From Gradient Clipping to Normalization for Heavy Tailed SGD

Florian Hübler[*]
ETH Zurich

Ilyas Fatkhullin[*]
ETH Zurich

Niao He
ETH Zurich

Abstract

Recent empirical evidence indicates that many machine learning applications involve heavy-tailed gradient noise, which challenges the standard assumptions of bounded variance in stochastic optimization. Gradient clipping has emerged as a popular tool to handle this heavy-tailed noise, as it achieves good performance both theoretically and practically. However, our current theoretical understanding of non-convex gradient clipping has three main shortcomings. First, the theory hinges on large, increasing clipping thresholds, which are in stark contrast to the small constant clipping thresholds employed in practice. Second, clipping thresholds require knowledge of problem-dependent parameters to guarantee convergence. Lastly, even with this knowledge, current sample complexity upper bounds for the method are sub-optimal in nearly all parameters. To address these issues and motivated by practical observations, we make the connection of gradient clipping to its close relative — Normalized `SGD` (`NSGD`) — and study its convergence properties. First, we establish a parameter-free sample complexity for `NSGD` of $\mathcal{O}\left(\varepsilon^{-\frac{2p}{p-1}}\right)$ to find an ε-stationary point, only assuming a finite p-th central moment of the noise, $p \in (1, 2]$. Furthermore, we prove the tightness of this result, by providing a matching algorithm-specific lower bound. In the setting where all problem parameters are known, we show this complexity is improved to $\mathcal{O}\left(\varepsilon^{-\frac{3p-2}{p-1}}\right)$, matching the previously known lower bound for all first-order methods in all problem dependent parameters. Finally, we establish high-probability convergence of `NSGD` with a mild logarithmic dependence on the failure probability. Our work complements the studies of gradient clipping under heavy-tailed noise, improving the sample complexities of existing algorithms and offering an alternative mechanism to achieve high-probability convergence.

1 INTRODUCTION

We study the stochastic optimization problem

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

where $F\colon \mathbb{R}^d \to \mathbb{R}$ is a potentially non-convex, L-smooth objective function, and ξ is a random variable with an unknown distribution $\mathcal{D}$. Such problems are pervasive in machine learning applications, where obtaining exact gradients is often infeasible, necessitating reliance on stochastic gradients (Bottou et al., 2018).

Traditionally, stochastic gradient methods rely on the assumption that the variance of the gradient noise is bounded. Under this assumption, it is well established that first-order algorithms require at least $\Omega\left(\Delta_1 L\varepsilon^{-2} + \Delta_1 L\sigma^2\varepsilon^{-4}\right)$ stochastic gradient oracle queries in the worst case to find an ε-stationary point, i.e., $x \in \mathbb{R}^d$ with $\mathbb{E}\left[\|\nabla F(x)\|\right] \le \varepsilon$ (Arjevani et al., 2023). Here Δ_1 denotes the initialization gap $\Delta_1 \ge F(x_1) - \inf_{x\in\mathbb{R}^d} F(x)$, L the smoothness-parameter and σ^2 the variance. Stochastic Gradient Descent (`SGD`) with an appropriately chosen step-size achieves this optimal sample complexity (Ghadimi and Lan, 2013).

However, new insights in machine learning suggest that the bounded variance (BV) assumption may be overly restrictive. Empirical evidence from fields such as image classification (Simsekli et al., 2019; Battash et al., 2024), training large language models (LLMs) (Zhang et al., 2020; Ahn et al., 2024), and policy optimization in reinforcement learning (RL) (Garg et al., 2021) indicates that stochastic gradients often follow heavy-tailed distributions. These findings challenge

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

[*] These authors contributed equally to this work.

the standard assumption, suggesting a shift towards weaker noise models which only assume boundedness of the p-th central moment of the gradient noise for some $p \in (1, 2]$, i.e.

$$\mathbb{E}\left[\|\nabla f(x,\xi) - \nabla F(x)\|^p\right] \leq \sigma^p \qquad \text{(p-BCM)}$$

with $\sigma = \sigma_p \geq 0$, where p denotes the tail index. Specifically, the aforementioned works verify the tail index of stochastic gradients using statistical tests (e.g., Mohammadi et al. (2015)) and find $p < 2$. Even when the bounded variance assumption holds, the resulting constant $\sigma(2)$ can be prohibitively large compared to $\sigma(p)$ for some $p < 2$.

While SGD achieves the optimal sample complexity under finite variance, empirical evidence suggests that adaptive algorithms become crucial in the presence of heavy tailed noise (Zhang et al., 2020). The vast majority of works[1] which are able to prove convergence under (p-BCM) employ the gradient clipping mechanism (Zhang et al., 2020; Gorbunov et al., 2020; Cutkosky and Mehta, 2021; Sadiev et al., 2023; Gorbunov et al., 2024; Li and Liu, 2023; Nguyen et al., 2023a; Kornilov et al., 2024; Liu et al., 2024b). This mechanism replaces the stochastic gradient in optimization algorithms by its clipped counterpart

$$\widehat{\nabla} f(x_t, \xi_t) = \min\left\{1, \frac{\gamma_t}{\|\nabla f(x_t,\xi_t)\|}\right\} \nabla f(x_t, \xi_t), \quad (2)$$

where $\{\gamma_t\}_{t\geq 1}$ is the sequence of clipping thresholds and $\xi_t \overset{\text{i.i.d.}}{\sim} \mathcal{D}$.

Perhaps, the most popular scheme is Clip-SGD,[2] which updates the iterates as $x_{t+1} = x_t - \eta_t \widehat{\nabla} f(x_t, \xi_t)$, where $\{\eta_t\}_{t\geq 1}$ is a predefined sequence of step-sizes.

1.1 Drawbacks of Gradient Clipping Theory

Despite its popularity in the literature, we want to outline several drawbacks of current clipping theory.

Misalignment between theoretical and practical insights. Existing theoretical analyses of Clip-SGD (and its variants) under the (p-BCM) assumption hinge on using a large, p-dependent sequence of increasing clipping thresholds (e.g., $\gamma_t = \gamma \cdot t^{\frac{1}{3p-2}}$) (Zhang et al., 2020; Cutkosky and Mehta, 2021; Li and Liu, 2023; Nguyen et al., 2023b,a). This choice of clipping thresholds is based on the following two ideas. First, clipping allows one to control the variance of the clipped gradient estimator $\widehat{\nabla} f(x_t, \xi_t)$, even in cases where the original gradient oracle has infinite variance. Second, it ensures that the probability of gradients being clipped decreases over time as γ_t increases, thereby reducing the bias introduced by clipping and facilitating convergence. However, this theoretical recommendation contradicts common practice for clipping in machine learning, where small, constant thresholds (e.g., $\gamma_t \equiv 0.25$) are typically used instead (Merity et al., 2018; Zhang et al., 2022; Touvron et al., 2023; Liu et al., 2024a).

In contrast, one can observe that the clipping thresholds commonly used in practice lead to an *increasing* probability of clipping gradients, eventually resulting in gradients being clipped at every iteration. This observation runs counter to theoretical insights, which suggest clipping is becoming less frequent as training progresses. Specifically, we observe this phenomenon in language modelling tasks in Section 4, and notice the same effect on simpler, synthetic examples in Appendix F. This aggressive clipping behaviour essentially transforms Clip-SGD into a variant of Normalized SGD:

$$x_{t+1} = x_t - \eta_t \frac{g_t}{\|g_t\|}, \qquad \text{(NSGD)}$$

where $g_t = \nabla f(x_t, \xi_t)$ in this case. However, it should be noted that unlike Clip-SGD, NSGD only requires tuning a single parameter η_t, highlighting its simplicity in comparison.

Need for tuning. To our knowledge, all existing convergence results for gradient clipping under (p-BCM) require knowledge of all problem parameters to set the clipping thresholds $\{\gamma_t\}_{t\geq 1}$ and other hyper-parameters of the underlying algorithms. As these problem-dependent parameters are not known in practice, this necessitates an extensive hyper-parameter tuning. In particular, for Clip-SGD, there are 2 hyper-parameters which potentially require tuning. In Appendix F we observe that even in simple scenarios, tuning both parameters may be needed to match the performance of NSGD, which only has 1 parameter. Furthermore, in Section 4 we empirically observe that even while requiring extensive hyper-parameter tuning, Clip-SGD is not able to outperform vanilla NSGD in language modeling tasks.

Suboptimal sample complexities. None of the existing convergence analysis of non-convex Clip-SGD (and its variants) achieve the sample complexity lower

[1] Except, for example, (Wang et al., 2021), which studies convergence of SGD in the strongly convex case under additional p-positive definiteness assumption on the Hessian.

[2] Many variants and modifications of Clip-SGD exist including its combinations with Nesterov's acceleration (Gorbunov et al., 2020), normalization (Cutkosky and Mehta, 2021), zero-order (Kornilov et al., 2024) and coordinate-wise variants (Zhang et al., 2020), but gradient clipping is the key building block of these methods.

bound by Zhang et al. (2020), $\Omega\left(\frac{\Delta_1 L}{\varepsilon^2} + \frac{\Delta_1 L}{\varepsilon^2}\left(\frac{\sigma}{\varepsilon}\right)^{\frac{p}{p-1}}\right)$, in all problem parameters, even when problem parameters are known. In particular, prior to this work, the optimal heavy-tailed sample complexity remained an open question.

1.2 Our Contributions

Our work seeks to remove the drawbacks listed above by diving into the convergence analysis of NSGD with different gradient estimators[3] under heavy tailed noise. We summarize our contributions as follows:

1. We prove in-expectation convergence of NSGD using either mini-batches or momentum under the (p-BCM) assumption for $p \in (1, 2]$ in two settings.

   a) Without any knowledge of problem specific parameters, we show that the algorithms require at most $\mathcal{O}\left(\frac{\Delta_1^4 + L^4}{\varepsilon^4} + \left(\frac{\sigma}{\varepsilon}\right)^{\frac{2p}{p-1}}\right)$ stochastic gradient oracle queries to reach an ε-stationary point in expectation, providing the first parameter-free heavy-tailed convergence guarantee. Furthermore, we construct an algorithm specific lower bound showing that this sample complexity is tight for NSGD with polynomial step-size and batch-size.

   b) When problem parameters are known, we achieve an improved sample complexity of at most $\mathcal{O}\left(\frac{\Delta_1 L}{\varepsilon^2} + \frac{\Delta_1 L}{\varepsilon^2}\left(\frac{\sigma}{\varepsilon}\right)^{\frac{p}{p-1}}\right)$, improving the previously best known heavy-tailed sample complexity. This sample complexity exactly matches the mini-max lower bound in all parameters for the class of first-order algorithms under our assumptions.

To our knowledge, NSGD is the first algorithm which achieves either a) or b) in the heavy tail regime $p < 2$.

2. We provide a high probability convergence guarantee for minibatch-NSGD, removing the need for clipping, thereby extending our understanding of high-probability guarantees under heavy-tailed noise. The sample complexity in this case corresponds to the same complexity as its in-expectation counterpart with a mild multiplicative polylog $(1/\delta)$ factor.

1.3 Related Work

Gradient clipping is widely used to stabilize the training in various fields of machine learning (Pascanu et al., 2013; Schulman et al., 2017; Zhang et al., 2020). Recently a number of works provide convergence guarantees for Clip-SGD and its variants in different settings, e.g., (Nazin et al., 2019; Gorbunov et al., 2020; Davis et al., 2021; Gorbunov et al., 2024; Liu and Zhou, 2023; Puchkin et al., 2024) to name a few. However, the results in the non-convex stochastic setting are relatively scarce. In particular, Zhang et al. (2020) study in-expectation and Sadiev et al. (2023); Nguyen et al. (2023a) investigate high probability convergence of Clip-SGD under (p-BCM). All above mentioned works use increasing (iteration dependent) clipping parameters, e.g., $\gamma_t = \gamma \cdot t^{\frac{1}{3p-2}}$, and derive suboptimal convergence rates, see Section 3.1 for a more detailed discussion. A momentum version of Clip-SGD was analyzed in (Mai and Johansson, 2021) assuming the bounded second moment of stochastic gradients. However, their proof crucially relies on setting the clipping threshold larger than the expected gradient norm. Recently, Koloskova et al. (2023) offer a new analysis of Clip-SGD with constant clipping threshold under BV setting. However, their proof crucially relies on bounded variance and seems challenging to extend to (p-BCM) setting. It is worth mentioning that gradient clipping is also used to tackle heavy tailed noise in bandits and RL literature, e.g., (Bubeck et al., 2013; Cayci and Eryilmaz, 2024) and in online learning (Zhang and Cutkosky, 2022). Moreover, Clip-SGD is the key mechanism to ensure differential privacy (Abadi et al., 2016; Sha et al., 2024).

Normalized SGD was first proposed by Nesterov (1984, 2018) and analyzed in the deterministic convex case. Later the analysis was extended to smooth (Levy, 2017) and stochastic (Hazan et al., 2015) settings. In the non-convex case, Cutkosky and Mehta (2020) show how to remove large mini-batch requirement for NSGD by incorporating Polyak's momentum. Later, Yang et al. (2024) derive a tight lower bound for NSGD without momentum and Hübler et al. (2024) study the parameter agnosticity of momentum NSGD under a relaxed smoothness assumption. In a different line of works, Levy (2016) study the ability of NSGD to escape from saddle points. However, all above mentioned works make strong noise assumptions such as BV. The most closely related to our work are (Cutkosky and Mehta, 2021; Liu et al., 2023), which study variants of NSGD under heavy tailed noise. Unfortunately, these works use both normalization and gradient clipping with increasing clipping parameter, which necessitates tuning their clipping thresholds. Moreover, Cutkosky and Mehta (2021) assume bounded non-central moment assumption, i.e., $\mathbb{E}\left[\|\nabla f(x,\xi)\|^p\right] \leq G^p$, which is stronger than our (p-BCM). This assumption is relaxed in (Liu et al., 2023) to (p-BCM) at the cost of imposing an additional (almost sure) individual smoothness assumption for each $f(x,\xi)$. Another line of work assumes that the noise

[3] Our results in the main body are stated for minibatch-NSGD, the corresponding results for NSGD with momentum can be found in Appendix D.

distribution has a probability density function (PDF) that is symmetric and strictly positive in a neighborhood of zero (Polyak and Tsypkin, 1979; Jakovetić et al., 2023; Armacki et al., 2023). Under this assumption, they study SGD type methods with general nonlinearities, which include gradient clipping and normalization as a special case. Compared to these works, we work with a different (p-BCM) assumption.

More recently, the role of normalization was investigated for sharpness aware minimization (Dai et al., 2024), and the variants of `NSGD` showed an impressive empirical and theoretical success in more structured non-convex problems in RL (Fatkhullin et al., 2023; Barakat et al., 2023; Ganesh et al., 2024). However, these works are also restricted to benign BV noise assumption. Some recent works also make connections with `SignSGD` algorithm (Bernstein et al., 2018; Karimireddy et al., 2019; Crawshaw et al., 2022), which applies a coordinate-wise normalization. Indeed, the convergence analysis of `SignSGD` and `NSGD` are closely related and our techniques can be extended to its sign variants (Liu et al., 2019; Sun et al., 2023).

In a concurrent work, Liu and Zhou (2024) also derive similar in-expectation upper bounds for `NSGD-M` under heavy tailed noise, albeit with more general (σ, σ_1)-affine variant of (p-BCM) and (L, L_1)-smoothness. In comparison to their work, we additionally study tightness of our rates designing a non-trivial algorithm-specific lower bound, establish high probability convergence of `NSGD` and provide insights on the convergence measure of `NSGD`. While Liu and Zhou (2024) also note the parameter-free convergence of `NSGD-M`, their result is parameter-free only in the special case $\sigma_1 = L_1 = 0$, the setting which recovers our assumptions.

2 PRELIMINARIES

Let us introduce basic notations, definitions and assumptions needed in the upcoming analysis.

Notation. We adopt the common conventions $\mathbb{N} = \{0, 1, \dots\}$, $[n] = \{1, 2, \dots, n\}$ and that empty sums and products are given by their corresponding neutral element. Throughout this paper, $d \in \mathbb{N}_{\geq 1}$ denotes the dimension of the variable to be optimized, $F \colon \mathbb{R}^d \to \mathbb{R}$ the objective and $\nabla f(\cdot, \cdot)$ the stochastic gradient oracle. Unless stated otherwise, $L \geq 0$ denotes the L-smoothness parameter, and $\eta_t > 0$ is the stepsize. We use the standard $\mathcal{O}(\cdot), \Omega(\cdot), \omega(\cdot)$ complexity notations (Howell, 2008), $\widetilde{\mathcal{O}}(\cdot)$ additionally hides poly-logarithmic factors.

Problem Setup. Since solving $\min_{x \in \mathbb{R}^d} F(x)$ to global optimality is computationally intractable (Nemirovskij and Yudin, 1983), our goal instead is to find an ε-stationary point, i.e., $x \in \mathbb{R}^d$ such that $\|\nabla F(x)\| \leq \varepsilon$ in expectation or with high probability. Furthermore, we assume the access to first order information is limited to a (potentially noisy) gradient oracle, $\nabla f(\cdot, \xi)$ of ∇F, where ξ is a random variable. The sample complexity is defined as the number of calls the algorithm makes to this oracle to find an ε-stationary point.

Throughout the paper we work under the following standard assumptions.

Assumption 1 (Lower Boundedness)**.** *The objective function F is lower bounded by $F^* > -\infty$ and we denote $\Delta_1 \geq F(x_1) - F^*$, an upper bound on the initialization gap.*

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

Instead of the classical bounded variance assumption, we adopt the weaker concept of the bounded p-th central moment, as discussed in the introduction.

Assumption 3 (p-BCM)**.** *The gradient oracle is unbiased and has a finite p-th central moment, i.e. there exists $\sigma_p \geq 0$ such that, for all $x \in \mathbb{R}^d$,*

i) $\mathbb{E}\left[\nabla f(x, \xi)\right] = \nabla F(x)$, *and*

ii) $\mathbb{E}\left[\|\nabla f(x, \xi) - \nabla F(x)\|^p\right] \leq \sigma_p^p$.

In this work, we focus on the case $p \in (1, 2]$. It is worth noting that, by Jensen's inequality, any oracle satisfying (p-BCM) also satisfies the assumption for all $p' \leq p$, with $\sigma_{p'} \leq \sigma_p$. Notably, ($p$-BCM) is weaker than the bounded variance assumption, and it is possible for $\sigma_{p'}$ to be much smaller than σ_p. We will omit the subscript throughout the work to improve readability, though the dependence of σ on p remains important to keep in mind.

3 MAIN RESULTS

In this section, we present our convergence results for normalized stochastic gradient methods under the (p-BCM) assumption. In order to guarantee a consistent presentation, we will present the results for `minibatch-NSGD`,[4] i.e., `NSGD` with the mini-batch gradient estimator

$$g_t = \frac{1}{B_t} \sum_{j=1}^{B_t} \nabla f\Big(x_t, \xi_t^{(j)}\Big), \tag{3}$$

[4] We furthermore present the results for known horizon (T-dependent) parameters. Note that all convergence guarantees also hold for decaying (t-dependent) parameters at the mild cost of a multiplicative $\log(T)$ term.

where $\xi_t^{(1)}, \dots, \xi_t^{(B_t)}$ are independent copies of ξ_t. All results presented in this section (except for Corollary 5) are also derived for `NSGD` with momentum, i.e., `NSGD` with the momentum gradient estimator

$$\begin{aligned} g_1 &= \nabla f(x_1, \xi_1), \\ g_t &= \beta_t g_{t-1} + (1-\beta_t)\nabla f(x_t, \xi_t) \end{aligned}$$

for a momentum parameter $(\beta_t)_{t\geq 1}$, and are presented in Appendix D. Furthermore, for vanilla `NSGD` (i.e. using $g_t = \nabla f(x_t, \xi_t)$), the results imply convergence to a σ-neighbourhood, in line with corresponding algorithm specific lower bound (Yang et al., 2024, Theorem 3).

In Section 3.1 we first examine the convergence of `minibatch-NSGD` for unknown problem-parameters, providing a parameter-free convergence guarantee under the (p-BCM) assumption. Afterwards, we examine the performance for optimally tuned parameters. In Section 3.2, we derive a high-probability convergence result for `minibatch-NSGD`. Finally, in Section 3.3, we examine the importance of different convergence measures for our analysis.

3.1 Normalized `SGD` can Handle Heavy Tailed Noise

We first theoretically confirm the robustness of `minibatch-NSGD`, by providing a parameter-free convergence guarantee. This is in stark contrast to current `Clip-SGD` analyses, which hinge on the knowledge of all parameters.

Proposition 1 (Simplified)**.** *Assume (Lower Boundedness), (L-smoothness) and (p-BCM) with* $p \in (1, 2]$*. Let* $q \geq 0, \eta, B > 0$ *and* $r \in (0, 1)$*. Then the iterates generated by* `minibatch-NSGD` *with parameters* $\eta_t \equiv \eta T^{-r}$ *and* $B_t \equiv \lceil \max\{1, BT^q\} \rceil$ *satisfy*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\|\nabla_t\|\right] \leq \frac{\Delta_1}{\eta T^{1-r}} + \frac{\eta L}{2T^r} + \frac{4\sigma}{\max\{1, BT^q\}^{\frac{p-1}{p}}},$$

where $\nabla_t := \nabla F(x_t)$*. The sample complexity is bounded by* $\mathcal{O}\left(\left(\frac{\Delta_1}{\varepsilon}\right)^{\frac{1+q}{1-r}} + \left(\frac{L}{\varepsilon}\right)^{\frac{1+q}{r}} + \left(\frac{\sigma}{\varepsilon}\right)^{\frac{p(1+q)}{q(p-1)}} \right)$.

This result characterises the sample complexity of `minibatch-NSGD` under the (p-BCM) assumption for different orders of step-sizes and batch-sizes. Recall that any stochastic gradient oracle satisfying (p-BCM), also satisfies the assumption for all $p' \leq p$ with $\sigma_{p'} \leq \sigma_p$. In particular, it is possible that $\sigma_{p'} \ll \sigma_p$ and applying Proposition 1 with p' may yield a smaller sample complexity. Hence the result also implies a potentially better sample complexity bound for a specific oracle by taking the infimum over all $p' \in (1, p]$ of our result.

The proof of Proposition 1 can be found in Appendix C.1.1 and follows a similar structure to the case when $p = 2$, though it demands additional attention to the noise term. Notably, we employ a vectorized version of the von Bahr and Esseen inequality (von Bahr and Esseen, 1965) (see Lemma 10 in Appendix B), which provides a more general foundation compared to the ad-hoc approach using additional gradient clipping Cutkosky and Mehta (2021), who analyzed `NSGD` with momentum and gradient clipping. In the special case of $p = 2$, our Proposition 1 can recover the previous rates for `NSGD` in (Cutkosky and Mehta, 2020). The extended result, including the dependence on η and B, can be found in Equations (15) and (16).

Tightness of Proposition 1. While lower bounds on the sample complexity for general first-order algorithms are well-established (Arjevani et al., 2023; Zhang et al., 2021), there are no algorithm-specific lower bounds specifying the optimal oracle complexity of `minibatch-NSGD` with general parameters. As a consequence, it is unclear whether the parameter-dependence — in particular the dependence on r and q — in Proposition 1 is tight. To address this, we establish an algorithm-specific lower bound, demonstrating that Proposition 1 is indeed tight in all parameters.

Theorem 2 (Simplified)**.** *Under the setting of Proposition 1, there exists a function* F *that satisfies (Lower Boundedness), (L-smoothness), and an oracle* $\nabla f(\cdot, \cdot)$ *that satisfies (p-BCM) such that* `minibatch-NSGD` *with parameters* $\eta_t \equiv \eta T^{-r}$ *and* $B_t \equiv \lceil \max\{1, BT^q\} \rceil$ *requires at least*

$$\Omega\left(\left(\frac{\Delta_1}{\varepsilon}\right)^{\frac{1+q}{1-r}} + \left(\frac{L}{\varepsilon}\right)^{\frac{1+q}{r}} + \left(\frac{\sigma}{\varepsilon}\right)^{\frac{p(q+1)}{q(p-1)}} \right)$$

samples to generate an iterate with $\mathbb{E}\left[\|\nabla F(x_t)\|\right] \leq \varepsilon$.

The extended result, which includes the dependence on η and B, and its proof can be found in Appendix E.2. Its proof is based on two key ideas. First, in the deterministic setting, we construct a hard function that exactly satisfies (L-smoothness) and (Lower Boundedness), penalizing excessively small and large step sizes within a single function, see Figure 1. This yields an iteration complexity lower bound of $\Omega\left(\left({}^{\Delta_1}/_{\varepsilon}\right)^{1/(1-r)} + \left({}^{L}/_{\varepsilon}\right)^{1/r} \right)$. Second, we construct an oracle that points in the opposite direction of the true gradient with maximal probability, while adhering to the (p-BCM) assumption. This oracle leads to a lower bound on the required batchsize, which in turn implies an iteration complexity lower bound of $\Omega\left(\left({}^{\sigma}/_{\varepsilon}\right)^{\frac{p}{q(p-1)}} \right)$. Combining these iteration complexity lower bounds with the samples per iteration results in Theorem 2.

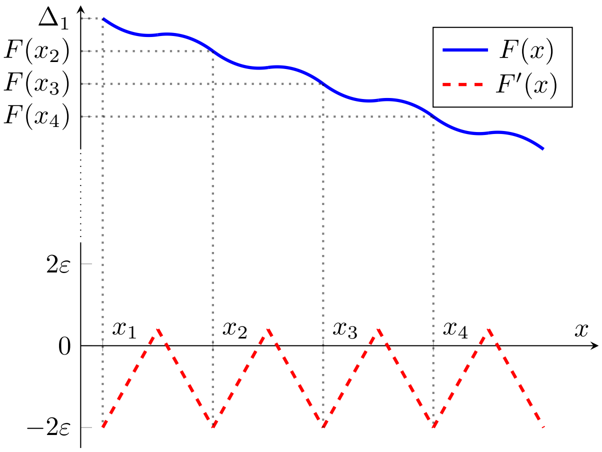


Figure 1: Plot of the hard instance function and its derivative used in Theorem 2. Dotted lines mark the iterates generated by the algorithm and their function values. Δ_1 denotes the initial suboptimality $F(x_1)$, and ε is the target accuracy.

Parameter-free convergence. When considering the parameters $r = 1/2$ and $q = 1$, Proposition 1 implies the sample complexity

$$\mathcal{O}\left(\frac{\Delta_1^4 + L^4}{\varepsilon^4} + \left(\frac{\sigma}{\varepsilon}\right)^{\frac{2p}{p-1}}\right), \tag{4}$$

without requiring knowledge of any problem-dependent parameters, including the tail index p. It turns out, that this choice of step-size and batch-size parameters is the best parameter-free choice possible, in the sense that (4) cannot be *uniformly* improved for all $p \in (1, 2]$. More precisely, (4) is tight for all $p \in (1, 2]$, as can be seen by plugging $r = 1/2$ and $q = 1$ into Theorem 2. Furthermore, while a different choice of r and q may improve the sample complexity for *some* p, the complexity would get strictly worse for $p = 2$. That is any other choice of $(r, q) \neq (1/2, 1)$ implies a sample complexity lower bound of $\omega(\varepsilon^{-4})$, which is strictly worse than the $\mathcal{O}(\varepsilon^{-4})$ we get from (4) for $p = 2$.

Optimal sample complexity with tuning. For algorithms with knowledge of problem parameters, Zhang et al. (2020) provide a sample complexity lower bound for our setting of

$$\Omega\left(\frac{\Delta_1 L}{\varepsilon^2} + \frac{\Delta_1 L}{\varepsilon^2}\left(\frac{\sigma}{\varepsilon}\right)^{\frac{p}{p-1}}\right). \tag{5}$$

To the best of our knowledge, there are no upper bounds exactly matching this lower bound, leaving the tightness of (5) an open question. The following result closes this question, by improving the sample complexity of (4) — when given access to problem-parameters — to tightly match the lower bound in all parameters.

Corollary 3 (Optimal Sample Complexity)**.** *Assume (Lower Boundedness), (L-smoothness) and (p-BCM) with $p \in (1, 2]$. Then the iterates generated by* **`minibatch-NSGD`** *with parameters $\eta_t \equiv \sqrt{\Delta_1/LT}$ and $B_t \equiv \left\lceil \max\left\{1, \left(\frac{\sigma^2 T}{\Delta_1 L}\right)^{\frac{p}{2p-2}}\right\}\right\rceil$ satisfy*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\|\nabla F(x_t)\|\right] \leq 6\frac{\sqrt{\Delta_1 L}}{\sqrt{T}}.$$

In particular the sample complexity is bounded by $\mathcal{O}\left(\frac{\Delta_1 L}{\varepsilon^2} + \frac{\Delta_1 L}{\varepsilon^2}\left(\frac{\sigma}{\varepsilon}\right)^{\frac{p}{p-1}}\right)$.

For comparison, Zhang et al. (2020) derived in-expectation convergence for `Clip-SGD` with a sample complexity of[5]

$$\mathcal{O}\left(\frac{\Delta_1 L \sigma^{\frac{p^2}{p-1}}}{\varepsilon^2} + \frac{(\Delta_1 L \sigma^p)^{\frac{3p-2}{2p-2}} + \sigma^{\frac{3p-2}{p-1}}}{\varepsilon^{\frac{3p-2}{p-1}}}\right)$$

which is suboptimal in all parameters besides ε.

3.2 Convergence with High-Probability

While in-expectation results guarantee small gradient norms given sufficiently many optimization runs, computational constraints often preclude running enough procedures. Therefore, results of the form *with probability at least $1 - \delta$, a single optimization run achieves a certain gradient norm*, often called in-probability results, are more desirable. While the Markov inequality can convert in-expectation guarantees to in-probability guarantees, the poor polynomial dependence on $1/\delta$ renders these results impractical.

Therefore, the gold standard are so called high-probability results with a mild polylog $(1/\delta)$ dependence. To achieve such results, existing literature relies on either light tail noise assumptions (e.g., Ghadimi and Lan (2013); Li and Orabona (2020); Madden et al. (2024); Li and Liu (2022); Fatkhullin and He (2024)), or the gradient clipping mechanism (e.g., Cutkosky and Mehta (2021); Sadiev et al. (2023); Nguyen et al. (2023a)). In contrast, Sadiev et al. (2023) prove that vanilla `SGD` (without clipping) cannot achieve a better δ dependence than $\Omega(1/\sqrt{\delta})$ under heavy-tailed noise.

The following Theorem provides a unified high-probability guarantee for `NSGD` with a general gradient estimator. The result will imply high-probability convergence for `minibatch-NSGD`, and high-probability convergence to a σ-neighbourhood for vanilla `NSGD`.

[5] We ignore non-leading terms and simplify the rate in their favour.

Theorem 4 (High-Probability)**.** *Let $\delta \in (0,1)$. Assume (Lower Boundedness), (L-smoothness) and $\infty > \sigma_t \coloneqq \mathbb{E}\left[\|g_t - \nabla F(x_t)\| \,|\, \mathcal{F}_{t-1}\right]$, where $\mathcal{F}_{t-1} \coloneqq \sigma(g_1, \ldots, g_{t-1})$. Additionally let $\eta_T^{\max} \coloneqq \max_{t \in [T]} \eta_t$ and $C_T \coloneqq \max_{t \in [T]} \eta_t \sum_{\tau=1}^{t-1} \eta_\tau$. Then, with probability at least $1 - \delta$, the iterates generated by* `NSGD` *satisfy*

$$\sum_{t=1}^{T} w_t \|\nabla_t\| \leq \frac{2\Delta_1 + L\sum_{t=1}^{T} \eta_t^2 + 4\sum_{t=1}^{T} \eta_t \sigma_t}{\sum_{\tau=1}^{T} \eta_\tau} + \frac{12(\eta_T^{\max} \|\nabla F(x_1)\| + C_T L) \log(1/\delta)}{\sum_{\tau=1}^{T} \eta_\tau},$$

where $w_t \coloneqq \frac{\eta_t}{\sum_{\tau=1}^{T} \eta_\tau}$ and $\nabla_t \coloneqq \nabla F(x_t)$.

The main idea behind the proof is to reduce the statement to lower bounding the expected cosine between g_t and $\nabla F(x_t)$. Since the cosine is bounded within $[-1, 1]$, we can apply a high-probability concentration inequality on it and obtain the mild $\log(1/\delta)$ dependence. We would like to point out that our proof technique for establishing this high probability result significantly deviates from the existing high probability analysis of methods using gradient clipping. The formal proof can be found in Appendix C.2.

Note that Theorem 4 does not make any unbiasedness or decreasing stepsize assumptions. Furthermore, when comparing this result with its in-expectation counterpart (see Proposition 14), the bound can be interpreted as a concentration inequality around the expected value. Crucially, compared to (Cutkosky and Mehta, 2021), our algorithm does not require the additional clipping mechanism, effectively reducing the need to tune an additional parameter and aligning theory with practice.

We next apply Theorem 4 to `minibatch-NSGD` with optimal parameters. A parameter-free version can be found in Appendix C.2.3. To the best of our knowledge, we are the first work to show a high-probability result without requiring strong noise assumptions or clipping.[6]

Corollary 5. *Assume (Lower Boundedness), (L-smoothness) and (p-BCM) with $p \in (1, 2]$. Then the iterates generated by* `minibatch-NSGD` *with parameters $\eta_t \equiv \sqrt{\frac{\Delta_1}{LT}}$ and $B_t \equiv \left\lceil \max\left\{1, \left(\frac{\sigma^2 T}{\Delta_1 L}\right)^{\frac{p}{2p-2}}\right\}\right\rceil$ satisfy*

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla F(x_t)\| \leq (11 + 30\log(1/\delta))\frac{\sqrt{\Delta_1 L}}{\sqrt{T}}$$

with probability at least $1 - \delta$. This corresponds to a $\widetilde{\mathcal{O}}\left(\frac{\Delta_1 L}{\varepsilon^2} + \frac{\Delta_1 L}{\varepsilon^2}\left(\frac{\sigma}{\varepsilon}\right)^{\frac{p}{p-1}}\right)$ sample complexity.

This result is optimal in Δ_1, L, σ and ε. We are not aware of any lower bounds specifying the optimal δ-dependence. For comparison, Nguyen et al. (2023a) derived high-probability convergence of `Clip-SGD` with a sample complexity of

$$\widetilde{\mathcal{O}}\left(\frac{(\Delta_1 L)^{\frac{3p-2}{4p-4}}}{\varepsilon^{\frac{6p-4}{2p-1}}} + \frac{\left(\frac{\sigma^{2p}}{(\Delta_1 L)^{2-p}}\right)^{\frac{3p-2}{p-1}} + (\Delta_1 L \sigma^2)^{\frac{3p-2}{4p-4}}}{\varepsilon^{\frac{3p-2}{p-1}}}\right),$$

which is suboptimal in all parameters besides ε in the stochastic case. In the deterministic case, even the dependence on ε appears suboptimal. In contrast, our result is noise adaptive in the sense that, for $\sigma = 0$, the optimal deterministic iteration complexity is obtained. In particular, this result closes open questions posed by Liu et al. (2023).

While Theorem 4 can be applied to show high-probability convergence of vanilla `NSGD` to a σ-neighbourhood, technical difficulties prevent us from extending it to `NSGD` with momentum. We investigate this empirically in Section 4, and describe the technical challenges in Appendix D.3.

3.3 Can we Improve the Convergence Measure of Normalized `SGD`?

One may observe that the convergence of `NSGD` above is stated in terms of the average gradient norm, which is different from the average of *squared* gradient norm that is commonly used in non-convex optimization. By Jensen's inequality it is straightforward to see that

$$\sqrt{\frac{1}{T}\sum_{t=1}^{T} \|\nabla F(x_t)\|^2} \geq \frac{1}{T}\sum_{t=1}^{T} \|\nabla F(x_t)\|, \qquad (6)$$

This raises a natural question whether this different convergence measure is a limitation of our analysis or an intrinsic property of the algorithm. The following result shows that this is indeed an intrinsic property of the algorithm, by providing a lower bound on the second moment of the gradient norm.

Theorem 6. *There exists an L-smooth function $F \colon \mathbb{R} \to \mathbb{R}$ such that if* `minibatch-NSGD` *with parameters as in Corollary 3 is run from any initial point $x_1 > 0$ for $T \geq 18$ iterations, then we have*

$$\sqrt{\frac{1}{T}\sum_{t=1}^{T} \|\nabla F(x_t)\|^2} \geq \frac{2}{3}\frac{\sqrt{\Delta_1 L}}{\sqrt{T}} \cdot T^{1/4}, \quad \textit{while}$$

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla F(x_t)\| \leq 6\frac{\sqrt{\Delta_1 L}}{\sqrt{T}} \quad \textit{(by Corollary 3).}$$

The above result implies that even if we select the optimal step-size parameter (to minimize the upper

[6] Note that some works, e.g., (Armacki et al., 2023), provide high-probability guarantees for `NSGD` under different noise assumptions. Specifically, it is important for their technique to assume the existence of a symmetric PDF.

bound on $\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(x_t)\|$), `minibatch-NSGD` does not achieve optimal rates in terms of the stronger measure $\sqrt{\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(x_t)\|^2}$. Specifically, the convergence rate for the latter measure is worse at least by a factor of $\Theta(T^{1/4})$. Our Theorem 20 implies that the step-size $\eta_t = \sqrt{\frac{\Delta_1}{LT}}$ is essentially the only order of step-size to attain the optimal convergence rate in terms of $\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(x_t)\|$ (up to a numerical constant). Combined with inequality (6) and Theorem 6, it implies that there is no predefined constant step-size for `NSGD` that can guarantee optimal convergence when the rate is measured by $\sqrt{\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(x_t)\|^2}$.

4 EXPERIMENTS

In this section, we present experiments designed to empirically motivate and validate the theoretical findings of this paper. Since heavy tails have prominently been observed in language modelling tasks (Zhang et al., 2020), our experiments target this task.

Experimental Setup. We conduct training on the Penn Treebank (PTB) (Marcus et al., 1993) and WikiText-2 (Merity et al., 2017) datasets using the AWD-LSTM architecture (Merity et al., 2018). Hyperparameters of the model and batchsizes were chosen according to (Merity et al., 2018). To observe the exact optimization behaviour of algorithms, the averaging mechanism of the model was disabled. Additional licensing and compute information can be found in Appendix F.

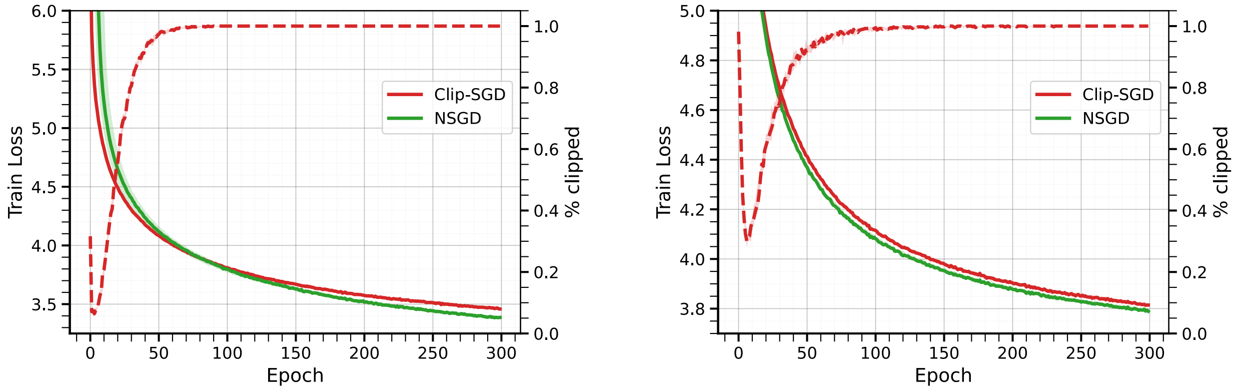
In order to examine the behaviour of `Clip-SGD` and compare it to `NSGD`, we tuned their respective parameters using a course grid search in a 50 epoch training. For `NSGD` we considered the stepsizes $\eta_t = \eta t^{-r}$ and tuned η and r. For `Clip-SGD` we considered the same stepsizes and additionally tuned the clipping threshold γ. The parameters resulting on the above described tuning scheme on the PTB dataset were $(\eta, r, \gamma) = (50, 0.1, 0.25)$ for `Clip-SGD` and $(\eta, r) = (50, 0.25)$ for `NSGD`. It should be noted that the observed optimal clipping threshold $\gamma = 0.25$ is in line with the previous empirical work by Merity et al. (2018) that introduced the AWD-LSTM. The resulting parameters on the WikiText-2 dataset were $(\eta, r, \gamma) = (30, 0, 0.15)$ for `Clip-SGD` and $(\eta, r) = (15, 0.1)$ for `NSGD`. The final training was then carried out for 300 epochs on the seeds $0, 1970, 2000, 2024$ and 2112.

Motivation and Validation. Figure 2 shows the behaviour of `Clip-SGD` and `NSGD` with their corresponding tuned parameters on both datasets. Dashed lines represent the proportion of stochastic gradients that got clipped by `Clip-SGD` per epoch. We want to discuss two observations. First, perhaps surprisingly, we observe on both datasets that the percentage of events when gradients are clipped increases for `Clip-SGD`, contradicting theoretical insights as discussed in Section 1.1. Interestingly, `Clip-SGD` eventually clips at every iteration, becoming equivalent to `NSGD` after a certain number of epochs. Second, it can be noted that both algorithms perform similarly when measured with their corresponding training loss, depicted with solid lines, despite `NSGD` having one less parameter and hence requiring substantially less time to tune.

High probability convergence. In this set of experiments, we verify high probability convergence of `NSGD`, and `NSGD-M` on a strongly convex quadratic function under heavy tailed noise. We run each algorithm for $k = 10^5$ times with default parameters: $\eta_t = 1/\sqrt{t}$ for `NSGD` and $\eta_t = \sqrt{\alpha_t/t}$, $\alpha_t = 1/\sqrt{t}$ for `NSGD-M` see (21) in Appendix D. Figure 3 (left) visualizes the convergence behaviour by selecting the median along with δ and $1-\delta$ quantiles of the algorithm runs based on the average gradient norm at $T = 100$, where $\delta := 1-10^{-4}$. We observe that the δ quantile run of `NSGD-M` deviates significantly from the median compared to that of `NSGD` from its own median. Figure 3 (right) plots the average gradient norm at $T = 100$ corresponding to different values of quantiles δ. `NSGD-M` shows a super-linear dependence on $\log(1/\delta)$, indicating that it may not achieve high probability convergence as effectively as `NSGD` even in the presence of BV noise. This suggests that extending high probability bounds to `NSGD-M` is more challenging due to momentum's impact on the dynamics. We refer to Appendix D.3 for a more detailed discussion on theoretical challenges of showing high-probability bound for `NSGD-M`, and to Appendix F.3 for additional experiments including comparison with light tail noise and `SGD`.

5 CONCLUSION

This work analyzes Normalized `SGD` under heavy-tailed noise. Our theoretical analysis reveals several interesting insights. First, we extend our understanding of high-probability convergence under heavy tailed noise, providing the first such guarantee with an algorithm that does not require gradient clipping. Second, we tightly characterize the optimal sample complexity in all parameters under the (p-BCM) assumption. Lastly, our results for parameter-free `NSGD` suggest the robustness of the algorithm to misspecification of its parameters. Additionally, our algorithm-specific lower bounds in Theorem 2 and Theorem 6 allow additional insights into the behavior of `NSGD`.



(a) PTB

(b) WikiText-2

Figure 2: All plots consider Clip-SGD and NSGD with tuned parameters. Solid lines represent the training loss and correspond to the left y-axis. The dashed line correspond to the right y-axis, and represent the percentage of clipped gradients by Clip-SGD in an epoch. Shaded areas represent the minimal and maximal value within 5 seeds, the line the median.

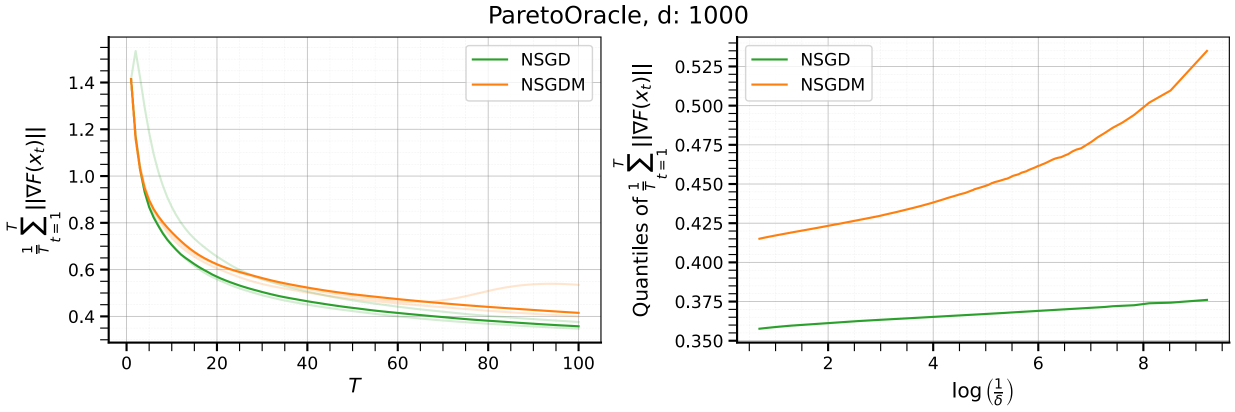


Figure 3: Verifying high probability convergence of NSGD and NSGD-M. We use $f(x,\xi) = \frac{1}{2}\|x\|_2^2 + \langle x, \xi \rangle$, where ξ is a random vector with i.i.d. components drawn from a symmetrized Pareto distribution with tail index $p = 2.5$. For NSGD-M, the quantiles of average gradient norm (left plot) exhibit a super-linear dependence on $\log(1/\delta)$ indicating the lack of high probability convergence.

Several open questions arise from this work for future research. For instance, it remains unclear whether our high-probability result can be extended to NSGD with momentum or variance-reduced gradient estimators. More importantly, it remains open whether the sample complexity that is optimal for parameter-dependent algorithms (5) can be achieved by any algorithm without knowledge of problem parameters.

Acknowledgements

The work is supported by ETH AI Center Doctoral Fellowship and Swiss National Science Foundation (SNSF) Project Funding No. 200021-207343.

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.

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.

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

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.

A. Barakat, I. Fatkhullin, and N. He. Reinforcement learning with general utilities: Simpler variance reduction and large state-action space. In *International Conference on Machine Learning*, pages 1753–1800, 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.

J. Bernstein, Y.-X. Wang, K. Azizzadenesheli, and A. Anandkumar. Signsgd: Compressed optimisation for non-convex problems. In *International Conference on Machine Learning*, pages 560–569. PMLR, 2018.

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

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, Nov 2020. ISSN 1436-4646.

S. Cayci and A. Eryilmaz. Provably robust temporal difference learning for heavy-tailed rewards. *Advances in Neural Information Processing Systems*, 36, 2024.

Y. Cherapanamjeri, N. Tripuraneni, P. Bartlett, and M. Jordan. Optimal mean estimation without a variance. In *Conference on Learning Theory*, pages 356–357. PMLR, 2022.

D. C. Cox. Note on a martingale inequality of pisier. *Mathematical Proceedings of the Cambridge Philosophical Society*, 92(1):163–165, 1982. doi: 10.1017/S0305004100059818.

M. Crawshaw, M. Liu, F. Orabona, W. Zhang, and Z. Zhuang. Robustness to unbounded smoothness of generalized signsgd. *Advances in Neural Information Processing Systems*, 35:9955–9968, 2022.

A. Cutkosky and H. Mehta. Momentum improves normalized SGD. In *International Conference on Machine Learning*, volume 119, 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.

Y. Dai, K. Ahn, and S. Sra. The crucial role of normalization in sharpness-aware minimization. *Advances in Neural Information Processing Systems*, 36, 2024.

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

R. Durrett. *Probability: Theory and Examples*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 5 edition, 2019.

I. Fatkhullin and N. He. Taming nonconvex stochastic mirror descent with general bregman divergence. In *International Conference on Artificial Intelligence and Statistics*, pages 3493–3501, 2024.

I. Fatkhullin, A. Barakat, A. Kireeva, and N. He. Stochastic policy gradient methods: Improved sample complexity for Fisher-non-degenerate policies. In *International Conference on Machine Learning*, pages 9827–9869, 2023.

S. Ganesh, W. U. Mondal, and V. Aggarwal. Variance-reduced policy gradient approaches for infinite horizon average reward markov decision processes. *arXiv preprint arXiv:2404.02108*, 2024.

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

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

I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

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

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

E. Hazan, K. Levy, and S. Shalev-Shwartz. Beyond convexity: Stochastic quasi-convex optimization. *Advances in Neural Information Processing Systems*, 28, 2015.

R. Howell. On Asymptotic Notation with Multiple Variables. *Tech. Rep.*, 2008.

F. Hübler, J. Yang, X. Li, and N. He. Parameter-Agnostic Optimization under Relaxed Smoothness. In *International Conference on Artificial Intelligence and Statistics*, pages 4861–4869, 2024.

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

S. P. Karimireddy, Q. Rebjock, S. Stich, and M. Jaggi. Error feedback fixes SignSGD and other gradient compression schemes. In *International Conference on Machine Learning*, 2019.

A. Koloskova, H. Hendrikx, and S. U. Stich. Revisiting gradient clipping: Stochastic bias and tight convergence guarantees. In *International Conference on Machine Learning*, pages 17343–17363. PMLR, 2023.

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

K. Levy. Online to offline conversions, universality and adaptive minibatch sizes. *Advances in Neural Information Processing Systems*, 30, 2017.

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

S. Li and Y. Liu. High probability guarantees for nonconvex stochastic gradient descent with heavy tails. In *International Conference on Machine Learning*, pages 12931–12963, 2022.

S. Li and Y. Liu. High Probability Analysis for Non-Convex Stochastic Optimization with Clipping. In *ECAI 2023*, pages 1406–1413. IOS Press, 2023.

X. Li and F. Orabona. A high probability analysis of adaptive sgd with momentum. *arXiv preprint arXiv:2007.14294*, 2020.

A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, et al. Deepseek-V3 Technical Report. *arXiv preprint arXiv:2412.19437*, 2024a.

L. Liu, Y. Wang, and L. Zhang. High-Probability Bound for Non-Smooth Non-Convex Stochastic Optimization with Heavy Tails. In *International Conference on Machine Learning*, pages 32122–32138, 2024b.

S. Liu, P.-Y. Chen, X. Chen, and M. Hong. Signsgd via zeroth-order oracle. In *International Conference on Learning Representations*, 2019.

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

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

Z. Liu, J. Zhang, and Z. Zhou. Breaking the Lower Bound with (Little) Structure: Acceleration in Non-Convex Stochastic Optimization with Heavy-Tailed Noise. In *Conference on Learning Theory*, pages 2266–2290, 2023.

L. Madden, E. Dall'Anese, and S. Becker. High probability convergence bounds for non-convex stochastic gradient descent with sub-weibull noise. *Journal of Machine Learning Research*, 25(241):1–36, 2024.

V. V. Mai and M. Johansson. Stability and convergence of stochastic gradient clipping: Beyond lipschitz continuity and smoothness. In *International Conference on Machine Learning*, pages 7325–7335. PMLR, 2021.

M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini. Building a large annotated corpus of English: the penn treebank. *Comput. Linguist.*, 19(2):313–330, jun 1993. ISSN 0891-2017.

S. Merity, C. Xiong, J. Bradbury, and R. Socher. Pointer Sentinel Mixture Models. In *International Conference on Learning Representations*, 2017.

S. Merity, N. S. Keskar, and R. Socher. Regularizing and Optimizing LSTM Language Models. In *International Conference on Learning Representations*, 2018.

M. Mohammadi, A. Mohammadpour, and H. Ogata. On estimating the tail index and the spectral measure of multivariate α-stable distributions. *Metrika*, 78(5):549–561, 2015.

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

A. Nemirovskij and D. Yudin. Problem complexity and method efficiency in optimization. *SIAM Review*, 1983.

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

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

T. D. Nguyen, T. H. Nguyen, A. Ene, and H. Nguyen. Improved convergence in high probability of clipped gradient methods with heavy tailed noise. In *Advances in Neural Information Processing Systems*, volume 36, pages 24191–24222, 2023a.

T. D. Nguyen, T. H. Nguyen, A. Ene, and H. L. Nguyen. High probability convergence of clipped-sgd under heavy-tailed noise. *arXiv preprint arXiv:2302.05437*, 2023b.

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

G. Pisier. Martingales with values in uniformly convex spaces. *Israel Journal of Mathematics*, 20 (3):326–350, Sep 1975. ISSN 1565-8511. doi: 10.1007/BF02760337. URL https://doi.org/10.1007/BF02760337.

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

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

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

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

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

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

T. Sun, Q. Wang, D. Li, and B. Wang. Momentum ensures convergence of signsgd under weaker assumptions. In *International Conference on Machine Learning*, pages 33077–33099. PMLR, 2023.

H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*, 2023.

B. von Bahr and C.-G. Esseen. Inequalities for the r-th absolute moment of a sum of random variables, $1 \leq r \leq 2$. *The Annals of Mathematical Statistics*, pages 299–303, 1965.

H. Wang, M. Gurbuzbalaban, L. Zhu, U. Simsekli, and M. A. Erdogdu. Convergence Rates of Stochastic Gradient Descent under Infinite Noise Variance. In *Advances in Neural Information Processing Systems*, volume 34, pages 18866–18877, 2021.

J. Yang, X. Li, I. Fatkhullin, and N. He. Two sides of one coin: the limits of untuned SGD and the power of adaptive methods. *Advances in Neural Information Processing Systems*, 36, 2024.

J. Zhang and A. Cutkosky. Parameter-free regret in high probability with heavy tails. *Advances in Neural Information Processing Systems*, 35:8000–8012, 2022.

J. Zhang, S. P. Karimireddy, A. Veit, S. Kim, S. Reddi, S. Kumar, and S. Sra. Why are adaptive methods good for attention models? *Advances in Neural Information Processing Systems*, 33:15383–15393, 2020.

J. Zhang, C. Ni, C. Szepesvari, M. Wang, et al. On the convergence and sample efficiency of variance-reduced policy gradient method. *Advances in Neural Information Processing Systems*, 34:2228–2240, 2021.

S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, et al. Opt: Open Pre-trained Transformer Language Models. *arXiv preprint arXiv:2205.01068*, 2022.

S.-Y. Zhao, Y.-P. Xie, and W.-J. Li. On the convergence and improvement of stochastic normalized gradient descent. *Science China Information Sciences*, (3):132103, 2021.

Contents

A SUMMARY OF TECHNICAL CONTRIBUTIONS

In the following we want to summarise the main technical contributions of our work to simplify using the developed technique in other proofs. We group the contributions by topic, so other authors can focus on the area they are interested in.

In-Expectation Upper-Bounds under (p-BCM). With Lemma 10 in combination with Lemma 12 we provide rigorous tools to control the expected deviation of momentum and mini-batch gradient estimators from the true gradient. Examples of applications can be found in (12)-(14) for the mini-batch gradient estimator and in (23)-(24) for the momentum estimator.

In-Expectation Lower-Bounds. We provide a hard function and oracle that is able to tightly characterize the parameter-dependence of NSGD in Lemma 19 and Theorem 21. Very similar constructions can be used to derive a similar result for SGD and possibly other algorithms.

High-Probability Upper-Bounds under (p-BCM). We lay out a different convergence analysis of NSGD that hinges on controlling $\phi_t := \frac{g_t^\top \nabla F(x_t)}{\|g_t\|\|\nabla F(x_t)\|}$ instead of $\|g_t - \nabla F(x_t)\|$, which is used in previous analyses. This allows to prove high-probability guarantees due to the boundedness of ϕ_t. In contrast, Cutkosky and Mehta (2021) require an additionally gradient clipping to guarantee concentration of $\|g_t - \nabla F(x_t)\|$ with high probability.

Difference between Convergence Measures. We construct a function which explicitly showcases that the convergence rate of NSGD deteriorates when different convergence measures are used. Similar constructions could potentially be applied to other algorithms and convergence measures.

B TECHNICAL RESULTS

This section contains various technical results required for our analysis. We start with two lemmas that arise due to the normalization in NSGD. Slightly different formulations of these were used by Zhao et al. (2021).

Lemma 7. *For all $a, b \in \mathbb{R}^d$ with $b \neq 0$ we have*

$$\frac{a^\top b}{\|b\|} \geq \|a\| - 2\|a-b\|.$$

Proof. We calculate

$$\frac{a^\top b}{\|b\|} = \frac{(a-b)^\top b}{\|b\|} + \|b\| \geq -\|a-b\| + \|b\| \geq \|a\| - 2\|a-b\|,$$

where we used Cauchy-Schwarz in the first, and $\|a\| \leq \|a-b\| + \|b\|$ in the second inequality. □

Lemma 8 (Expected Angle Bound)**.** *Let $(\Omega, \mathfrak{A}, \mathbb{P})$ be a probability space and $X\colon \Omega \to \mathbb{R}^d$ a random vector. Furthermore let $\mu \in \mathbb{R}^d \setminus \{0\}, \sigma := \mathbb{E}[\|X-\mu\|]$ and suppose that $X \neq 0$ almost surely. Then it holds that*

$$\mathbb{E}\left[\frac{\mu^\top X}{\|\mu\|\|X\|}\right] \geq 1 - 2\frac{\sigma}{\|\mu\|}.$$

Proof. We apply Lemma 7 with $a \leftarrow \mu$ and $b \leftarrow X$ to derive

$$\frac{\mu^\top X}{\|X\|} \geq \|\mu\| - 2\|\mu - X\|.$$

Dividing both sides by $\|\mu\|$ and taking expectations yields the claim. □

The next lemma shows that t-dependent parameters have the *same* (up to constants) behavior as constant, T-dependent, parameters in NSGD.

Lemma 9 (see Hübler et al. (2024, Lemma 10))**.** *Let $q \in (0,1), r \in [0,1]$ and $T \in \mathbb{N}_{\geq 2}$. Then*

$$\sum_{t=1}^{T} t^{-r} \prod_{\tau=t+1}^{T} \left(1 - \tau^{-q}\right) \leq 2\exp\left(\frac{1}{1-q}\right)(T+1)^{q-r}.$$

To control the error of momentum or mini-batch gradient estimator, we use a von Bahr and Esseen type inequality stated in Lemma 10. This result was initially proved by von Bahr and Esseen (1965) for $d = 1$. Later, it was extended to Banach-Spaces (Pisier, 1975, Proposition 2.4) and optimal constants were derived (Cox, 1982). An alternative proof was rediscovered by Cherapanamjeri et al. (2022) for the one dimensional i.i.d. case and extended to higher dimensions by Kornilov et al. (2024). The extension in the latter work has some inaccuracies, which we fix below.

Lemma 10. *Let $p \in [1,2]$, and $X_1, \ldots, X_n \in \mathbb{R}^d$ be a martingale difference sequence (MDS), i.e., $\mathbb{E}[X_j \mid X_{j-1}, \ldots, X_1] = 0$ a.s. for all $j = 1, \ldots, n$ satisfying*

$$\mathbb{E}[\|X_j\|^p] < \infty \qquad \text{for all } j = 1, \ldots, n.$$

Define $S_n := \sum_{j=1}^{n} X_j$, then

$$\mathbb{E}[\|S_n\|^p] \leq 2\sum_{j=1}^{n} \mathbb{E}[\|X_j\|^p].$$

Proof. The claim is true for $d = 1$, see (von Bahr and Esseen, 1965, Equation (4)). Namely, let $y_1, \ldots, y_n \in \mathbb{R}$ be a MDS, i.e., $\mathbb{E}[y_j \mid y_{j-1}, \ldots, y_1] = 0$ a.s. for all $j = 1, \ldots, n$, satisfying

$$\mathbb{E}[|y_j|^p] < \infty \qquad \text{for all } j = 1, \ldots, n. \tag{7}$$

Then

$$\mathbb{E}\left[\left|\sum_{j=1}^{n} y_j\right|^p\right] \leq 2\sum_{j=1}^{n}\mathbb{E}\left[|y_j|^p\right]. \tag{8}$$

Following (Kornilov et al., 2024), define $g \sim \mathcal{N}(0, I)$ and $y_j := g^\top X_j$, where g is independent of X_j. We need to verify that $y_1, \ldots, y_n$ defined this way is indeed a MDS. Define the sigma algebra $\mathcal{H}_1 := \sigma(y_{j-1}, \ldots, y_1)$ and $\mathcal{H}_2 := \sigma(X_{j-1}, \ldots, X_1, g)$. Observe that $\mathcal{H}_1 \subset \mathcal{H}_2$ and by the tower property

$$\mathbb{E}\left[g^\top X_j|\mathcal{H}_1\right] = \mathbb{E}\left[\mathbb{E}\left[g^\top X_j|\mathcal{H}_2\right]|\mathcal{H}_1\right] = \mathbb{E}\left[g^\top \mathbb{E}\left[X_j|\mathcal{H}_2\right]|\mathcal{H}_1\right] = 0,$$

where the second equality holds because g is $\mathcal{H}_2$-measurable and the last equality holds by independence of X_j and g, and the assumption that $\mathbb{E}\left[X_j \mid X_{j-1}, \ldots, X_1\right] = 0$ a.s. Next, we need to verify that $\mathbb{E}\left[|y_j|^p\right] < \infty$. We know that $g^\top a \sim \mathcal{N}(0, \|a\|^2)$ for any vector $a \in \mathbb{R}^d$. Therefore, $\mathbb{E}\left[|y_j|^p|X_j\right] = \mathbb{E}\left[|g^\top X_j|^p|X_j\right] = C(p)\left\|X_j\right\|^p$, with $C(p) := 2^{p/2}\frac{\Gamma\left(\frac{p+1}{2}\right)}{\sqrt{\pi}}$, where we used the p-th absolute moment of normal distribution applied to a random variable $g^\top X_j$ given X_j. Taking full expectation, we get

$$\mathbb{E}\left[|y_j|^p\right] = C(p)\,\mathbb{E}\left[\|X_j\|^p\right] < \infty. \tag{9}$$

We have verified that the sequence $y_1, \ldots, y_n$ is a MDS and property (7) holds, thus we are ready to apply (8). Using the p-th moment of normal distribution applied to $g^\top S_n$ given S_n, we have

$$C(p)\,\mathbb{E}\left[\|S_n\|^p\right] = \mathbb{E}\left[\mathbb{E}\left[|g^\top S_n|^p|S_n\right]\right] = \mathbb{E}\left[\left|\sum_{j=1}^{n} y_j\right|^p\right] \leq 2\sum_{j=1}^{n}\mathbb{E}\left[|y_j|^p\right],$$

where in the last step we used (8). It remains to use (9) to conclude the proof.

□

We use the following martingale concentration inequality for our high-probability guarantees, see e.g., (Li and Orabona, 2020, Lemma 1).

Lemma 11. *Let $(\mathcal{F}_t)_{t\in\mathbb{N}}$ be a Filtration and $(D_t)_{t\in\mathbb{N}}$ a Martingale Difference Sequence with respect to $(\mathcal{F}_t)_{t\in\mathbb{N}}$. Furthermore, for each $t \in \mathbb{N}_{\geq 1}$, let σ_t be $\mathcal{F}_{t-1}$-measurable and assume that $\mathbb{E}\left[\exp\left(\frac{D_t^2}{\sigma_t^2}\right)\middle|\mathcal{F}_{t-1}\right] \leq e$. Then, for all $T \in \mathbb{N}$,*

$$\forall\, \lambda > 0, \delta \in (0,1)\colon \mathbb{P}\left(\sum_{t=1}^{T} D_t \leq \frac{3}{4}\lambda\sum_{t=1}^{T}\sigma_t^2 + \frac{1}{\lambda}\log\left(\frac{1}{\delta}\right)\right) \geq 1-\delta.$$

In order to apply Lemma 10, we require the following lemma on conditional expectations.

Lemma 12 (c.f. Durrett (2019, Example 4.1.7))**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and X, Y be independent random variables mapping to measurable spaces (E_1, Σ_1) and (E_2, Σ_2) respectively. Furthermore let $h\colon E_1 \times E_2 \to \mathbb{R}^d$ be a (Lebesgue-)measurable function with $\mathbb{E}\left[\|h(X,Y)\|\right] < \infty$. Then*

$$\mathbb{E}\left[h(X,Y)\mid X\right] \overset{a.s.}{=} g(X), \qquad \textit{where} \qquad g(x) := \mathbb{E}\left[h(x,Y)\right].$$

Proof. First note that, by Fubini's Theorem, g is $\Sigma_1/\mathcal{B}^d$ measurable and hence $g(X)$ is $\sigma(X)/\mathcal{B}^d$ measurable. Therefore it suffices to show that

$$\mathbb{E}\left[h(X,Y)1_A\right] = \mathbb{E}\left[g(X)1_A\right]$$

for all $A \in \sigma(X)$. First note that, by definition of $\sigma(X) = \left\{X^{-1}(C)\colon C \in \Sigma_1\right\}$, there exists $B \in \Sigma_1$ with $A = X^{-1}(B)$. Next, by independence of X and Y, their joint induced measure is a product measure $\mu \times \nu$ on $E_1 \times E_2$. Combining, we get

$$\mathbb{E}\left[h(X,Y)1_A\right] = \int_A h(X(\omega), Y(\omega))d\mathbb{P}(\omega) = \int_{E_1\times E_2} h(x,y)1_B(x)d(\mu\times\nu)(x,y).$$

By our assumption $\mathbb{E}\left[\|h(X,Y)\|\right] < \infty$ we know that h is $\mu \times \nu$ integrable and Fubini's Theorem hence yields

$$\int_{E_1 \times E_2} h(x,y) 1_B(x) d(\mu \times \nu)(x,y) = \int_{E_1} \int_{E_2} h(x,y) d\nu(y) 1_B(x) d\mu(x) = \int_{E_1} g(x) 1_B(x) d\mu(x) = \mathbb{E}\left[g(X) 1_A\right].$$

This completes the proof. □

C UPPER-BOUNDS FOR NSGD

This section contains the proofs that are missing in the main part of the paper. Throughout this section we denote the iterates generated by `NSGD` with $(x_t)_{t \in \mathbb{N}}$. Furthermore, we denote the natural filtration of our iterates by $\mathcal{F}_t \coloneqq \sigma(g_1, \ldots, g_t)$.

We start by deriving a descent lemma. While such descent lemmas are well studied for `NSGD` in the literature — to the best of our knowledge — none highlight the importance of the cosine between g_t and $\nabla F(x_t)$. As this term will play a crucial role in our high-probability result, we will provide our version of the descent lemma and its proof below.

Lemma 13 (Descent Lemma)**.** *Assume (Lower Boundedness) and (L-smoothness). Furthermore let*

$$\phi_t \coloneqq \frac{\nabla F(x_t)^\top g_t}{\|\nabla F(x_t)\| \, \|g_t\|}$$

denote the cosine between g_t and $\nabla F(x_t)$. Then the iterates of `NSGD` satisfy

$$\sum_{t=1}^T \eta_t \phi_t \, \|\nabla F(x_t)\| \leq \Delta_1 + \frac{L}{2} \sum_{t=1}^T \eta_t^2.$$

Proof. By the definition of x_{t+1}, (L-smoothness) implies

$$F(x_{t+1}) - F(x_t) \leq -\eta_t \nabla F(x_t)^\top \frac{g_t}{\|g_t\|} + \frac{L}{2}\eta_t^2 = -\eta_t \frac{\nabla F(x_t)^\top g_t}{\|\nabla F(x_t)\| \, \|g_t\|} \|\nabla F(x_t)\| + \frac{L}{2}\eta_t^2.$$

Summing up over $t \in [T]$ and telescoping now yields

$$F^* - F(x_1) \leq F(x_{T+1}) - F(x_1) \leq -\sum_{t=1}^T \eta_t \phi_t \, \|\nabla F(x_t)\| + \frac{L}{2} \sum_{t=1}^T \eta_t^2,$$

where we used (Lower Boundedness) in the first inequality. This completes the proof. □

Thus, if we could guarantee that the angle between the gradient oracle and true gradient remains bounded away from zero, we would be done. Since this can however, even in expectation, not be guaranteed, we need a more detailed analysis to prove our results.

C.1 In-Expectation Upper-Bounds

To prove our in-expectation results, we start with a unified analysis for normalized algorithms. This result does not specify the exact gradient estimator, allowing to incorporate different gradient estimators and noise assumptions afterward. This result was first derived by Cutkosky and Mehta (2020) in a slightly different formulation.

Proposition 14 (c.f. Cutkosky and Mehta (2020, Lemma 2))**.** *Assume (Lower Boundedness), (L-smoothness) and $\infty > \sigma_t \coloneqq \mathbb{E}\left[\|g_t - \nabla F(x_t)\|\right]$. Then the iterates $(x_t)_{t \in \mathbb{N}_{\geq 1}}$ generated by `NSGD` satisfy*

$$\sum_{t=1}^T \frac{\eta_t}{\sum_{\tau=1}^T \eta_\tau} \mathbb{E}\left[\|\nabla F(x_t)\|\right] \leq \frac{\Delta_1 + \frac{L}{2}\sum_{t=1}^T \eta_t^2 + 2\sum_{t=1}^T \eta_t \sigma_t}{\sum_{\tau=1}^T \eta_\tau}.$$

Note that for constant parameters $\eta_t \equiv \eta$ and $\sigma_t \equiv \sigma$ this result reduces to

$$\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[\|\nabla F(x_t)\|\right] \leq \frac{\Delta_1}{\eta T} + \frac{\eta L}{2} + 2\sigma. \tag{10}$$

We provide a slightly different proof of Proposition 14 when compared to (Cutkosky and Mehta, 2020) below.

Proof. Let $\phi_t \coloneqq \frac{\nabla F(x_t)^\top g_t}{\|\nabla F(x_t)\| \|g_t\|}$ denote the cosine between $\nabla F(x_t)$ and g_t. Then, by Lemma 13, we have

$$\sum_{t=1}^{T} \eta_t \phi_t \left\|\nabla F(x_t)\right\| \le \Delta_1 + \frac{L}{2} \sum_{t=1}^{T} \eta_t^2. \tag{11}$$

Next we apply Lemma 7 to get

$$\mathbb{E}\left[\phi_t \left\|\nabla F(x_t)\right\|\right] \ge \mathbb{E}\left[\left\|\nabla F(x_t)\right\| - 2\left\|g_t - \nabla F(x_t)\right\|\right] \ge \mathbb{E}\left[\left\|\nabla F(x_t)\right\|\right] - 2\sigma_t,$$

where we applied our assumption $\sigma_t \ge \mathbb{E}\left[\|g_t - \nabla F(x_t)\|\right]$ in the last inequality. Therefore, by taking expectations in (11), we get

$$\sum_{t=1}^{T} \eta_t \mathbb{E}\left[\|\nabla F(x_t)\|\right] \le \Delta_1 + \frac{L}{2} \sum_{t=1}^{T} \eta_t^2 + 2 \sum_{t=1}^{T} \eta_t \sigma_t.$$

Dividing by $\sum_{\tau=1}^{T} \eta_\tau$ yields the claim. □

C.1.1 Proof of Proposition 1

Now we are ready to prove Proposition 1.

Proof of Proposition 1. To shorten the notation we write $\bar{\eta} \coloneqq \eta T^{-r}$ and $\bar{B} \coloneqq \lceil \max\{1, BT^q\} \rceil$. Remember, that we are considering the mini-batch gradient-estimator

$$g_t = \frac{1}{\bar{B}} \sum_{j=1}^{\bar{B}} \nabla f\left(x_t, \xi_t^{(j)}\right).$$

We start by controlling the (conditional) expected deviation of g_t from $\nabla F(x_t)$ using Lemma 10. Let $x \in \mathbb{R}^d$ and define $X_j(x) \coloneqq \nabla f\left(x, \xi_t^{(j)}\right) - \nabla F(x)$ for all $j \in \left[\bar{B}\right]$. Now note that $X_1(x), \ldots, X_{\bar{B}}(x)$ are independent random variables with mean zero and hence a Martingale Difference Sequence (MDS). Furthermore note that $\mathbb{E}\left[\|X_j(x)\|^p\right] \le \sigma^p$ by (p-BCM) and we can hence apply Lemma 10 to get

$$g(x) \coloneqq \mathbb{E}\left[\left\|\sum_{j=1}^{\bar{B}} X_j(x)\right\|^p\right] \le 2 \sum_{j=1}^{\bar{B}} \mathbb{E}\left[\|X_j(x)\|^p\right] \le 2\bar{B}\sigma^p. \tag{12}$$

Next we calculate

$$\begin{aligned} \mathbb{E}\left[\|g_t - \nabla F(x_t)\| \,|\, x_t\right] &= \mathbb{E}\left[\left\|\frac{1}{\bar{B}} \sum_{j=1}^{\bar{B}} \left(\nabla f(x_t, \xi_t^{(j)}) - \nabla F(x_t)\right)\right\| \,\middle|\, x_t\right] \\ &\le \frac{1}{\bar{B}} \mathbb{E}\left[\left\|\sum_{j=1}^{\bar{B}} \left(\nabla f(x_t, \xi_t^{(j)}) - \nabla F(x_t)\right)\right\|^p \,\middle|\, x_t\right]^{1/p} \end{aligned} \tag{13}$$

where we applied Jensen in the last inequality. Next define

$$Y = \left(\xi_t^{(1)}, \ldots, \xi_t^{(\bar{B})}\right) \quad \text{and} \quad h(x_t, Y) = \left\|\sum_{j=1}^{\bar{B}} \left(\nabla f(x_t, \xi_t^{(j)}) - \nabla F(x_t)\right)\right\|^p.$$

and note that x_t and Y are independent. Hence we may apply Lemma 12 which yields

$$\mathbb{E}\left[\|g_t - \nabla F(x_t)\| \,|\, x_t\right] \overset{(13)}{\le} \frac{1}{\bar{B}} \mathbb{E}\left[h(x_t, Y) \,|\, x_t\right]^{1/p} \overset{\text{Lem. 12}}{=} \frac{1}{\bar{B}} g(x_t)^{1/p} \overset{(12)}{\le} 2\frac{\sigma}{\bar{B}^{\frac{p-1}{p}}} \tag{14}$$

almost surely. By the tower property we get $\mathbb{E}\left[\|g_t - \nabla F(x_t)\|\right] = \mathbb{E}\left[\mathbb{E}\left[\|g_t - \nabla F(x_t)\| \,|\, x_t\right]\right] \leq 2\sigma \bar{B}^{-\frac{p-1}{p}}$ and plugging into (10) yields

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla F(x_t)\|\right] \leq \frac{\Delta_1}{\bar{\eta}T} + \frac{\bar{\eta}L}{2} + \frac{4\sigma}{\bar{B}^{\frac{p-1}{p}}}.$$

Using the definitions of $\bar{\eta}$ and $\bar{B}$ we get

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla F(x_t)\|\right] \leq \frac{\Delta_1}{\eta T^{1-r}} + \frac{\eta L}{2T^r} + \frac{4\sigma}{\lceil \max\{1, BT^q\}\rceil^{\frac{p-1}{p}}} \leq \frac{\Delta_1}{\eta T^{1-r}} + \frac{\eta L}{2T^r} + \frac{4\sigma}{\max\{1, BT^q\}^{\frac{p-1}{p}}} \tag{15}$$

This implies an iteration complexity of

$$\mathcal{O}\left(\left(\frac{\Delta_1}{\eta\varepsilon}\right)^{\frac{1}{1-r}} + \left(\frac{\eta L}{\varepsilon}\right)^{\frac{1}{r}} + \frac{1}{B^{1/q}}\left(\frac{\sigma}{\varepsilon}\right)^{\frac{p}{q(p-1)}}\right)$$

and hence a sample complexity of

$$\mathcal{O}\left(\left(\frac{\Delta_1}{\eta\varepsilon}\right)^{\frac{1}{1-r}} + \left(\frac{\eta L}{\varepsilon}\right)^{\frac{1}{r}} + B\left(\frac{\Delta_1}{\eta\varepsilon}\right)^{\frac{1+q}{1-r}} + B\left(\frac{\eta L}{\varepsilon}\right)^{\frac{1+q}{r}} + \frac{1}{B^{\frac{1}{q}}}\left(\frac{\sigma}{\varepsilon}\right)^{\frac{p(1+q)}{q(p-1)}}\right). \tag{16}$$

This completes the proof. □

C.1.2 Proof of Corollary 3

Finally we provide a slightly more refined analysis to prove Corollary 3.

Proof of Corollary 3. Applying Proposition 1 to our choice of parameters yields

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla F(x_t)\|\right] \leq 2\frac{\sqrt{\Delta_1 L}}{\sqrt{T}} + \frac{4\sigma}{\max\left\{1, \left(\frac{\sigma^2 T}{\Delta_1 L}\right)^{\frac{p}{2p-2}}\right\}^{\frac{p-1}{p}}}.$$

We proceed with a case distinction.
Case 1: $\frac{\sigma^2 T}{\Delta_1 L} \leq 1$**.** In this case we get $\sigma \leq \frac{\sqrt{\Delta_1 L}}{\sqrt{T}}$ and hence

$$4\sigma\left(\max\left\{1, \left(\frac{\sigma^2 T}{\Delta_1 L}\right)^{\frac{p}{2p-2}}\right\}\right)^{-\frac{p-1}{p}} = 4\sigma \leq 4\frac{\sqrt{\Delta_1 L}}{\sqrt{T}}.$$

Case 2: $\frac{\sigma^2 T}{\Delta_1 L} > 1$**.** We calculate

$$4\sigma\left(\max\left\{1, \left(\frac{\sigma^2 T}{\Delta_1 L}\right)^{\frac{p}{2p-2}}\right\}\right)^{-\frac{p-1}{p}} = 4\sigma\left(\frac{\sigma^2 T}{\Delta_1 L}\right)^{-\frac{1}{2}} = 4\frac{\sqrt{\Delta_1 L}}{\sqrt{T}}.$$

This implies an iteration complexity of $\mathcal{O}\big(\Delta_1 L\varepsilon^{-2}\big)$ and hence a sample complexity of

$$\mathcal{O}\left(\Delta_1 L\varepsilon^{-2} \cdot \left\lceil \max\left\{1, \left(\frac{\sigma^2}{\varepsilon^2}\right)^{\frac{p}{2p-2}}\right\}\right\rceil\right) = \mathcal{O}\left(\frac{\Delta_1 L}{\varepsilon^2} + \frac{\Delta_1 L}{\varepsilon^2}\left(\frac{\sigma}{\varepsilon}\right)^{\frac{p}{p-1}}\right).$$

□

C.2 High-Probability Upper-Bounds

This subsection contains the proofs for our high-probability results. We start off with the proof of Theorem 4. The proof hinges on the observation that $\frac{\nabla F(x_t)^\top g_t}{\|\nabla F(x_t)\|\|g_t\|} \in [-1, 1]$ is bounded and hence concentrates well. This will allow us to apply Lemma 11 to get the mild $\log\left({}^1\!/\!_\delta\right)$ dependence.

C.2.1 Proof of Theorem 4

Proof of Theorem 4. Let $\phi_t \coloneqq \frac{\nabla F(x_t)^\top g_t}{\|\nabla F(x_t)\|\|g_t\|}$ denote the cosine between $\nabla F(x_t)$ and g_t. Then, by Lemma 13, we have

$$\sum_{t=1}^{T} \eta_t \phi_t \|\nabla F(x_t)\| \le \Delta_1 + \frac{L}{2}\sum_{t=1}^{T} \eta_t^2.$$

Next, we use the fact that ϕ_t is bounded and hence sharply concentrates around its (conditional) expectation. Formally, let $\psi_t \coloneqq \mathbb{E}[\phi_t \,|\, \mathcal{F}_{t-1}]$ and note that $D_t \coloneqq -\eta_t(\phi_t - \psi_t)\|\nabla F(x_t)\|$ is a martingale difference sequence with respect to $(\mathcal{F}_t)_{t\in\mathbb{N}}$. Furthermore, noting that

$$\exp\left(\frac{D_t^2}{4\eta_t^2\|\nabla F(x_t)\|^2}\right) = \exp\left(\frac{(\phi_t - \psi_t)^2}{4}\right) \le e$$

implies that we may apply Lemma 11 with $\sigma_t^2 = 4\eta_t^2\|\nabla F(x_t)\|^2$ (note the abuse of notation, this σ_t is unrelated to σ_t defined in the statement). Doing so yields, for all $\lambda > 0$,

$$\sum_{t=1}^{T} \eta_t(\psi_t - 3\lambda\eta_t\|\nabla F(x_t)\|)\|\nabla F(x_t)\| \le \Delta_1 + \frac{L}{2}\sum_{t=1}^{T}\eta_t^2 + \frac{1}{\lambda}\log(1/\delta)$$

with probability at least $1-\delta$. Using (L-smoothness) we get $\|\nabla F(x_t)\| \le \|\nabla F(x_1)\| + L\sum_{\tau=1}^{t-1}\eta_\tau$ and hence choosing $\lambda \coloneqq \frac{1}{6\left(\eta_T^{\max}\|\nabla F(x_1)\| + C_T L\right)}$ yields, with probability at least $1-\delta$,

$$\sum_{t=1}^{T} \eta_t\left(\psi_t - \frac{1}{2}\right)\|\nabla F(x_t)\| \le \Delta_1 + \frac{L}{2}\sum_{t=1}^{T}\eta_t^2 + 6(\eta_T^{\max}\|\nabla F(x_1)\| + C_T L)\log(1/\delta). \tag{17}$$

Finally we are left with the challenge of guaranteeing that ψ_t is *large enough*. Therefore we use Lemma 7 to get $\psi_t\|\nabla F(x_t)\| = \mathbb{E}\left[\frac{\nabla F(x_t)^\top g_t}{\|g_t\|} \,\middle|\, \mathcal{F}_{t-1}\right] \le \|\nabla F(x_t)\| - 2\mathbb{E}[\|g_t - \nabla F(x_t)\| \,|\, \mathcal{F}_{t-1}] = \|\nabla F(x_t)\| - 2\sigma_t$ and hence

$$\frac{1}{2}\sum_{t=1}^{T}\eta_t\|\nabla F(x_t)\| \le \Delta_1 + \frac{L}{2}\sum_{t=1}^{T}\eta_t^2 + 2\sum_{t=1}^{T}\eta_t\sigma_t + 6(\eta_T^{\max}\|\nabla F(x_1)\| + C_T L)\log(1/\delta).$$

Dividing by $\frac{1}{2}\sum_{\tau=1}^{T}\eta_\tau$ yields the claim. □

C.2.2 Proof of Corollary 5

Next we apply Theorem 4 to derive the high-probability result for tuned `minibatch-NSGD`.

Proof of Corollary 5. To shorten the notation we write $\eta_t \equiv \eta$ and $B_t \equiv B$. First note that x_t is $\sigma(x_t) \subseteq \mathcal{F}_{t-1}$ measurable and $\xi_t^{(1)}, \ldots, \xi_t^{(B)}$ are independent of $\mathcal{F}_{t-1}$. In particular we have $\mathbb{E}[\|g_t - \nabla F(x_t)\| \,|\, \mathcal{F}_{t-1}] = \mathbb{E}[\|g_t - \nabla F(x_t)\| \,|\, x_t]$ and hence may apply (14) to get

$$\mathbb{E}[\|g_t - \nabla F(x_t)\| \,|\, \mathcal{F}_{t-1}] = \mathbb{E}[\|g_t - \nabla F(x_t)\| \,|\, x_t] \overset{(14)}{\le} \frac{2\sigma}{B^{\frac{p-1}{p}}}, \tag{18}$$

Plugging into Theorem 4 now yields

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(x_t)\| &\le \frac{2\Delta_1}{\eta T} + \eta L + 8\sigma B^{-\frac{p-1}{p}} + 12\left(\frac{\|\nabla F(x_1)\|}{T} + \eta L\right)\log(1/\delta)\\ &= 2\frac{\sqrt{\Delta_1 L}}{\sqrt{T}} + \frac{\sqrt{\Delta_1 L}}{\sqrt{T}} + 8\sigma B^{-\frac{p-1}{p}} + 12\left(\frac{\|\nabla F(x_1)\|}{T} + \frac{\sqrt{\Delta_1 L}}{\sqrt{T}}\right)\log(1/\delta) \qquad (19)\\ &\le (3 + 30\log(1/\delta))\frac{\sqrt{\Delta_1 L}}{\sqrt{T}} + 8\sigma B^{-\frac{p-1}{p}},\end{aligned}$$

where we used $\|\nabla F(x_1)\| \leq \sqrt{2\Delta_1 L}$ in the last inequality. We now proceed with a case distinction.
Case 1: $B = 1$**.** This implies $\sigma \leq \sqrt{\frac{\Delta_1 L}{T}}$ and hence

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(x_t)\| \leq (11 + 30\log(1/\delta))\frac{\sqrt{\Delta_1 L}}{\sqrt{T}}.$$

Case 2: $B = \left(\frac{\sigma^2 T}{\Delta_1 L}\right)^{\frac{p}{2p-2}}$**.** In this case we have $\sigma B^{\frac{1-p}{p}} = \sqrt{\frac{\Delta_1 L}{T}}$ and plugging into (19) yields

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(x_t)\| \leq (11 + 30\log(1/\delta))\frac{\sqrt{\Delta_1 L}}{\sqrt{T}}.$$

This finishes the convergence result. To prove the oracle complexity, note that each iteration requires 1 and $\left(\frac{\sigma^2 T}{\Delta_1 L}\right)^{\frac{p}{2p-2}}$ oracle calls in Case 1 and 2 respectively. To reach an ε-stationary point, $\mathcal{O}\left(\Delta_1 L \varepsilon^{-2}\log(1/\delta)^2\right)$ iterations are required. Plugging into the oracle complexity per iteration yields the second claim. □

C.2.3 Parameter-Free High-Probability

Similar to Proposition 1, we can also derive a parameter-free high-probability result for `minibatch-NSGD`.

Corollary 15. *Assume (Lower Boundedness), (L-smoothness) and (p-BCM) with* $p \in (1, 2]$*. Furthermore let* $\eta, B, q > 0$ *and* $\delta, r \in (0, 1)$*. Then the iterates generated by* `minibatch-NSGD` *with parameters* $\eta_t \equiv \eta T^{-r}$ *and* $B_t \equiv \lceil \max\{1, BT^q\}\rceil$ *satisfy, with probability at least* $1 - \delta$*,*

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(x_t)\| \leq \frac{2\Delta_1}{\eta T^{1-r}} + \frac{\eta L}{T^r}(1 + 12\log(1/\delta)) + 17\frac{\sqrt{\Delta_1 L}}{T}\log(1/\delta) + \frac{8\sigma}{\max\{1, BT^q\}^{\frac{p-1}{p}}}$$

In particular, the sample complexity is bounded by $\widetilde{\mathcal{O}}\left(\left(\frac{\Delta_1}{\varepsilon}\right)^{\frac{1+q}{1-r}} + \left(\frac{L}{\varepsilon}\right)^{\frac{1+q}{r}} + \left(\frac{\sigma}{\varepsilon}\right)^{\frac{p(1+q)}{q(p-1)}}\right)$.

Proof. To shorten the notation we write $\eta_t \equiv \bar{\eta}$ and $B_t \equiv \bar{B}$. First we apply (18) to get $\mathbb{E}\left[\|g_t - \nabla F(x_t)\| \,|\, \mathcal{F}_{t-1}\right] \leq \frac{2\sigma}{B^{\frac{p-1}{p}}}$ and plugging into Theorem 4 yields

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(x_t)\| &\leq \frac{2\Delta_1}{\bar{\eta}T} + \bar{\eta}L + 8\sigma\bar{B}^{-\frac{p-1}{p}} + 12\left(\frac{\|\nabla F(x_1)\|}{T} + \bar{\eta}L\right)\log(1/\delta) \\
&\leq \frac{2\Delta_1}{\eta T^{1-r}} + \frac{\eta L}{T^r}(1 + 12\log(1/\delta)) + \frac{8\sigma}{\lceil \max\{1, BT^q\}\rceil^{\frac{p-1}{p}}} + 17\frac{\sqrt{\Delta_1 L}}{T}\log(1/\delta) \qquad (20) \\
&\leq \frac{2\Delta_1}{\eta T^{1-r}} + \frac{\eta L}{T^r}(1 + 12\log(1/\delta)) + 17\frac{\sqrt{\Delta_1 L}}{T}\log(1/\delta) + \frac{8\sigma}{\max\{1, BT^q\}^{\frac{p-1}{p}}}
\end{aligned}$$

where we used $\|\nabla F(x_1)\| \leq \sqrt{2\Delta_1 L}$ in the second inequality. This corresponds to an iteration complexity of

$$\widetilde{\mathcal{O}}\left(\left(\frac{\Delta_1}{\eta\varepsilon}\right)^{\frac{1}{1-r}} + \left(\frac{\eta L}{\varepsilon}\right)^{\frac{1}{r}} + \frac{1}{B^{\frac{1}{q}}}\left(\frac{\sigma}{\varepsilon}\right)^{\frac{p}{q(p-1)}}\right),$$

where we used $\frac{\sqrt{\Delta_1 L}}{\varepsilon} \leq \left(\frac{\sqrt{\Delta_1}}{\eta\varepsilon}\right)^{\frac{1}{1-r}} + \left(\frac{\eta\sqrt{L}}{\varepsilon}\right)^{\frac{1}{r}} = \mathcal{O}\left(\left(\frac{\Delta_1}{\eta\varepsilon}\right)^{\frac{1}{1-r}} + \left(\frac{\eta L}{\varepsilon}\right)^{\frac{1}{r}}\right)$ by Young's inequality. Finally, this implies a sample complexity of

$$\widetilde{\mathcal{O}}\left(\left(\frac{\Delta_1}{\eta\varepsilon}\right)^{\frac{1}{1-r}} + \left(\frac{\eta L}{\varepsilon}\right)^{\frac{1}{r}} + B\left(\frac{\Delta_1}{\eta\varepsilon}\right)^{\frac{1+q}{1-r}} + B\left(\frac{\eta L}{\varepsilon}\right)^{\frac{1+q}{r}} + \frac{1}{B^{\frac{1}{q}}}\left(\frac{\sigma}{\varepsilon}\right)^{\frac{p(1+q)}{q(p-1)}}\right).$$

□

D UPPER-BOUNDS FOR NSGD WITH MOMENTUM

In this section we discuss the version of our results for NSGD with momentum (`NSGD-M`), i.e., `NSGD` with the gradient estimator

$$g_t = \beta_t g_{t-1} + (1-\beta_t)\nabla f(x_t,\xi_t), \tag{21}$$

where $g_0 = 0$. Throughout this section we use the notation $\alpha_t := 1-\beta_t$ an $\beta_{a:b} := \prod_{\kappa=a}^{b} \beta_\kappa$.

We first derive a deviation bound for g_t from $\nabla F(x_t)$, similar to (14) but for the momentum estimator, generalising the bound in (Cutkosky and Mehta, 2020) to $p<2$.

Lemma 16. *Let $\beta_1 = 0$ and assume (L-smoothness), (p-BCM) with $p \in (1,2]$. Then the iterates generated by `NSGD-M` satisfy*

$$\mathbb{E}\left[\|g_t - \nabla F(x_t)\|\right] \le L\sum_{\tau=2}^{t} \eta_{\tau-1}\beta_{\tau:t} + 2\sigma\left(\sum_{\tau=1}^{t}\left(\beta_{(\tau+1):t}(1-\beta_\tau)\right)^p\right)^{1/p}.$$

Proof. To simplify notation we first define

$$\begin{aligned}\mu_t &:= g_t - \nabla F(x_t),\\ \varepsilon_t &:= \nabla f(x_t,\xi_t) - \nabla F(x_t),\\ S_t &:= \nabla F(x_{t-1}) - \nabla F(x_t).\end{aligned}$$

Now we calculate

$$\begin{aligned} g_t &= \beta_t g_{t-1} + (1-\beta_t)\nabla f(x_t,\xi_t)\\ &= \beta_t(\nabla F(x_{t-1}) + \mu_{t-1}) + (1-\beta_t)(\varepsilon_t + \nabla F(x_t))\\ &= \nabla F(x_t) + (1-\beta_t)\varepsilon_t + \beta_t S_t + \beta_t\mu_{t-1}\end{aligned}$$

and unrolling yields

$$\mu_t = \beta_{2:t}\gamma_1 + \sum_{\tau=2}^{t}\beta_{(\tau+1):t}\alpha_\tau\varepsilon_\tau + \sum_{\tau=2}^{t}\beta_{\tau:t}S_\tau = \sum_{\tau=1}^{t}\beta_{(\tau+1):t}\alpha_\tau\varepsilon_\tau + \sum_{\tau=2}^{t}\beta_{\tau:t}S_\tau,$$

where we used $\beta_1 = 0$ in the second equality. Therefore

$$\mathbb{E}\left[\|\mu_t\|\right] \le \mathbb{E}\left[\left\|\sum_{\tau=1}^{t}\beta_{(\tau+1):t}\alpha_\tau\varepsilon_\tau\right\|\right] + \sum_{\tau=2}^{t}\beta_{\tau:t}\mathbb{E}\left[\|S_\tau\|\right] \le \mathbb{E}\left[\left\|\sum_{\tau=1}^{t}\beta_{(\tau+1):t}\alpha_\tau\varepsilon_\tau\right\|^p\right]^{1/p} + \sum_{\tau=2}^{t}\beta_{\tau:t}\mathbb{E}\left[\|S_\tau\|\right], \tag{22}$$

where we applied Jensen in the second inequality. The second sum can be upper bounded by $L\sum_{\tau=2}^{t}\eta_{\tau-1}\beta_{\tau:t}$. To control the first sum we want to apply Lemma 10.

Therefore, to simplify notation, let $C_\tau := \beta_{(\tau+1):t}\alpha_\tau$ and $X_\tau := C_\tau\varepsilon_\tau$. To check whether $X_1,\dots,X_t$ satisfies the assumptions of Lemma 10, first note that, for all $\tau \in [t]$,

$$\mathbb{E}\left[X_\tau \mid X_1,\dots,X_{\tau-1}\right] = C_\tau\mathbb{E}\left[\nabla f(x_\tau,\xi_\tau) - \nabla F(x_\tau) \mid X_1,\dots,X_{\tau-1}\right] \tag{23}$$

and furthermore, as x_τ is $\sigma(X_1,\dots,X_{\tau-1})$ measurable and ξ_τ independent of $X_1,\dots,X_{\tau-1}$, we have

$$\mathbb{E}\left[\nabla f(x_\tau,\xi_\tau) - \nabla F(x_\tau) \mid X_1,\dots,X_{\tau-1}\right] = \mathbb{E}\left[\nabla f(x_\tau,\xi_\tau) - \nabla F(x_\tau) \mid x_\tau\right] = 0,$$

where we applied our unbiasedness assumption in conjunction with Lemma 12 in the last equality. By a similar argument, using (p-BCM), we get

$$\mathbb{E}\left[\|X_\tau\|^p\right] = C_\tau^p\mathbb{E}\left[\mathbb{E}\left[\|\nabla f(x_\tau,\xi_\tau) - \nabla F(x_\tau)\|^p \mid x_\tau\right]\right] \le C_\tau^p\sigma^p < \infty.$$

Hence we may apply Lemma 10 to get

$$\mathbb{E}\left[\left\|\sum_{\tau=1}^{t}\beta_{(\tau+1):t}\alpha_\tau\varepsilon_\tau\right\|^p\right]^{1/p} \leq \left(2\sum_{\tau=1}^{t}C_\tau^p\sigma^p\right)^{1/p} \leq 2\sigma\left(\sum_{\tau=1}^{t}\left(\beta_{(\tau+1):t}\alpha_\tau\right)^p\right)^{1/p} \tag{24}$$

Combining these bounds with (22) yields

$$\mathbb{E}\left[\|\mu_t\|\right] \overset{(22)}{\leq} \mathbb{E}\left[\left\|\sum_{\tau=1}^{t}\beta_{(\tau+1):t}\alpha_\tau\varepsilon_\tau\right\|^p\right]^{1/p} + \sum_{\tau=2}^{t}\beta_{\tau:t}\mathbb{E}\left[\|S_\tau\|\right] \leq 2\sigma\left(\sum_{\tau=1}^{t}\left(\beta_{(\tau+1):t}\alpha_\tau\right)^p\right)^{1/p} + L\sum_{\tau=2}^{t}\eta_{\tau-1}\beta_{\tau:t}$$

and hence the claim. □

D.1 Parameter-Free

Next we derive the `NSGD-M` counterpart to the parameter-free result (4). Additionally, the result is phrased for decreasing stepsizes, outlining how results can be extended to those.

Corollary 17 (Parameter-Agnostic Convergence)**.** *Let $T \geq 3$ and assume (Lower Boundedness), (L-smoothness) and (p-BCM) with $p \in (1,2]$. Then the iterates generated by* `NSGD` *with $g_t = \beta_t g_{t-1} + (1-\beta_t)\nabla f(x_t, \xi_t)$ and parameters $\beta_t = 1 - t^{-1/2}$ and $\eta_t = \eta t^{-3/4}$ satisfy*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla F(x_t)\|\right] \leq \frac{\frac{\Delta_1}{\eta} + 120\eta L\log(T) + 120\sigma\frac{4p}{2-p}\left(T^{\frac{2-p}{4p}} - 1\right)}{T^{\frac{1}{4}}}.$$

In particular, this corresponds to a rate of convergence of $\widetilde{\mathcal{O}}\Big((\Delta_1 + L)T^{-1/4} + \sigma T^{-\frac{p-1}{2p}}\Big)$ and hence a sample complexity of $\widetilde{\mathcal{O}}\Big(\frac{\Delta_1^4 + L^4}{\varepsilon^4} + \left(\frac{\sigma}{\varepsilon}\right)^{\frac{2p}{p-1}}\Big)$.

The proof follows similar steps as in (Cutkosky and Mehta, 2020), but requires additional attention to the noise term to handle the case $p < 2$.

Proof. To shorten notation we define $r := 3/4, q := 1/2$, and hence $\eta_t = \eta t^{-r}, \beta_t = 1 - t^{-q}$. Furthermore let $\sigma_t := \mathbb{E}\left[\|g_t - \nabla F(x_t)\|\right]$. From Proposition 14 we get

$$\begin{aligned}\sum_{t=1}^{T}\frac{\eta_t}{\sum_{\tau=1}^{T}\eta_\tau}\mathbb{E}\left[\|\nabla F(x_t)\|\right] &\leq \left(\sum_{t=1}^{T}\eta_t\right)^{-1}\left(\Delta_1 + \frac{L}{2}\sum_{t=1}^{T}\eta_t^2 + 2\sum_{t=1}^{T}\eta_t\sigma_t\right)\\ &\leq T^{r-1}\left(\frac{\Delta_1}{\eta} + \frac{3}{2}\eta L + 2\sum_{t=1}^{T}t^{-r}\sigma_t\right),\end{aligned} \tag{25}$$

where we used $\sum_{t=1}^{T}\eta_t \geq \eta T^{1-r}$ and $\sum_{t=1}^{T}\eta_t^2 \leq 3\eta^2$ in the second inequality. To control the third term, we apply Lemma 16 and Lemma 9 to get

$$\begin{aligned}\sum_{t=1}^{T}t^{-r}\sigma_t &\leq 4\exp\left(\frac{1}{1-q}\right)\sum_{t=1}^{T}\left(\sigma t^{-r-q\frac{p-1}{p}} + \eta L t^{-2r+q}\right)\\ &= 4e^2\sum_{t=1}^{T}\left(\sigma t^{-\frac{5p-2}{4p}} + \eta L t^{-1}\right).\\ &\leq 4e^2\left(\sigma\sum_{t=1}^{T}t^{-\frac{5p-2}{4p}} + \eta L(1 + \log(T))\right).\end{aligned}$$

In order to bound $\sum_{t=1}^{T}t^{-\frac{5p-2}{4p}}$ we note that $\frac{5p-2}{4p} = 1$ iff $p = 2$ and hence

$$\sum_{t=1}^{T}t^{-\frac{5p-2}{4p}} \leq 1 + \int_1^T t^{-\frac{5p-2}{4p}}dt \leq \begin{cases}1 + \log(T), & \text{if } p = 2\\ 1 + \frac{1}{1-\frac{5p-2}{4p}}\left(T^{1-\frac{5p-2}{4p}} - 1\right), & \text{otherwise.}\end{cases}$$

Now note that, due to L'Hôspital, $\lim_{q\to 1}\frac{1}{1-q}\left(T^{1-q}-1\right)=\log(T)$ and hence we can unify the cases by writing the second expression and using continuous extensions. Plugging into (25) yields

$$\begin{aligned}\sum_{t=1}^{T}\frac{\eta_t}{\sum_{\tau=1}^{T}\eta_\tau}\mathbb{E}\left[\|\nabla F(x_t)\|\right] &\le T^{r-1}\left(\frac{\Delta_1}{\eta}+8e^2\eta L(1+\log(T))+8e^2\sigma\left(1+\frac{4p}{2-p}\left(T^{\frac{2-p}{4p}}-1\right)\right)\right)\\ &\le T^{-1/4}\left(\frac{\Delta_1}{\eta}+120\eta L\log(T)+120\sigma\frac{4p}{2-p}\left(T^{\frac{2-p}{4p}}-1\right)\right),\end{aligned}$$

where we used that $\frac{4p}{2-p}\left(T^{\frac{2-p}{4p}}-1\right)\ge 1$ for $T\ge 3$ in the last inequality. The other statements follow from the observation $\lim_{q\to 1}\frac{1}{1-q}\left(T^{1-q}-1\right)=\log(T)$. □

D.2 Optimal Sample Complexity

Finally we provide the `NSGD-M` version of Corollary 3.

Corollary 18 (Optimal Oracle Complexity)**.** *Assume (Lower Boundedness), (L-smoothness) and (p-BCM) with $p\in(1,2]$. Then the iterates generated by* `NSGD-M` *with parameters $\beta_1:=0,\beta_t\equiv\beta:=1-\min\left\{1,\left(\frac{\Delta_1 L}{\sigma^2 T}\right)^{\frac{p}{3p-2}}\right\}$ for $t\ge 2$ and $\eta_t\equiv\sqrt{\frac{\Delta_1(1-\beta)}{LT}}$ satisfy*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla F(x_t)\|\right]\le 6\frac{\sqrt{\Delta_1 L}}{\sqrt{T}}+6\left(\frac{\Delta_1 L\sigma^{\frac{p}{p-1}}}{T}\right)^{\frac{p-1}{3p-2}}.$$

In particular, this corresponds to an oracle complexity of $\mathcal{O}\left(\frac{\Delta_1 L}{\varepsilon^2}+\frac{\Delta_1 L}{\varepsilon^2}\left(\frac{\sigma}{\varepsilon}\right)^{\frac{p}{p-1}}\right)$.

Proof. To shorten the notation we write $\eta_t\equiv\eta,\beta_t\equiv\beta$ and $\alpha:=1-\beta$. Combining (10) with Lemma 16 yields

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla F(x_t)\|\right] &\le \frac{\Delta_1}{\eta T}+\frac{\eta L}{2}+2\sigma\alpha^{\frac{p-1}{p}}+\frac{2L\eta}{\alpha}\\ &=\sqrt{\frac{\Delta_1 L}{\alpha T}}+\frac{\sqrt{\Delta_1 L\alpha}}{2\sqrt{T}}+2\sigma\alpha^{\frac{p-1}{p}}+2\sqrt{\frac{\Delta_1 L}{\alpha T}} \qquad (26)\\ &\le 4\sqrt{\frac{\Delta_1 L}{\alpha T}}+2\sigma\alpha^{\frac{p-1}{p}}.\end{aligned}$$

Case 1: $\alpha=1$**.** This implies $\sigma\le\sqrt{\frac{\Delta_1 L}{T}}$ and hence

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla F(x_t)\|\right]\le 6\sqrt{\frac{\Delta_1 L}{T}}.$$

Case 2: $\alpha=\left(\frac{\Delta_1 L}{\sigma^2 T}\right)^{\frac{p}{3p-2}}$**.** Plugging into (26) yields

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla F(x_t)\|\right]\le 4\sigma^{\frac{p}{3p-2}}\left(\frac{\Delta_1 L}{T}\right)^{\frac{p-1}{3p-2}}+2\sigma^{\frac{p}{3p-2}}\left(\frac{\Delta_1 L}{T}\right)^{\frac{p-1}{3p-2}}=6\sigma^{\frac{p}{3p-2}}\left(\frac{\Delta_1 L}{T}\right)^{\frac{p-1}{3p-2}}.$$

Therefore we get

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla F(x_t)\|\right]\le 6\max\left\{\sqrt{\frac{\Delta_1 L}{T}},\sigma^{\frac{p}{3p-2}}\left(\frac{\Delta_1 L}{T}\right)^{\frac{p-1}{3p-2}}\right\}$$

and hence the claim. □

Additionally, this result recover those in Cutkosky and Mehta (2020)[7] with improved constants when $p=2$.

[7] Note that the authors did not use $\beta_1=0$, resulting in an additional term. However this term is not leading and hence does not affect the oracle complexity.

D.3 Technical Difficulties of Proving High Probability Convergence

In the previous section we showed that Equation (4) and Corollary 3 also hold for NSGD with momentum. For Theorem 5 on the other hand, while it still holds for time-varying and constant parameters, we were not able to prove the result for NSGD-M. We shortly want to discuss the technical difficulty of extending Corollary 5 to the momentum version.

The proof of Theorem 4 hinges on two parts: Firstly, one shows that the angle ϕ_t sharply concentrates around its conditional expectation $\psi_t = \mathbb{E}\left[\phi_t \,|\, \mathcal{F}_{t-1}\right]$. This step only requires the boundedness of ϕ_t and is hence applicable for both the minibatch and momentum version of NSGD. In the next step however, we have to lower bound ψ_t. Our current proof technique — and to some extend intuition — tells us that such lower bounds involves the term

$$\mathbb{E}\left[\|g_t - \nabla F(x_t)\| \,|\, \mathcal{F}_{t-1}\right]. \tag{27}$$

In the case of minibatch NSGD, g_t only depends on randomness sampled in iteration t, and (27) can hence be upper bounded by a constant as seen in (14). However, in the case of NSGD with momentum, g_t consists of random samples from all previous iterations. This results in (27) being a random variable instead, and it is not clear how to uniformly control it. Our empirical evidence indicates that an extension to NSGD-M might not be possible since quantiles of average gradient norms of NSGD-M behave super-linearly in $\log(1/\delta)$, see Section 4 and appendix F.3 for more details.

E LOWER-BOUNDS FOR NSGD

In this section we prove that Proposition 1 is tight and the sample complexity achieved in Equation (4) is optimal, in the sense that no other choice of parameters can lead to a uniformly better guarantee for `minibatch-NSGD` when problem-dependent parameters are unknown. To do so, we first derive a lower bound for the deterministic setting which might be of independent interest. Afterwards, we will equip this hard function with a stochastic gradient oracle to prove the lower bound for the stochastic setting.

E.1 Deterministic Setting

The following Lemma derives a lower bound for `NGD` with arbitrary stepsizes. We will use it afterwards to show optimality of our polynomial stepsize order.

Lemma 19 (Lower Bounds for Deterministic Setting)**.** *Let $\mathcal{F}_{\Delta_1,L}$ be the set of functions that satisfy (Lower Boundedness) and (L-smoothness) and let $\varepsilon > 0$. Denote `NGD` with stepsizes $\eta_t \geq 0$ as $A_{(\eta_t)}$. Then there exists a function $F \in \mathcal{F}_{\Delta_1,L}$ such that the iterates of $A_{(\eta_t)}$ satisfy*

$$\|\nabla F(x_t)\| > \varepsilon$$

for all $t \in [T^*]$, where $T^* \coloneqq \inf \left\{ T \in \mathbb{N} \;\middle|\; \varepsilon > \frac{\Delta_1 - \frac{\varepsilon}{L}}{2\sum_{t=1}^T \eta_t} + \frac{L\sum_{t=1}^T \eta_t^2}{8\sum_{t=1}^T \eta_t} \right\}$.

The proof extends ideas from (Hübler et al., 2024) by also including the (*L*-smoothness) assumption into the function construction.

Proof. We first define

$$g_\eta \colon [0,\eta] \to \mathbb{R}, x \mapsto \begin{cases} -2\varepsilon + Lx, & x \leq \frac{\eta}{2} \\ -2\varepsilon + \eta L - Lx, & x > \frac{\eta}{2}, \end{cases}$$

which will correspond to the gradient of our constructed function between two consecutive points. To formalise this idea, define

$$A_t \coloneqq -2\eta_t\varepsilon + \frac{\eta_t^2 L}{4} = \int_0^\eta g_{\eta_t}(x)dx$$

and T^* according to the statement. Furthermore, let $\tau_t \coloneqq \sum_{\kappa=1}^{t-1} \eta_\kappa$. Then we define our hard function via its derivative

$$F'(x) \coloneqq \begin{cases} -2\varepsilon, & x < 0 \\ g_{\eta_t}(x - \tau_t), & x \in [\tau_t, \tau_{t+1}),\, t \in [T^* - 1] \\ -2\varepsilon + L(x - \tau_{T^*}), & x \in (\tau_{T^*}, \tau_{T^*} + \frac{2\varepsilon}{L}] \\ 0, & \text{otherwise} \end{cases} \tag{28}$$

and $F(x) \coloneqq \Delta_1 + \int_0^x F'(\lambda)d\lambda$. Note that, by definition of F, we have

$$F(\tau_t) = \Delta_1 + \sum_{\tau=1}^{t-1} A_t = \Delta_1 - 2\varepsilon \sum_{\tau=1}^{t-1} \eta_\tau + \frac{L}{4}\sum_{\tau=1}^{t-1} \eta_\tau^2$$

and hence $F(\tau_t) > \frac{\varepsilon}{L}$ for all $t \leq T^*$[8] by our definition of T^*. In particular we have $\inf_x F(x) \geq 0$ and hence $F(0) - \inf_x F(x) \leq \Delta_1$. Furthermore F is L-smooth by definition and hence $F \in \mathcal{F}_{\Delta_1,L}$. Next we will show that `NGD`, when started at $x_1 = 0$, produces the iterates $x_t = \tau_t$. Therefore note that $\tau_1 = 0 = x_1$ and, assuming $x_t = \tau_t, t < T^*$, we have $x_{t+1} = x_t - \operatorname{sgn}(F'(x_t))\eta_t = x_t + \eta_t = \tau_{t+1}$. By induction we hence get $|F'(x_t)| = |F'(\tau_t)| \equiv 2\varepsilon$ for all $t \in [T^*]$, which completes the proof. □

[8] Due to the shift $t-1$ in the sum. Otherwise, if we would start the algorithm at x_0, $t \leq T^* - 1$.

The following Theorem is a consequence of Lemma 19 and implies that $r = 1/2$ is the (only) optimal choice of polynomial stepsize decay in the deterministic setting. To formulate the result, we will use the complexity definition introduced by Carmon et al. (2020), i.e., for an algorithm A, a function class $\mathcal{F}$ and $\varepsilon > 0$ we define

$$T_\varepsilon(A, \mathcal{F}) \coloneqq \sup_{F \in \mathcal{F}} \inf \left\{ t \in \mathbb{N} \;\middle|\; \|\nabla F(x_t)\| \le \varepsilon,\ (x_t)_{t \in \mathbb{N}} = A(F) \right\}.$$

Theorem 20 (Optimality in Deterministic Setting)**.** *Let $\mathcal{F}_{\Delta_1, L}$ be the set of functions that satisfy (Lower Boundedness) and (L-smoothness). Suppose $\varepsilon \le \frac{\Delta_1 L}{2}$ and $\varepsilon \le \frac{\sqrt{\Delta_1 L}}{3}$. Furthermore let $\eta, r > 0$ and denote* `NGD` *with decaying stepsizes $\eta_t = \eta\, t^{-r}$ as A_d^r, and with constant stepsizes $\eta_t \equiv \eta\, T^{-r}$ as A_c^r. Then $A^r \in \{A_d^r, A_c^r\}$ satisfies*

$$T_\varepsilon(A^r, \mathcal{F}_{\Delta_1, L}) \ge \left(\frac{\eta L}{4\varepsilon}\right)^{\frac{1}{r}} + \left(\frac{(1-r)\Delta_1}{8\eta\varepsilon}\right)^{\frac{1}{1-r}}$$

for $r \in (0,1)$. For $r \ge 1$ and small enough ε we have $T_\varepsilon(A^r, \mathcal{F}_{\Delta_1, L}) \ge \exp\left(\Delta_1/(8\eta\varepsilon)\right)$.

Proof. First note that the definition of T_ε starts with x_0 instead of x_1. For the sake of consistency with other works, we will also apply this convention in our result by denoting $(x_0, x_1, \ldots) \leftarrow (x_1, x_2, \ldots)$. We first consider constant stepsizes $\eta_t \equiv \eta T^{-r}$. Setting $x_0 = 0$ and applying Lemma 19 yields

$$\begin{aligned} T^* &= \inf \left\{ T \in \mathbb{N} \;\middle|\; \varepsilon > \frac{\Delta_1 - \frac{\varepsilon}{L}}{2\eta T^{1-r}} + \frac{\eta L}{8 T^r} \right\} \\ &\ge \inf \left\{ T \in \mathbb{N} \;\middle|\; \varepsilon > \frac{\Delta_1}{4\eta T^{1-r}} + \frac{\eta L}{8 T^r} \right\}, \end{aligned}$$

where we used our assumption $\varepsilon \le \frac{\Delta_1 L}{2}$ in the last line. For $r \in (0,1)$ we calculate

$$\begin{aligned} \varepsilon > \frac{\Delta_1}{4\eta T^{1-r}} + \frac{\eta L}{8T^r} &\Rightarrow \varepsilon > \max\left\{ \frac{\Delta_1}{4\eta T^{1-r}}, \frac{\eta L}{8T^r} \right\} \\ &\Leftrightarrow T > \max\left\{ \left(\frac{\Delta_1}{4\eta\varepsilon}\right)^{\frac{1}{1-r}}, \left(\frac{\eta L}{8\varepsilon}\right)^{\frac{1}{r}} \right\}. \end{aligned}$$

In particular we have $T^* \ge \max\left\{ \left(\frac{\Delta_1}{4\eta\varepsilon}\right)^{\frac{1}{1-r}}, \left(\frac{\eta L}{8\varepsilon}\right)^{\frac{1}{r}} \right\}$ and hence

$$T_\varepsilon(A_c, \mathcal{F}_{\Delta_1, L}) \ge \max\left\{ \left(\frac{\Delta_1}{4\eta\varepsilon}\right)^{\frac{1}{1-r}}, \left(\frac{\eta L}{8\varepsilon}\right)^{\frac{1}{r}} \right\}.$$

For $r \ge 1$ we have

$$\begin{aligned} \varepsilon > \frac{\Delta_1}{4\eta T^{1-r}} + \frac{\eta L}{8T^r} &\Rightarrow \varepsilon > \max\left\{ \frac{\Delta_1 T^{r-1}}{4\eta}, \frac{\eta L}{8T^r} \right\} \\ &\Leftrightarrow T > \left(\frac{\eta L}{8\varepsilon}\right)^{\frac{1}{r}} \text{ and } T^{r-1} > \frac{4\eta\varepsilon}{\Delta_1} \end{aligned}$$

In the case $r = 1$ this implies $T_\varepsilon(A_c^r, \mathcal{F}_{\Delta_1, L}) \ge \infty$ for all $\varepsilon < \frac{\Delta_1}{4\eta}$. For $r > 1$ the same holds for $\varepsilon < (\eta L)^\alpha \left(\frac{\Delta_1}{\eta}\right)^\beta$, where $\alpha \coloneqq \frac{r-1}{2r-1}$ and $\beta \coloneqq \frac{r}{2r-1}$.

Next we consider decreasing stepsizes $\eta = \eta\, t^{-r}$. Therefore note that, for $q \ne 1$,

$$\frac{\eta\Big((T+1)^{1-q} - 1\Big)}{1-q} = \eta \int_1^{T+1} t^{-q} dt \le \eta \sum_{t=1}^{T} t^{-q} \le \eta \left(1 + \int_1^T t^{-q} dt\right) = \eta\left(1 + \frac{T^{1-q} - 1}{1-q}\right).$$

In particular we have

$$2\varepsilon \sum_{t=1}^{T} \eta_t \leq 2\varepsilon\eta\left(1 + \frac{T^{1-r}-1}{1-r}\right) \text{ and}$$
$$\frac{L}{4}\sum_{t=1}^{T} \eta_t^2 \geq \frac{\eta^2 L}{4(1-2r)}\left((T+1)^{1-2r} - 1\right).$$

We first consider $r \in (0,1) \setminus \{1/2\}$. Plugging into Lemma 19 yields

$$T^* \geq \inf\left\{T \in \mathbb{N} \;\middle|\; \varepsilon > \frac{\Delta_1(1-r)}{4\eta T^{1-r}} + \frac{\eta L(1-r)\left(T^{1-2r}-1\right)}{8(1-2r)T^{1-r}}\right\}.$$

By our assumptions on ε we get $T^* \geq 2$ and hence can assume $T \geq 2$ which in turn implies $\frac{T^{1-2r}-1}{1-2r} \geq \frac{T^{1-2r}}{2}$. Therefore we get

$$T^* \geq \inf\left\{T \in \mathbb{N} \;\middle|\; \varepsilon > \frac{\Delta_1(1-r)}{4\eta T^{1-r}} + \frac{\eta L(1-r)}{16 T^{r}}\right\}$$
$$\geq \max\left\{\left(\frac{\Delta_1(1-r)}{4\eta\varepsilon}\right)^{\frac{1}{1-r}}, \left(\frac{\eta L(1-r)}{16\varepsilon}\right)^{\frac{1}{r}}\right\}.$$

Finally we have to consider the edge cases $r \in \{1/2, 1\}$. Let $r = 1/2$, then

$$\sum_{t=1}^{T} \eta_t \leq \eta\left(2\sqrt{T} - 1\right) \leq 2\eta\sqrt{T},$$
$$\sum_{t=1}^{T} \eta_t^2 \geq \eta^2 \int_1^{T+1} t^{-1} dt \geq \eta^2 \log(T)$$

and hence

$$T^* \geq \inf\left\{T \in \mathbb{N} \;\middle|\; \varepsilon > \frac{\Delta_1}{8\eta\sqrt{T}} + \frac{\eta L \log(T)}{16\sqrt{T}}\right\}$$
$$\geq \inf\left\{T \in \mathbb{N} \;\middle|\; \varepsilon > \frac{\Delta_1(1-r)}{4\eta T^{r}} + \frac{\eta L(1-r)}{16 T^{r}}\right\}$$

and we can hence proceed as before. Note that a more careful analysis can additionally show tightness of the $\log(T)$ dependence. Now let $r = 1$ and note that

$$\sum_{t=1}^{T} \eta_t = \eta \sum_{t=1}^{T} t^{-1} \leq \eta(1 + \log(T)).$$

In particular

$$T^* \geq \inf\left\{T \in \mathbb{N} \;\middle|\; \varepsilon > \frac{\Delta_1}{4\eta(1+\log(T))}\right\} \geq \exp\left(\frac{\Delta_1}{4\eta\varepsilon} - 1\right)$$

and hence $T_\varepsilon(A_d^r, \mathcal{F}_{\Delta_1,L}) \geq e^{\frac{\Delta_1}{8\eta\varepsilon}}$ for $\varepsilon \leq \frac{\Delta_1}{8\eta}$. □

E.2 Stochastic Setting

Finally we will extend the above result to the stochastic setting.

Theorem 21. *Let $\mathcal{F}_{\Delta_1,L}$ be the set of functions that satisfy (Lower Boundedness) and (L-smoothness). Furthermore let $\mathcal{O}_{\sigma,p}$ denote the set of stochastic gradient oracles that satisfy (p-BCM). Suppose $\varepsilon \leq \frac{\Delta_1 L}{2}$ and let $\eta, B, q > 0, r \in (0,1)$. Let $\mathcal{A}$ denote* `NGD` *with parameters $\eta_t \equiv \eta T^{-r}, B_t \equiv \max\{1, BT^q\}$ and the mini-batch gradient estimator $g_t = \frac{1}{B_t}\sum_{j=1}^{B_t} \nabla f(x_t, \xi_t^{(j)})$, where $\xi_t^{(1)}, \ldots, \xi_t^{(B_t)} \overset{i.i.d.}{\sim} \xi_t$. Then there exists a function $F \in \mathcal{F}_{\Delta_1,L}$ and oracle $\nabla f(\cdot,\cdot) \in \mathcal{O}_{\sigma,p}$ such that $\mathcal{A}$ requires at least*

$$m_\varepsilon^{\mathbb{E}} \geq \max\left\{ \left(\frac{\Delta_1}{6\eta\varepsilon}\right)^{\frac{1}{1-r}}, \left(\frac{\eta L}{12\varepsilon}\right)^{\frac{1}{r}}, B\left(\frac{\Delta_1}{6\eta\varepsilon}\right)^{\frac{1+q}{1-r}}, B\left(\frac{\eta L}{12\varepsilon}\right)^{\frac{1+q}{r}}, \frac{1}{B^{\frac{1}{q}}}\left(\frac{\sigma}{28\varepsilon}\right)^{\frac{p(q+1)}{q(p-1)}}. \right\}$$

oracle calls to generate an iterate with $\mathbb{E}\left[\|\nabla F(x_t)\|\right] \leq \varepsilon$.

When comparing to the corresponding upper bound (16) we can see that both bounds are tight in all parameters. This is due to $\frac{1}{n}\sum_{i=1}^n a_i \leq \max\{a_1, \ldots, a_n\} \leq \sum_{i=1}^n a_i$, hence the maximum and sum notation are equivalent up to constants.

Proof. The idea behind the proof is the following. We will again use a very similar construction to (28) and add a noise oracle on top. The goal of this noise oracle will be to point in the wrong direction with the highest possible probability, effectively slowing down the progress we make even more. As this noise oracle may however lead to iterates going below x_1, we need to slightly modify the construction of F.

Construction of the hard function F. To this end, let $\bar{\eta} := \eta T^{-r}$, $\tau_k := k\bar{\eta}$ for $k \in \mathbb{Z}$,

$$\begin{aligned} T_d^* &:= \inf\left\{ T \in \mathbb{N} \;\middle|\; \varepsilon > \frac{\Delta_1 - \frac{\varepsilon}{L}}{3T\bar{\eta}} + \frac{L\bar{\eta}}{12} \right\} \\ &\geq \max\left\{ \left(\frac{\Delta_1}{6\eta\varepsilon}\right)^{\frac{1}{1-r}}, \left(\frac{\eta L}{12\varepsilon}\right)^{\frac{1}{r}} \right\}. \end{aligned}$$

where we used $\varepsilon \leq \frac{\Delta_1 L}{2}$ in the last inequality as before, and define

$$F'(x) := \begin{cases} g_{\bar{\eta}}(x - \tau_t), & x \in [\tau_t, \tau_{t+1}),\ t+1 \leq T_d^* \\ -2\varepsilon + L\big(x - \tau_{T_d^*}\big), & x \in \big(\tau_{T_d^*}, \tau_{T_d^* *} + \frac{3\varepsilon}{L}\big] \\ 0, & \text{otherwise}, \end{cases} \tag{29}$$

where

$$g_\eta \colon [0, \eta] \to \mathbb{R} \quad \text{is given by} \quad x \mapsto \begin{cases} -3\varepsilon + Lx, & x \leq \frac{\eta}{2} \\ -3\varepsilon + \eta L - Lx, & x > \frac{\eta}{2}. \end{cases}$$

As before, we define the hard function as $F(x) := \Delta_1 + \int_0^x F'(t)dt$. In the following we will denote the derivative of F using ∇F to align with the gradient oracle notation. Now firstly note that we can use the deterministic Lemma 19 to rule out any stepsize that satisfies $\bar{\eta} \geq \frac{8\varepsilon}{L}$: In this case we would have

$$\frac{\Delta_1 - \frac{\varepsilon}{L}}{2T\bar{\eta}} + \frac{LT\bar{\eta}^2}{8T\bar{\eta}} = \frac{\Delta_1 - \frac{\varepsilon}{L}}{2T\bar{\eta}} + \frac{L\bar{\eta}}{8} \geq 0 + \varepsilon,$$

where we used $\varepsilon \leq \frac{\Delta_1 L}{2}$ in the last inequality. Hence we may assume $\eta T^{-r} \leq \frac{8\varepsilon}{L}$. Under this assumption[9] we again have $F(0) - \inf_x F(x) \leq \Delta_1$ and that F is L-smooth.

Construction of the noise oracle $\nabla f(x, \xi)$. We will construct the mentioned oracle, which aims to point in the wrong direction with the maximal probability. This oracle construction follows a similar idea as (Yang et al., 2024, Theorem 3). To construct the oracle, let $\rho > 0$ to be defined later and define $\alpha := \frac{p}{p-1}$,

$$\delta(x) := \min\left\{ 1, \left(\frac{2(1+\rho)\,\|\nabla F(x)\|}{\sigma}\right)^{\alpha} \right\}.$$

[9] Note that for $\eta T^{-r} \geq \frac{12\varepsilon}{L}$ we would get $\lim_{x\to-\infty} F(x) = -\infty$ for F defined by (29).

This will be the probability of the oracle returning *the correct* direction. Now let

$$\nabla f(x,\xi) := \begin{cases} -\rho\nabla F(x), & \xi \geq \delta(x) \\ \left(1+\frac{(1-\delta(x))(1+\rho)}{\delta(x)}\right)\nabla F(x), & \xi < \delta(x) \end{cases}$$

and $\xi \sim \mathrm{Unif}\left([0,1]\right)$. Straightforward calculations yield

$$\begin{aligned}\mathbb{E}\left[\nabla f(x,\xi)\right] &= \nabla F(x)\left((1-\delta(x))(-\rho)+\delta(x)\left(1+\frac{(1-\delta(x))(1+\rho)}{\delta(x)}\right)\right)\\ &= \nabla F(x)((1-\delta(x))(-\rho)+\delta(x)+(1-\delta(x))(1+\rho))\\ &= \nabla F(x)\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}\left[\|\nabla f(x,\xi)-\nabla F(x)\|^p\right] &= (1-\delta(x))(1+\rho)^p\left\|\nabla F(x)\right\|^p+\delta(x)\left(\frac{(1-\delta(x))(1+\rho)}{\delta(x)}\left\|\nabla F(x)\right\|\right)^p\\ &= (1+\rho)^p\left\|\nabla F(x)\right\|^p\left((1-\delta(x))+\frac{(1-\delta(x))^p}{\delta(x)^{p-1}}\right)\\ &\leq (1+\rho)^p\left\|\nabla F(x)\right\|^p\left(1+\frac{1}{\delta(x)^{p-1}}\right)\\ &\leq \frac{2(1+\rho)^p\left\|\nabla F(x)\right\|^p}{\delta(x)^{p-1}} \leq \sigma^p.\end{aligned}$$

In particular $\nabla f(\cdot,\cdot) \in \mathcal{O}_{\sigma,p}$.

The behaviour of NSGD on the constructed function and oracle. Finally we are able to show the lower bound by analysing the behaviour of NSGD on our constructed objects. Firstly it is clear that we can upper bound the stochastic progress by the deterministic progress, i.e. $x_{t+1} \leq \bar{\eta}t$. In particular, with the same arguments as in Lemma 19 and theorem 20, we get that $\|\nabla F(x_t)\| > \varepsilon$ for all $t \in [T_d^*]$ and we hence can lower bound the iteration complexity by $T \geq T_d^*$. We next differentiate two cases.
Case 1: $\max\left\{1, B(T_d^*)^q\right\} \geq \frac{\sigma^\alpha}{2(7\varepsilon)^\alpha}$**.** In this case nothing else needs to be done and we can lower bound the number of oracle calls required to find an ε-stationary point by

$$T_d^* \cdot \max\left\{1, B(T_d^*)^q\right\} = \max\left\{T_d^*, B\left(\frac{\Delta_1}{6\eta\varepsilon}\right)^{\frac{1+q}{1-r}}, B\left(\frac{\eta L}{12\varepsilon}\right)^{\frac{1+q}{r}}\right\}.$$

Case 2: $\max\left\{1, B(T_d^*)^q\right\} < \frac{\sigma^\alpha}{2(7\varepsilon)^\alpha}$**.** In this case we will make use of the gradient oracle to construct a stronger lower bound $T_s^* > T_d^*$. Therefore first note that, due to the constant stepsize and $x_1 = 0$, the iterations x_t always stay on the lattice $\Gamma = \bar{\eta}\mathbb{Z}$. Furthermore, by the construction of F, we have that

$$\forall\, x \in \Gamma\colon \nabla F(x) \in [-3\varepsilon, 0],$$

which, for $\rho \leq \frac{1}{6}$, in particular implies

$$\forall\, x \in \Gamma\colon \delta(x) \leq \left(\frac{6(1+\rho)\varepsilon}{\sigma}\right)^\alpha \leq \left(\frac{7\varepsilon}{\sigma}\right)^\alpha =: \delta. \tag{30}$$

Now define the random variable $\zeta_t := 1_{\{g_t<0\}}$, where $1_A(\omega) = \begin{cases} 1, & \omega \in A \\ 0, & \text{o.w.}\end{cases}$, and compute

$$x_{t+1} = x_t + \bar{\eta}(2\zeta_t - 1) = \bar{\eta}\sum_{\tau=1}^{t}(2\zeta_\tau - 1) =: 2\bar{\eta}S_t - \bar{\eta}t, \tag{31}$$

where $S_t \coloneqq \sum_{\tau=1}^{t} \zeta_\tau$. Furthermore, let $\bar{B} = \max\{1, BT^q\}$ and $\rho \le \min\{1/6, 1/\bar{B}\}$, then we have

$$\mathbb{P}(g_t < 0) = \mathbb{P}\left(\frac{1}{\bar{B}} \sum_{j=1}^{\bar{B}} \nabla f(x_t, \xi_t^{(j)}) < 0\right) = 1 - (1 - \delta(x_t))^{\bar{B}} \overset{(30)}{\le} 1 - (1-\delta)^{\bar{B}}. \tag{32}$$

By definition $(1-\delta)^{\bar{B}} = \exp\left(\bar{B} \log(1-\delta)\right)$ and

$$\log(1-\delta) \ge \frac{-\delta}{1-\delta} = -\frac{(7\varepsilon)^\alpha}{\sigma^\alpha - (7\varepsilon)^\alpha} \ge -\frac{2(7\varepsilon)^\alpha}{\sigma^\alpha}$$

where we used $\log(1+x) \ge \frac{x}{1+x}$ for $x \in (-1, 0]$ in the first, and $1 < \frac{\sigma^\alpha}{2(7\varepsilon)^\alpha}$ by our case 2 assumption in the last inequality. Plugging into (32) yields

$$\mathbb{P}(g_t < 0) \le 1 - \exp\left(-\bar{B}\frac{2(7\varepsilon)^\alpha}{\sigma^\alpha}\right) \le \min\left\{1, \bar{B}\frac{2(7\varepsilon)^\alpha}{\sigma^\alpha}\right\}, \tag{33}$$

where we used $e^x \ge \max\{0, 1+x\}$ in the last inequality. Next up let $T \in \mathbb{N}$ such that $\bar{B} \le \frac{\sigma^\alpha}{4(7\varepsilon)^\alpha}$, then

$$\mathbb{P}(g_t < 0) \overset{(33)}{\le} \min\left\{1, \bar{B}\frac{2(7\varepsilon)^\alpha}{\sigma^\alpha}\right\} \le \frac{1}{2}$$

and ζ_t is hence a Bernoulli random variable with probability at most $1/2$. In particular, we have $\operatorname{median}(S_t) \le \lfloor t/2 \rfloor$ and hence

$$\frac{1}{2} \le \mathbb{P}\left(S_t \le \frac{t}{2}\right) = \mathbb{P}(2\bar{\eta} S_t \le \bar{\eta} t) \overset{(31)}{=} \mathbb{P}(x_{t+1} \le 0).$$

Finally note that all x in $(\Gamma \cap (-\infty, 0])$ satisfy $\nabla F(x) = -3\varepsilon$ and hence

$$\mathbb{E}\left[\|\nabla F(x_t)\|\right] = 3\varepsilon \mathbb{P}(x_t \le 0) + \mathbb{E}\left[\|\nabla F(x_t)\| \, 1_{\{x_t > 0\}}\right] > \varepsilon + 0$$

for all $t \in [T]$. Summing up the results in *Case 2*, we so far proved the auxiliary result that for any T with $\max\{1, BT^q\} \le \frac{\sigma^\alpha}{4(7\varepsilon)^\alpha}$ all iterates $t \in [T]$ satisfy $\mathbb{E}\left[\|\nabla F(x_t)\|\right] > \varepsilon$. By the assumption of case 2, this implies

$$\forall\, T \le T_s^* \colon \forall\, t \in [T] \colon \mathbb{E}\left[\|\nabla F(x_t)\|\right] > \varepsilon,$$

where $T_s^* \coloneqq \left(\frac{\sigma^\alpha}{4B(7\varepsilon)^\alpha}\right)^{1/q} > T_d^*$ with $\alpha = \frac{p}{p-1}$. In particular, we can lower bound the number of oracle calls required to reach an expected ε-stationary point by

$$T_s^* \cdot B (T_s^*)^q = \frac{1}{B^{\frac{1}{q}}}\left(\frac{\sigma}{28\varepsilon}\right)^{\frac{p(q+1)}{q(p-1)}}.$$

Combining. Finally we are able to combine everything into our lower bound. Therefore, let

$$T^* \coloneqq \max\left\{T_d^*, B\left(\frac{\Delta_1}{6\eta\varepsilon}\right)^{\frac{1+q}{1-r}}, B\left(\frac{\eta L}{12\varepsilon}\right)^{\frac{1+q}{r}}, \frac{1}{B^{\frac{1}{q}}}\left(\frac{\sigma}{28\varepsilon}\right)^{\frac{p(q+1)}{q(p-1)}}\right\}$$

and note that for $T \le T^*$, one of the above cases applies, showing

$$\forall\, t \in [T] \colon \mathbb{E}\left[\|\nabla F(x_t)\|\right] > \varepsilon.$$

This completes the proof. □

E.3 Lower-Bound on the Convergence Measure

Proof. In this proof, we consider a slightly more general step-size $\gamma = \sqrt{\frac{\Delta_1}{L}}\frac{1}{T^a}$ for any $a < 1$. Take $F(x) = \frac{1}{2}x^2$ and $x_1 > 0$ (w.l.g.), then the step-size is $\gamma = \frac{x_0}{T^a}$. Denote by $N := \sqrt{2}T^a$. For the first $\lceil N \rceil$ iterations the update rule is $x_t = x_1 - \gamma(t-1) = x_1(1 + \frac{1}{N} - \frac{t}{N}) \geq 0$, for $t = 0, \ldots, \lceil N \rceil$. We compute for $T \geq \lceil N \rceil = \lceil \sqrt{2}T^a \rceil$

$$\begin{aligned}
\sum_{t=1}^{T} \|\nabla F(x_t)\|^2 &\geq \sum_{t=1}^{\lceil N \rceil} \|\nabla F(x_t)\|^2 = x_1^2 \sum_{t=1}^{\lceil N \rceil} \left(1 + \frac{1}{N} - \frac{t}{N}\right)^2 \\
&= x_1^2 \lceil N \rceil \left(1 + \frac{1}{N}\right)^2 - 2x_1^2 \left(1 + \frac{1}{N}\right) \frac{1}{N} \sum_{t=1}^{\lceil N \rceil} t + \frac{x_1^2}{N^2} \sum_{t=1}^{\lceil N \rceil} t^2 \\
&= x_1^2 \lceil N \rceil \left(1 + \frac{1}{N}\right)^2 - 2x_1^2 \left(1 + \frac{1}{N}\right) \frac{1}{N} \frac{\lceil N \rceil(\lceil N \rceil + 1)}{2} \\
&\qquad + \frac{x_1^2}{N^2}\left(\frac{\lceil N \rceil^3}{3} + \frac{\lceil N \rceil^2}{2} + \frac{\lceil N \rceil}{6}\right) \\
&\geq x_1^2 \frac{\lceil N \rceil (N+1)(N - \lceil N \rceil)}{N^2} + \frac{x_1^2 \lceil N \rceil^3}{3\,N^2} \\
&\geq -\frac{x_1^2 \lceil N \rceil^2}{N^2} + \frac{x_1^2 \lceil N \rceil^3}{3\,N^2} \\
&\geq \frac{x_1^2 \lceil N \rceil^3}{6\,N^2}\left(2 - \frac{6}{\lceil N \rceil}\right) \geq \frac{x_1^2 \lceil N \rceil^3}{6\,N^2} \geq \frac{x_1^2 N}{6} = \frac{\sqrt{2}L\Delta_1 T^a}{3},
\end{aligned}$$

where in the second inequality we dropped the last two terms, in the third inequality we used $\lceil N \rceil - N \leq 1$ and $N + 1 \leq \lceil N \rceil^2$ for $N \geq 6$. The forth inequality holds by the assumption $N \geq 6$. It remains to divide both sides by T and verify that in case $a = 1/2$, the assumption $T \geq 18$ implies the assumed conditions $T \geq \lceil N \rceil \geq 6$. Rearranging, we get

$$\sqrt{\mathbb{E}\left[\|\nabla F(\bar{x}_T)\|^2\right]} \geq \sqrt{\frac{\sqrt{2}L\Delta_1}{3T}} \cdot T^{1/4}.$$

Noting that $\sqrt{\frac{\sqrt{2}}{3}} \geq \frac{2}{3}$ completes the proof. □

F ADDITIONAL EXPERIMENTS AND DETAILS

In this section we provide additional information and experiments for Section 4.

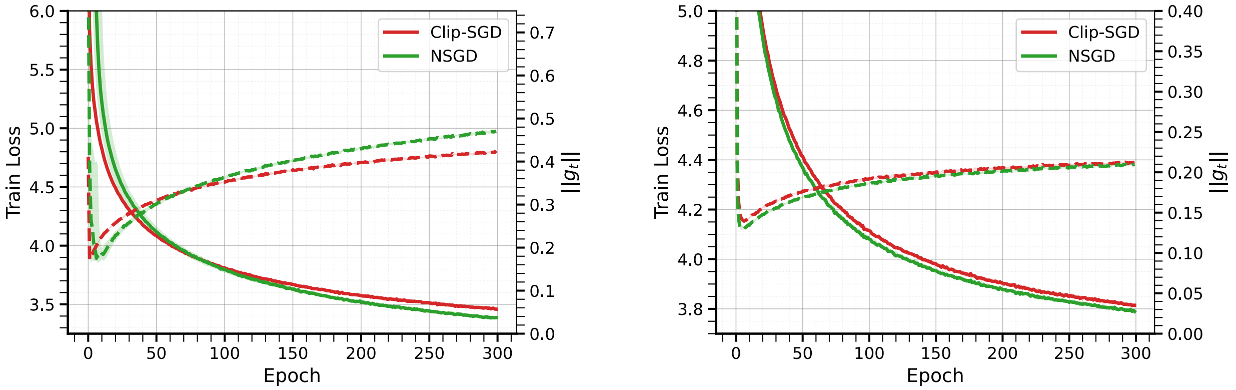
F.1 Additional Details on Language Modelling

Additional Details All experiments were carried out on Nvidia RTX 3090 GPUs in an internal cluster. The total compute including preliminary experiments were approximately 380 GPU hours. Roughly 200 of these were required for preliminary experiments and parameter-tuning, 180 for the final experiments.

The AWD-LSTM (Merity et al., 2018) is released under a BSD 3-Clause License, the Penn Treebank dataset (Marcus et al., 1993) under the LDC User Agreement for Non-Members and the WikiText-2 dataset (Merity et al., 2017) under the Creative Commons BY-SA 3.0 license.

The below experiments all follow the general structure outlined in Section 4.

Reasons for the Clipping Behaviour. We additionally want to understand why the percentage of clipped gradients increases over time. Therefore Figure 4 examine the average minibatch-gradient norm per epoch, i.e. $\frac{1}{E}\sum_{t=t_0}^{t_0+E-1}\|g_t\|$ where the epoch consists of E mini-batches and starts at iteration t_0. The plot shows that, while the training loss decreases, the stochastic gradient norms increase. This observation is in line with previous observations on different tasks (Goodfellow et al., 2016, Chapter 8). Therefore, while surprising at first, the increase in stochastic gradient norms is able to explain the increasing clipping percentage in hindsight.



(a) PTB

(b) WikiText-2

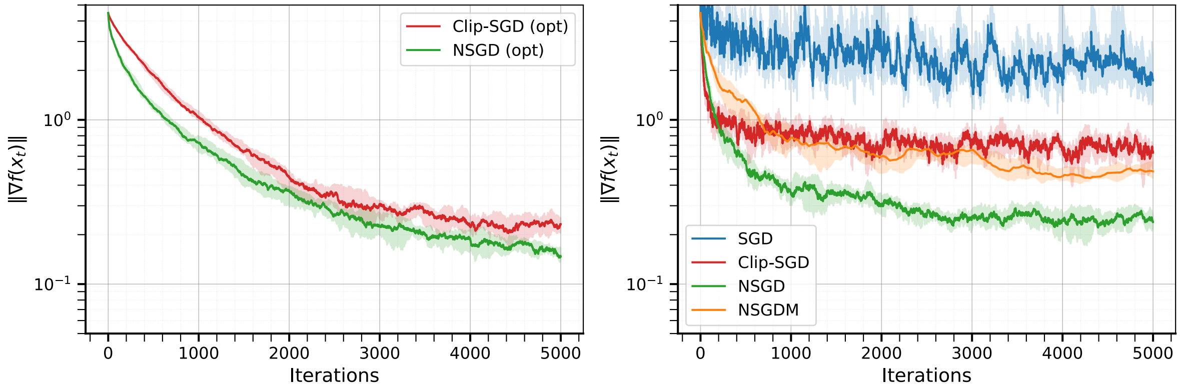
Figure 4: All plots consider `Clip-SGD` and `NSGD` with tuned parameters. Solid lines represent the training loss and correspond to the left y-axis. The dashed line corresponds to the right y-axis, and represents the average mini-batch gradient norm in an epoch. Shaded areas represent the minimal and maximal value within 5 seeds, the line the median.

F.2 Illustrating Drawbacks of Gradient Clipping Theory

In the two sets of experiments below we compare several algorithms with and without parameter tuning on a simple quadratic model to better understand the influence of heavy tailed noise and the necessity of parameter tuning.

Comparison with tuned parameters. In this set of experiments, we compare `NSGD` using step-sizes $\eta_t = \eta/\sqrt{t}$ (Yang et al., 2024) and `Clip-SGD` with $\eta_t = \eta/\sqrt{t}$, $\gamma_t = \gamma \cdot \sqrt[4]{t}$ (Zhang et al., 2020; Nguyen et al., 2023a) and tune the pair (η, γ) over a grid. We report the optimal parameters, η for `NSGD` and the pair (η, γ) for `Clip-SGD` in Table 1. Convergence plots for different noise distributions are presented in Figures 5a, 6a and 7a.

We observe that in both heavy tailed and light tailed settings, when parameters are tuned, the two algorithms exhibit comparable convergence rate. However, for all three noise distributions we tested, the optimal parameter for `NSGD` appeared to be $\eta = 0.5$, while `Clip-SGD` required different parameters to reach similar performance.



(a) Tuned `Clip-SGD` vs. `NSGD`. When parameters are tuned, clipping is triggered at every iteration.

(b) Comparison *without parameter tuning.*

Figure 5: Performance of different algorithms on a quadratic problem $f(x, \xi) = \frac{1}{2} \|x\|_2^2 + \langle x, \xi \rangle$, $d = 10$, where ξ is a random vector with i.i.d. components drawn from a symmetrized Pareto distribution with tail index $p = 1.5$.

This illustrates that `Clip-SGD` can be more sensitive to misspecification of its parameters compared to `NSGD`. Moreover, `Clip-SGD` requires two parameters for tuning compared to only one parameter for `NSGD`.

Table 1: Comparison of Tuned Parameters for `NSGD` and `Clip-SGD` under Different Noise Distributions

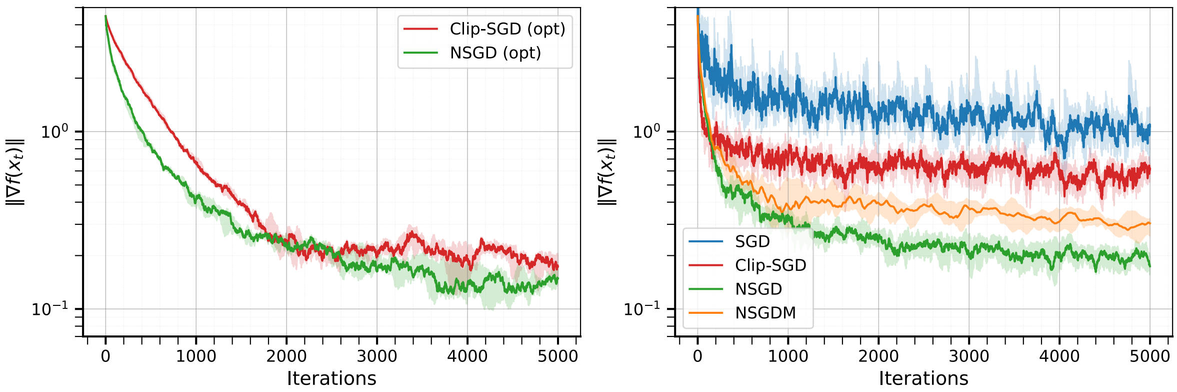
Noise Distribution	**Algorithm**	**Optimal η**	**Optimal γ**
Heavy tailed ($p = 1.5$) (Figure 5a)	`NSGD` `Clip-SGD`	0.5 0.1	– 1
Heavy tailed ($p = 1.8$) (Figure 6a)	`NSGD` `Clip-SGD`	0.5 100	– 0.001
Light tailed (Figure 7a)	`NSGD` `Clip-SGD`	0.5 5	– 0.1

Comparison without parameter tuning. In Figures 5b, 6b and 7b, we compare several adaptive algorithms with default parameter sequences, i.e., $\gamma = \eta = 1$. Specifically, we use $\eta_t = 1/\sqrt{t}$ for `SGD` and `NSGD`; $\eta_t = \sqrt{\alpha_t/t}$, $\alpha_t = 1/\sqrt{t}$ (where α_t is momentum sequence) (Yang et al., 2024), and $\eta_t = 1/\sqrt{t}$, $\gamma_t = \sqrt[4]{t}$ for `Clip-SGD`. This order of step-sizes is selected based on the theory for each algorithm under BV setting, where this order is known to give an asymptotically optimal convergence rate as $T \to \infty$. We observe that the performance of `Clip-SGD` significantly degrades when γ and η are not tuned, while the performance of `NSGD`remains nearly the same.

We also see that under heavier tailed noise such as Pareto distribution, the performance of untuned `SGD` and `Clip-SGD` degrades substantially compared to the light tail noise setting, confirming the sensitivity of these algorithms to different noise distributions. On the other hand, `NSGD` and its momentum variant (see (21) in Appendix D) are more stable and converge to a smaller neighbourhood around the optimal solution even under heavy tailed noise.

F.3 Verifying High Probability Bounds for SGD and NSGD(M)

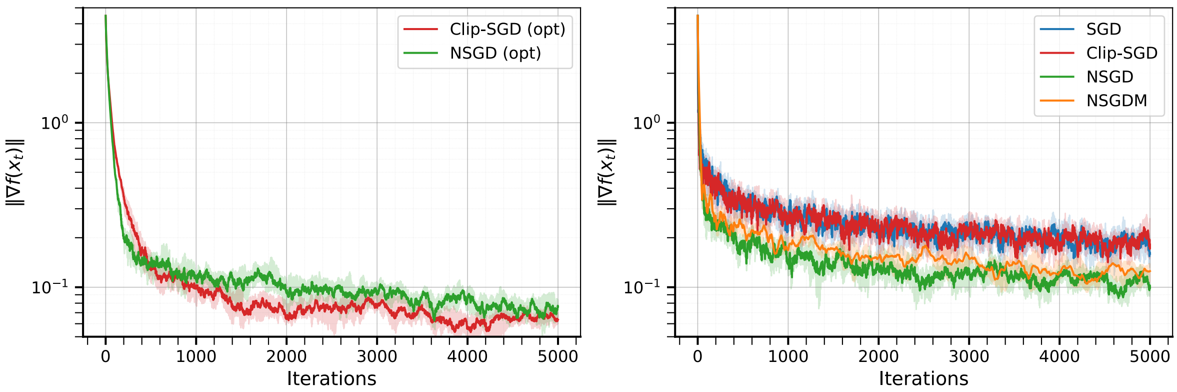
In this section, we conduct experiments to verify high probability convergence for three algorithms: `SGD`, `NSGD`, `NSGD-M`. High probability convergence refers to the convergence rate of the average gradient norm depending linearly on $\log(1/\delta)$ as demonstrated in our Corollary 5, where $\delta \in (0, 1)$ is a failure probability. Previous theoretical results have shown that vanilla `SGD` does not exhibit high probability convergence. In particular, Sadiev et al. (2023) demonstrated that under a bounded variance setting, `SGD` fails to achieve this property, mainly due to noise injected in the final iteration. In contrast, for adaptive methods such as `Clip-SGD` (Nguyen et al., 2023a) or `NSGD` Corollary 5, one can establish a high probability convergence with a mild linear dependence on $\log(1/\delta)$. However, extending our result to `NSGD-M` is challenging due to correlation issues introduced by momentum. Thus, the primary objective of this section is to empirically investigate whether `NSGD-M` might still



(a) Tuned Clip-SGD vs. NSGD. When parameters are tuned, clipping is triggered at every iteration.

(b) Comparison *without parameter tuning.*

Figure 6: Performance of different algorithms on a quadratic problem $f(x,\xi) = \frac{1}{2}\|x\|_2^2 + \langle x, \xi \rangle$, $d = 10$, where ξ is a random vector with i.i.d. components drawn from a symmetrized Pareto distribution with tail index $p = 1.8$.



(a) Tuned Clip-SGD vs. NSGD. When parameters are tuned, clipping is triggered at every iteration.

(b) Comparison *without parameter tuning.*

Figure 7: Performance of different algorithms on a quadratic problem $f(x,\xi) = \frac{1}{2}\|x\|_2^2 + \langle x, \xi \rangle$, $d = 10$, where ξ is distributed according standard Normal distribution.

exhibit high probability convergence similar to NSGD.

To achieve this, we evaluate the performance of the three algorithms on a simple quadratic function $F(x) = \frac{1}{2}\|x\|^2$, $x \in \mathbb{R}^d$ using dimensions $d = 1$ and $d = 1000$. We introduce three types of noise during training: (standard) Normal, (component-wise) Pareto with $p = 1.5$ and $p = 2.5$, to simulate both light-tailed and heavy-tailed noise environments. Each algorithm is run $k = 10^5$ times over $T = 100$ iterations, and the convergence criterion is the average gradient norm over T iterations. We present two plots for each set of experiments. The left plot visualizes the convergence behavior by selecting the median, δ and $1 - \delta$ quantiles (where $\delta := 10^{-4}$) of the algorithm runs based on the average gradient norm at $T = 100$. These quantile-based trajectories are plotted against iteration $t = 1, \ldots, T$. The right plot shows the quantiles of average gradient norm at $T = 100$ plotted against $\log(1/\delta)$. For algorithms with high probability convergence, this plot should have a sub-linear dependence on $\log(1/\delta)$.

Light tailed noise. Our results reveal that for the Normal noise distribution, which has light tails, all three algorithms exhibit sub-linear curves Figures 8 and 9, indicating high probability convergence. However, for Pareto noise Figures 10 to 13 (particularly with $p = 1.5$ and $p = 2.5$), which corresponds to infinite and finite variance regimes respectively, SGD exhibits a super-linear curve, confirming its lack of high probability convergence, consistent with theoretical predictions.

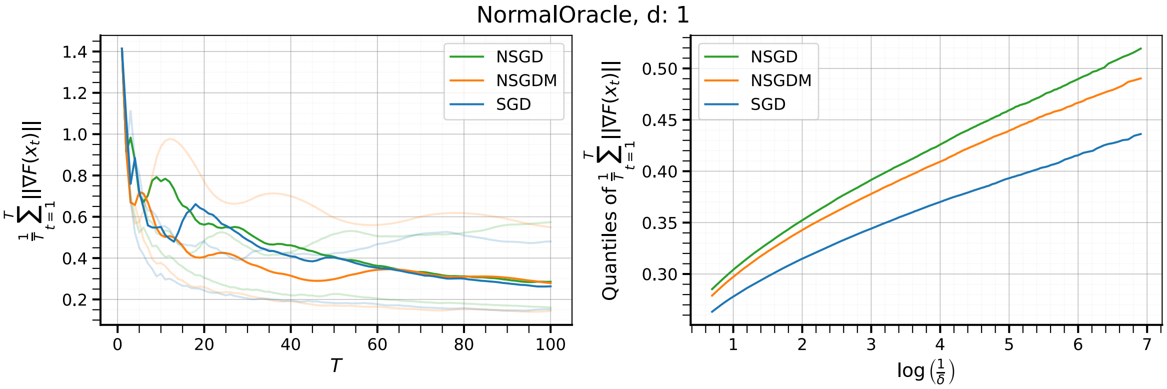


Figure 8: Light tail noise, $\xi_t \sim \mathcal{N}(0, I)$.

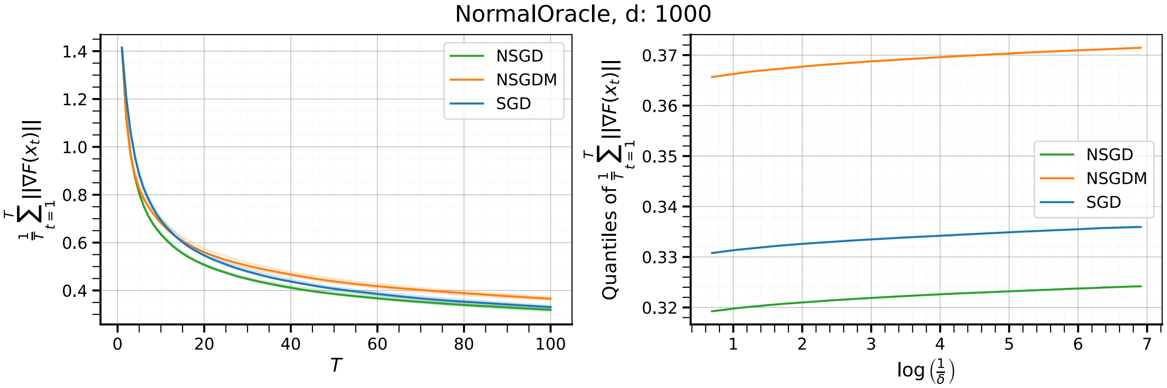


Figure 9: Light tail noise, $\xi_t \sim \mathcal{N}(0, I)$.

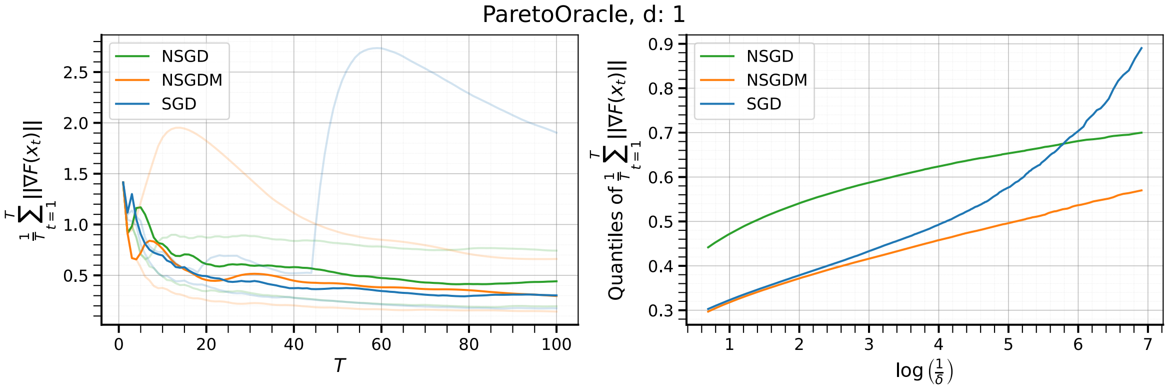


Figure 10: Heavy tailed noise with finite variance. Pareto with $p = 2.5$.

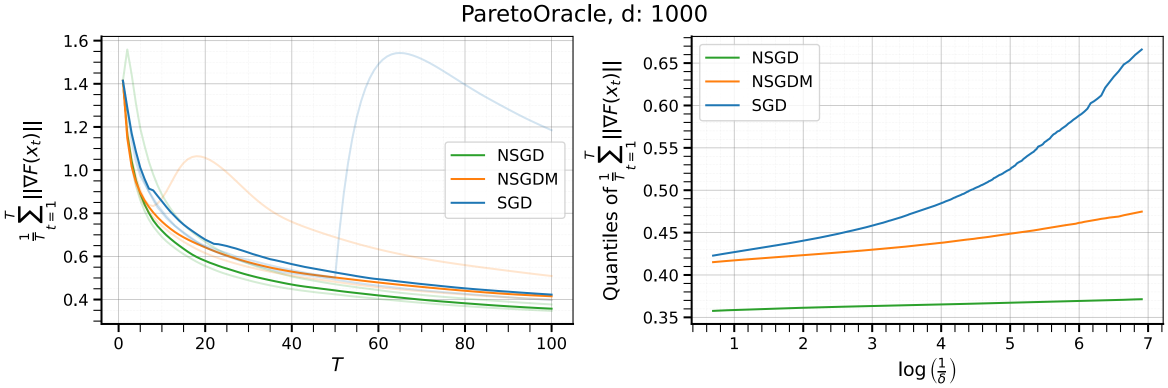


Figure 11: Heavy tailed noise with finite variance. Pareto with $p = 2.5$. See Figure 3 for the same experiment, but without SGD on the plot.

Heavy tailed noise. Most importantly, we observe that while both NSGD and NSGD-M exhibit similar qualitative behaviors when the noise has light tails, Figures 8 and 9; the quantile dependence on $\log(1/\delta)$ can be super-linear under heavy tailed noise Figures 3 and 14, strongly suggesting that the momentum version of NSGD (NSGD-M) may not possess a high probability convergence as NSGD.

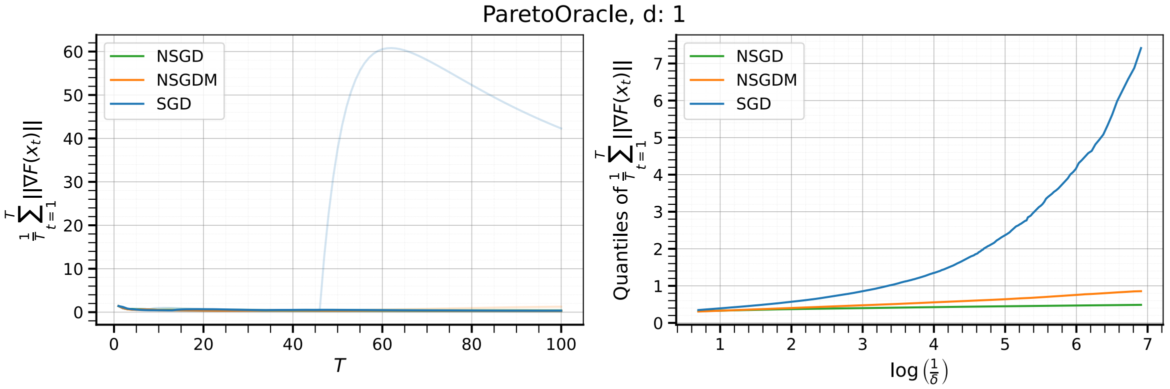


Figure 12: Heavy tailed noise with infinite variance. Pareto with $p = 1.5$.

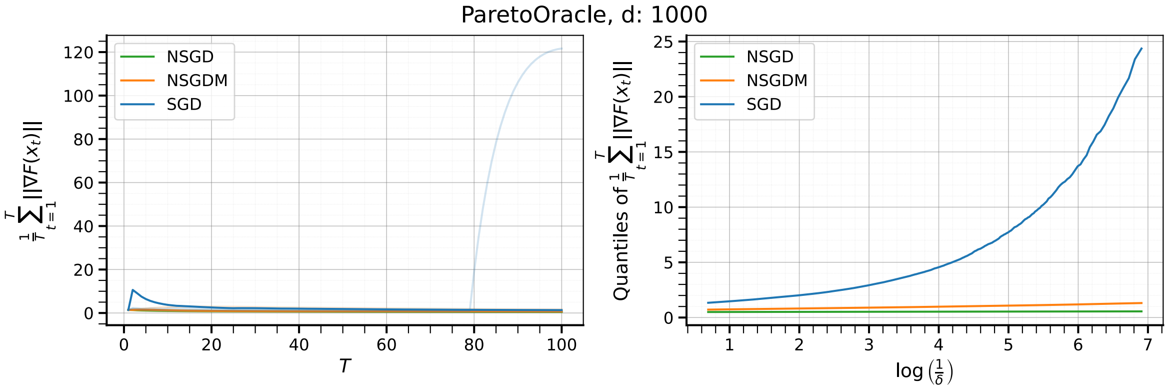


Figure 13: Heavy tailed noise with infinite variance. Pareto with $p = 1.5$. See Figure 14 for the same experiment, but without SGD on the plot.

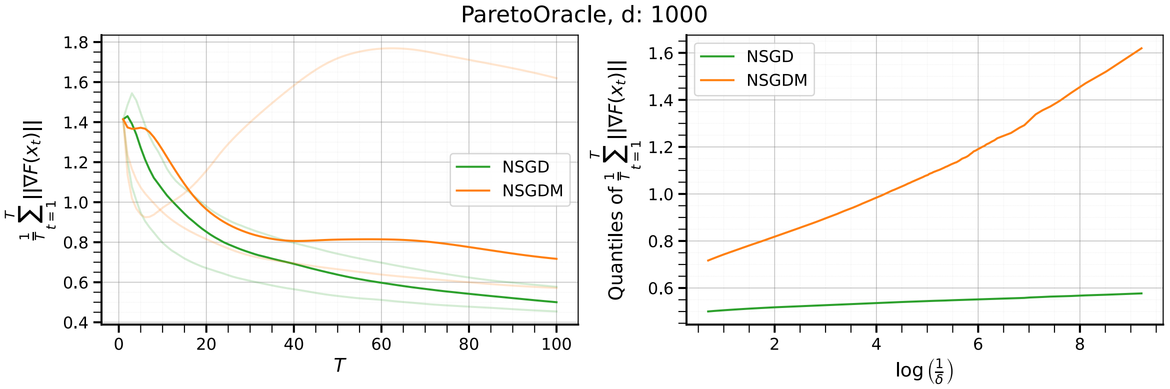


Figure 14: Heavy tailed noise with infinite variance. Pareto with $p = 1.5$.