothness L_h ,
 228 which can be potentially infinite as we discussed in the previous section. We refer to Table 1 for the
 229 summary of existing results.

230 To design an optimal second-order algorithm without these limitations, we take an inspiration from
 231 the recent developments in reinforcement learning [Salehkaleybar et al., 2022, Fatkhullin et al.,
 232 2023]. Their algorithms called SHARP and (N)HAR-PG are variants of SGDHess in [Tran and
 233 Cutkosky, 2021] with optional normalization step. We adapt one variant of their method to our
 234 general stochastic optimization setting and present it in Algorithm 1. This algorithm computes a
 235 random interpolation point \hat{x}_t uniformly distributed between consecutive iterates x_{t-1} and x_t . Then

²However, in the noiseless case their complexity has better $\tilde{\mathcal{O}}(\varepsilon^{-3/2})$ iteration complexity compared to $\tilde{\mathcal{O}}(\varepsilon^{-2})$ for SGDHess.

Algorithm 1 NSGDHess (Normalized SGD with Hessian correction)

1: **Input:** Starting point $x_0 \in \mathbb{R}^d$, a vector $g_0 \in \mathbb{R}^d$, a stepsize $\gamma > 0$, momentum parameters $\alpha > 0$, the number of iterations T .
 2: $x_1 = x_0 - \gamma \frac{g_0}{\|g_0\|}$
 3: **for** $t = 1, 2, \dots, T - 1$ **do**
 4: Sample $q_t \sim \mathcal{U}([0, 1])$
 5: $\hat{x}_t = q_t x_t + (1 - q_t) x_{t-1}$
 6: Sample $\xi_t, \hat{\xi}_t \sim \mathcal{D}$ independently
 7: $g_t = (1 - \alpha) \left(g_{t-1} + \nabla^2 f(\hat{x}_t, \hat{\xi}_t)(x_t - x_{t-1}) \right) + \alpha \nabla f(x_t, \xi_t)$
 8: $x_{t+1} = x_t - \gamma \frac{g_t}{\|g_t\|}$
 9: **end for**
 10: **Output:**

236 it evaluates a stochastic gradient and stochastic Hessian-vector product using independent samples ξ_t
 237 and $\hat{\xi}_t$. Finally, a recursive momentum is constructed before applying the normalization step.³ Now
 238 we are ready to state the convergence guarantee for Algorithm 1.

Theorem 2. Suppose Assumptions 1, 2, 3 and 4 hold. Let the initial gradient estimate be given by

$$g_0 = \frac{1}{B_{\text{init}}} \sum_{j=1}^{B_{\text{init}}} \nabla f(x_0, \xi_{0,j}), \text{ where } B_{\text{init}} = \max \left\{ 1, \left(\frac{\sigma}{\varepsilon} \right)^{\frac{p}{p-1}}, \left(\frac{\sigma}{\varepsilon} \right)^{\frac{p}{2p-1}} \right\}.$$

239 Set stepsize as $\gamma = \sqrt{\frac{\Delta \alpha^{1/p}}{(L + \sigma_h)T}}$, momentum parameter as $\alpha = \min \{1, \alpha_{\text{eff}}\}$, where $\alpha_{\text{eff}} =$
 240 $\max \left\{ \left(\frac{\varepsilon_0}{T\sigma} \right)^{\frac{p}{2p-1}}, \left(\frac{\Delta(L + \sigma_h)}{T\sigma^2} \right)^{\frac{p}{2p-1}} \right\}$, $\varepsilon_0 = 2\sigma/B_{\text{init}}^{\frac{p-1}{p}}$. Then, Algorithm 1 guarantees that
 241 $\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} [\|\nabla F(x_t)\|] \leq \varepsilon$ with the total sample complexity

$$\mathcal{O} \left(\frac{\Delta(L + \sigma_h)}{\varepsilon^2} + \frac{\Delta(L + \sigma_h)}{\varepsilon^2} \left(\frac{\sigma}{\varepsilon} \right)^{\frac{1}{p-1}} + \frac{\sigma}{\varepsilon} \left(\frac{\sigma}{\varepsilon} \right)^{\frac{1}{p-1}} \right). \quad (4)$$

242 The proof is deferred to the Appendix E. We also investigate different choices of g_0 and their influence
 243 on the total sample complexity rate in Appendix E. We find that the initial batch size to estimate g_0 is
 244 not necessary for convergence, but helps to slightly improve the total sample complexity.

245 **Discussion:** Comparing this result with Theorem 1, we observe that our upper bound (4) matches our
 246 lower bound—the first term in (3) in terms of the target accuracy ε . Moreover, when $\Delta(L + \sigma_h) \geq \sigma\varepsilon$,
 247 we have a tight upper bound, which exactly matches the lower bound from previous section in the
 248 leading term (up to a numerical constant). Moreover, since our upper bound does not depend on
 249 constant L_h , it explains our observations in Section 3 regarding the limit case $L_h \rightarrow \infty$, and answers
 250 the raised question affirmatively. We refer to Table 1 for more technical comparison with prior work.

251 5 Hessian Clipping for High-probability Convergence

252 In Section 4 we provide an in-expectation guarantee for Algorithm 1. While in-expectation guarantees
 253 are useful in the case when we are allowed to run method multiple times to analyze average-case
 254 behavior, it can be impractical. In many real-world scenarios only a single run of the methods can be
 255 performed, due to computational or time constraints. Thus, our main goal in this section is to conduct
 256 high-probability analysis for Algorithm 1 or its modification.

257 Since we work in the unbounded variance case, which corresponds to heavy-tailed noise, it is
 258 important to ensure that the analyzed method is robust to such noise. Following works of Gorbunov
 259 et al. [2020], Cutkosky and Mehta [2021], Sadiev et al. [2023], Nguyen et al. [2023], we propose
 260 a new algorithm called Normalized SGD with momentum and Hessian clipping. It is presented in

³Normalization step for this method is important for our analysis under p -BCM, however, it can be removed when $p = 2$ or when additional clipping is used as in the next section.

Algorithm 2, where we incorporate gradient clipping and Hessian clipping compared to Algorithm 1. Formally, a clipping operator (clipping for short) is defined as

$$\text{clip}(v, \lambda) = \min \left\{ 1, \frac{\lambda}{\|v\|} \right\} v \quad \text{for any } v \neq 0 \text{ from } \mathbb{R}^d, \quad (5)$$

where $\lambda > 0$ is called clipping level/threshold.

Algorithm 2 Clip NSGDHess (Normalized SGD with Hessian correction and clipping)

1: **Input:** Starting point $x_0 \in \mathbb{R}^d$, a vector $g_0 = 0 \in \mathbb{R}^d$, a stepsize $\gamma > 0$, momentum parameters $\alpha > 0$, clipping levels $\lambda > 0, \bar{\lambda}_h > 0$, the number of iterations T .
2: $x_1 = x_0$ and $g_0 = 0$
3: **for** $t = 1, 2, \dots, T - 1$ **do**
4: Sample $q_t \sim \mathcal{U}([0, 1])$
5: $\hat{x}_t = q_t x_t + (1 - q_t) x_{t-1}$
6: Sample $\xi_t, \hat{\xi}_t \sim \mathcal{D}$ independently
7: $g_t = (1 - \alpha) \left(g_{t-1} + \gamma \text{clip} \left(\gamma^{-1} \nabla^2 f(\hat{x}_t, \hat{\xi}_t)(x_t - x_{t-1}), \bar{\lambda}_h \right) \right) + \alpha \text{clip}(\nabla f(x_t, \xi_t), \lambda)$
8: $x_{t+1} = x_t - \gamma \frac{g_t}{\|g_t\|}$
9: **end for**
10: **Output:**

As in Algorithm 1, we use normalization in the gradient update (line 8), which allows the next point x_{t+1} to stay in the ball with center at x_t with radius γ . The main difference between Algorithms 1 and 2 is in the momentum term. To enhance the robustness we clip not only gradient, but also the Hessian-vector product. As we mentioned above, gradient clipping is a common approach, but Hessian clipping is new and we it allows us to provide the first high-probability guarantees for second-order stochastic optimization without light-tail noise assumptions. It is worth to draw attention to the following term:

$$\gamma \text{clip} \left(\gamma^{-1} \nabla^2 f(\hat{x}_t, \hat{\xi}_t)(x_t - x_{t-1}), \bar{\lambda}_h \right). \quad (6)$$

Observe that clipping the entire Hessian based on its operator norm would have been very costly since it typically requires $\mathcal{O}(d^2)$ arithmetic operations (e.g., using power iteration [Golub and Van Loan, 2013] or Lanczos algorithm [Lanczos, 1950, Saad, 2011]). Instead of clipping the stochastic Hessian, we clip the Hessian-vector product, which is computationally tractable and only required $\mathcal{O}(d)$ operations. Thus the computation of this clipped Hessian-vector product can be easily implemented via backpropagation and the subsequent “vector” clipping (5). Another reason why we prefer to use this form of clipping comes from the analysis: we want to preserve the following useful property:

$$\begin{aligned} \mathbb{E}_{q_t, \hat{\xi}_t} \left[\nabla^2 f(\hat{x}_t, \hat{\xi}_t)(x_t - x_{t-1}) \right] &= \mathbb{E}_{q_t} \left[\mathbb{E}_{\hat{\xi}_t} \left[\nabla^2 f(\hat{x}_t, \hat{\xi}_t) \right] (x_t - x_{t-1}) \right] \\ &= \mathbb{E}_{q_t} \left[\nabla^2 F(\hat{x}_t)(x_t - x_{t-1}) \right] \\ &= \int_0^1 \nabla^2 F(qx_t + (1 - q)x_{t-1})(x_t - x_{t-1}) dq \\ &= \nabla F(x_t) - \nabla F(x_{t-1}), \end{aligned} \quad (7)$$

which would have been invalidated if we were to use the direct Hessian matrix clipping. Instead, we prove that the expectation of (6) will approximately be equal to (7) when $\bar{\lambda}_h$ is selected properly. By smoothness and gradient step, we have $\|\nabla F(x_t) - \nabla F(x_{t-1})\| \leq L\|x_t - x_{t-1}\| = L\gamma$, meaning the expectation is bounded. To provide analysis, we need to rewrite clipped Hessian-vector product as follows

$$\gamma \text{clip} \left(\gamma^{-1} \nabla^2 f(\hat{x}_t, \hat{\xi}_t)(x_t - x_{t-1}), \bar{\lambda}_h \right) = \text{clip} \left(\nabla^2 f(\hat{x}_t, \hat{\xi}_t)(x_t - x_{t-1}), \gamma \bar{\lambda}_h \right),$$

where we denote $\lambda_h = \gamma \bar{\lambda}_h$. The former formulation in the above formula is useful for implementation, since as we will see that $\lambda_h \sim \mathcal{O}(\gamma \alpha^{-1/p})$ and $\lambda \sim \mathcal{O}(\alpha^{-1/p})$. Thus, λ and $\bar{\lambda}_h$ are of the same order $\sim \mathcal{O}(\alpha^{-1/p}) = \mathcal{O}(T^{\frac{1}{2p-1}})$, which makes it easier to tune these 2 parameters together in practice. The latter formulation will be used in the high-probability analysis. Now we are ready to state the main theorem of this section.

Theorem 3. Let Assumptions 1, 2, 3, and 4 hold, and define the initial optimality gap as $\Delta_1 := F(x_0) - F_*$. Let $\delta \in (0, 1]$ and $T \geq 1$ be such that $\log(8T/\delta) \geq 1$. Suppose Algorithm 2 is executed with the following settings: momentum parameter $\alpha = T^{-\frac{p}{2p-1}}$, clipping levels $\lambda = \max \left\{ 4\sqrt{L\Delta_1}, \frac{\sigma}{\alpha^{1/p}} \right\}$, $\lambda_h = \frac{2\gamma(L+\sigma_h)}{\alpha^{1/p}}$, stepsize γ satisfying:

$$\gamma \leq \mathcal{O} \left(\min \left\{ \sqrt{\frac{\Delta_1}{LT}}, \alpha \sqrt{\frac{\Delta_1}{L}}, \frac{\sqrt{\Delta_1/L}}{\alpha T \log(T/\delta)}, \frac{\Delta_1}{\sigma \alpha^{\frac{p-1}{p}} T \log(T/\delta)}, \sqrt{\frac{\Delta_1 \alpha^{1/p}}{(L + \sigma_h) T \log(T/\delta)}} \right\} \right).$$

Then, with probability at least $1 - \delta$, the output of the algorithm satisfies: $\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla F(x_t)\| \leq \frac{2\Delta_1}{\gamma T}$. As a result, the gradient norm converges at the rate:

$$\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla F(x_t)\| = \mathcal{O} \left(\frac{\sqrt{\Delta_1(L + \sigma_h)} + \sigma}{T^{\frac{p-1}{2p-1}}} \log \left(\frac{T}{\delta} \right) \right)$$

with high probability.

The complete proof of Theorem 3 is in Appendix F. Disregarding parameters $L, \sigma, \sigma_h, \Delta_1$ in the choices of the stepsize and the clipping level, we see that momentum parameter and stepsize decrease with the same rate $\sim \mathcal{O}(1/T^{\frac{p}{2p-1}})$, similarly to Algorithm 1. Also, as we mentioned above clipping levels λ and λ_h has the same rate and increase with $\sim \mathcal{O}(T^{\frac{1}{2p-1}})$.

Corollary 1. In the setting of Theorem 3 Algorithm 2 ensures that $\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla F(x_t)\| \leq \varepsilon$ with probability at least $1 - \delta$. To achieve this, the algorithm requires at most

$$\mathcal{O} \left(\left((\sqrt{\Delta_1(L + \sigma_h)} + \sigma) / \varepsilon \log 1/\delta \varepsilon \right)^{\frac{2p-1}{p-1}} \right) \text{ first/second-order oracle calls.}$$

According to the results of Theorem 3 and Corollary 1, we have shown that Algorithm 2 has a near-optimal convergence rate in the dependence of ε . Moreover, the complexity has the logarithmic dependence on $1/\delta$, which is better than a polynomial dependence that can be achieved by directly translating the in-expectation result of Theorem 2 to high probability using Markov's inequality. Unfortunately, the dependence on $\Delta_1, L, \sigma, \sigma_h$ does not matches with lower bounds (3).

Comparison with Liu et al. [2023b]. Under additional strong assumptions, recently Liu et al. [2023b] analyze a momentum-based variance-reduced first-order method and obtain rates similar those in our Corollary 1. However, they assume individual smoothness (i.e. $\|\nabla f(x, \xi) - \nabla f(y, \xi)\| \leq \ell \|x - y\|$ for all $x, y \in \mathbb{R}^d$ and all realizations of random variable ξ). However, the individual smoothness is a very strong assumption which implies the bounded stochastic Hessian, i.e. $\|\nabla^2 f(x, \xi)\| \leq \ell$ a.s., for all $x \in \mathbb{R}^d$. We also note that under individual smoothness fast p -independent rates can be obtained, e.g., Lei et al. [2019] achieve $\tilde{\mathcal{O}}(\varepsilon^{-4})$ complexity for SGD under this individual smoothness.

6 Limitations and Future Work

While our proposed method with Hessian clipping has many desirable properties such as near-optimal sample complexity, batch-free and high-probability convergence under mild statistical assumptions, it also has several limitations. First, it requires tuning 4 extra parameters compared to, e.g., vanilla NSGD [Hübler et al., 2024]. It would be interesting to investigate if such extensive tuning can be removed for second-order methods. Second, while our sample complexity is near optimal, it does not exactly match our lower bound in all important parameters. Moreover, while we mainly pay attention to the leading stochastic terms in the complexity, our methods do not recover optimal deterministic complexities of second-order methods. Perhaps a different analysis technique for clipping and some algorithmic modifications (like Newton type preconditioning [Chayti et al., 2024]) are required to exactly match the upper and lower bounds. Third, high probability upper bounds are derived for a fixed confidence level δ rather than uniformly for all $\delta \in (0, 1)$ (as e.g., in [Liu et al., 2023a, Hübler et al., 2024]) and the algorithm's parameters depend on this choice. This is limiting as it doesn't necessarily imply high probability convergence for a fixed parameter choice. Finally, while we make an important progress in the fundamental understanding of second-order stochastic optimization, this results should be interpreted with caution. In practice, the overhead of second-order methods (computing Hessian information or even Hessian-vector products) and the challenge of tuning additional hyperparameters (such as clipping thresholds for both gradients and Hessians) can be significant.

References

- M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.
- A. Agafonov, D. Kamzolov, A. Gasnikov, A. Kavis, K. Antonakopoulos, V. Cevher, and M. Takáč. Advancing the lower bounds: An accelerated, stochastic, second-order method with optimal adaptation to inexactness. *arXiv preprint arXiv:2309.01570*, 2023.
- N. Agarwal and E. Hazan. Lower bounds for higher-order convex optimization. In *Conference On Learning Theory*, pages 774–792. PMLR, 2018.
- K. Ahn, X. Cheng, M. Song, C. Yun, A. Jadbabaie, and S. Sra. Linear attention is (maybe) all you need (to understand transformer optimization). In *International Conference on Learning Representations*, 2024.
- K. Antonakopoulos, A. Kavis, and V. Cevher. Extra-newton: A first approach to noise-adaptive accelerated second-order methods. *Advances in Neural Information Processing Systems*, 35: 29859–29872, 2022.
- Y. Arjevani, Y. Carmon, J. C. Duchi, D. J. Foster, N. Srebro, and B. Woodworth. Lower bounds for non-convex stochastic optimization. *arXiv preprint arXiv:1912.02365*, 2019a.
- Y. Arjevani, O. Shamir, and R. Shiff. Oracle complexity of second-order methods for smooth convex optimization. *Mathematical Programming*, 178(1):327–360, 2019b.
- Y. Arjevani, Y. Carmon, J. C. Duchi, D. J. Foster, A. Sekhari, and K. Sridharan. Second-order information in non-convex stochastic optimization: Power and limitations. In *Conference on Learning Theory*, pages 242–299. PMLR, 2020.
- A. Armacki, P. Sharma, G. Joshi, D. Bajovic, D. Jakovetic, and S. Kar. High-probability convergence bounds for nonlinear stochastic gradient descent under heavy-tailed noise. *arXiv preprint arXiv:2310.18784*, 2023.
- B. Battash, L. Wolf, and O. Lindenbaum. Revisiting the noise model of stochastic gradient descent. In *International Conference on Artificial Intelligence and Statistics*, pages 4780–4788. PMLR, 2024.
- A. A. Bennett. Newton’s method in general analysis. *Proceedings of the National Academy of Sciences*, 2(10):592–598, 1916.
- S. Bubeck, N. Cesa-Bianchi, and G. Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013.
- Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford. Lower bounds for finding stationary points i. *Mathematical Programming*, 184(1):71–120, 2020.
- S. Cayci and A. Eryilmaz. Provably robust temporal difference learning for heavy-tailed rewards. *Advances in Neural Information Processing Systems*, 36, 2024.
- E. M. Chayti, N. Doikov, and M. Jaggi. Improving stochastic cubic newton with momentum. *arXiv preprint arXiv:2410.19644*, 2024.
- A. Cutkosky and H. Mehta. Momentum improves normalized SGD. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2260–2268. PMLR, 2020.
- A. Cutkosky and H. Mehta. High-probability bounds for non-convex stochastic optimization with heavy tails. *Advances in Neural Information Processing Systems*, 34:4883–4895, 2021.
- A. Cutkosky and F. Orabona. Momentum-based variance reduction in non-convex sgd. *Advances in neural information processing systems*, 32, 2019.

372 D. Davis, D. Drusvyatskiy, L. Xiao, and J. Zhang. From low probability to high confidence in
373 stochastic convex optimization. *The Journal of Machine Learning Research*, 22(1):2237–2274,
374 2021.

375 J. E. Dennis, Jr and J. J. Moré. Quasi-newton methods, motivation and theory. *SIAM review*, 19(1):
376 46–89, 1977.

377 N. Doikov, K. Mishchenko, and Y. Nesterov. Super-universal regularized newton method. *SIAM*
378 *Journal on Optimization*, 34(1):27–56, 2024.

379 I. Fatkhullin, A. Barakat, A. Kireeva, and N. He. Stochastic policy gradient methods: Improved
380 sample complexity for Fisher-non-degenerate policies. In *Proceedings of the 40th International*
381 *Conference on Machine Learning*, volume 202, pages 9827–9869, 2023.

382 S. Garg, J. Zhanson, E. Parisotto, A. Prasad, Z. Kolter, Z. Lipton, S. Balakrishnan, R. Salakhutdinov,
383 and P. Ravikumar. On proximal policy optimization’s heavy-tailed gradients. In *International*
384 *Conference on Machine Learning*, pages 3610–3619. PMLR, 2021.

385 S. Ghadimi, H. Liu, and T. Zhang. Second-order methods with cubic regularization under inexact
386 information. *arXiv preprint arXiv:1710.05782*, 2017.

387 G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 4th edition,
388 2013.

389 E. Gorbunov, M. Danilova, and A. Gasnikov. Stochastic optimization with heavy-tailed noise via
390 accelerated gradient clipping. *Advances in Neural Information Processing Systems*, 33:15042–
391 15053, 2020.

392 E. Gorbunov, M. Danilova, I. Shibaev, P. Dvurechensky, and A. Gasnikov. High-probability com-
393 plexity bounds for non-smooth stochastic convex optimization with heavy-tailed noise. *Journal of*
394 *Optimization Theory and Applications*, pages 1–60, 2024a.

395 E. Gorbunov, A. Sadiev, M. Danilova, S. Horváth, G. Gidel, P. Dvurechensky, A. Gasnikov, and
396 P. Richtárik. High-probability convergence for composite and distributed stochastic minimization
397 and variational inequalities with heavy-tailed noise. In *International Conference on Machine*
398 *Learning*, pages 15951–16070, 2024b.

399 R. M. Gower, N. Loizou, X. Qian, A. Sailanbayev, E. Shulgin, and P. Richtárik. SGD: General
400 analysis and improved rates. In *International Conference on Machine Learning*, pages 5200–5209.
401 PMLR, 2019.

402 F. Hübler, I. Fatkhullin, and N. He. From gradient clipping to normalization for heavy tailed sgd.
403 *arXiv preprint arXiv:2410.13849*, 2024.

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

407 L. V. Kantorovich. Functional analysis and applied mathematics. *Uspekhi Matematicheskikh Nauk*, 3
408 (6):89–185, 1948.

409 A. Khaled and P. Richtárik. Better theory for SGD in the nonconvex world. *arXiv preprint*
410 *arXiv:2002.03329*, 2020.

411 N. Kornilov, O. Shamir, A. Lobanov, D. Dvinskikh, A. Gasnikov, I. Shibaev, E. Gorbunov, and
412 S. Horváth. Accelerated zeroth-order method for non-smooth stochastic convex optimization
413 problem with infinite variance. *Advances in Neural Information Processing Systems*, 36, 2024.

414 C. Lanczos. An iteration method for the solution of the eigenvalue problem of linear differential and
415 integral operators. *Journal of Research of the National Bureau of Standards*, 45:255–282, 1950.

416 Y. Lei, T. Hu, G. Li, and K. Tang. Stochastic gradient descent for nonconvex learning without
417 bounded gradient assumptions. *IEEE Transactions on Neural Networks and Learning Systems*,
418 2019.

419 Z. Liu and Z. Zhou. Stochastic nonsmooth convex optimization with heavy-tailed noises. *arXiv*
420 *preprint arXiv:2303.12277*, 2023.

421 Z. Liu and Z. Zhou. Nonconvex stochastic optimization under heavy-tailed noises: Optimal conver-
422 gence without gradient clipping. *arXiv preprint arXiv:2412.19529*, 2024.

423 Z. Liu, T. D. Nguyen, T. H. Nguyen, A. Ene, and H. Nguyen. High probability convergence of
424 stochastic gradient methods. In *International Conference on Machine Learning*, pages 21884–
425 21914. PMLR, 2023a.

426 Z. Liu, J. Zhang, and Z. Zhou. Breaking the lower bound with (little) structure: Acceleration in non-
427 convex stochastic optimization with heavy-tailed noise. In *Proceedings of Thirty Sixth Conference*
428 *on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 2266–2290.
429 PMLR, 12–15 Jul 2023b.

430 A. V. Nazin, A. S. Nemirovsky, A. B. Tsybakov, and A. B. Juditsky. Algorithms of robust stochastic
431 optimization based on mirror descent method. *Automation and Remote Control*, 80:1607–1627,
432 2019.

433 Y. Nesterov. A method for unconstrained convex minimization problem with the rate of convergence
434 $O(1/k^2)$. In *Doklady AN USSR*, volume 269, pages 543–547, 1983.

435 Y. Nesterov. Accelerating the cubic regularization of newton’s method on convex problems. *Mathe-*
436 *matical Programming*, 112(1):159–181, 2008.

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

439 T. D. Nguyen, A. Ene, and H. L. Nguyen. Improved convergence in high probability of clipped
440 gradient methods with heavy tails. *arXiv preprint arXiv:2304.01119*, 2023.

441 R. Pascanu, T. Mikolov, and Y. Bengio. On the difficulty of training recurrent neural networks. In
442 *International Conference on Machine Learning*, pages 1310–1318. Pmlr, 2013.

443 B. T. Polyak and Y. Z. Tsypkin. Adaptive estimation algorithms: convergence, optimality, stability.
444 *Avtomatika i telemekhanika*, (3):71–84, 1979.

445 N. Puchkin, E. Gorbunov, N. Kutuzov, and A. Gasnikov. Breaking the heavy-tailed noise barrier
446 in stochastic optimization problems. In *International Conference on Artificial Intelligence and*
447 *Statistics*, pages 856–864. PMLR, 2024.

448 Y. Saad. *Numerical Methods for Large Eigenvalue Problems*. SIAM, 2011.

449 A. Sadiev, M. Danilova, E. Gorbunov, S. Horváth, G. Gidel, P. Dvurechensky, A. Gasnikov, and
450 P. Richtárik. High-probability bounds for stochastic optimization and variational inequalities: the
451 case of unbounded variance. In *International conference on machine learning*, 2023.

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

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

456 H. Sha, Y. Cao, Y. Liu, Y. Wu, R. Liu, and H. Chen. Clip body and tail separately: High probability
457 guarantees for DPSGD with heavy tails. *arXiv preprint arXiv:2405.17529*, 2024.

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

461 H. Tran and A. Cutkosky. Better SGD using Second-order Momentum. *arXiv preprint*
462 *arXiv:2103.03265*, 2021.

463 N. Tripuraneni, M. Stern, C. Jin, J. Regier, and M. I. Jordan. Stochastic cubic regularization for fast
464 nonconvex optimization. *Advances in neural information processing systems*, 31, 2018.

- 465 H. Wang, M. Gurbuzbalaban, L. Zhu, U. Simsekli, and M. A. Erdogdu. Convergence Rates of
466 Stochastic Gradient Descent under Infinite Noise Variance. In *Advances in Neural Information*
467 *Processing Systems*, volume 34, pages 18866–18877, 2021.
- 468 J. Yang, X. Li, I. Fatkhullin, and N. He. Two sides of one coin: the limits of untuned sgd and the
469 power of adaptive methods. *Advances in Neural Information Processing Systems*, 36:74257–74288,
470 2023.
- 471 J. Zhang and A. Cutkosky. Parameter-free regret in high probability with heavy tails. *Advances in*
472 *Neural Information Processing Systems*, 35:8000–8012, 2022.
- 473 J. Zhang, S. P. Karimireddy, A. Veit, S. Kim, S. Reddi, S. Kumar, and S. Sra. Why are adaptive
474 methods good for attention models? *Advances in Neural Information Processing Systems*, 33:
475 15383–15393, 2020.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: yes, we derive new sample complexity upper and lower bounds, the formal proofs are in the appendix

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: yes, there is limitations sections in the end of main part

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: Yes, all formal proofs are in the appendix, assumptions are clearly stated.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Yes, we describe the details of our experiments, although our experiments use very simple toy illustration.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: Our work’s focus is advancing theory of stochastic optimization, so this question is not applicable. Our one numerical illustration is a very toy experiment involving 10 dimensional quadratic problem with synthetic heavy-tailed noise.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Yes, we provide all necessary information to reproduce our toy illustration.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Yes, we include error bars as it is important to illustrate the confidence levels of optimization algorithms.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [NA]

Justification: The illustration is toy and can be run on any device without advanced compute resources.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: Yes, we follow the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: Our work advances theory of optimization, it has not immediate societal impact.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: There are no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: We don't use existing assets.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset’s creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: We don’t release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: we don’t use crowdsourcing.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: the work does not involve crowdsourcing.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

790 Question: Does the paper describe the usage of LLMs if it is an important, original, or
791 non-standard component of the core methods in this research? Note that if the LLM is used
792 only for writing, editing, or formatting purposes and does not impact the core methodology,
793 scientific rigorousness, or originality of the research, declaration is not required.

794 Answer: [NA]

795 Justification: We don't use LLM for method development.

796 Guidelines:

- 797 • The answer NA means that the core method development in this research does not
798 involve LLMs as any important, original, or non-standard components.
- 799 • Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>)
800 for what should or should not be described.

801	Contents	
802	1 Introduction	1
803	2 Notations and Assumptions	4
804	3 Lower Complexity Bounds for SOSO under p-BCM	4
805	4 Near-optimal Second-Order Method under p-BCM	6
806	5 Hessian Clipping for High-probability Convergence	7
807	6 Limitations and Future Work	9
808	A Additional Related Work	22
809	B About Normalized SGD and Different Variants of Momentum	22
810	C Technical Lemmas	24
811	D Missing Proofs for Section 3	25
812	E Missing Proofs for Section 4	28
813	F Missing Proofs for Section 5	33
814	F.1 Analysis: Optimization Part	33
815	F.2 Analysis: Statistical Part	34

Table 1: Summary of sample complexities of stochastic second-order (SOSO) methods for finding an ε -stationary point, in high probability of in expectation, i.e., number of stochastic gradient and Hessian evaluations to find \bar{x} with $\|\nabla F(\bar{x})\| \leq \varepsilon$. The column “ p ” indicates the range of moments in Assumptions 3, 4 for which the result holds, that is when $p = 2$, the corresponding result holds only under bounded variance. When $p = \infty$ it means that at least for stochastic gradient or stochastic Hessian bounded noise assumption is required, which corresponds to all moments being bounded. We are not aware of any prior works for SOSO for $p < 2$. The column “HP?” denotes whether the high probability guarantee with polylogarithmic dependence on the inverse of failure probability $1/\delta$ is available.

Algorithm	Sample Complexity	p	HP?
SCN [Tripuraneni et al., 2018]	$\frac{\Delta\sqrt{L_h}\sigma^2}{\varepsilon^{7/2}}$	∞	✓ ⁽¹⁾
SGD with HVP-RVR [Arjevani et al., 2020]	$\frac{\Delta\sigma\sigma_h}{\varepsilon^3} + \frac{\Delta\sqrt{L_h}\sigma_1}{\varepsilon^{5/2}} + \frac{\Delta L}{\varepsilon^2}$	2	✗
SN ⁽²⁾ with HVP-RVR [Arjevani et al., 2020]	$\frac{\Delta\sigma\sigma_h}{\varepsilon^3} + \frac{\Delta\sqrt{L_h}\sigma_1}{\varepsilon^{5/2}} + \frac{\Delta\sigma_h}{\varepsilon^2}$	2	✗
N-SGDHess [Tran and Cutkosky, 2021]	$\frac{\sigma^3}{\varepsilon^3} + \frac{\Delta\sigma\sigma_h}{\varepsilon^3} + \frac{\Delta\sqrt{L_h}\sigma_1}{\varepsilon^{5/2}} + \frac{\Delta\sigma_h}{\varepsilon^2}$	2	✗
SCN with IT-HB [Chayti et al., 2024]	$\frac{\Delta\sqrt{L_h}}{\varepsilon^{3/2}} + \frac{\Delta L_h^{1/4}\sigma_h^{1/2}}{\varepsilon^{7/4}} + \frac{\Delta\sqrt{L}\sigma^2}{\varepsilon^{7/2}}$	∞	✗
Lower Bounds Theorem 1	$\min\left\{\frac{\Delta\sigma_h}{\varepsilon^2}, \frac{\Delta L\sigma}{\varepsilon^3}, \frac{\Delta\sqrt{L_h}\sigma}{\varepsilon^{5/2}}\right\} \left(\frac{\sigma}{\varepsilon}\right)^{\frac{1}{p-1}}$	(1, 2]	
NSGDMHess Theorem 2	$\frac{\Delta(L+\sigma_h)}{\varepsilon^2} + \left(\frac{\Delta(L+\sigma_h)}{\varepsilon^2} + \frac{\sigma}{\varepsilon}\right) \left(\frac{\sigma}{\varepsilon}\right)^{\frac{1}{p-1}}$	(1, 2]	✗
Clip NSGDMHess Theorem 3	$\left(\frac{\sqrt{\Delta_1(L+\sigma_h)} + \sigma}{\varepsilon}\right)^{2+\frac{1}{p-1}}$	(1, 2]	✓

⁽¹⁾ Tripuraneni et al. [2018] provide analysis under stronger assumptions, which implies the noise has light tails.

⁽²⁾ SN = Subsampled Newton.

A Additional Related Work

Gradient clipping and normalization. Gradient clipping has been applied successfully also in zero-order optimization [Kornilov et al., 2024], bandit and RL literature [Bubeck et al., 2013, Cayci and Eryilmaz, 2024], online learning [Zhang and Cutkosky, 2022] and differential privacy [Abadi et al., 2016, Sha et al., 2024]. Some other works, which use normalization and gradient clipping under heavy-tailed noise include [Armacki et al., 2023, Puchkin et al., 2024, Liu and Zhou, 2024].

B About Normalized SGD and Different Variants of Momentum

In this section we describe a general framework of momentum terms. There is a lot of literature about it especially for convex optimization, since Nesterov [1983] showed that an appropriate use of momentum with GD accelerates convergence. But we want to emphasize that in our work we consider momentum in the context of stochastic non-convex optimization. The general update rules is defined as follows

$$x_{t+1} = x_t - \gamma_t \frac{g_t}{\|g_t\|},$$

where g_t is a momentum term. Different selections of g_t leads to different types of methods. Let us consider several of them.

NSGD is the well-known method. if $g_t = \nabla f(x_t, \xi_t)$, we immediately get NSGD. As it was shown that NSGD converges only to the neighborhood, which radius is equal to σ [Yang et al., 2023, Hübler et al., 2024]. One of the possible way to fix it is mini-batching, allowing to reduce the variance. But if we want to use only one sample per iteration, this strategy cannot be applied.

834 **NSGDM** is the solution of the above problem and was proposed and analyzed by Cutkosky and
 835 Mehta [2020]. The momentum term is defined as

$$g_t = (1 - \alpha)g_{t-1} + \alpha \nabla f(x_t, \xi_t).$$

836 In their work we showed that NSGDM converges to ε -stationary point and has sample complexity,
 837 which is equal to $\mathcal{O}(\varepsilon^{-4})$, which is optimal according to Arjevani et al. [2019a]. Later, Hübler
 838 et al. [2024] extended this result to the case of bounded p -th moment deriving $\mathcal{O}(\varepsilon^{-\frac{3p-2}{p-1}})$ sample
 839 complexity.

840 **NSGDM with clipping** was proposed directly for the general set up with bounded p -th moment of
 841 stochastic gradient. The momentum term is modified as follows

$$g_t = (1 - \alpha)g_{t-1} + \alpha \text{clip}(\nabla f(x_t, \xi_t), \lambda).$$

842 Cutkosky and Mehta [2021] provided high-probability convergence guarantees, and the sample
 843 complexity of this method is $\mathcal{O}(\varepsilon^{-\frac{3p-2}{p-1}})$, which is optimal and matches with lower bounds from the
 844 paper of Zhang et al. [2020]. It is worth to mention that the clipping operator is not necessary and
 845 normalization of momentum term is sufficient to guarantee convergence under Assumption 3 [Hübler
 846 et al., 2024, Liu and Zhou, 2024].

847 **NIGT** is NSGD with momentum and implicit gradient transportation proposed by Cutkosky and
 848 Mehta [2020]. They modified momentum term with replacing the point, at which stochastic gradient
 849 is computed:

$$g_t = (1 - \alpha)g_{t-1} + \alpha \nabla f(y_t, \xi_t), \quad \text{where } y_t = x_t + \frac{1 - \alpha}{\alpha} x_{t-1}.$$

850 Thanks to this modification, assuming that the objective function has Lipschitz continuous Hessian
 851 Cutkosky and Mehta [2020] proved NIGT has sample complexity $\mathcal{O}(\varepsilon^{-7/2})$, which is better than
 852 NSGDM.

853 **NIGT with clipping** was motivated for high-probability analysis under bounded p -th moment of
 854 stochastic gradient by Cutkosky and Mehta [2021]. The main difference from NIGT is only a clipping
 855 operator, which is needed from theoretical perspective. According to Cutkosky and Mehta [2021],
 856 NIGT with clipping has sample complexity $\mathcal{O}(\varepsilon^{-\frac{5p-3}{2p-2}})$, which is optimal according to Theorem 1.

857 **NSGD with MVR** is other version of NSGD with momentum variance reduction (MVR).
 858 STORM/MVR was proposed by Cutkosky and Orabona [2019] to improve the sample complexity
 859 compared to SGD.

860 The above considered methods use only first-order information. What if we try to exploit second-order
 861 information to improve sample complexity?

862 **NSGD-Hess** is Normalized SGD with Hessian-corrected momentum proposed by Tran and Cutkosky
 863 [2021]. They introduced stochastic Hessian-vector product into momentum term:

$$g_t = (1 - \alpha)(g_{t-1} + \nabla^2 f(x_t, \xi_t) \cdot (x_t - x_{t-1})) + \alpha \nabla f(y_t, \xi_t),$$

864 **Refined NSGD-Hess.** Our refined momentum modifies the point where the stochastic Hessian is
 865 evaluated and uses another fresh sample $\hat{\xi}_t \sim \mathcal{D}$ for the Hessian-vector product computation.

$$\begin{aligned} q_t &\sim \mathcal{U}([0, 1]), \\ \hat{x}_t &= q_t x_t + (1 - q_t) x_{t-1}, \\ g_t &= (1 - \alpha) \left(g_{t-1} + \nabla^2 f(\hat{x}_t, \hat{\xi}_t) \cdot (x_t - x_{t-1}) \right) + \alpha \nabla f(x_t, \xi_t). \end{aligned}$$

866 C Technical Lemmas

867 We list the following technical lemmas, which are useful for our analysis in the subsequent sections.

868 **Lemma 1** (Lemma 10 from [Hübler et al. \[2024\]](#)). *Let $p \in [1; 2]$, and $X_1, \dots, X_n \in \mathbb{R}^d$ be a*
 869 *martingale difference sequence, i.e. $\mathbb{E}[X_j | X_{j-1}, \dots, X_1] = 0$ a.s. for all $j = 1, \dots, n$ satisfying*

$$\mathbb{E}[\|X_j\|^p] < \infty \quad \text{for all } i = 1, \dots, n.$$

870 Define $S_n \stackrel{\text{def}}{=} \sum_{j=1}^n X_j$, then

$$\mathbb{E}[\|S_n\|^p] \leq 2 \sum_{i=j}^n \mathbb{E}[\|X_j\|^p].$$

871 **Lemma 2** (Lemma 10 from [Cutkosky and Mehta \[2021\]](#)). *Let $X_1, \dots, X_n \in \mathbb{R}^d$ be a sequence of*
 872 *random vectors. Define the sequence of real numbers w_1, \dots, w_n recursively*

873 1. $w_0 = 0$

874 2. If $\sum_{i=1}^{j-1} X_i \neq 0$, then we set:

$$w_j = \text{sign} \left(\sum_{i=1}^{j-1} w_i \right) \frac{\left\langle \sum_{i=1}^{j-1} X_i, X_j \right\rangle}{\left\| \sum_{i=1}^{j-1} X_i \right\|}.$$

875 3. If $\sum_{i=1}^{j-1} X_i = 0$, set $w_j = 0$.

876 Then $|w_j| \leq \|X_j\|$ for all $j = 1, \dots, n$, and

$$\|S_n\| \leq \left| \sum_{j=1}^n w_j \right| + \sqrt{\max_{j \in [n]} \|X_j\|^2 + \sum_{j=1}^n \|X_j\|^2}. \quad (8)$$

877 **Lemma 3** (Bernstein inequality). *Let $X_1, \dots, X_n \in \mathbb{R}^d$ be a martingale difference sequence,*
 878 *i.e. $\mathbb{E}[X_j | X_{j-1}, \dots, X_1] = 0$ a.s. for all $j = 1, \dots, n$. Assume that conditional variances*
 879 $\sigma_j^2 \stackrel{\text{def}}{=} \mathbb{E}[X_j^2 | X_{j-1}, \dots, X_1]$ *exist and are bounded, and assume that there exists deterministic*
 880 *constant $c > 0$ such that $|X_j| \leq c$ almost surely for all $j = 1, \dots, n$. Define $S_n \stackrel{\text{def}}{=} \sum_{i=1}^n X_j$, then*
 881 *for all $b > 0, G > 0$*

$$\mathbb{P} \left\{ |S_n| > b \text{ and } \sum_{j=1}^n \sigma_j^2 \leq G \right\} \leq 2 \exp \left(-\frac{b^2}{2G + 2cb/3} \right), \quad (9)$$

882 **Lemma 4** (Lemma 5.1 from [Sadiev et al. \[2023\]](#)). *Let $\lambda > 0$ and $X \in \mathbb{R}^d$ be a random vector and*
 883 $\tilde{X} \stackrel{\text{def}}{=} \text{clip}(X, \lambda)$. *Then, $\|\tilde{X} - \mathbb{E}[\tilde{X}]\| \leq 2\lambda$. Moreover, if for some $\sigma \geq 0$ and $p \in (1; 2]$ we have*
 884 $\mathbb{E}[X] = x \in \mathbb{R}^d$, $\mathbb{E}[\|X - x\|^p] \leq \sigma^p$, *and $\|x\| \leq \lambda/2$, then*

$$\left\| \mathbb{E}[\tilde{X}] - x \right\| \leq \frac{2^p \sigma^p}{\lambda^{p-1}}, \quad \mathbb{E} \left[\left\| \tilde{X} - x \right\|^2 \right] \leq 18 \lambda^{2-p} \sigma^p, \quad \mathbb{E} \left[\left\| \tilde{X} - \mathbb{E}[\tilde{X}] \right\|^2 \right] \leq 18 \lambda^{2-p} \sigma^p.$$

D Missing Proofs for Section 3

In this section we provide lower bounds for the certain class of functions and the oracle class. Our prove is inspired by the paper of [Arjevani et al. \[2020\]](#).

Auxiliary Lemmas. Now we state two helpful lemmas to prove lower bound. The first lemma is about the number of iteration to reveal all coordinates, when zero-respecting algorithms use the information received from the oracle forming *probability- ρ zero-chain*.

Definition 2. A collection of derivative estimators $\nabla^1 f(x, \xi), \nabla^2 f(x, \xi), \dots, \nabla^q f(x, \xi)$ for a function F forms a *probability- ρ zero-chain* if

$$\mathbb{P} \left\{ \exists x \mid \text{prog}(\nabla^1 f(x, \xi), \dots, \nabla^q f(x, \xi)) = \text{prog}_{\frac{1}{4}}(x) + 1 \right\} \leq \rho$$

and

$$\mathbb{P} \left\{ \exists x \mid \text{prog}(\nabla^1 f(x, \xi), \dots, \nabla^q f(x, \xi)) = \text{prog}_{\frac{1}{4}}(x) + i \right\} = 0, \quad i > 1.$$

The second lemma is about the main properties of the *worst* function from the class $\mathcal{F}(\Delta, L_{1:q})$.

Lemma 5 ([\[Arjevani et al., 2020\]](#)). Let $\nabla^1 f(x, \xi), \dots, \nabla^q f(x, \xi)$ be a collection of probability- ρ zero-chain derivative estimators for $F : \mathbb{R}^T \rightarrow \mathbb{R}$, and let O_F^q be an oracle with $O_F^p = (\nabla^r f(x, \xi))_{r \in [q]}$. Let $\{x_{A[O_F^q]}^{(t)}\}$ be a sequence of queries produced by a zero-respecting algorithm A interacting with O_F^q . Then, with probability at least $1 - \delta$

$$\text{prog} \left(x_{A[O_F^q]}^{(t)} \right) < T, \quad \text{for all } t \leq \frac{T - \log \frac{1}{\delta}}{2\rho}.$$

Lemma 6 ([\[Carmon et al. 2020\]](#)). Let $h : \mathbb{R}^T \rightarrow \mathbb{R}$ be the following function:

$$h(x) = -\Psi(1)\Phi(x_1) + \sum_{i=2}^T (\Psi(-x_{i-1})\Phi(-x_i) - \Psi(x_{i-1})\Phi(x_i)),$$

where

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

Then the function h satisfies the following properties:

- $h(0) - \inf_x h(x) \leq \Delta_0 T$, where $\Delta_0 = 12$.
- For $q \geq 1$, the q -th order derivatives of h are ℓ_q -Lipschitz continuous, where $\ell_q \leq \exp\left(\frac{5}{2}q \log q + cq\right)$ for a numerical constant $c < \infty$.
- For all $x \in \mathbb{R}^T$, $q \in \mathbb{N}$ and $i \in [T]$, we have $\|\nabla_i^q h(x)\|_{op} \leq \ell_{q-1}$, where $\ell_0 = 23$.
- For all $x \in \mathbb{R}^T$ and $q \in \mathbb{N}$, $\text{prog}(\nabla^q h(x)) \leq \text{prog}_{\frac{1}{2}}(x) + 1$.
- For all $x \in \mathbb{R}^T$, if $\text{prog}_1(x) < T$, then $\|\nabla h(x)\| \geq |\nabla_{\text{prog}_1(x)+1} h(x)| > 1$.

Lemma 7. Let the derivative estimator is defined as follows for each $r \in [q]$

$$[\nabla^r h(x, \xi)]_i \stackrel{\text{def}}{=} \left(1 + \mathbb{I} \left\{ i > \text{prog}_{\frac{1}{4}}(x) \right\} \left(\frac{\xi}{\rho} - 1 \right) \right) \cdot \nabla_i^r h(x),$$

where $\xi \sim \text{Bernoulli}(\rho)$. Then, $\{\nabla^r h(x, \xi)\}_{r \in [q]}$ forms probability- ρ zero-chain, and for any $p \in [1, 2]$ and for each $r \in [q]$ the derivative estimator satisfies

$$\mathbb{E} [\nabla^r h(x, z)] = \nabla^r h(x), \quad \text{and} \quad \mathbb{E} \left[\|\nabla^r h(x, \xi) - \nabla^r h(x)\|_{op}^p \right] \leq \frac{2\ell_{r-1}^p}{\rho^{p-1}} \quad \text{for all } x \in \mathbb{R}^d.$$

911 *Proof.* First of all, it is easy to show that the proposed estimator is unbiased, i.e.

$$\mathbb{E}[\nabla^r h(x, \xi)] = (1 + 0) \cdot \nabla^r h(x) = \nabla^r h(x).$$

912 According to Lemma 6, we have $\text{prog}(\nabla^r h(x)) < \text{prog}_{\frac{1}{2}}(x) + 1 \leq \text{prog}_{\frac{1}{4}}(x) + 1$, where the last
913 inequality is true because prog_α is non-increasing with respect to α . Then, by Lemma 6 we have

$$[\nabla^r h(x, \xi)]_i = \nabla_i^r h(x) = 0, \quad \forall i > \text{prog}_{\frac{1}{4}}(x) + 1.$$

914 Therefore, due to that $\xi \sim \text{Bernoulli}(\rho)$ we obtain

$$\mathbb{P}\left\{\exists x \mid \text{prog}(\nabla^1 f(x, z), \dots, \nabla^q f(x, z)) = \text{prog}_{\frac{1}{4}}(x) + 1\right\} \leq \rho,$$

915 i.e. $\{\nabla^r h(x, \xi)\}_{r \in [q]}$ forms probability- ρ zero-chain.

916 Now we bound p -th moment of $\nabla^r h(x, \xi) - \nabla^r h(x)$ for any $p \in [1; 2]$. Denoting $i_x = \text{prog}_{\frac{1}{4}}(x) + 1$,
917 for any $r \in [q]$ we have

$$\begin{aligned} \mathbb{E}\left[\|\nabla^r h(x, \xi) - \nabla^r h(x)\|_{\text{op}}^p\right] &= \mathbb{E}\left[\left|\mathbb{I}\left\{i_x > \text{prog}_{\frac{1}{4}}(x)\right\}\left(\frac{\xi}{\rho} - 1\right)\right|^p \|\nabla_{i_x}^r h(x)\|_{\text{op}}^p\right] \\ &= \mathbb{E}\left[\left|\frac{\xi}{\rho} - 1\right|^p\right] \|\nabla_{i_x}^r h(x)\|_{\text{op}}^p \\ &\leq \ell_{r-1}^p \mathbb{E}\left[\left|\frac{\xi}{\rho} - 1\right|^p\right] \\ &= \ell_{r-1}^p \frac{(1-\rho)^p \rho + (1-\rho)\rho^p}{\rho^p} \\ &\leq \ell_{r-1}^p \frac{2}{\rho^{p-1}}. \end{aligned}$$

918 □

919 **Main Theorem.** Now we are ready to state and prove the main theorem in this section, i.e. Theorem 1.

920 *Proof.* First of all, we fix all parameters: $q \in \mathbb{N}$, $\Delta, L_{1:q}, \sigma_{1:q} > 0$ and $\varepsilon > 0$. Next, we rescale our
921 function h as follows: $h^*(x) = \nu h(\beta x)$, where $\nu > 0$ and $\beta > 0$. According to Lemma 6, selecting
922 $T = \left\lfloor \frac{\Delta}{\nu \Delta_0} \right\rfloor$, we have for any $r \in [q-1]$

$$\begin{aligned} h^*(0) - \inf_x h^*(x) &= \nu \left(h^*(0) - \inf_x h^*(\beta x) \right) \leq \nu \Delta_0 T \leq \Delta; \\ \|\nabla^{r+1} h^*(x)\|_{\text{op}} &= \nu \beta^{r+1} \|\nabla^{r+1} h(\beta x)\| \leq \nu \beta^{r+1} \ell_r; \\ \|\nabla h^*(x)\| &= \nu \beta \|\nabla h(\beta x)\| \geq \nu \beta \|\nabla h(x)\| \geq \nu \beta, \quad \forall x : \text{prog}_1(x) < T. \end{aligned}$$

923 Since $\{\nabla^r h^*(x, \xi) = \nu \beta^r \nabla^r h(\beta x, \xi)\}_{r \in [p]}$ forms a probability- ρ zero-chain, then by Lemma 6 with
924 probability at least $\delta = \frac{1}{2}$ we have

$$\mathbb{E}\left[\left\|\nabla h^*(x_{A[O_h^q]}^t)\right\|\right] = \nu \beta \mathbb{E}\left[\left\|\nabla h^*(\beta x_{A[O_h^q]}^t)\right\|\right] \geq \frac{\nu \beta}{2}, \quad \forall t \leq \frac{T-1}{2\rho}$$

925 The p -th central moment of the scaled derivatives estimators is bounded as

$$\begin{aligned} \mathbb{E}\left[\|\nabla^r h^*(x, \xi) - \nabla^r h^*(x)\|_{\text{op}}^p\right] &\leq \nu^p \beta^{rp} \mathbb{E}\left[\|\nabla^r h(\beta x, \xi) - \nabla^r h(\beta x)\|_{\text{op}}^p\right] \\ &\leq \frac{2\nu^p \beta^{rp} \ell_{r-1}^p}{\rho^{p-1}}, \end{aligned}$$

926 where in the last inequality we applied Lemma 7. Thus, we have the set of constraints for parameters
927 ν and β for all $r \in [q]$:

$$\begin{cases} \nu \Delta_0 T \leq \Delta; \\ \nu \beta^{r+1} \ell_r \leq L_r; \\ \frac{\nu \beta}{2} \geq \varepsilon; \\ \nu^p \beta^{rp} \ell_{r-1}^p \frac{2}{\rho^{p-1}} \leq \sigma_r^p. \end{cases}$$

928 We will resolve the system of inequalities step by step. The first step is to set $\nu = \frac{2\varepsilon}{\beta}$. For $r = 1$, we
 929 have

$$\nu^p \beta^p \ell_0^p \cdot \frac{2}{\rho^{p-1}} \leq \sigma_1^p \Rightarrow \rho = \min \left\{ \left(\frac{\varepsilon}{\sigma_1} \right)^{\frac{p}{p-1}} \cdot 2^{\frac{p+1}{p-1}} \ell_0^{\frac{p}{p-1}}, 1 \right\}.$$

930 To set β , we need to find its specific value, which can satisfy the following constraints: for any
 931 $r \in \{2, \dots, q\}$ and any $r' \in [q]$

$$\beta^{r+1} \leq \frac{L_r}{\nu \ell_r} = \frac{\beta L_r}{2\varepsilon \ell_r}, \quad \nu^p \beta^{pr} \ell_{r-1}^p \frac{2}{\rho^{p-1}} \stackrel{(*)}{\leq} \beta^{p(r-1)} \left(\frac{\sigma_1 \ell_{r-1}}{\ell_0} \right)^p \leq \sigma_r^p,$$

932 where in the inequality $(*)$ we used $\frac{2}{\rho^{p-1}} \leq \frac{\sigma_1^p}{\nu^p \beta^p \ell_0^p}$. Therefore, we obtain

$$\beta = \min_{r' \in [q]; r \in \{2, \dots, q\}} \min \left\{ \left(\frac{\ell_0 \sigma_r}{\ell_{r-1} \sigma_1} \right)^{\frac{1}{r-1}}, \left(\frac{L_{r'}}{2\varepsilon \ell_{r'}} \right)^{\frac{1}{r'}} \right\}.$$

933 Then, assuming $T \geq 3$ and $\left(\frac{\varepsilon}{\sigma_1} \right)^{\frac{p}{p-1}} \cdot 2^{\frac{p+1}{p-1}} \ell_0^{\frac{p}{p-1}} \leq 1$, we get

$$\begin{aligned} \frac{T-1}{2\rho} &= \frac{1}{2\rho} \left(\left\lfloor \frac{\Delta\beta}{2\Delta_0\varepsilon} \right\rfloor - 1 \right) \geq \frac{\Delta\beta}{8\rho\Delta_0\varepsilon} \\ &\geq \left(\frac{\sigma_1}{2\varepsilon\ell_0} \right)^{\frac{p}{p-1}} \cdot \frac{\Delta}{4\Delta_0\varepsilon} \min_{r' \in [q]; r \in \{2, \dots, q\}} \min \left\{ \left(\frac{\ell_0 \sigma_r}{\ell_{r-1} \sigma_1} \right)^{\frac{1}{r-1}}, \left(\frac{L_{r'}}{2\varepsilon \ell_{r'}} \right)^{\frac{1}{r'}} \right\} \\ &\geq \Omega(1) \cdot \frac{\Delta}{\varepsilon} \left(\frac{\sigma_1}{\varepsilon} \right)^{\frac{p}{p-1}} \min \left\{ \min_{r \in \{2, \dots, q\}} \left(\frac{\sigma_r}{\sigma_1} \right)^{\frac{1}{r-1}}, \min_{r' \in [q]} \left(\frac{L_{r'}}{\varepsilon} \right)^{\frac{1}{r'}} \right\}. \end{aligned}$$

934 Since h and h^* are functions of T arguments, the dimension of the problem is equal to T . This
 935 concludes the proof. \square

E Missing Proofs for Section 4

In this section we study Normalized SGD with Hessian correction (NSGD-Hess) (see Algorithm 1).

Auxiliary Lemmas. To show the convergence guarantees for Algorithm 1, we state and prove auxiliary lemmas. The first one is well-known *Descent Lemma* for Normalized SGD. The proof for this lemma can be found in Cutkosky and Mehta [2020], Hübler et al. [2024], but for completeness we restate and reprove it. The second one is devoted to the bounds on *error term* from *Descent Lemma*.

Lemma 8. *Let Assumptions 1 and 2 hold. Then for any selection of stepsize $\gamma > 0$ the iterates $\{x_t\}_{t=0}^T$ generated by Algorithm 1 satisfy*

$$\gamma \sum_{t=0}^{T-1} \|\nabla F(x_t)\| + \Delta_T \leq \Delta_0 + 2\gamma \sum_{t=0}^{T-1} \|\hat{e}_t\| + \frac{\gamma^2 L T}{2}. \quad (10)$$

where functional gap is defined as $\Delta_t \stackrel{\text{def}}{=} F(x_t) - F_*$, error term is defined as $\hat{e}_t \stackrel{\text{def}}{=} g_t - \nabla F(x_t)$.

Proof. According to the update rule for x_t and Assumption 2, we have

$$\begin{aligned} F(x_{t+1}) &\leq F(x_t) + \langle \nabla F(x_t), x_{t+1} - x_t \rangle + \frac{L}{2} \|x_{t+1} - x_t\|^2 \\ &= F(x_t) - \gamma \left\langle \nabla F(x_t), \frac{g_t}{\|g_t\|} \right\rangle + \frac{\gamma^2 L}{2} \\ &= F(x_t) - \gamma \|g_t\| - \gamma \left\langle \nabla F(x_t) - g_t, \frac{g_t}{\|g_t\|} \right\rangle + \frac{\gamma^2 L}{2} \\ &\leq F(x_t) - \gamma \|g_t\| + \gamma \|\nabla F(x_t) - g_t\| + \frac{\gamma^2 L}{2} \\ &\leq F(x_t) - \gamma \|\nabla F(x_t)\| + 2\gamma \|\nabla F(x_t) - g_t\| + \frac{\gamma^2 L}{2}. \end{aligned}$$

Using notation for the functional gap and the error term, we get

$$\gamma \|\nabla F(x_t)\| + \Delta_{t+1} \leq \Delta_t + 2\gamma \|\hat{e}_t\| + \frac{\gamma^2 L}{2}.$$

Summing over t from 0 to $T - 1$, we obtain

$$\gamma \sum_{t=0}^{T-1} \|\nabla F(x_t)\| + \Delta_T \leq \Delta_0 + 2\gamma \sum_{t=0}^{T-1} \|\hat{e}_t\| + \frac{\gamma^2 L T}{2}. \quad (11)$$

□

Next, we derive the lemma for error control for the Hessian corrected momentum estimator (lines 4–7 in Algorithm 1). Similar lemma for the special case $p = 2$ appeared previously in [Salehkaleybar et al., 2022, Fatkhullin et al., 2023]. For the case $p < 2$, similar recursion was derived by Hübler et al. [2024] for first-order momentum. Now we extend this idea to the case of second-order momentum.

Lemma 9. *Let Assumptions 2, 3 and 4 hold. Then for all $t \geq 0$, we have*

$$\mathbb{E} [\|\hat{e}_t\|] \leq (1 - \alpha)^t \mathbb{E} [\|\hat{e}_0\|] + 2\alpha^{\frac{p-1}{p}} \sigma + 12\gamma(L + \sigma_h)\alpha^{-1/p}, \quad (12)$$

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

Proof. Define $e_t = \nabla f(x_t, \xi_t) - \nabla F(x_t)$, $\hat{S}_{t+1} = \nabla^2 f(\hat{x}_{t+1}, \hat{\xi}_{t+1}) \cdot (x_{t+1} - x_t) - \nabla F(x_{t+1}) + \nabla F(x_t)$. We have $\mathbb{E}[e_t] = 0$, $\mathbb{E}[\|e_t\|^p] \leq \sigma^p$, $\mathbb{E}[\hat{S}_t] = 0$, and

$$\begin{aligned} \mathbb{E} [\|\hat{S}_t\|^p] &= \mathbb{E} [\|\nabla F(x_t) - \nabla F(x_{t+1}) + \nabla^2 f(\hat{x}_{t+1}, \hat{\xi}_{t+1}) \cdot (x_{t+1} - x_t)\|^p] \\ &\leq 3\mathbb{E} [\|\nabla F(x_t) - \nabla F(x_{t+1})\|^p] + 3\mathbb{E} [\|\nabla^2 F(\hat{x}_{t+1}) \cdot (x_{t+1} - x_t)\|^p] \\ &\quad + 3\mathbb{E} [\|\nabla^2 f(\hat{x}_{t+1}, \hat{\xi}_{t+1}) - \nabla^2 F(\hat{x}_{t+1})\| \cdot \|x_{t+1} - x_t\|^p] \\ &\leq 6L^p \gamma^p + 3\sigma_h^p \gamma^p. \end{aligned}$$

957 By the update rule for the gradient estimator:

$$\hat{e}_t = g_t - \nabla F(x_t) = (1 - \alpha)\hat{e}_{t-1} + \alpha e_t + (1 - \alpha)\hat{S}_t.$$

958 Unrolling the recursion, we have

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

959 Then, taking the norm and the total expectation, we get

$$\mathbb{E} [\|\hat{e}_t\|] \leq (1 - \alpha)^t \mathbb{E} [\|\hat{e}_0\|] + \mathbb{E} \left[\left\| \alpha \sum_{j=0}^{t-1} (1 - \alpha)^{t-j-1} e_{j+1} \right\| \right] + \mathbb{E} \left[\left\| \sum_{j=0}^{t-1} (1 - \alpha)^{t-j} \hat{S}_{j+1} \right\| \right]. \quad (13)$$

960 To continue the proof, we need to bound the last two terms from the previous inequality. By Jensen's
961 inequality, we obtain

$$\begin{aligned} \mathbb{E} \left[\left\| \alpha \sum_{j=0}^{t-1} (1 - \alpha)^{t-j-1} e_{j+1} \right\| \right] &\leq \left(\mathbb{E} \left[\left\| \alpha \sum_{j=0}^{t-1} (1 - \alpha)^{t-j-1} e_{j+1} \right\|^p \right] \right)^{1/p} \\ &\stackrel{(*)}{\leq} \left(2\alpha^p \sum_{j=0}^{t-1} (1 - \alpha)^{p(t-j-1)} \mathbb{E} [\|e_{j+1}\|^p] \right)^{1/p} \\ &\stackrel{(**)}{\leq} \left(2\alpha^p \sum_{j=0}^{t-1} (1 - \alpha)^{p(t-j-1)} \sigma^p \right)^{1/p} \\ &\stackrel{(***)}{\leq} 2 \left(\alpha^{p-1} \sigma^p \right)^{1/p} = 2\alpha^{\frac{p-1}{p}} \sigma, \end{aligned} \quad (14)$$

962 where in $(*)$ we used Lemma 1, in $(**)$ we used $\mathbb{E} [\|e_{j+1}\|^p] \leq \sigma^p$ and in $(***)$ we used the
963 following inequality

$$\sum_{j=0}^{t-1} (1 - \alpha)^{p(t-j-1)} \leq \sum_{j=0}^{t-1} (1 - \alpha)^{t-j-1} \leq \sum_{j=0}^{\infty} (1 - \alpha)^j = \frac{1}{\alpha}.$$

964 We bound the third term in the same way as we did for the second term:

$$\begin{aligned} \mathbb{E} \left[\left\| \sum_{j=0}^{t-1} (1 - \alpha)^{t-j} \hat{S}_{j+1} \right\| \right] &\leq \left(\mathbb{E} \left[\left\| \sum_{j=0}^{t-1} (1 - \alpha)^{t-j} \hat{S}_{j+1} \right\|^p \right] \right)^{1/p} \\ &\stackrel{(*)}{\leq} \left(2 \sum_{j=0}^{t-1} (1 - \alpha)^{p(t-j)} \mathbb{E} [\|\hat{S}_{j+1}\|^p] \right)^{1/p} \\ &\stackrel{(**)}{\leq} \left(12 \sum_{j=0}^{t-1} (1 - \alpha)^{p(t-j)} \gamma^p (L^p + \sigma_h^p) \right)^{1/p} \\ &\stackrel{(***)}{\leq} 12 \left(\alpha^{-1} \gamma^p (L^p + \sigma_h^p) \right)^{1/p} \leq 12(L + \sigma_h) \gamma \alpha^{-1/p}, \end{aligned} \quad (15)$$

965 where in $(*)$ we used Lemma 1, in $(**)$ we used $\mathbb{E} [\|\hat{S}_{j+1}\|^p] \leq 6L^p \gamma^p + 3\sigma^p \gamma^p$, and in $(***)$ we
966 used the following inequality

$$\sum_{j=0}^{t-1} (1 - \alpha)^{p(t-j)} \leq \sum_{j=0}^{t-1} (1 - \alpha)^{t-j} \leq \sum_{j=0}^{\infty} (1 - \alpha)^j = \alpha^{-1}.$$

967 Plugging (14) and (15) into (13), we obtain

$$\mathbb{E} [\|\hat{e}_t\|] \leq (1 - \alpha)^t \mathbb{E} [\|\hat{e}_0\|] + 2\alpha^{\frac{p-1}{p}} \sigma + 12\gamma(L + \sigma_h) \alpha^{-1/p},$$

968 which concludes the proof. \square

969 **Main Theorem.** Now we are ready to state and to prove the main convergence theorem for Algo-
 970 rithm 1.

971 **Theorem 4.** Let Assumptions 1, 2, 3 and 4 hold, and stepsize $\gamma = \sqrt{\frac{\Delta\alpha^{1/p}}{(L+\sigma_h)T}}$, momentum parameter
 972 $\alpha = \min\{1, \alpha_{\text{eff}}\}$, where $\alpha_{\text{eff}} = \max\left\{\left(\frac{\mathcal{E}_0}{T\sigma}\right)^{\frac{p}{2p-1}}, \left(\frac{\Delta(L+\sigma_h)}{T\sigma^2}\right)^{\frac{p}{2p-1}}\right\}$. Then iterates $\{x_t\}_{t=0}^{T-1}$ of
 973 Algorithm 1 satisfy

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} [\|\nabla F(x_t)\|] = \mathcal{O} \left(\sqrt{\frac{\Delta(L+\sigma_h)}{T}} + \sigma \left(\frac{\Delta(L+\sigma_h)}{T\sigma^2} \right)^{\frac{p-1}{2p-1}} + \frac{\mathcal{E}_0}{T} + \sigma \left(\frac{\mathcal{E}_0}{T\sigma} \right)^{\frac{p-1}{2p-1}} \right),$$

974 where \mathcal{E}_0 is defined as some upper bound on $\mathbb{E} [\|g_0 - \nabla F(x_0)\|]$

975 *Proof.* Applying Lemma 9, we have

$$\begin{aligned} \sum_{t=0}^{T-1} \mathbb{E} [\|\hat{e}_t\|] &\leq \sum_{t=0}^{T-1} (1-\alpha)^t \mathbb{E} [\|\hat{e}_0\|] + 2\alpha^{\frac{p-1}{p}} \sigma T + 12\gamma(L+\sigma_h)\alpha^{-1/p} T \\ &\leq \frac{\mathbb{E} [\|\hat{e}_0\|]}{\alpha} + 2\alpha^{\frac{p-1}{p}} \sigma T + 12\gamma(L+\sigma_h)\alpha^{-1/p} T. \end{aligned}$$

976 Therefore, according to Lemma 8, we obtain

$$\begin{aligned} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} [\|\nabla F(x_t)\|] &\leq \frac{\Delta}{\gamma T} + \frac{2}{T} \sum_{t=0}^{T-1} \mathbb{E} [\|\hat{e}_t\|] + \frac{\gamma L}{2} \\ &\leq \frac{\Delta}{\gamma T} + \frac{\gamma L}{2} + \frac{2\mathbb{E} [\|\hat{e}_0\|]}{\alpha T} + 4\alpha^{\frac{p-1}{p}} \sigma + 24\gamma(L+\sigma_h)\alpha^{-1/p} \\ &\leq \frac{\Delta}{\gamma T} + \frac{2\mathbb{E} [\|\hat{e}_0\|]}{\alpha T} + 4\alpha^{\frac{p-1}{p}} \sigma + 25\gamma(L+\sigma_h)\alpha^{-1/p} \\ &\leq \mathcal{O} \left(\sqrt{\frac{\Delta(L+\sigma_h)}{\alpha^{1/p} T}} + \frac{\mathbb{E} [\|\hat{e}_0\|]}{\alpha T} + \alpha^{\frac{p-1}{p}} \sigma \right), \end{aligned}$$

977 where in the last inequality we took $\gamma = \sqrt{\frac{\Delta\alpha^{1/p}}{(L+\sigma_h)T}}$. Denoting $\mathbb{E} [\|\hat{e}_0\|] \leq \mathcal{E}_0$ and taking momentum
 978 parameter $\alpha = \min\{1, \alpha_{\text{eff}}\}$, where $\alpha_{\text{eff}} = \max\left\{\left(\frac{\mathcal{E}_0}{T\sigma}\right)^{\frac{p}{2p-1}}, \left(\frac{\Delta(L+\sigma_h)}{T\sigma^2}\right)^{\frac{p}{2p-1}}\right\}$, we have

$$\begin{aligned} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} [\|\nabla F(x_t)\|] &= \mathcal{O} \left(\sqrt{\frac{\Delta(L+\sigma_h)}{\alpha^{1/p} T}} + \frac{\mathcal{E}_0}{\alpha T} + \alpha^{\frac{p-1}{p}} \sigma \right) \\ &= \mathcal{O} \left(\sqrt{\frac{\Delta(L+\sigma_h)}{T}} + \frac{\mathcal{E}_0}{T} + \sigma \left(\frac{\Delta(L+\sigma_h)}{T\sigma^2} \right)^{\frac{p-1}{2p-1}} + \sigma \left(\frac{\mathcal{E}_0}{T\sigma} \right)^{\frac{p-1}{2p-1}} \right). \end{aligned}$$

979 □

980 Now we investigate how different choices of initial estimator g_0 affect the total sample complexity
 981 bound.

982 **Corollary 2.** Let all assumptions of Theorem 4 hold and the step-size and momentum parameters are
 983 set according to this theorem statement.

984 1. If we set $g_0 = \nabla F(x_0)$, then $\mathcal{E}_0 = 0$, and the total sample complexity of Algorithm 1 is
 985 equal to

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

986 2. If we set $g_0 = 0$, then $\mathcal{E}_0 = \sqrt{2\Delta L}$, and the total sample complexity of Algorithm 1 is equal
987 to

$$\mathcal{O}\left(\frac{\Delta(L + \sigma_h)}{\varepsilon^2} + \frac{\Delta(L + \sigma_h)}{\varepsilon^2} \left(\frac{\sigma}{\varepsilon}\right)^{\frac{1}{p-1}} + \frac{\sqrt{\Delta L}\sigma}{\varepsilon^2} \left(\frac{\sigma}{\varepsilon}\right)^{\frac{1}{p-1}}\right).$$

988 3. If we set $g_0 = \frac{1}{B_{\text{init}}} \sum_{j=1}^{B_{\text{init}}} \nabla f(x_0, \xi_{0,j})$ with $B_{\text{init}} = \max\left\{1, \left(\frac{\sigma}{\varepsilon}\right)^{\frac{p}{p-1}}, \left(\frac{\sigma}{\varepsilon}\right)^{\frac{p}{2p-1}}\right\}$, then
989 $\mathcal{E}_0 = 2\sigma/B_{\text{init}}^{\frac{p-1}{p}}$, and the total sample complexity of Algorithm 1 is equal to

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

990 *Proof.* The first case is ideal, when we can have access to the full gradient but once: set $g_0 = \nabla F(x_0)$,
991 then $\hat{e}_0 = 0$ and $\mathcal{E}_0 = 0$. Thus we have

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

992 In other words, we can guarantee $\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla F(x_t)\|] \leq \varepsilon$ after

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

993 iterations of Algorithm 1.

994 The second case is that we select g_0 as a zero vector. This choice implies that by smoothness of F ,
995 we have

$$\mathbb{E}[\|\hat{e}_0\|] = \|\nabla F(x_0)\| \leq \sqrt{2L\Delta} = \mathcal{E}_0.$$

996 Plugging the obtained value of \mathcal{E}_0 into the result of Theorem 2, we have

$$\begin{aligned} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla F(x_t)\|] &= \mathcal{O}\left(\sqrt{\frac{\Delta(L + \sigma_h)}{T}} + \sigma \left(\frac{\Delta(L + \sigma_h)}{T\sigma^2}\right)^{\frac{p-1}{2p-1}} + \frac{\sqrt{\Delta L}}{T} + \sigma \left(\frac{\sqrt{\Delta L}}{T\sigma}\right)^{\frac{p-1}{2p-1}}\right) \\ &= \mathcal{O}\left(\sqrt{\frac{\Delta(L + \sigma_h)}{T}} + \sigma \left(\frac{\Delta(L + \sigma_h)}{T\sigma^2}\right)^{\frac{p-1}{2p-1}} + \sigma \left(\frac{\sqrt{\Delta L}}{T\sigma}\right)^{\frac{p-1}{2p-1}}\right) \end{aligned}$$

997 which implies that the total sample complexity is

$$\mathcal{O}\left(\frac{\Delta(L + \sigma_h)}{\varepsilon^2} + \frac{\Delta(L + \sigma_h)}{\varepsilon^2} \left(\frac{\sigma}{\varepsilon}\right)^{\frac{1}{p-1}} + \frac{\sqrt{\Delta L}\sigma}{\varepsilon^2} \left(\frac{\sigma}{\varepsilon}\right)^{\frac{1}{p-1}}\right).$$

998 The third case is that we set $g_0 = \frac{1}{B_{\text{init}}} \sum_{j=1}^{B_{\text{init}}} \nabla f(x_0, \xi_{0,j})$. By Lemma 1, we have

$$\begin{aligned} \mathbb{E}[\|\hat{e}_0\|] &\leq (\mathbb{E}[\|\hat{e}_0\|^p])^{1/p} = (\mathbb{E}[\|g_0 - \nabla F(x_0)\|^p])^{1/p} \\ &\leq \frac{2}{B_{\text{init}}} \left(\sum_{j=0}^{B_{\text{init}}} \mathbb{E}[\|\nabla f(x_0, \xi_{0,j}) - \nabla F(x_0)\|^p] \right)^{1/p} \\ &\leq \frac{2}{B_{\text{init}}} \left(\sum_{j=0}^{B_{\text{init}}} \sigma^p \right)^{1/p} = \frac{2\sigma}{B_{\text{init}}^{\frac{p-1}{p}}} = \mathcal{E}_0. \end{aligned}$$

999 Plugging the obtained value of \mathcal{E}_0 into the result of Theorem 2, we have

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla F(x_t)\|] = \mathcal{O}\left(\sqrt{\frac{\Delta(L + \sigma_h)}{T}} + \sigma \left(\frac{\Delta(L + \sigma_h)}{T\sigma^2}\right)^{\frac{p-1}{2p-1}} + \frac{\sigma}{TB_{\text{init}}^{\frac{p-1}{p}}} + \sigma \left(\frac{1}{TB_{\text{init}}^{\frac{p-1}{p}}}\right)^{\frac{p-1}{2p-1}}\right),$$

1000 from which we have that the total sample complexity is

$$\mathcal{O} \left(\frac{\Delta(L + \sigma_h)}{\varepsilon^2} + \frac{\Delta(L + \sigma_h)}{\varepsilon^2} \left(\frac{\sigma}{\varepsilon} \right)^{\frac{1}{p-1}} + \frac{\sigma^2}{B_{\text{init}}^{\frac{p}{p-1}} \varepsilon^2} \left(\frac{\sigma}{\varepsilon} \right)^{\frac{1}{p-1}} + \frac{\sigma}{\varepsilon B_{\text{init}}^{\frac{p}{p-1}}} + B_{\text{init}} \right).$$

1001 Taking $B_{\text{init}} = \max \left\{ 1, \left(\frac{\sigma}{\varepsilon} \right)^{\frac{p}{p-1}}, \left(\frac{\sigma}{\varepsilon} \right)^{\frac{p}{2p-1}} \right\}$, we have that the total sample complexity is

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

1002

□

1003 F Missing Proofs for Section 5

1004 Now we start the high-probability convergence analysis for Algorithm 2. We divide our analysis into
 1005 two parts: Optimization part, where we prove descent lemma, and High-Probability part, where we
 1006 use concentration inequality to bound several terms from Descent Lemma. Finally, combining results
 1007 from both parts, we prove the main results of this section via induction. The idea is based on work of
 1008 Sadiev et al. [2023] and Liu et al. [2023b].

1009 F.1 Analysis: Optimization Part

1010 We start with *Descent Lemma*.

1011 **Lemma 10.** *Let Assumptions 1, 2 hold. Then Algorithm 2 with stepsize $\gamma > 0$ and momentum*
 1012 *parameter $\alpha \in (0, 1)$ generates iterates $\{x_t\}_{t=0}^T$ satisfying the following inequality*

$$\begin{aligned} \gamma \sum_{t=0}^{T-1} \|\nabla F(x_t)\| + \Delta_T &\leq \Delta_1 + \frac{\gamma^2 LT}{2} + \frac{3\gamma}{\alpha} \sqrt{L\Delta_1} \\ &\quad + 2\gamma\alpha \sum_{t=1}^{T-1} \left\| \sum_{j=1}^t (1-\alpha)^{t-j} \theta_j \right\| + 2\gamma \sum_{t=1}^{T-1} \left\| \sum_{j=1}^t (1-\alpha)^{t-j+1} \omega_j \right\|, \end{aligned}$$

1013 for any $j \in [T-1]$ vectors θ_j and ω_j are defined as follows

$$\theta_j \stackrel{\text{def}}{=} \text{clip}(\nabla f(x_j, \xi_j), \lambda) - \nabla F(x_j), \quad (16)$$

$$\omega_j \stackrel{\text{def}}{=} \text{clip}(\nabla^2 f(\hat{x}_t, \hat{\xi}_t)(x_t - x_{t-1}), \lambda_h) - (\nabla F(x_j) - \nabla F(x_{j-1})). \quad (17)$$

1014 *Proof.* First, we notice that Lemma 8 holds for Algorithm 2 and can be proven in the same way as
 1015 for Algorithm 1, since the gradient updates in the both methods are identical except for x_1 , and the
 1016 update rule for momentum term does not play any role in the proof of Lemma 8. Therefore, the
 1017 iterates $\{x_t\}_{t=0}^T$ of Algorithm 2 satisfy

$$\gamma \sum_{t=0}^{T-1} \|\nabla F(x_t)\| + \Delta_T \leq \Delta_1 + 2\gamma \sum_{t=0}^{T-1} \|\hat{e}_t\| + \frac{\gamma^2 LT}{2}, \quad (18)$$

1018 where $\Delta_0 = \Delta_1$, because $x_0 = x_1$. Next we bound $\|\hat{e}_t\|$ in almost the same way as we did in
 1019 Lemma 9. By update rule for momentum parameter, we have

$$\begin{aligned} \hat{e}_t &= g_t - \nabla F(x_t) \\ &= (1-\alpha) \left(g_{t-1} + \text{clip}(\nabla^2 f(\hat{x}_t, \hat{\xi}_t)(x_t - x_{t-1}), \lambda_h) \right) \\ &\quad + \alpha \text{clip}(\nabla f(x_t, \xi_t), \lambda) - \nabla F(x_t) \\ &\stackrel{(16),(17)}{=} (1-\alpha) \hat{e}_{t-1} + \alpha \theta_t + (1-\alpha) \omega_t \\ &= (1-\alpha)^t \hat{e}_0 + \alpha \sum_{j=1}^t (1-\alpha)^{t-j} \theta_j + \sum_{j=1}^t (1-\alpha)^{t-j+1} \omega_j. \end{aligned}$$

1020 Taking the norm, we obtain the following bound

$$\begin{aligned} \|\hat{e}_t\| &\leq (1-\alpha)^t \|\hat{e}_0\| + \alpha \left\| \sum_{j=1}^t (1-\alpha)^{t-j} \theta_j \right\| + \left\| \sum_{j=1}^t (1-\alpha)^{t-j+1} \omega_j \right\| \\ &\leq (1-\alpha)^t \sqrt{2L\Delta_1} + \alpha \left\| \sum_{j=1}^t (1-\alpha)^{t-j} \theta_j \right\| + \left\| \sum_{j=1}^t (1-\alpha)^{t-j+1} \omega_j \right\|, \end{aligned} \quad (19)$$

1021 where in the last inequality we used the following chain of inequalities

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

1022 Plugging (19) into (18), we acquire

$$\begin{aligned}
\gamma \sum_{t=0}^{T-1} \|\nabla F(x_t)\| + \Delta_T &\leq \Delta_1 + 2\gamma \sum_{t=0}^{T-1} \|\hat{e}_t\| + \frac{\gamma^2 LT}{2} \\
&\leq \Delta_1 + \frac{\gamma^2 LT}{2} + 2\gamma \sum_{t=0}^{T-1} (1-\alpha)^t \sqrt{2L\Delta_1} \\
&\quad + 2\gamma\alpha \sum_{t=1}^{T-1} \left\| \sum_{j=1}^t (1-\alpha)^{t-j} \theta_j \right\| + 2\gamma \sum_{t=1}^{T-1} \left\| \sum_{j=1}^t (1-\alpha)^{t-j+1} \omega_j \right\| \\
&\leq \Delta_1 + \frac{\gamma^2 LT}{2} + \frac{3\gamma}{\alpha} \sqrt{L\Delta_1} \\
&\quad + 2\gamma\alpha \sum_{t=1}^{T-1} \left\| \sum_{j=1}^t (1-\alpha)^{t-j} \theta_j \right\| + 2\gamma \sum_{t=1}^{T-1} \left\| \sum_{j=1}^t (1-\alpha)^{t-j+1} \omega_j \right\|,
\end{aligned}$$

1023 where in the last inequality we used

$$2\gamma \sum_{t=0}^{T-1} (1-\alpha)^t \sqrt{2L\Delta_1} \leq 2\gamma \sum_{t=0}^{\infty} (1-\alpha)^t \sqrt{2L\Delta_1} \leq \frac{2\sqrt{2}\gamma}{1-(1-\alpha)} \sqrt{L\Delta_1} \leq \frac{3\gamma}{\alpha} \sqrt{L\Delta_1}.$$

1024

□

1025 F.2 Analysis: Statistical Part

1026 According to Lemma 10, we have two new terms

$$\left\| \sum_{j=1}^t (1-\alpha)^{t-j} \theta_j \right\| \quad \text{and} \quad \left\| \sum_{j=1}^t (1-\alpha)^{t-j+1} \omega_j \right\|.$$

1027 To bound both of them, we use the same idea as in the work of [Gorbunov et al. \[2020\]](#), [Sadiev et al. \[2023\]](#), [Liu et al. \[2023b\]](#): introduce unbiased and biased parts of θ_t and ω_t , i.e. for any $t \in [T]$

$$\theta_t = \theta_t^u + \theta_t^b, \quad \text{where} \quad \theta_t^u \stackrel{\text{def}}{=} \text{clip}(\nabla f(x_t, \xi_t), \lambda) - \mathbb{E}_{\xi_t} [\text{clip}(\nabla f(x_t, \xi_t), \lambda)], \quad (20)$$

$$\theta_t^b \stackrel{\text{def}}{=} \mathbb{E}_{\xi_t} [\text{clip}(\nabla f(x_t, \xi_t), \lambda)] - \nabla F(x_t); \quad (21)$$

$$\omega_t = \omega_t^u + \omega_t^b, \quad \text{where} \quad \omega_t^u \stackrel{\text{def}}{=} \text{clip}(\nabla^2 f(\hat{x}_t, \hat{\xi}_t)(x_t - x_{t-1}), \lambda_h) - \mathbb{E}_{q_t, \xi_t} [\text{clip}(\nabla^2 f(\hat{x}_t, \hat{\xi}_t)(x_t - x_{t-1}), \lambda_h)] \quad (22)$$

$$\omega_t^b \stackrel{\text{def}}{=} \mathbb{E}_{q_t, \xi_t} [\text{clip}(\nabla^2 f(\hat{x}_t, \hat{\xi}_t)(x_t - x_{t-1}), \lambda_h)] - (\nabla F(x_t) - \nabla F(x_{t-1})). \quad (23)$$

1029 Under Assumption 3 for any $\lambda \geq 2\|\nabla F(x_t)\|$ Lemma 4 implies

$$\mathbb{E}_{\xi_t} [\|\theta_t^u\|^2] \leq 18\lambda^{2-p}\sigma^p \quad \text{and} \quad \|\theta_t^b\| \leq 2^p\lambda^{1-p}\sigma^p, \quad \text{for all } t \in [T]. \quad (24)$$

1030 It is worth to mention we have already shown in Lemma 8 that under Assumptions 2 and 4 the vector

1031 $\hat{S}_t = \nabla^2 f(\hat{x}_t, \hat{\xi}_t) \cdot (x_t - x_{t-1}) - \nabla F(x_t) + \nabla F(x_{t-1})$ has zero expectation $\mathbb{E}_{q_t, \xi_t} [\hat{S}_t] = 0$, since

$$\begin{aligned}
\mathbb{E}_{q_t, \xi_t} [\nabla^2 f(\hat{x}_t, \hat{\xi}_t)(x_t - x_{t-1})] &= \mathbb{E}_{q_t} [\mathbb{E}_{\hat{\xi}_t} [\nabla^2 f(\hat{x}_t, \hat{\xi}_t)] (x_t - x_{t-1})] \\
&= \mathbb{E}_{q_t} [\nabla^2 F(\hat{x}_t)(x_t - x_{t-1})] \\
&= \int_0^1 \nabla^2 F(qx_t + (1-q)x_{t-1})(x_t - x_{t-1}) dq \\
&= \nabla F(x_t) - \nabla F(x_{t-1}).
\end{aligned}$$

1032 The p -central moment of \hat{S}_t is bounded, i.e. $\mathbb{E}_{q_t, \xi_t} \left[\left\| \hat{S}_t \right\|^p \right] \leq 6L^p \gamma^p + 3\sigma_h^p \gamma^p$. Then, according to

1033 Lemma 4, for all $t \in [T]$ we have

$$\mathbb{E}_{q_t, \xi_t} \left[\left\| \omega_j^u \right\|^2 \right] \leq 18\lambda_h^{2-p} \gamma^p (6L^p + 3\sigma_h^p) \quad \text{and} \quad \left\| \omega_t^b \right\| \leq 2^p \lambda_h^{1-p} \gamma^p (6L^p + 3\sigma_h^p), \quad (25)$$

1034 for any $2 \left\| \nabla F(x_t) - \nabla F(x_{t-1}) \right\| \leq 2L\gamma \leq \lambda_h$.

Lemma 11. *Let Assumption 3 hold. For any $\delta' \in (0, 1/2]$ and any $t \in [T]$, if clipping level satisfy*

$$\lambda \geq \max \left\{ 2 \left\| \nabla F(x_j) \right\|, \frac{\sigma}{\alpha^{1/p}} \right\},$$

1035 *then with probability at least $1 - 2\delta'$*

$$\left\| \sum_{j=1}^t (1-\alpha)^{t-j} \theta_j \right\| \leq 22\lambda \log \frac{2}{\delta'}.$$

1036 *Proof.* We start with bounding $\left\| \sum_{j=1}^t (1-\alpha)^{t-j} \theta_j \right\|$.

1037 **Upper bound for** $\left\| \sum_{j=1}^t (1-\alpha)^{t-j} \theta_j \right\|$ By (20) and (21), we have

$$\left\| \sum_{j=1}^t (1-\alpha)^{t-j} \theta_j \right\| \leq \left\| \sum_{j=1}^t (1-\alpha)^{t-j} \theta_j^u \right\| + \underbrace{\left\| \sum_{j=1}^t (1-\alpha)^{t-j} \theta_j^b \right\|}_{\textcircled{4}}.$$

1038 Denote $Y_j^t \stackrel{\text{def}}{=} \left\| (1-\alpha)^{t-j} \theta_j^u \right\|^2 - \mathbb{E}_{\xi_j} \left[\left\| (1-\alpha)^{t-j} \theta_j^u \right\|^2 \right]$, and $|X_j^t| \leq \left\| (1-\alpha)^{t-j} \theta_j^u \right\|$ for any

1039 $j \in [t]$

$$V_j^t \stackrel{\text{def}}{=} \begin{cases} 0, & \text{if } j = 0; \\ \text{sign} \left(\sum_{i=1}^{j-1} V_i^t \right) \frac{\left\langle \sum_{i=1}^{j-1} (1-\alpha)^{t-i} \theta_i^u, (1-\alpha)^{t-j} \theta_j^u \right\rangle}{\left\| \sum_{i=1}^{j-1} (1-\alpha)^{t-i} \theta_i^u \right\|}, & \text{if } j \neq 0 \text{ and } \sum_{i=1}^{j-1} (1-\alpha)^{t-i} \theta_i^u \neq 0; \\ 0, & \text{if } j \neq 0 \text{ and } \sum_{i=1}^{j-1} (1-\alpha)^{t-i} \theta_i^u = 0. \end{cases}$$

1040 Then to bound $\left\| \sum_{j=1}^t (1-\alpha)^{t-j} \theta_j^u \right\|$ we use Lemma 2 and obtain

$$\begin{aligned} \left\| \sum_{j=1}^t (1-\alpha)^{t-j} \theta_j^u \right\| &\leq \left| \sum_{j=1}^t V_j^t \right| + \sqrt{\max_{j \in [t]} \left\| (1-\alpha)^{t-j} \theta_j^u \right\|^2 + \sum_{j=1}^t \left\| (1-\alpha)^{t-j} \theta_j^u \right\|^2} \\ &\leq \left| \sum_{j=1}^t V_j^t \right| + \sqrt{2 \sum_{j=1}^t \left\| (1-\alpha)^{t-j} \theta_j^u \right\|^2} \\ &\leq \underbrace{\left| \sum_{j=1}^t V_j^t \right|}_{\textcircled{1}} + \underbrace{\left(2 \cdot \sum_{j=1}^t Y_j^t + 2 \cdot \sum_{j=1}^t \mathbb{E}_{\xi_j} \left[\left\| (1-\alpha)^{t-j} \theta_j^u \right\|^2 \right] \right)^{1/2}}_{\textcircled{3}}. \end{aligned}$$

1041 **Upper bound for ①.** The sequence X_1^t, \dots, X_t^t is martingale difference sequence, since, by definition

1042 of θ_j^t and X_j^t , for all $j \in [t]$ we have $\mathbb{E} \left[X_j^t \mid X_{j-1}^t, \dots, X_1^t \right] = \mathbb{E}_{\xi_j} \left[X_j^t \right] = 0$. Also, according to

1043 Lemma 2, we have $|X_j^t| \leq \left\| (1-\alpha)^{t-j} \theta_j^u \right\|$ for all $j \in [t]$. Using Lemma 4, we obtain that

$$|X_j^t| \leq \left\| (1-\alpha)^{t-j} \theta_j^u \right\| \leq \left\| \theta_j^u \right\| \leq 2\lambda, \quad \text{where } c_1 \stackrel{\text{def}}{=} 2\lambda. \quad (26)$$

1044 Denoting $\sigma_j^2 \stackrel{\text{def}}{=} \mathbb{E}[(X_j^t)^2 \mid X_{j-1}^t, \dots, X_1^t] = \mathbb{E}_{\xi_j}[(X_j^t)^2]$, Lemma 3 implies

$$\mathbb{P} \left\{ |\mathbb{Q}| > b_1 \text{ and } \sum_{j=1}^t \sigma_j^2 \leq G_1 \log \frac{2}{\delta'} \right\} \leq 2 \exp \left(-\frac{b_1^2}{2G_1 \log \frac{2}{\delta'} + \frac{2c_1 b_1}{3}} \right) = \delta',$$

1045 where the last identity is true, if we set $b_1 = \left(\frac{1}{3}c_1 + \sqrt{\frac{1}{9}c_1 + 2G_1} \right) \log \frac{2}{\delta'}$. To define G_1 , we need
 1046 to bound $\sum_{j=1}^t \sigma_j^2$:

$$\begin{aligned} \sum_{j=1}^t \sigma_j^2 &= \sum_{j=1}^t \mathbb{E}_{\xi_j}[(X_j^t)^2] \leq \sum_{j=1}^t \mathbb{E}_{\xi_j}[\|(1-\alpha)^{t-j}\theta_j^u\|^2] = \sum_{j=1}^t (1-\alpha)^{2(t-j)} \mathbb{E}_{\xi_j}[\|\theta_j^u\|^2] \\ &\stackrel{(24)}{\leq} \sum_{j=1}^t (1-\alpha)^{2(t-j)} \cdot 18\lambda^{2-p}\sigma^p \leq \frac{18\lambda^{2-p}\sigma^p}{1-(1-\alpha)^2} \leq \frac{18\lambda^{2-p}\sigma^p}{\alpha} = G_1. \end{aligned} \quad (27)$$

1047 where we assumed $\|\nabla F(x_j)\| \leq \frac{\lambda}{2}$ for any $j \in [t]$.

1048 **Upper bound for ②.** The sequence Y_1^t, \dots, Y_t^t forms a martingale difference sequence, since the
 1049 definition of Y_j^t leads to $\mathbb{E}[Y_j^t \mid Y_{j-1}^t, \dots, Y_1^t] = \mathbb{E}_{\xi_j}[Y_j^t] = 0$ for all $j \in [t]$. Also, according to
 1050 Lemma 4, we obtain that

$$|Y_j^t| \leq \|(1-\alpha)^{t-j}\theta_j^u\|^2 + \mathbb{E}_{\xi_j}[\|(1-\alpha)^{t-j}\theta_j^u\|^2] \leq 4\lambda^2 + 4\lambda^2 = 8\lambda^2, \quad (28)$$

1051 where we define $c_2 \stackrel{\text{def}}{=} 8\lambda^2$. Denoting the conditional variance of Y_j^t as $\tilde{\sigma}_j^2 \stackrel{\text{def}}{=} \mathbb{E}[(Y_j^t)^2 \mid Y_{j-1}^t, \dots, Y_1^t] = \mathbb{E}_{\xi_j}[(Y_j^t)^2]$, we can bound $\tilde{\sigma}_j^2$ easily

$$\begin{aligned} \tilde{\sigma}_j^2 &\stackrel{(28)}{\leq} 8\lambda^2 \cdot \mathbb{E}[\|(1-\alpha)^{t-j}\theta_j^u\|^2 - \mathbb{E}_{\xi_j}[\|(1-\alpha)^{t-j}\theta_j^u\|^2]] \\ &\leq 16\lambda^2 \mathbb{E}_{\xi_j}[\|(1-\alpha)^{t-j}\theta_j^u\|^2]. \end{aligned}$$

1053 Then, Lemma 3 implies

$$\mathbb{P} \left\{ |\mathbb{Q}| > b_2 \text{ and } \sum_{j=1}^t \tilde{\sigma}_j^2 \leq G_2 \log \frac{2}{\delta'} \right\} \leq 2 \exp \left(-\frac{b_2^2}{2G_2 \log \frac{2}{\delta'} + \frac{2c_2 b_2}{3}} \right) = \delta',$$

1054 where the last identity is true, if we set $b_2 = \left(\frac{1}{3}c_2 + \sqrt{\frac{1}{9}c_2 + 2G_2} \right) \log \frac{2}{\delta'}$. To define G_2 , we need
 1055 to bound $\sum_{j=1}^t \tilde{\sigma}_j^2$:

$$\begin{aligned} \sum_{j=1}^t \tilde{\sigma}_j^2 &\leq 16\lambda^2 \sum_{j=1}^t \mathbb{E}_{\xi_j}[\|(1-\alpha)^{t-j}\theta_j^u\|^2] \\ &= 16\lambda^2 \sum_{j=1}^t (1-\alpha)^{2(t-j)} \mathbb{E}_{\xi_j}[\|\theta_j^u\|^2] \\ &\stackrel{(27)}{\leq} 16\lambda^2 \cdot \frac{18\lambda^{2-p}\sigma^p}{\alpha} = \frac{16 \cdot 18\lambda^{4-p}\sigma^p}{\alpha} = G_2, \end{aligned}$$

1056 where we assumed $\|\nabla F(x_j)\| \leq \frac{\lambda}{2}$ for any $j \in [t]$.

1057 **Upper bound for ③.** Thanks to the proof of the bound on ③ (see (27)), we have already shown what
 1058 we need: with probability 1

$$\textcircled{3} = \sum_{j=1}^t \mathbb{E}_{\xi_j}[\|(1-\alpha)^{t-j}\theta_j^u\|^2] \stackrel{(27)}{\leq} \frac{18\lambda^{2-p}\sigma^p}{\alpha}.$$

1059 **Upper bound on ④.** Assuming $\|\nabla F(x_j)\| \leq \frac{\lambda}{2}$ for any $j \in [t]$, with probability 1 we have

$$\textcircled{4} = \left\| \sum_{j=1}^t (1-\alpha)^{t-j} \theta_j^b \right\| \leq \sum_{j=1}^t (1-\alpha)^{t-j} \|\theta_j^b\| \stackrel{(27)}{\leq} 4\lambda^{1-p} \sigma^p \sum_{j=1}^t (1-\alpha)^{t-j} \leq \frac{4\lambda^{1-p} \sigma^p}{\alpha}.$$

1060 To sum up, we introduce event $E_{\textcircled{1},t}$ as follows

$$E_{\textcircled{1},t} \stackrel{\text{def}}{=} \left\{ |\textcircled{1}| \leq b_1 \text{ or } \sum_{j=1}^t \sigma_j^2 > G_1 \log \frac{2}{\delta'} \right\}, \quad (29)$$

1061 where $c_1 = 2\lambda$, $G_1 = \frac{18\lambda^{2-p}\sigma^p}{\alpha}$, $b_1 = \left(\frac{1}{3}c_1 + \sqrt{\frac{1}{9}c_1^2 + 2G_1}\right) \log \frac{2}{\delta'}$. We have shown that

1062 $\mathbb{P}\{E_{\textcircled{1},t}\} \geq 1 - \delta'$. The bound on b_1

$$\begin{aligned} b_1 &= \left(\frac{1}{3}c_1 + \sqrt{\frac{1}{9}c_1^2 + 2G_1}\right) \log \frac{2}{\delta'} \leq \left(\frac{2}{3}c_1 + \sqrt{2G_1}\right) \log \frac{2}{\delta'} \\ &= \left(\frac{4}{3}\lambda + \sqrt{36\frac{\lambda^{2-p}\sigma^p}{\alpha}}\right) \log \frac{2}{\delta'} = \lambda \left(\frac{4}{3} + 6\sqrt{\frac{1}{\alpha} \left(\frac{\sigma}{\lambda}\right)^p}\right) \log \frac{2}{\delta'}. \end{aligned}$$

1063 Also we define event $E_{\textcircled{2},t}$ as follows

$$E_{\textcircled{2},t} = \left\{ |\textcircled{2}| \leq b_2 \text{ or } \sum_{j=1}^t \tilde{\sigma}_j^2 > G_2 \log \frac{2}{\delta'} \right\}, \quad (30)$$

1064 where $c_2 = 8\lambda^2$, $G_2 = \frac{16 \cdot 18\lambda^{4-p}\sigma^p}{\alpha}$, $b_2 = \left(\frac{1}{3}c_2 + \sqrt{\frac{1}{9}c_2^2 + 2G_2}\right) \log \frac{2}{\delta'}$. We have shown that

1065 $\mathbb{P}\{E_{\textcircled{2},t}\} \geq 1 - \delta'$. We adjust the bound on b_2 :

$$\begin{aligned} b_2 &= \left(\frac{1}{3}c_2 + \sqrt{\frac{1}{9}c_2^2 + 2G_2}\right) \log \frac{2}{\delta'} \leq \left(\frac{2}{3}c_2 + \sqrt{2G_2}\right) \log \frac{2}{\delta'} \\ &= \left(\frac{16}{3}\lambda^2 + \sqrt{\frac{16 \cdot 36\lambda^{4-p}\sigma^p}{\alpha}}\right) \log \frac{2}{\delta'} = \lambda^2 \left(\frac{16}{3} + 24\sqrt{\frac{1}{\alpha} \left(\frac{\sigma}{\lambda}\right)^p}\right) \log \frac{2}{\delta'}. \end{aligned}$$

1066 Thus, the event $E_{\textcircled{1},t} \cap E_{\textcircled{2},t}$ implies

$$\begin{aligned} \left\| \sum_{j=1}^t (1-\alpha)^{t-j} \theta_j \right\| &\leq \left\| \sum_{j=1}^t (1-\alpha)^{t-j} \theta_j^u \right\| + \left\| \sum_{j=1}^t (1-\alpha)^{t-j} \theta_j^b \right\| \\ &\leq |\textcircled{1}| + \sqrt{2 \cdot \textcircled{2} + 2 \cdot \textcircled{3}} + \textcircled{4} \\ &\leq \lambda \left(\frac{4}{3} + 6\sqrt{\frac{1}{\alpha} \left(\frac{\sigma}{\lambda}\right)^p}\right) \log \frac{2}{\delta'} + 4\lambda \cdot \frac{1}{\alpha} \left(\frac{\sigma}{\lambda}\right)^p \\ &\quad + \sqrt{2\lambda^2 \left(\frac{16}{3} + 24\sqrt{\frac{1}{\alpha} \left(\frac{\sigma}{\lambda}\right)^p}\right) \log \frac{2}{\delta'} + \frac{36\lambda^{2-p}\sigma^p}{\alpha}}. \end{aligned}$$

1067 Taking $\lambda \geq \frac{\sigma}{\alpha^{1/p}}$, we have the event $E_{\textcircled{1},t} \cap E_{\textcircled{2},t}$ implies

$$\left\| \sum_{j=1}^t (1-\alpha)^{t-j} \theta_j \right\| \leq \lambda \left(\frac{4}{3} + 6 + 4 + \sqrt{\frac{32}{3} + 84}\right) \log \frac{2}{\delta'} \leq 22\lambda \log \frac{2}{\delta'}.$$

1068 This concludes the proof. \square

Lemma 12. *Let Assumptions 2 and 4 hold. For any $\delta'' \in (0, 1/2]$ and any $t \in [T]$, if clipping level satisfy*

$$\lambda_h \geq \max \left\{ 2\gamma L, \frac{\gamma(L + \sigma_h)}{\alpha^{1/p}} \right\},$$

1069 *and with probability at least $1 - 2\delta''$*

$$\left\| \sum_{j=1}^t (1 - \alpha)^{t-j+1} \omega_j \right\| \leq 59\lambda_h \log \frac{2}{\delta''}.$$

1070 *Proof.* By (22) and (23), we have

$$\left\| \sum_{j=1}^t (1 - \alpha)^{t-j+1} \omega_j \right\| \leq \left\| \sum_{j=1}^t (1 - \alpha)^{t-j+1} \omega_j^u \right\| + \underbrace{\left\| \sum_{j=1}^t (1 - \alpha)^{t-j+1} \omega_j^b \right\|}_{\textcircled{8}}. \quad (31)$$

1071 Denote $Z_j^t \stackrel{\text{def}}{=} \|(1 - \alpha)^{t-j+1} \omega_j^u\|^2 - \mathbb{E}_{q_j, \xi_j} [\|(1 - \alpha)^{t-j+1} \omega_j^u\|^2]$, and $|W_j^t| \leq \|(1 - \alpha)^{t-j+1} \omega_j^u\|$
 1072 for any $j \in [t]$

$$W_j^t \stackrel{\text{def}}{=} \begin{cases} 0, & \text{if } j = 0; \\ \text{sign} \left(\sum_{i=1}^{j-1} W_i^t \right) \frac{\left\langle \sum_{i=1}^{j-1} (1 - \alpha)^{t-i+1} \omega_i^u, (1 - \alpha)^{t-j+1} \omega_j^u \right\rangle}{\left\| \sum_{i=1}^{j-1} (1 - \alpha)^{t-i+1} \omega_i^u \right\|}, & \text{if } j \neq 0 \text{ and } \sum_{i=1}^{j-1} (1 - \alpha)^{t-i+1} \omega_i^u \neq 0; \\ 0, & \text{if } j \neq 0 \text{ and } \sum_{i=1}^{j-1} (1 - \alpha)^{t-i+1} \omega_i^u = 0. \end{cases}$$

1073 Then to bound $\left\| \sum_{j=1}^t (1 - \alpha)^{t-j+1} \omega_j^u \right\|$ we use Lemma 2 and obtain

$$\begin{aligned} \left\| \sum_{j=1}^t (1 - \alpha)^{t-j} \omega_j^u \right\| &\leq \left| \sum_{j=1}^t W_j^t \right| + \sqrt{\max_{j \in [t]} \|(1 - \alpha)^{t-j} \omega_j^u\|^2 + \sum_{j=1}^t \|(1 - \alpha)^{t-j} \omega_j^u\|^2} \\ &\leq \left| \sum_{j=1}^t W_j^t \right| + \sqrt{2 \sum_{j=1}^t \|(1 - \alpha)^{t-j+1} \omega_j^u\|^2} \\ &\leq \underbrace{\left| \sum_{j=1}^t W_j^t \right|}_{\textcircled{5}} + \left(\underbrace{2 \cdot \sum_{j=1}^t Z_j^t}_{\textcircled{6}} + \underbrace{2 \cdot \sum_{j=1}^t \mathbb{E}_{q_j, \xi_j} [\|(1 - \alpha)^{t-j+1} \omega_j^u\|^2]}_{\textcircled{7}} \right)^{1/2}. \end{aligned}$$

1074 **Upper bound for ⑤.** The sequence W_1^t, \dots, W_t^t is martingale difference sequence, since, by
 1075 definition of ω_j^t and W_j^t , for all $j \in [t]$ we have $\mathbb{E} [W_j^t \mid W_{j-1}^t, \dots, W_1^t] = \mathbb{E}_{q_j, \xi_j} [W_j^t] = 0$. Also,
 1076 according to Lemma 2, we have $|W_j^t| \leq \|(1 - \alpha)^{t-j} \omega_j^u\|$ for all $j \in [t]$. Using Lemma 4, we obtain
 1077 that

$$|W_j^t| \leq \|(1 - \alpha)^{t-j} \omega_j^u\| \leq \|\omega_j^u\| \leq 2\lambda_h, \quad \text{where } c_3 \stackrel{\text{def}}{=} 2\lambda_h. \quad (32)$$

1078 Denoting $\sigma_j^2 \stackrel{\text{def}}{=} \mathbb{E} [(W_j^t)^2 \mid W_{j-1}^t, \dots, W_1^t] = \mathbb{E}_{\xi_j} [(W_j^t)^2]$, Lemma 3 implies

$$\mathbb{P} \left\{ |\textcircled{5}| > b_3 \text{ and } \sum_{j=1}^t \bar{\sigma}_j^2 \leq G_3 \log \frac{2}{\delta''} \right\} \leq 2 \exp \left(- \frac{b_3^2}{2G_3 \log \frac{2}{\delta''} + \frac{2c_3 b_3}{3}} \right) = \delta'',$$

1079 where the last identity is true, if we set $b_3 = \left(\frac{1}{3}c_3 + \sqrt{\frac{1}{9}c_3 + 2G_3}\right) \log \frac{2}{\delta''}$. To define G_3 , we need
 1080 to bound $\sum_{j=1}^t \bar{\sigma}_j^2$:

$$\begin{aligned}
 \sum_{j=1}^t \bar{\sigma}_j^2 &= \sum_{j=1}^t \mathbb{E}_{q_j, \xi_j} [(W_j^t)^2] \stackrel{(32)}{\leq} \sum_{j=1}^t (1-\alpha)^{2(t-j+1)} \mathbb{E}_{q_j, \xi_j} [\|\omega_j^u\|^2] \\
 &\stackrel{(25)}{\leq} \sum_{j=1}^t (1-\alpha)^{2(t-j+1)} \cdot 18\lambda_h^{2-p} (6L^p + 3\sigma_h^p) \gamma^p \leq \frac{18\lambda_h^{2-p} (6L^p + 3\sigma_h^p) \gamma^p}{1 - (1-\alpha)^2} \\
 &= \frac{18\lambda_h^{2-p} (6L^p + 3\sigma_h^p) \gamma^p}{\alpha(2-\alpha)} \leq \frac{18\lambda_h^{2-p} (6L^p + 3\sigma_h^p) \gamma^p}{\alpha} = G_3.
 \end{aligned} \tag{33}$$

1081 **Upper bound for ⑥.** The sequence Z_1^t, \dots, Z_t^t forms a martingale difference sequence, since the
 1082 definition of Z_j^t leads to $\mathbb{E}[Z_j^t | Z_{j-1}^t, \dots, Z_1^t] = \mathbb{E}_{q_j, \xi_j}[Z_j^t] = 0$ for all $j \in [t]$. Also, according
 1083 to Lemma 4, we obtain that

$$|Z_j^t| \leq \|(1-\alpha)^{t-j+1} \omega_j^u\|^2 + \mathbb{E}_{q_j, \xi_j} [\|(1-\alpha)^{t-j+1} \omega_j^u\|^2] \leq 4\lambda_h^2 + 4\lambda_h^2 = 8\lambda_h^2, \tag{34}$$

1084 where we define $c_4 \stackrel{\text{def}}{=} 8\lambda_h^2$. Denoting the conditional variance of Z_j^t as $\hat{\sigma}_j^2 \stackrel{\text{def}}{=} \mathbb{E}[(Z_j^t)^2 | Z_{j-1}^t, \dots, Z_1^t] = \mathbb{E}_{q_j, \xi_j}[(Z_j^t)^2]$, we can bound $\hat{\sigma}_j^2$ easily
 1085

$$\begin{aligned}
 \hat{\sigma}_j^2 &\stackrel{(34)}{\leq} 8\lambda_h^2 \cdot \mathbb{E}_{q_j, \xi_j} [\|(1-\alpha)^{t-j+1} \omega_j^u\|^2 - \mathbb{E}_{q_j, \xi_j} [\|(1-\alpha)^{t-j+1} \omega_j^u\|^2]] \\
 &\leq 16\lambda_h^2 \mathbb{E}_{q_j, \xi_j} [\|(1-\alpha)^{t-j+1} \omega_j^u\|^2].
 \end{aligned} \tag{35}$$

1086 Then, Lemma 3 implies

$$\mathbb{P} \left\{ |\textcircled{6}| > b_4 \text{ and } \sum_{j=1}^t \hat{\sigma}_j^2 \leq G_4 \log \frac{2}{\delta'} \right\} \leq 2 \exp \left(-\frac{b_4^2}{2G_4 \log \frac{2}{\delta'} + \frac{2c_4 b_4}{3}} \right) = \delta'',$$

1087 where the last identity is true, if we set $b_4 = \left(\frac{1}{3}c_4 + \sqrt{\frac{1}{9}c_4 + 2G_4}\right) \log \frac{2}{\delta''}$. To define G_4 , we need
 1088 to bound $\sum_{j=1}^t \hat{\sigma}_j^2$:

$$\begin{aligned}
 \sum_{j=1}^t \hat{\sigma}_j^2 &\stackrel{(35)}{\leq} 16\lambda_h^2 \sum_{j=1}^t \mathbb{E}_{q_j, \xi_j} [\|(1-\alpha)^{t-j+1} \omega_j^u\|^2] = 16\lambda_h^2 \sum_{j=1}^t (1-\alpha)^{2(t-j+1)} \mathbb{E}_{q_j, \xi_j} [\|\omega_j^u\|^2] \\
 &\leq 16\lambda_h^2 \cdot \frac{18\lambda_h^{2-p} (6L^p + 3\sigma_h^p) \gamma^p}{\alpha} = \frac{16 \cdot 18\lambda_h^{4-p} (6L^p + 3\sigma_h^p) \gamma^p}{\alpha} = G_4.
 \end{aligned}$$

1089 **Upper bound for ⑦.** Thanks to the proof of the bound on ④ (see (33)), we have already shown what
 1090 we need: with probability 1

$$\textcircled{7} = \sum_{j=1}^t \mathbb{E}_{\xi_j} [\|(1-\alpha)^{t-j+1} \omega_j^u\|^2] \stackrel{(33)}{\leq} \frac{18\lambda_h^{2-p} (6L^p + 3\sigma_h^p) \gamma^p}{\alpha}.$$

1091 **Upper bound on ⑧.** By definition of ω_j^b , with probability 1 we have

$$\begin{aligned}
 \textcircled{8} &\stackrel{(31)}{=} \left\| \sum_{j=1}^t (1-\alpha)^{t-j+1} \omega_j^b \right\| \leq \sum_{j=1}^t (1-\alpha)^{t-j+1} \|\omega_j^b\| \\
 &\stackrel{(25)}{\leq} 4\lambda_h^{1-p} (6L^p + 3\sigma_h^p) \gamma^p \sum_{j=1}^t (1-\alpha)^{t-j+1} \leq \frac{4\lambda_h^{1-p} (6L^p + 3\sigma_h^p) \gamma^p}{\alpha}.
 \end{aligned}$$

1092 To sum up, we introduce event $E_{\textcircled{5},t}$ as follows

$$E_{\textcircled{5},t} = \left\{ |\textcircled{5}| \leq b_3 \text{ or } \sum_{j=1}^t \bar{\sigma}_j^2 > G_3 \log \frac{2}{\delta''} \right\}, \quad (36)$$

1093 where $c_3 = 2\lambda_h$, $G_3 = \frac{18\lambda_h^{2-p}(6L^p+3\sigma_h^p)\gamma^p}{\alpha}$, $b_3 = \left(\frac{1}{3}c_3 + \sqrt{\frac{1}{9}c_3^2 + 2G_3}\right) \log \frac{2}{\delta''}$. We have shown
 1094 that $\mathbb{P}\{E_{\textcircled{5},t}\} \geq 1 - \delta''$. We adjust the bound on b_3 :

$$\begin{aligned} b_3 &= \left(\frac{1}{3}c_3 + \sqrt{\frac{1}{9}c_3^2 + 2G_3}\right) \log \frac{2}{\delta'} \leq \left(\frac{2}{3}c_3 + \sqrt{2G_3}\right) \log \frac{2}{\delta''} \\ &= \left(\frac{4}{3}\lambda_h + \sqrt{36\frac{\lambda_h^{2-p}(6L^p+3\sigma_h^p)\gamma^p}{\alpha}}\right) \log \frac{2}{\delta''} \\ &= \lambda_h \left(\frac{4}{3} + 6\sqrt{\frac{(6L^p+3\sigma_h^p)\gamma^p}{\alpha\lambda_h^p}}\right) \log \frac{2}{\delta''}. \end{aligned}$$

1095 Also we introduce event $E_{\textcircled{6},t}$ as follows

$$E_{\textcircled{6},t} = \left\{ |\textcircled{6}| \leq b_4 \text{ or } \sum_{j=1}^t \hat{\sigma}_j^2 > G_4 \log \frac{2}{\delta''} \right\}, \quad (37)$$

1096 where $c_4 = 8\lambda_h^2$, $G_4 = \frac{16 \cdot 18\lambda_h^{4-p}(6L^p+3\sigma_h^p)\gamma^p}{\alpha}$, $b_4 = \left(\frac{1}{3}c_4 + \sqrt{\frac{1}{9}c_4^2 + 2G_4}\right) \log \frac{2}{\delta''}$. We have shown
 1097 that $\mathbb{P}\{E_{\textcircled{6},t}\} \geq 1 - \delta''$. We adjust the bound on b_4 :

$$\begin{aligned} b_4 &= \left(\frac{1}{3}c_4 + \sqrt{\frac{1}{9}c_4^2 + 2G_4}\right) \log \frac{2}{\delta'} \leq \left(\frac{2}{3}c_4 + \sqrt{2G_4}\right) \log \frac{2}{\delta''} \\ &= \left(\frac{16}{3}\lambda_h^2 + \sqrt{\frac{16 \cdot 36\lambda_h^{4-p}(6L^p+3\sigma_h^p)\gamma^p}{\alpha}}\right) \log \frac{2}{\delta''} \\ &= \lambda_h^2 \left(\frac{16}{3} + 24\sqrt{\frac{(6L^p+3\sigma_h^p)\gamma^p}{\alpha\lambda_h^p}}\right) \log \frac{2}{\delta''}. \end{aligned}$$

1098 Thus, the event $E_{\textcircled{5},t} \cap E_{\textcircled{6},t}$ implies

$$\begin{aligned} \left\| \sum_{j=1}^t (1-\alpha)^{t-j+1} \omega_j \right\| &\leq \left\| \sum_{j=1}^t (1-\alpha)^{t-j+1} \omega_j^u \right\| + \left\| \sum_{j=1}^t (1-\alpha)^{t-j+1} \omega_j^b \right\| \\ &\leq |\textcircled{5}| + \sqrt{2 \cdot |\textcircled{6}| + 2 \cdot |\textcircled{7}| + |\textcircled{8}|} \\ &\leq \lambda_h \left(\frac{4}{3} + 6\sqrt{\frac{(6L^p+3\sigma_h^p)\gamma^p}{\alpha\lambda_h^p}}\right) \log \frac{2}{\delta''} + 4\lambda_h \cdot \frac{(6L^p+3\sigma_h^p)\gamma^p}{\alpha\lambda_h^p} \\ &\quad + \sqrt{2\lambda_h^2 \left(\frac{16}{3} + 24\sqrt{\frac{(6L^p+3\sigma_h^p)\gamma^p}{\alpha\lambda_h^p}}\right) \log \frac{2}{\delta''} + 36\lambda_h^2 \frac{(6L^p+3\sigma_h^p)\gamma^p}{\alpha\lambda_h^p}}. \end{aligned}$$

1099 Assuming that $\lambda_h \geq \frac{\gamma(L+\sigma_h)}{\alpha^{1/p}}$, we have the event $E_{\textcircled{5},t} \cap E_{\textcircled{6},t}$ implies

$$\begin{aligned} \left\| \sum_{j=1}^t (1-\alpha)^{t-j+1} \omega_j \right\| &\leq \lambda_h \left(\frac{4}{3} + 6\sqrt{6} + 24 + \sqrt{\frac{32}{3} + 48\sqrt{6} + 216}\right) \log \frac{2}{\delta''} \\ &\leq 59\lambda_h \log \frac{2}{\delta''}. \end{aligned}$$

1100

□

1101 **Main Theorem** Now we are ready to state and prove the main results of this section: high-
 1102 probability convergence guarantees for Algorithm 2.

1103 **Theorem 5.** Let Assumptions 1, 2, 3 and 4 hold, and $\Delta_1 = \Delta_0 = F(x_0) - F_*$. Then if the
 1104 parameters of Algorithm 2: stepsize

$$\gamma \leq \min \left\{ \frac{1}{2} \sqrt{\frac{\Delta_1}{LT}}, \frac{\alpha}{12} \sqrt{\frac{\Delta_1}{L}}, \frac{1}{1408\alpha T \log \frac{8T}{\delta}} \sqrt{\frac{\Delta_1}{L}}, \frac{\Delta_1}{352\sigma\alpha^{\frac{p-1}{p}} T \log \frac{8T}{\delta}}, \sqrt{\frac{\Delta_1\alpha^{1/p}}{968(L+\sigma_h)T \log \frac{8T}{\delta}}} \right\}, \quad (38)$$

1105 momentum parameter and clipping levels

$$\alpha = \frac{1}{T^{\frac{p}{2p-1}}}, \quad \lambda = \max \left\{ 4\sqrt{L\Delta_1}, \frac{\sigma}{\alpha^{1/p}} \right\}, \quad \lambda_h = \frac{2\gamma(L+\sigma_h)}{\alpha^{1/p}} \quad (39)$$

1106 for some $T \geq 1$ and $\delta \in (0, 1]$ such that $\log \frac{8T}{\delta} \geq 1$. Then after T iterations of Algorithm 2 the
 1107 iterates with probability at least $1 - \delta$ satisfy

$$\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla F(x_t)\| \leq \frac{2\Delta_1}{\gamma T}. \quad (40)$$

1108 Moreover, if we set (38) as identity, then the iterates generated by Algorithm 2 after T iterations with
 1109 probability at least $1 - \delta$ satisfy

$$\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla F(x_t)\| = \mathcal{O} \left(\frac{\max \left\{ \sqrt{\Delta_1(L+\sigma_h)}, \sigma \right\}}{T^{\frac{p-1}{2p-1}}} \log \frac{T}{\delta} \right), \quad (41)$$

1110 implying that to achieve $\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla F(x_t)\| \leq \varepsilon$ with probability at least $1 - \delta$ Algorithm 2 requires

$$T = \mathcal{O} \left(\left(\frac{\max \left\{ \sqrt{\Delta_1(L+\sigma_h)}, \sigma \right\}}{\varepsilon} \right)^{\frac{2p-1}{p-1}} \log^{\frac{2p-1}{p-1}} \left(\frac{1}{\delta} \left(\frac{\max \left\{ \sqrt{\Delta_1(L+\sigma_h)}, \sigma \right\}}{\varepsilon} \right)^{\frac{2p-1}{p-1}} \right) \right) \quad (42)$$

1111 iterations/samples.

1112 *Proof.* We remind in the proofs of Lemma 11 and Lemma 12 we have shown

$$\begin{aligned} E_{\textcircled{1},t} &\stackrel{(29)}{=} \left\{ |\textcircled{1}| \leq b_1 \text{ or } \sum_{j=1}^t \sigma_j^2 > G_1 \log \frac{2}{\delta'} \right\}, & E_{\textcircled{2},t} &\stackrel{(30)}{=} \left\{ |\textcircled{2}| \leq b_2 \text{ or } \sum_{j=1}^t \tilde{\sigma}_j^2 > G_2 \log \frac{2}{\delta'} \right\}, \\ E_{\textcircled{5},t} &\stackrel{(36)}{=} \left\{ |\textcircled{5}| \leq b_3 \text{ or } \sum_{j=1}^t \bar{\sigma}_j^2 > G_3 \log \frac{2}{\delta''} \right\}, & E_{\textcircled{6},t} &\stackrel{(37)}{=} \left\{ |\textcircled{6}| \leq b_4 \text{ or } \sum_{j=1}^t \hat{\sigma}_j^2 > G_4 \log \frac{2}{\delta''} \right\}, \end{aligned}$$

where

$$\begin{aligned} c_1 &= 2\lambda, \quad G_1 = \frac{18\lambda^{2-p}\sigma^p}{\alpha}, \quad b_1 \leq \lambda \left(\frac{4}{3} + 6\sqrt{\frac{1}{\alpha} \left(\frac{\sigma}{\lambda} \right)^p} \right) \log \frac{2}{\delta'}, \\ c_2 &= 8\lambda^2, \quad G_2 = \frac{16 \cdot 18\lambda^{4-p}\sigma^p}{\alpha}, \quad b_2 \leq \lambda^2 \left(\frac{16}{3} + 24\sqrt{\frac{1}{\alpha} \left(\frac{\sigma}{\lambda} \right)^p} \right) \log \frac{2}{\delta'}, \\ c_3 &= 2\lambda_h, \quad G_3 = \frac{18\lambda_h^{2-p}(6L^p + 3\sigma_h^p)\gamma^p}{\alpha}, \quad b_3 \leq \lambda_h \left(\frac{4}{3} + 6\sqrt{\frac{(6L^p + 3\sigma_h^p)\gamma^p}{\alpha\lambda^p}} \right) \log \frac{2}{\delta''}, \\ c_4 &= 8\lambda_h^2, \quad G_4 = \frac{16 \cdot 18\lambda_h^{4-p}(6L^p + 3\sigma_h^p)\gamma^p}{\alpha}, \quad b_4 \leq \lambda_h^2 \left(\frac{16}{3} + 24\sqrt{\frac{(6L^p + 3\sigma_h^p)\gamma^p}{\alpha\lambda_h^p}} \right) \log \frac{2}{\delta''}. \end{aligned}$$

Moreover, according to Lemma 11 and Lemma 12, we have

$$\mathbb{P}\{E_{\textcircled{1},t}\} \geq 1 - \delta', \quad \mathbb{P}\{E_{\textcircled{2},t}\} \geq 1 - \delta', \quad \mathbb{P}\{E_{\textcircled{5},t}\} \geq 1 - \delta'', \quad \mathbb{P}\{E_{\textcircled{6},t}\} \geq 1 - \delta''.$$

1113 The idea of the proof is based on technique form work of [Gorbunov et al. \[2020\]](#), [Sadiev et al. \[2023\]](#),
 1114 [Liu et al. \[2023b\]](#): via mathematical induction we plan to show that for any $\tau \in \{0, 1, \dots, T-1\}$
 1115 the event

$$G_\tau = E_\tau \cap E_{1,\tau} \cap E_{2,\tau} \cap E_{3,\tau} \cap E_{4,\tau}$$

1116 holds with probability at least $1 - \frac{\tau\delta}{T}$, where event E_τ is defined as

$$E_\tau = \left\{ \gamma \sum_{s=0}^t \|\nabla F(x_s)\| + \Delta_{t+1} \leq 2\Delta_1, \quad \forall t \leq \tau \right\},$$

1117 and

$$E_{1,\tau} = \bigcap_{t=1}^{\tau} E_{\textcircled{1},t}, \quad E_{2,\tau} = \bigcap_{t=1}^{\tau} E_{\textcircled{2},t}, \quad E_{3,\tau} = \bigcap_{t=1}^{\tau} E_{\textcircled{3},t}, \quad E_{4,\tau} = \bigcap_{t=1}^{\tau} E_{\textcircled{4},t}.$$

1118 **Basis of induction.** Obviously, we have $G_0 = E_0 = \{\Delta_0 = \Delta_1 \leq 2\Delta_1\}$ holds with probability
 1119 $1 - \frac{\tau\delta}{T} \stackrel{\tau=0}{=} 1$.

1120 **Step of induction.** Assume that the induction hypothesis is true for $\tau-1$: $\mathbb{P}\{G_{\tau-1}\} \geq 1 - (\tau-1)\delta/T$.
 1121 One needs to prove $\mathbb{P}\{G_\tau\} \geq 1 - \tau\delta/T$. We want to mention that

$$E_{\tau-1} \cap E_{1,\tau} \cap E_{2,\tau} \cap E_{3,\tau} \cap E_{4,\tau} = G_{\tau-1} \cap E_{\textcircled{1},\tau} \cap E_{\textcircled{2},\tau} \cap E_{\textcircled{3},\tau} \cap E_{\textcircled{4},\tau}.$$

1122 Then, we have

$$\begin{aligned} \mathbb{P}\{E_{\tau-1} \cap E_{1,\tau} \cap E_{2,\tau} \cap E_{3,\tau} \cap E_{4,\tau}\} &= \mathbb{P}\{G_{\tau-1} \cap E_{\textcircled{1},\tau} \cap E_{\textcircled{2},\tau} \cap E_{\textcircled{3},\tau} \cap E_{\textcircled{4},\tau}\} \\ &= 1 - \mathbb{P}\{\overline{G_{\tau-1} \cap E_{\textcircled{1},\tau} \cap E_{\textcircled{2},\tau} \cap E_{\textcircled{3},\tau} \cap E_{\textcircled{4},\tau}}\} \\ &\geq 1 - \mathbb{P}\{\overline{G_{\tau-1}}\} - \mathbb{P}\{\overline{E_{\textcircled{1},\tau}}\} - \mathbb{P}\{\overline{E_{\textcircled{2},\tau}}\} \\ &\quad - \mathbb{P}\{\overline{E_{\textcircled{3},\tau}}\} - \mathbb{P}\{\overline{E_{\textcircled{4},\tau}}\} \\ &\geq 1 - \frac{(\tau-1)\delta}{T} - 2\delta' - 2\delta'' \\ &= 1 - \frac{\tau\delta}{T} + \left(\frac{\delta}{T} - 2\delta' - 2\delta''\right) \\ &= 1 - \frac{\tau\delta}{T}, \end{aligned}$$

1123 where we have selected $\delta' = \delta'' = \frac{\delta}{4T}$. The event $E_{\tau-1}$ implies for any $t \leq \tau$

$$\Delta_t \leq \gamma \sum_{s=0}^{t-1} \|\nabla F(x_s)\| + \Delta_t \leq 2\Delta_1.$$

1124 Therefore, we have $\|\nabla F(x_t)\| \leq \sqrt{2L\Delta_t} \leq 2\sqrt{L\Delta_1} \leq \frac{\lambda}{2}$. Assuming $E_{\tau-1} \cap E_{1,\tau} \cap E_{2,\tau} \cap E_{3,\tau} \cap$
 1125 $E_{4,\tau}$ happens, Lemma 10 implies

$$\begin{aligned} \gamma \sum_{t=0}^{\tau} \|\nabla F(x_t)\| + \Delta_{\tau+1} &\leq \Delta_1 + \frac{3\gamma}{\alpha} \sqrt{L\Delta_1} + \frac{\gamma^2 L(\tau+1)}{2} \\ &\quad + 2\gamma\alpha \sum_{t=1}^{\tau} \left\| \sum_{j=1}^t (1-\alpha)^{t-j} \theta_j \right\| + 2\gamma \sum_{t=1}^{\tau} \left\| \sum_{j=1}^t (1-\alpha)^{t-j+1} \omega_j \right\| \\ &\leq \Delta_1 + \frac{3\gamma}{\alpha} \sqrt{L\Delta_1} + \frac{\gamma^2 L(\tau+1)}{2} \\ &\quad + 44\gamma\alpha(\tau+1)\lambda \log \frac{8T}{\delta} + 118\gamma(\tau+1)\lambda_h \log \frac{8T}{\delta}. \end{aligned}$$

1126 By setting clipping levels (39) and stepsize (38), we have

$$\gamma \sum_{t=0}^{\tau} \|\nabla F(x_t)\| + \Delta_{\tau+1} \leq \Delta_1 + \frac{\Delta_1}{4} + \frac{\Delta_1}{4} + \frac{\Delta_1}{4} + \frac{\Delta_1}{4} = 2\Delta_1.$$

1127 Therefore, we obtain

$$\begin{aligned}\mathbb{P}\{G_\tau\} &= \mathbb{P}\{E_\tau \cap E_{1,\tau} \cap E_{2,\tau} \cap E_{3,\tau} \cap E_{4,\tau}\} \\ &= \mathbb{P}\{E_{\tau-1} \cap E_{1,\tau} \cap E_{2,\tau} \cap E_{3,\tau} \cap E_{4,\tau}\} \\ &\geq 1 - \frac{\tau\delta}{T}.\end{aligned}$$

1128 Thus, we have

$$\mathbb{P}\{E_T\} \geq \mathbb{P}\{G_T\} \geq 1 - \delta,$$

1129 or equivalently with probability at least $1 - \delta$

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

1130 Pugging (38) into the inequality above, we have

$$\begin{aligned}\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla F(x_t)\| &\leq \frac{2\Delta_1}{\gamma T} \\ &= \mathcal{O} \left(\max \left\{ \sqrt{\frac{L\Delta_1}{T}}, \frac{\sqrt{L\Delta_1}}{\alpha T}, \alpha \sqrt{L\Delta_1} \log \frac{T}{\delta}, \sigma \alpha^{\frac{p-1}{p}} \log \frac{T}{\delta}, \sqrt{\frac{\Delta_1(L + \sigma_h) \log \frac{T}{\delta}}{T\alpha^{1/p}}} \right\} \right).\end{aligned}$$

1131 Selecting $\alpha = 1/T^{-p/2p-1}$ (see (39)), we obtain

$$\begin{aligned}\frac{1}{T} \sum_{t=1}^T \|\nabla F(x_t)\| &= \mathcal{O} \left(\max \left\{ \sqrt{\frac{L\Delta_1}{T}}, \frac{\sqrt{L\Delta_1}}{T^{\frac{p-1}{2p-1}}}, \frac{\sqrt{L\Delta_1}}{T^{\frac{p-1}{2p-1}}} \log \frac{T}{\delta}, \frac{\sigma}{T^{\frac{p-1}{2p-1}}} \log \frac{T}{\delta}, \sqrt{\frac{\Delta_1(L + \sigma_h) \log \frac{T}{\delta}}{T^{2p-2/2p-1}}} \right\} \right) \\ &= \mathcal{O} \left(\frac{\max \left\{ \sqrt{\Delta_1(L + \sigma_h)}, \sigma \right\}}{T^{\frac{p-1}{2p-1}}} \log \frac{T}{\delta} \right).\end{aligned}$$

1132 **Discussion.** According to Theorem 3, the sample complexity of Algorithm 2 is

$$\tilde{\mathcal{O}} \left(\left(\frac{\max \left\{ \sqrt{\Delta_1(L + \sigma_h)}, \sigma \right\}}{\varepsilon} \right)^{\frac{2p-1}{p-1}} \right),$$

1133 which has the same dependence on ε as lower bounds do (see Theorem 1). However, the obtained
1134 upper bound has worst dependence on parameters $L, \sigma, \sigma_h, \Delta_1$ than lower bounds have. We suppose
1135 that this caused by the analysis technique, since Algorithm 1 has better convergence rate in terms
1136 of parameter dependence. This observation raises an interesting question: Is it possible to conduct
1137 high-probability analysis for Algorithm 1, or equivalently, Algorithm 2 without clipping?

1138 □