OPT2024: 16th Annual Workshop on Optimization for Machine Learning

From Gradient Clipping to Normalization for Heavy Tailed SGD

Florian Hübler FLORIAN.HUEBLER@INF.ETHZ.CH
Ilyas Fatkhullin ILYAS.FATKHULLIN@AI.ETHZ.CH
Niao He NIAO.HE@INF.ETHZ.CH
ETH Zurich, Switzerland

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 in this setting 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 sampling complexity upper bounds for the method are sub-optimal in nearly all parameters. To address these issues, we study convergence of Normalized SGD (NSGD). First, we establish a parameter-free sample complexity guarantee for NSGD of $\widetilde{\mathcal{O}}\left(\frac{\Delta_1^4+L^4}{\varepsilon^4}+\left(\frac{\sigma}{\varepsilon}\right)^{\frac{2p}{p-1}}\right)$ to find an ε-stationary point, where $p\in(1,2]$ is the tail index of heavy tailed noise distribution. In the setting where all problem parameters are known, we show this complexity can be improved to $\mathcal{O}\left(\frac{\Delta_1 L}{\varepsilon^2}+\frac{\Delta_1 L}{\varepsilon^2}\left(\frac{\sigma}{\varepsilon}\right)^{\frac{p}{p-1}}\right)$, 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.

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 possibly non-convex, L-smooth objective function and ξ is a random variable from an unknown distribution $\mathcal{D}$. Such problems are pervasive in machine learning [6], where exact gradients are often infeasible to obtain, necessitating reliance on stochastic gradients.

Traditionally, stochastic gradient methods assume bounded variance of the gradient noise. Under this assumption, it is well known that first order algorithms require at least $\Omega\big(L\Delta_1\sigma^2\varepsilon^{-4}\big)$ oracle calls in the worst case to find an ε-stationary point, i.e. $x\in\mathbb{R}^d$ with $\mathbb{E}\left[\|\nabla F(x)\|\right]\leq\varepsilon$ [2]. Here Δ_1 denotes the initialization gap and σ^2 the variance. Stochastic Gradient Descent (`SGD`) with an appropriately chosen step-size achieves this optimal oracle complexity [18].

However, recent observations in machine learning suggest that the bounded variance assumption may be too restrictive. Empirical evidence from image classification [53], training large language models [58], and policy optimization in reinforcement learning [17] indicates that stochastic gradients often follow heavy-tailed distributions. These findings challenge the standard assump-

tion, suggesting a shift toward models where only the p-th central moment of the gradient noise is bounded for some $p \in (1, 2]$, i.e. $\mathbb{E}\left[\|\nabla f(x, \xi) - \nabla F(x)\|^p\right] \leq \sigma^p$ with $\sigma = \sigma(p) \geq 0$. Specifically, the aforementioned works estimate the tail index of stochastic gradients using statistical tests (e.g., [41]) and find $p < 2$; another set of experiments assumes $p = 2$ and estimates the variance σ^2, which turns out to be too large to be a useful upper bound.

While SGD is optimal when the variance is finite and is often the method of choice when the noise is benign, the empirical evidence suggests that adaptive algorithms are crucial in the regime with heavy tailed noise [58]. All works which are able to prove convergence under these conditions employ the gradient clipping mechanism [11, 20, 21, 27, 30, 32, 45, 50, 58]. This mechanism replaces the gradient oracle in optimization algorithms by its clipped counterpart

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

Perhaps the most popular scheme is Clip-SGD, which iterates: $x_{t+1} = x_t - \eta_t \widehat{\nabla} f(x_t, \xi_t)$.

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 clipping under the (p-BCM) assumption hinge on using a large, p-dependent sequence of increasing clipping thresholds, e.g. $\gamma_t = \gamma \cdot t^{\frac{1}{3p-2}}$ [11, 30, 45, 46, 58]. This choice of clipping thresholds is based on the following two ideas. First, clipping allows 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 [40, 55, 59]. 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 Figure 1, and notice the same effect on simpler, synthetic examples in Figure 2. This aggressive clipping behaviour essentially transforms Clip-SGD into a variant of Normalized SGD:

$$x_{t+1} = x_t - \eta_t \frac{g_t}{\|g_t\|}, \tag{NSGD}$$

where $g_t = \nabla f(x_t, \xi_t)$ in this case. It is worth noting, however, that unlike Clip-SGD, NSGD only requires tuning a single parameter η, highlighting its simplicity in comparison. More details on the experiments can be found in Appendix B.

Need for tuning. To our knowledge, all existing convergence results for clipping require knowledge of all problem parameters to set the clipping thresholds $\{\gamma_t\}_{t \geq 1}$ and other hyper-parameters of the underlying algorithm. As these problem-dependent parameters are not known in practice, this corresponds to the need for extensive hyper-parameter tuning. In particular, for Clip-SGD, there are 2 hyper-parameters which potentially require tuning. In Appendix B we empirically observe

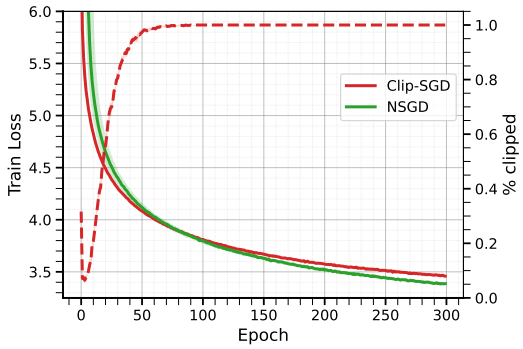

Figure 1: Language Modelling

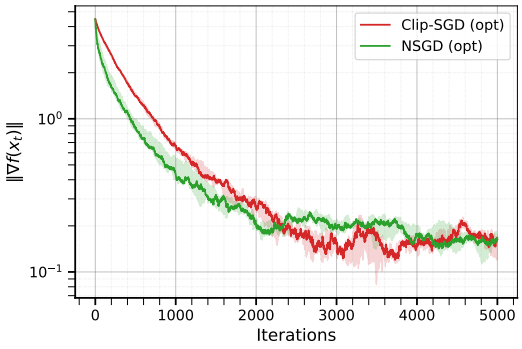

Figure 2: Toy Example

Both plots consider `Clip-SGD` and `NSGD` with tuned parameters. Solid lines correspond to the left y-axis. The dashed line shows the % of clipped iterations per epoch, corresponding to the right y-axis. Shaded areas represent the minimal and maximal value within 5 seeds around the median.

that even while requiring extensive hyper-parameter tuning, `Clip-SGD` is not able to outperform vanilla `NSGD` on language modelling tasks.

Suboptimal sample complexities. None of the existing convergence analysis of non-convex `Clip-SGD` (and its variants) achieve the sample complexity lower bound $\Omega\left(\frac{\Delta_1 L}{\varepsilon^2} + \frac{\Delta_1 L}{\varepsilon^2}\left(\frac{\sigma}{\varepsilon}\right)^{\frac{p}{p-1}}\right)$ [58] 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` under heavy tailed noise. We summarize our contributions as follows:

In Section 3.1 we demonstrate that `NSGD` converges for any tail index $p \in (1, 2]$ without any knowledge of problem specific parameters, achieving an $\widetilde{\mathcal{O}}\left(\frac{\Delta_1^4 + L^4}{\varepsilon^4} + (\sigma/\varepsilon)^{\frac{2p}{p-1}}\right)$ oracle complexity to find an ε-stationary point. When problem parameters are known, this sample complexity can be further improved to $\mathcal{O}\left(\frac{\Delta_1 L}{\varepsilon^2} + \frac{\Delta_1 L}{\varepsilon^2}\left(\frac{\sigma}{\varepsilon}\right)^{\frac{p}{p-1}}\right)$, which matches with the mini-max lower bound for the class of first-order algorithms under our assumptions [58, Appendix G]. To our knowledge, `NSGD` is the first algorithm to achieve either of these properties in the heavy tail regime $p < 2$.

In Section 3.2, we obtain the first high probability convergence guarantee with a mild $\log(1/\delta)$ dependence for `NSGD` without requiring additional clipping, which is novel even in the case $p = 2$.

1.3. Related Work

This section summarises closely related works. An extended version can be found in Appendix A.

Gradient Clipping. Gradient clipping is widely used to stabilize the training in various fields of machine learning [47, 51, 58]. Recently a number of works provide convergence guarantees for various algorithms when used with clipping in different settings [13, 20, 21, 34, 42, 49]. In the non-convex setting, Zhang et al. [58] study in-expectation and [45, 50] investigate high probability convergence of `Clip-SGD` under the (p-BCM) assumption.

Normalized SGD. NSGD was first proposed and analysed for deterministic convex functions by Nesterov [43, 44]. In the non-convex bounded variance setting, Cutkosky and Mehta [10] prove that incorporating Polyak's momentum into NSGD removes the necessity of large batchsizes. Yang et al. [57] furthermore prove parameter-free convergence of the algorithm and a lower bound for NSGD without momentum in this setting. In the non-convex heavy-tailed setting, [11, 35] study NSGD with momentum and extra gradient clipping. Cutkosky and Mehta [11] require the bounded non-central moment assumption, i.e., $\mathbb{E}\left[\|\nabla f(x,\xi)\|^p\right] \leq G^p$, which is stronger than the (p-BCM) assumption. Liu et al. [35] on the other hand require an additional almost sure individual smoothness assumption. Both works suffer from the limitations of clipping discussed in Section 1.1.

2. Preliminaries

Throughout this paper, $d \in \mathbb{N}_{\geq 1}$ denotes the dimension, $F\colon \mathbb{R}^d \to \mathbb{R}$ the objective and $\nabla f(\cdot,\cdot)$ the gradient oracle. Unless stated otherwise, $L \geq 0$ denotes the L-smoothness parameter and $\eta_t > 0$ the stepsizes. We use the common conventions $\mathbb{N} = \{0, 1, \ldots\}$, $[n] = \{1, 2, \ldots, n\}$ and that empty sums and products are given by their corresponding neutral element.

Building on established work in stochastic optimization [2, 18], we employ the following two standard assumptions in various results of this study.

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

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

Instead of the traditional 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 $\sigma \geq 0$ such that*

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

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

Note that for $p \in (1, 2)$, by Jensen's inequality, (p-BCM) is strictly weaker than the usual bounded variance assumption.

3. Main Result

In this section, we present our convergence results for normalized stochastic gradient methods (NSGD) under the (p-BCM) assumption. In Section 3.1 we examine the parameter-free, as well as the performance for optimally tuned parameters. In Section 3.2, we derive a high-probability convergence result for `minibatch-NSGD`.

3.1. Normalized SGD can Handle Heavy Tailed Noise

Firstly, we theoretically confirm the robustness of NSGD to misspecification of p and other problem parameters. This is in stark contrast to current Clip-SGD analyses, which heavily depends on the knowledge of all parameters. We state the following result for the momentum version of NSGD. The same result for the minibatch version, i.e. when considering NSGD with $g_t = \frac{1}{B}\sum_{j=1}^{B} \nabla f\left(x_t, \xi_t^{(j)}\right)$ where $\xi_t^{(1)}, \dots, \xi_t^{(B)} \overset{\text{i.i.d.}}{\sim} \xi_t$, can be found in Appendix E.

Theorem 1 (Parameter-Free 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 $\widetilde{\mathcal{O}}\left((\Delta_1 + L)T^{-1/4} + \sigma T^{-\frac{p-1}{2p}}\right)$.

We offer a few remarks on this result. Firstly, it might initially seem that the upper bound blows up as $p \to 2$. However, by L'Hôspital's rule, we have $\lim_{p\to 2} \frac{4p}{2-p}\left(T^{(2-p)/4p} - 1\right) = \log(T)$, making the second statement formally sound. Secondly, this result corresponds to an oracle complexity of $\widetilde{\mathcal{O}}\left(\left(\Delta_1^4 + L^4\right)\varepsilon^{-4} + (\sigma/\varepsilon)^{\frac{2p}{p-1}}\right)$. While this complexity does not match the lower bound for parameter-dependent algorithms, it is — to the best of our knowledge — the first fully parameter-agnostic convergence result under the (p-BCM) assumption. Finally, when plugging $p = 2$ in our result, the previous result for the bounded variance setting is reconstructed [57].

Given that the oracle complexity of the previous result is not optimal for parameter-dependent algorithms, it is intuitive to ask whether it can be improved when having knowledge of problem parameters. Therefore, the following corollary considers optimal parameters for NSGD-M with constant parameters. We emphasize that the same analysis can be carried out with decreasing stepsizes, at the mild cost of a multiplicative $\log(T)$ term.

Theorem 2 (Optimal Oracle Complexity) *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 $\beta_1 := 0, \beta_t \equiv \beta := 1 - \min\left\{1, \left(\frac{\Delta_1 L}{\sigma^2 T}\right)^{\frac{p}{3p-2}}\right\}$ for $t \geq 2$ and $\eta_t \equiv \sqrt{\frac{\Delta_1(1-\beta)}{LT}}$ satisfy*

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

This result implies an oracle complexity of $\mathcal{O}\left(\frac{\Delta_1 L}{\varepsilon^2} + \frac{\Delta_1 L}{\varepsilon^2}\left(\frac{\sigma}{\varepsilon}\right)^{\frac{p}{p-1}}\right)$ to reach an ε-stationary point in expectation, which is optimal in all problem parameters under the given assumptions for the class of first-order methods [58]. To the best of our knowledge, this is the first result to match this lower bound exactly in the non-convex case. For comparison, Zhang et al. [58] derived convergence

for Clip-SGD with a rate[1] of

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

which is suboptimal in all parameters besides T. In fact, as $p \to 1$, their upper bound blows up, whereas the optimal bound turns constant. In contrast, our result is optimal in all parameters and remains bounded as $p \to 1$. Additionally, our result recover those in [10][2] with improved constants when $p = 2$.

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, in-probability results of the form *with probability at least* $1-\delta$*, a single run achieves an* ε*-stationary point* are more desirable. While the Markov inequality can convert in-expectation results to in-probability results, the poor dependence on δ renders this result impractical. Therefore, the gold standard are so called high-probability results with a mild $\log(1/\delta)$ dependence.

To achieve such results, existing literature relies on either light tail noise assumptions [31, 36], or the gradient clipping mechanism [11, 20, 42, 45]. The following theorem confirms that NSGD also has a high-probability convergence guarantee. To the best of our knowledge, we are the first work to show such result for vanilla NSGD under weak noise assumptions.[3]

Theorem 3 *Assume (Lower Boundedness), (L-smoothness) and (p-BCM) with* $p \in (1,2]$*. Then the iterates generated by* NSGD *with* $g_t = \frac{1}{B_t}\sum_{i=1}^{B_t}\nabla f(x_t,\xi_t^{(i)})$*, where* $\xi_t^{(1)},\ldots,\xi_t^{(B_t)} \overset{i.i.d.}{\sim} \xi_t$*, and parameters* $\eta_t \equiv \sqrt{\frac{\Delta_1}{LT}}$ *and* $B_t \equiv \max\left\{1, \left(\frac{\sigma^2 T}{\Delta_1 L}\right)^{\frac{p}{2p-2}}\right\}$ *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 an oracle complexity of* $\widetilde{\mathcal{O}}\left(\frac{\Delta_1 L}{\varepsilon^2}+\frac{\Delta_1 L}{\varepsilon^2}\left(\frac{\sigma}{\varepsilon}\right)^{\frac{p}{p-1}}\right)$.

As high-probability guarantees imply in-expectation guarantees by integration, his result is again optimal in Δ_1, L, σ and ε. We are not aware of any lower bounds specifying the optimal δ-dependence. For comparison, for Clip-SGD Nguyen et al. [45] proved a rate of

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(x_t)\| \leq \widetilde{\mathcal{O}}\left(\left(\sigma^{\frac{p}{2(p-1)}}(\Delta_1 L)^{\frac{p-2}{4(p-1)}}+\left(\Delta_1 L\sigma^2\right)^{1/4}\right)T^{-\frac{p-1}{3p-2}}+(\Delta_1 L)^{1/4}T^{-\frac{p-1/2}{3p-2}}\right),$$

which is suboptimal in all parameters besides T in the stochastic case. In fact, as $p \to 1$, their upper bound blows up whenever $\sigma \geq \Delta_1 L$, whereas the optimal bound turns constant. In the deterministic case, even the dependence on T is suboptimal. In contrast, our result is noise adaptive in the sense

1. We ignore non-leading terms and simplify the rate in their favour.
2. 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.
3. Note that there are works such as [3] that provide high-probability guarantees for NSGD *under stronger noise assumptions*. Other works such as [11] require an additional clipping step on top of normalization.

that, for $\sigma = 0$, the optimal deterministic oracle complexity is obtained. More remarks can be found in Appendix D.3.

4. Conclusion

This work analyses 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 as Theorem 2 is first convergence result that tightly matches the corresponding lower bound [58, Appendix G]. Lastly, our results for parameter-free NSGD suggest the robustness of the algorithm to misspecification of its parameters.

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 can be achieved by any algorithm without knowledge of problem parameters.

Acknowledgements

The work is supported by ETH research grant, ETH AI Center Doctoral Fellowship and Swiss National Science Foundation (SNSF) Project Funding No. 200021-207343. We furthermore want to thank anonymous reviewers for helpful suggestions to improve the clarity of the work.

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[2] Yossi Arjevani, Yair Carmon, John C. Duchi, Dylan J. Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1-2):165–214, 2023.

[3] Aleksandar Armacki, Pranay Sharma, Gauri Joshi, Dragana Bajovic, Dusan Jakovetic, and Soummya Kar. High-probability Convergence Bounds for Nonlinear Stochastic Gradient Descent Under Heavy-tailed Noise. *arXiv preprint arXiv:2310.18784*, 2023.

[4] Anas Barakat, Ilyas Fatkhullin, and Niao He. Reinforcement learning with general utilities: Simpler variance reduction and large state-action space. In *International Conference on Machine Learning*, pages 1753–1800, 2023.

[5] Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. Signsgd: Compressed optimisation for non-convex problems. In *International Conference on Machine Learning*, pages 560–569. PMLR, 2018.

[6] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM review*, 60(2):223–311, 2018.

[7] Sébastien Bubeck, Nicolo Cesa-Bianchi, and Gábor Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013.

[8] Semih Cayci and Atilla Eryilmaz. Provably robust temporal difference learning for heavy-tailed rewards. *Advances in Neural Information Processing Systems*, 36, 2024.

[9] Michael Crawshaw, Mingrui Liu, Francesco Orabona, Wei Zhang, and Zhenxun Zhuang. Robustness to unbounded smoothness of generalized signsgd. *Advances in Neural Information Processing Systems*, 35:9955–9968, 2022.

[10] Ashok Cutkosky and Harsh Mehta. Momentum improves normalized SGD. In *International Conference on Machine Learning*, volume 119, pages 2260–2268. PMLR, 2020.

[11] Ashok Cutkosky and Harsh Mehta. High-probability bounds for non-convex stochastic optimization with heavy tails. *Advances in Neural Information Processing Systems*, 34:4883–4895, 2021.

[12] Yan Dai, Kwangjun Ahn, and Suvrit Sra. The crucial role of normalization in sharpness-aware minimization. *Advances in Neural Information Processing Systems*, 36, 2024.

[13] Damek Davis, Dmitriy Drusvyatskiy, Lin Xiao, and Junyu Zhang. From low probability to high confidence in stochastic convex optimization. *The Journal of Machine Learning Research*, 22(1):2237–2274, 2021.

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

[15] Ilyas Fatkhullin, Anas Barakat, Anastasia Kireeva, and Niao He. Stochastic policy gradient methods: Improved sample complexity for Fisher-non-degenerate policies. In *International Conference on Machine Learning*, pages 9827–9869, 2023.

[16] Swetha Ganesh, Washim Uddin Mondal, and Vaneet Aggarwal. Variance-reduced policy gradient approaches for infinite horizon average reward markov decision processes. *arXiv preprint arXiv:2404.02108*, 2024.

[17] Saurabh Garg, Joshua Zhanson, Emilio Parisotto, Adarsh Prasad, Zico Kolter, Zachary Lipton, Sivaraman Balakrishnan, Ruslan Salakhutdinov, and Pradeep Ravikumar. On proximal policy optimization's heavy-tailed gradients. In *International Conference on Machine Learning*, pages 3610–3619. PMLR, 2021.

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

[19] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[20] Eduard Gorbunov, Marina Danilova, and Alexander Gasnikov. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. *Advances in Neural Information Processing Systems*, 33:15042–15053, 2020.

[21] Eduard Gorbunov, Abdurakhmon Sadiev, Marina Danilova, Samuel Horváth, Gauthier Gidel, Pavel Dvurechensky, Alexander Gasnikov, and Peter 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.

[22] Elad Hazan, Kfir Levy, and Shai Shalev-Shwartz. Beyond convexity: Stochastic quasi-convex optimization. *Advances in Neural Information Processing Systems*, 28, 2015.

[23] Florian Hübler, Junchi Yang, Xiang Li, and Niao He. Parameter-Agnostic Optimization under Relaxed Smoothness. In *International Conference on Artificial Intelligence and Statistics*, pages 4861–4869, 2024.

[24] Dusan Jakovetić, Dragana Bajović, Anit Kumar Sahu, Soummya Kar, Nemanja Milosević, and Dusan 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.

[25] Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian Stich, and Martin Jaggi. Error feedback fixes SignSGD and other gradient compression schemes. In *International Conference on Machine Learning*, 2019.

[26] Anastasia Koloskova, Hadrien Hendrikx, and Sebastian U Stich. Revisiting gradient clipping: Stochastic bias and tight convergence guarantees. In *International Conference on Machine Learning*, pages 17343–17363. PMLR, 2023.

[27] Nikita Kornilov, Ohad Shamir, Aleksandr Lobanov, Darina Dvinskikh, Alexander Gasnikov, Innokentiy Shibaev, Eduard Gorbunov, and Samuel Horváth. Accelerated zeroth-order method for non-smooth stochastic convex optimization problem with infinite variance. *Advances in Neural Information Processing Systems*, 36, 2024.

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

[29] Kfir Y Levy. The power of normalization: Faster evasion of saddle points. *arXiv preprint arXiv:1611.04831*, 2016.

[30] Shaojie Li and Yong Liu. High Probability Analysis for Non-Convex Stochastic Optimization with Clipping. In *ECAI 2023*, pages 1406–1413. IOS Press, 2023.

[31] Xiaoyu Li and Francesco Orabona. A high probability analysis of adaptive sgd with momentum. *arXiv preprint arXiv:2007.14294*, 2020.

[32] Langqi Liu, Yibo Wang, and Lijun Zhang. High-Probability Bound for Non-Smooth Non-Convex Stochastic Optimization with Heavy Tails. In *International Conference on Machine Learning*, pages 32122–32138, 2024.

[33] Sijia Liu, Pin-Yu Chen, Xiangyi Chen, and Mingyi Hong. Signsgd via zeroth-order oracle. In *International Conference on Learning Representations*, 2019.

[34] Zijian Liu and Zhengyuan Zhou. Stochastic nonsmooth convex optimization with heavy-tailed noises. *arXiv preprint arXiv:2303.12277*, 2023.

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

[36] Liam Madden, Emiliano Dall'Anese, and Stephen 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.

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

[38] Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of English: the penn treebank. *Comput. Linguist.*, 19(2):313–330, jun 1993. ISSN 0891-2017.

[39] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer Sentinel Mixture Models. In *International Conference on Learning Representations*, 2017.

[40] Stephen Merity, Nitish Shirish Keskar, and Richard Socher. Regularizing and Optimizing LSTM Language Models. In *International Conference on Learning Representations*, 2018.

[41] Mohammad Mohammadi, Adel Mohammadpour, and Hiroaki Ogata. On estimating the tail index and the spectral measure of multivariate α-stable distributions. *Metrika*, 78(5):549–561, 2015.

[42] Alexander V Nazin, Arkadi S Nemirovsky, Alexandre B Tsybakov, and Anatoli B Juditsky. Algorithms of robust stochastic optimization based on mirror descent method. *Automation and Remote Control*, 80:1607–1627, 2019.

[43] Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

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

[45] Ta Duy Nguyen, Thien H Nguyen, Alina Ene, and Huy Nguyen. Improved convergence in high probability of clipped gradient methods with heavy tailed noise. In *Advances in Neural Information Processing Systems*, volume 36, pages 24191–24222, 2023.

[46] Ta Duy Nguyen, Thien Hang Nguyen, Alina Ene, and Huy Le Nguyen. High probability convergence of clipped-sgd under heavy-tailed noise. *arXiv preprint arXiv:2302.05437*, 2023.

[47] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *International Conference on Machine Learning*, pages 1310–1318. Pmlr, 2013.

[48] Boris Teodorovich Polyak and Yakov Zalmanovich Tsypkin. Adaptive estimation algorithms: convergence, optimality, stability. *Avtomatika i telemekhanika*, (3):71–84, 1979.

[49] Nikita Puchkin, Eduard Gorbunov, Nickolay Kutuzov, and Alexander Gasnikov. Breaking the heavy-tailed noise barrier in stochastic optimization problems. In *International Conference on Artificial Intelligence and Statistics*, pages 856–864. PMLR, 2024.

[50] Abdurakhmon Sadiev, Marina Danilova, Eduard Gorbunov, Samuel Horváth, Gauthier Gidel, Pavel Dvurechensky, Alexander Gasnikov, and Peter Richtárik. High-probability bounds for stochastic optimization and variational inequalities: the case of unbounded variance. In *International Conference on Machine Learning*, 2023.

[51] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[52] Haichao Sha, Yang Cao, Yong Liu, Yuncheng Wu, Ruixuan Liu, and Hong Chen. Clip body and tail separately: High probability guarantees for DPSGD with heavy tails. *arXiv preprint arXiv:2405.17529*, 2024.

[53] Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, pages 5827–5837. PMLR, 2019.

[54] Tao Sun, Qingsong Wang, Dongsheng Li, and Bao Wang. Momentum ensures convergence of signsgd under weaker assumptions. In *International Conference on Machine Learning*, pages 33077–33099. PMLR, 2023.

[55] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*, 2023.

[56] Bengt von Bahr and Carl-Gustav 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.

[57] Junchi Yang, Xiang Li, Ilyas Fatkhullin, and Niao 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.

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

[59] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open Pre-trained Transformer Language Models. *arXiv preprint arXiv:2205.01068*, 2022.

[60] Shen-Yi Zhao, Yin-Peng Xie, and Wu-Jun Li. On the convergence and improvement of stochastic normalized gradient descent. *Science China Information Sciences*, (3):132103, 2021.

Contents

1 Introduction **1**
1.1 Drawbacks of Gradient Clipping Theory . . . 2
1.2 Our Contributions . . . 3
1.3 Related Work . . . 3

2 Preliminaries **4**

3 Main Result **4**
3.1 Normalized SGD can Handle Heavy Tailed Noise . . . 5
3.2 Convergence with High-Probability . . . 6

4 Conclusion **7**

A Extended Related Work **12**

B Experiments **14**
B.1 Toy Example . . . 14
B.2 Language Modelling . . . 14

C Technical Results **17**

D Missing Proofs **19**
D.1 Proofs of Section 3.1 . . . 19
D.2 Proofs of Section 3.2 . . . 23
D.3 Remarks on Theorem 3 . . . 25

E Mini-Batch NSGD **26**

Appendix A. Extended Related Work

Gradient clipping is widely used to stabilize the training in various fields of machine learning [47, 51, 58]. Recently a number of works provide convergence guarantees for Clip-SGD and its variants in different settings, e.g., [13, 20, 21, 34, 42, 49] to name a few. However, the results in the non-convex stochastic setting are relatively scarce. In particular, Zhang et al. [58] study in-expectation and Nguyen et al. [45], Sadiev et al. [50] 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 [37] 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, [26] 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., [7, 8]. Moreover, Clip-SGD is the key mechanism to ensure differential privacy [1, 52].

Normalized SGD was first proposed by Nesterov [43, 44] and analyzed in the deterministic convex case. Later the analysis was extended to smooth [28] and stochastic [22] settings. In the non-convex case, Cutkosky and Mehta [10] show how to remove large mini-batch requirement for NSGD by incorporating Polyak's momentum. Later, Yang et al. [57] derive a tight lower bound for NSGD without momentum and Hübler et al. [23] study the parameter agnosticity of momentum NSGD under a relaxed smoothness assumption. In a different line of works, Levy [29] 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 [11, 35], 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 γ_t. Moreover, Cutkosky and Mehta [11] 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 [35] 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 distribution has a probability density that is symmetric and strictly positive in a neighborhood of zero [3, 24, 48]. Under this assumption, they study SGD type methods with general non-linearities, 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 [12], and the variants of NSGD showed an impressive empirical and theoretical success in more structured non-convex problems in RL [4, 15, 16]. However, these works are also restricted to benign BV noise assumption. Some recent works also make connections with SignSGD algorithm [5, 9, 25], 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 [33, 54].

Appendix B. Experiments

In this section, we present experiments designed to empirically motivate and validate the theoretical findings of this paper.

B.1. Toy Example

To better understand the behaviour of `Clip-SGD`, we conduct a small experiment on the function

$$f(x,\xi) = \frac{1}{2}\|x\|^2 + \langle x, \xi\rangle,$$

where $d = 10$ and ξ is a symmetrized Pareto vector with tail index $p = 1.8$. We pick the step-size for `Clip-SGD` according to theory (for $p = 2$), i.e., $\gamma_t = \gamma \cdot \sqrt{t}$, $\eta_t = \eta/\sqrt{t}$, and tune γ and η over a wide range of possible values. We find that the optimal choice for our problem is $\gamma = 0.001$, $\eta = 100$, which makes `Clip-SGD` trigger clipping at every iteration. This effectively reduces `Clip-SGD` to `NSGD`. For comparison we show the convergence of tuned `Clip-SGD` and `NSGD` (with $g_t = \nabla f(x_t, \xi_t)$, $\eta_t = \eta/\sqrt{t}$) in Figure 2 and find that indeed the two behave very similarly. Notice, however, that unlike `Clip-SGD`, `NSGD` has only one tuning parameter, the step-size η.

B.2. Language Modelling

Since heavy tails have prominently been observed in language modelling tasks [58], our experiments target this task.

Experimental Setup. We conduct training on the Penn Treebank (PTB) [38] and WikiText-2 [39] datasets using the AWD-LSTM architecture [40]. Hyperparameters of the model were chosen according to [40]. To observe the exact optimization behaviour of algorithms, the averaging mechanic of the model was disabled.

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. [40] 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 5 seeds $0, 1970, 2000, 2024, 2112$.

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 used for preliminary experiments and parameter-tuning, 180 for the final experiments.

The AWD-LSTM [40] is released under a BSD 3-Clause License, the Penn Treebank dataset [38] under the LDC User Agreement for Non-Members and the WikiText-2 dataset [39] under the Creative Commons BY-SA 3.0 license.

From Clipping to Normalization. Figures 3 and 4 show the behaviour of `Clip-SGD` and `NSGD` on both datasets with their corresponding tuned parameters. We want to discuss two observations.

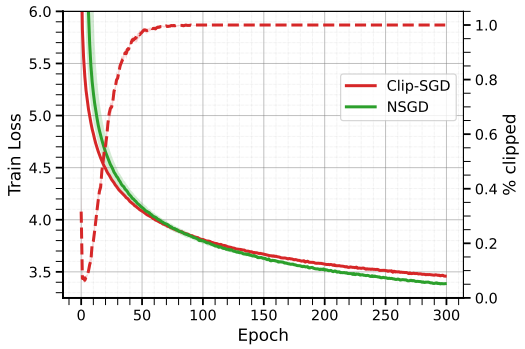

Figure 3: PTB

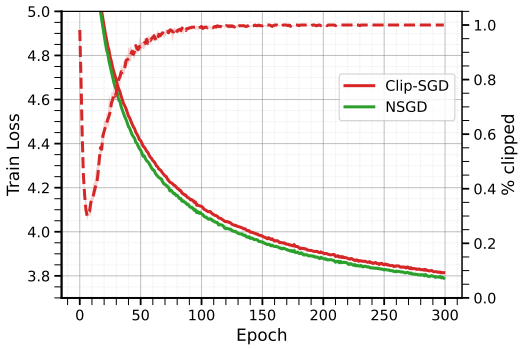

Figure 4: WikiText-2

Both 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 shows the percentage of clipped iterations per epoch by `Clip-SGD`, corresponding to the right y-axis. Shaded areas represent the minimal and maximal value within 5 seeds, the line the median.

Firstly, and maybe surprisingly, the dashed line represents the percentage of clipped stochastic gradients per epoch. One can see that in both cases, `Clip-SGD` clips every iteration after a certain epoch, becoming equivalent to `NSGD`. Secondly, it can be noted that both algorithms behave similar when measured with their corresponding training loss, depicted with solid lines. Both these observations suggest that the empirical behavior of `Clip-SGD` is very close to `NSGD` and the opposite of what its theory would suggest.

Reasons for the Clipping Behaviour. We first want to understand why a) the percentage of clipped gradients increases over time, before b) eventually clipping all iterations. Therefore Figure 5 and Figure 6 examine the average batch-gradient norm per epoch, i.e. $\frac{1}{B}\sum_{t=t_0}^{t_0+B-1}\|g_t\|$ where the epoch consists of B mini-batches and starts at iteration t_0. The plot suggests that, while the training loss decreases, the stochastic gradient norms increase — a phenomenon that has been well known for many years [19, Chapter 8]. Therefore, while surprising at first, the increasing clipping percentage was to be expected in hindsight, answering question a). The fact that the optimal clipping threshold is exactly at the scale of the is more surprising and, to the best of our knowledge, has not yet been observed in the literature.

To gain a deeper understanding of the clipping behavior, we fix the optimal stepsizes and compare different clipping threshold in Figure 7 and Figure 8.

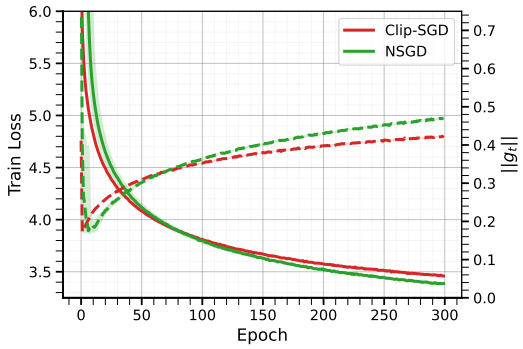

Figure 5: PTB

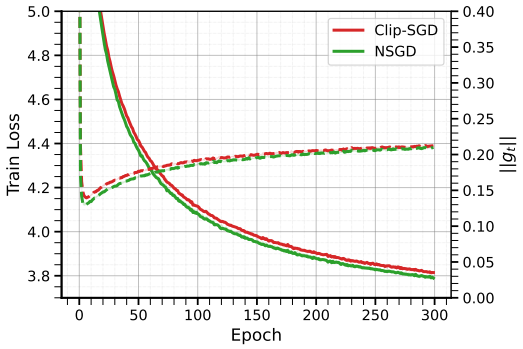

Figure 6: WikiText-2

Both plots consider `Clip-SGD` and `NSGD` with tuned parameters. Solid lines represent the training loss and correspond to the left y-axis. Dashed lines show the average stochastic gradient norm per epoch, corresponding to the right y-axis. Shaded areas represent the minimal and maximal value within 5 seeds, the line the median.

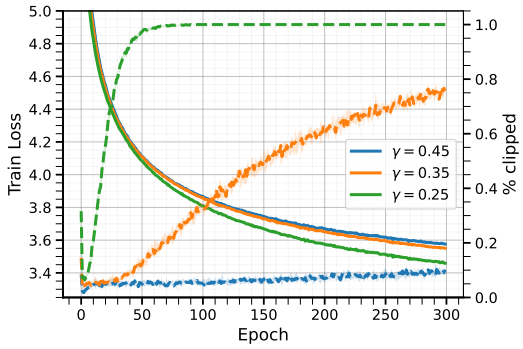

Figure 7: PTB dataset

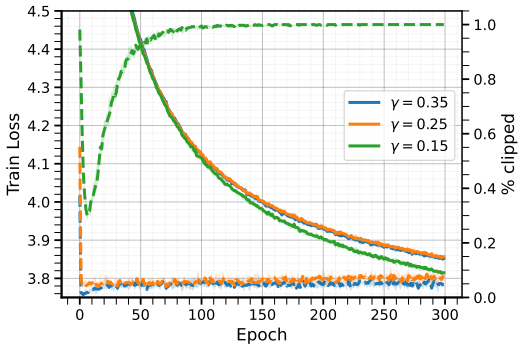

Figure 8: WikiText dataset

Comparison of different clipping thresholds for `Clip-SGD`. All parameters besides the clipping-threshold are fixed as described in Appendix B.

Appendix C. Technical Results

This section contains various technical results required for our results. We start with two lemmas that arise due to the normalization in NSGD.

Lemma 4 *For all $a, b \in \mathbb{R}^d$ with $a \neq 0$ 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 5 (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 $\mu \in \mathbb{R}^d \setminus \{0\}$, $\sigma \coloneqq \mathbb{E}[\|X - \mu\|]$ 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 4 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 6 (see [23, 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 the momentum gradient estimator, we will require the following inequality.

Lemma 7 (see [56]) *Let $X_1, \ldots, X_n \in \mathbb{R}^d$ be a sequence of random vectors, $S_n \coloneqq \sum_{j=1}^n X_j$ and $p \in [1,2]$. Assume $\mathbb{E}[\|X_j\|^p] < \infty$ and $\mathbb{E}[X_{j+1} \mid S_j] = 0$ a.s. for all $j \in [n]$. Then*

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

In order to apply Lemma 7, we require the following lemma on conditional expectations.

Lemma 8 (c.f. Durrett [14, 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*

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

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

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

This completes the proof. ■

Finally, we will require the following Martingale concentration inequality for high-probability guarantees.

Lemma 9 (see [31, Lemma 1]) *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 $T \in \mathbb{N}$,*

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

Appendix D. Missing Proofs

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). Furthermore, we denote the natural filtration of our iterates by $\mathcal{F}_t \coloneqq \sigma(g_1, \dots, g_t)$.

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 angle between g_t and $\nabla F(x_t)$. As this angle will play a crucial role in our high-probability result, we will provide our version of the descent lemma and its proof below.

Lemma 10 (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.

D.1. Proofs of Section 3.1

We start with a unified result for normalized algorithms. This method does not specify the exact gradient oracle, allowing to incorporate different gradient estimators and noise assumptions afterward.

Proposition 11 (c.f. [10, 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

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla F(x_t)\|\right] \leq \frac{\Delta_1}{\eta T} + \frac{\eta L}{2} + 2\sigma. \tag{3}$$

While the result stems from [10], we want to make a few remarks on this result. Firstly note that Proposition 11 does not require decreasing stepsizes or any other similar assumption. Secondly, the weights $w_t \coloneqq \frac{\eta_t}{\sum_{\tau=1}^{T}\eta_\tau}$ define a discrete probability distribution over the iterates of the method, allowing to sample from the iterates to find a point $\bar{x}_T$, which is an ε-stationary point in expectation. Lastly, Proposition 11 does not make any assumptions on the noise. In particular, the stochastic gradient oracle is not assumed to be unbiased. Instead the upper bound depends on the first central moment, which is typically controlled in one of three ways. Empirical studies have observed that noise often becomes more benign along the training trajectory by itself [58, Figure 2]. From an algorithmic standpoint, it is well known that noise can be controlled by using minibatches [60] or momentum [10] under the bounded variance assumption.

We provide a slightly different proof of Proposition 11 when compared to [10] below.

Proof of Proposition 11. 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 10, we have

$$\sum_{t=1}^{T}\eta_t\phi_t\left\|\nabla F(x_t)\right\| \leq \Delta_1 + \frac{L}{2}\sum_{t=1}^{T}\eta_t^2. \tag{4}$$

Next we apply Lemma 4 to get

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

where we applied our assumption $\sigma_t \geq \mathbb{E}\left[\|g_t - \nabla F(x_t)\|\right]$ in the last inequality. Therefore, by taking expectations in (4), we get

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

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

In the following lemma, we show that momentum is still able to improve NSGD when the noise assumption is relaxed from bounded variance to (p-BCM). In Appendix D.2, Equation (9), we will prove the same result for minibatches, effectively demonstrating that both methods still work under the relaxed noise assumption.

Lemma 12 *Let $\beta_1 = 0$ and assume (L-smoothness), (p-BCM) with $p \in (1, 2]$. Then the iterates generated by NSGD with $g_t = \beta_t g_{t-1} + (1-\beta_t)\nabla f(x_t, \xi_t)$ satisfy*

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

where $\beta_{a:b}$ denotes $\prod_{\kappa=a}^{b}\beta_\kappa$.

The proof of this Lemma mainly follows similar arguments as in the bounded variance setting but applies the more general von Bahr and Essen inequality [56] instead of the manual argument

by [10]. When using constant parameters, this lemma reduces to $\mathbb{E}\left[\|g_t - \nabla F(x_t)\|\right] \leq L\frac{\eta}{1-\beta} + \sigma(1-\beta)^{\frac{p-1}{p}}$. In particular, it recovers the known results for $p = 2$.

Proof (c.f. [10]). To simplify notation we first define

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

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

$$\mathbb{E}\left[\|\mu_t\|\right] \leq \mathbb{E}\left[\left\|\sum_{\tau=1}^{t}\beta_{(\tau+1):t}\alpha_\tau\varepsilon_\tau\right\|\right] + \sum_{\tau=2}^{t}\beta_{\tau:t}\mathbb{E}\left[\|S_\tau\|\right]. \tag{5}$$

The second sum can straight forward 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 7. Therefore let $X_\tau \coloneqq \beta_{(\tau+1):t}\alpha_\tau\varepsilon_\tau$ and note that this sequence does fulfil the required assumptions of the lemma. By applying Jensen's we hence get

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

Plugging these bounds into (5) yields

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

and hence the claim. ■

Next up, we provide the parameter-free convergence result for `NSGD-M`, i.e. `NSGD` with

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

The idea behind the proof follows similar steps as [10], does however require additional attention to the noise term.

Proof of Theorem 1. To shorten notation we define $r \coloneqq 3/4, q \coloneqq 1/2$, and hence $\eta_t = \eta t^{-r}, \beta_t = 1 - t^{-q}$. From Proposition 11 we get

$$\begin{aligned}\sum_{t=1}^{T} \frac{\eta_t}{\sum_{\tau=1}^{T}\eta_\tau}\mathbb{E}\left[\|\nabla F(x_t)\|\right] &\leq \left(\sum_{t=1}^{T}\eta_t\right)^{-1}\left(\Delta_1 + \frac{L}{2}\sum_{t=1}^{T}\eta_t^2 + 2\sum_{t=1}^{T}\eta_t\sigma_t\right)\\ &\leq T^{r-1}\left(\frac{\Delta_1}{\eta} + \frac{3}{2}\eta L + 2\sum_{t=1}^{T}t^{-r}\sigma_t\right),\end{aligned} \tag{6}$$

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 12 and Lemma 6 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 (6) yields

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

where we used that $\frac{4p}{2-p}\left(T^{\frac{2-p}{4p}}-1\right) \geq 1$ for $T \geq 3$ in the last inequality. The second statement follows from the observation $\lim_{q\to 1}\frac{1}{1-q}\left(T^{1-q}-1\right) = \log(T)$. ■

Finally we prove the oracle complexity of tuned NSGD.

Proof of Theorem 2. To shorten the notation we write $\eta_t \equiv \eta, \beta_t \equiv \beta$ and $\alpha \coloneqq 1-\beta$. Combining Proposition 11 with Lemma 12 yields

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla F(x_t)\|\right] &\leq \frac{\Delta_1}{\eta T}+\frac{\eta L}{2}+2\sigma\alpha^{\frac{p-1}{p}}+\frac{2L\eta}{\alpha}\\ &= \sqrt{\frac{\Delta_1 L}{\alpha T}}+\frac{\sqrt{\Delta_1 L\alpha}}{2\sqrt{T}}+2\sigma\alpha^{\frac{p-1}{p}}+2\sqrt{\frac{\Delta_1 L}{\alpha T}} \qquad (7)\\ &\leq 4\sqrt{\frac{\Delta_1 L}{\alpha T}}+2\sigma\alpha^{\frac{p-1}{p}}.\end{aligned}$$

We now make a case distinction. **Case 1:** $\alpha = 1$**.** This implies $\sigma \leq \sqrt{\frac{\Delta_1 L}{T}}$ and hence

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

Case 2: $\alpha = \left(\frac{\Delta_1 L}{\sigma^2 T}\right)^{\frac{p}{3p-2}}$**.** Plugging into (7) yields

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

Therefore we get

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

and hence the claim. ■

D.2. Proofs of Section 3.2

This subsection contains the proofs for our high-probability results. Similar to Proposition 11 we start off with a unified result that allows the usage of different oracles and noise assumptions. 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 9 to get the mild $\log(1/\delta)$ dependence. We would like to point out that this proof technique for establishing the high probability result significantly deviates from the existing high probability analysis of methods using gradient clipping.

Theorem 13 (High-Probability) *Let $\sigma_t \geq 0$ and $\delta \in (0,1)$. Assume (Lower Boundedness), (L-smoothness) and $\infty > \sigma_t \coloneqq \mathbb{E}\left[\|g_t - \nabla F(x_t)\| \mid \mathcal{F}_{t-1}\right]$. 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*

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

where $w_t \coloneqq \frac{\eta_t}{\sum_{\tau=1}^{T}\eta_\tau}$.

We want to make some remarks on Theorem 13. Note that, as in Proposition 11, we did not make any unbiasedness or decreasing stepsize assumptions. Furthermore, when comparing this result with Proposition 11, the bound can be interpreted as a concentration inequality around the expected value.

Proof Let $\phi_t \coloneqq \frac{\nabla F(x_t)^\top g_t}{\|\nabla F(x_t)\|\|g_t\|}$ denote the angle between $\nabla F(x_t)$ and g_t. Then, by Lemma 10, we have

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

Next, we use the fact that ϕ_t is bounded and hence sharply concentrates around its (conditional) expectation. Formally, let $\psi_t \coloneqq \mathbb{E}\left[\phi_t \,|\, \mathcal{F}_{t-1}\right]$ 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

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

implies that we may apply Lemma 9 with $\sigma_t^2 = 4\eta_t^2\|\nabla F(x_t)\|^2$. Doing so yields, for all $\lambda > 0$,

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

with probability at least $1-\delta$. Using (L-smoothness) we get $\|\nabla F(x_t)\| \leq \|\nabla F(x_1)\| + L\sum_{\tau=1}^{t-1}\eta_\tau$ 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 $1-\delta$,

$$\sum_{t=1}^{T}\eta_t\left(\psi_t - \frac{1}{2}\right)\|\nabla F(x_t)\| \leq \Delta_1 + \frac{L}{2}\sum_{t=1}^{T}\eta_t^2 + 6(\eta_T^{\max}\|\nabla F(x_1)\| + C_T L)\log(1/\delta). \tag{8}$$

Finally we are left with the challenge of guaranteeing that ψ_t is *large enough*. Therefore we use Lemma 4 to get $\psi_t\|\nabla F(x_t)\| \leq \|\nabla F(x_t)\| - 2\sigma_t$ and hence

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

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

Next we apply Theorem 13 to derive the high-probability result for tuned `Batch-NSGD` presented in the main paper.

Proof of Theorem 3 To shorten the notation we write $\eta_t \equiv \eta$ and $B_t \equiv B$. We first note that

$$\begin{aligned}
\mathbb{E}\left[\|g_t - \nabla F(x_t)\| \,|\, \mathcal{F}_{t-1}\right] &= \mathbb{E}\left[\left\|\frac{1}{B}\sum_{i=1}^{B}\left(\nabla f(x_t,\xi_t^{(i)}) - \nabla F(x_t)\right)\right\| \,\middle|\, \mathcal{F}_{t-1}\right] \\
&\leq \frac{1}{B}\mathbb{E}\left[\left\|\sum_{i=1}^{B}\left(\nabla f(x_t,\xi_t^{(i)}) - \nabla F(x_t)\right)\right\|^p \,\middle|\, \mathcal{F}_{t-1}\right]^{1/p} \\
&\leq \frac{1}{B}(2B\sigma^p)^{1/p} \leq 2B^{\frac{1-p}{p}}\sigma,
\end{aligned} \tag{9}$$

where we applied Lemma 7 in the second inequality. Combining (9) with Theorem 13 now yields

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(x_t)\| \\
&\leq \frac{2\Delta_1}{\eta T} + \eta L + 8\sigma B^{\frac{1-p}{p}} + 12\left(\frac{\|\nabla F(x_1)\|}{T} + \eta L\right)\log\left(1/\delta\right) \\
&= 2\sqrt{\frac{\Delta_1 L}{T}} + \frac{\sqrt{\Delta_1 L}}{\sqrt{T}} + 8\sigma B^{\frac{1-p}{p}} + 12\left(\frac{\|\nabla F(x_1)\|}{T} + \frac{\sqrt{\Delta_1 L}}{\sqrt{T}}\right)\log\left(1/\delta\right) \\
&\leq \left(3 + 30\log\left(1/\delta\right)\right)\sqrt{\frac{\Delta_1 L}{T}} + 8\sigma B^{\frac{1-p}{p}},
\end{aligned} \tag{10}$$

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

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

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 (10) yields

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

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, $T \geq \Delta_1 L \varepsilon^{-2}$ iterations are required. Plugging into the oracle complexity per iteration yields the second claim. ■

D.3. Remarks on Theorem 3

Our Theorems 1 and 2 hold for NSGD with either momentum or minibatches, as well as for time-varying (t-dependent) and "constant" (T-dependent) parameters. Our choice to present the momentum version with different types of parameters has the goal of showcasing different applications of Proposition 11. For Theorem 3 on the other hand, while it still holds for time-varying and constant parameters, we were not able to prove the result for NSGD with momentum. We shortly want to discuss the technical difficulty of extending Theorem 3 to the momentum version.

Technical difficulty of proving Theorem 3 for NSGD with momentum. The proof of Theorem 13 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

$$\mathbb{E}\left[\|g_t - \nabla F(x_t)\| \,|\, \mathcal{F}_{t-1}\right]. \tag{11}$$

In the case of minibatch NSGD, g_t only depends on randomness sampled in iteration t, and (11) can hence be upper bounded by a constant as seen in (9). However, in the case of NSGD with momentum, g_t consists of random samples from all previous iterations. This results in (11) being a random variable instead, and it is not clear how to uniformly control it.

Appendix E. Mini-Batch NSGD

In this section we discuss the version of our results in Section 3.1 for NSGD with mini-batches (minibatch−NSGD), i.e., NSGD with the gradient estimator

$$g_t = \frac{1}{B}\sum_{j=1}^{B} \nabla f\left(x_t, \xi_t^{(j)}\right), \tag{12}$$

where $\xi_t^{(1)}, \dots, \xi_t^{(B)} \overset{\text{i.i.d.}}{\sim} \xi_t$ are independent of all other random samples.

Proposition 14 (Mini-batch Version of Theorem 1) *Assume (Lower Boundedness), (L-smoothness) and (p-BCM) with $p \in (1, 2]$. Let $\eta, B, q > 0$ 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*

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

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

When plugging $B = q = 1, r = 1/2$ into Proposition 14 we get the sample complexity

$$\mathcal{O}\left(\frac{\Delta_1^4 + L^4}{\varepsilon^4} + \left(\frac{\sigma}{\varepsilon}\right)^{\frac{2p}{p-1}}\right), \tag{13}$$

which equals the complexity of Theorem 1.

Proof of Proposition 14. 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 7. 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 $X_1(x), \dots, X_{\bar{B}}(x)$ are independent random variables with mean zero and hence a Martingale Difference Sequence (MDS). Furthermore note that $\mathbb{E}\left[\|X_j(x)\|^p\right] \leq \sigma^p$ by (p-BCM) and we can hence apply Lemma 7 to get

$$g(x) \coloneqq \mathbb{E}\left[\left\|\sum_{j=1}^{\bar{B}} X_j(x)\right\|^p\right] \leq 2\sum_{j=1}^{\bar{B}} \mathbb{E}\left[\|X_j(x)\|^p\right] \leq 2\bar{B}\sigma^p. \tag{14}$$

Next we calculate

$$\begin{aligned}\mathbb{E}\left[\|g_t - \nabla F(x_t)\| \,|\, x_t\right] &= \mathbb{E}\left[\left\|\frac{1}{\bar{B}}\sum_{j=1}^{\bar{B}}\left(\nabla f(x_t,\xi_t^{(j)}) - \nabla F(x_t)\right)\right\| \middle| x_t\right] \\ &\leq \frac{1}{\bar{B}}\mathbb{E}\left[\left\|\sum_{j=1}^{\bar{B}}\left(\nabla f(x_t,\xi_t^{(j)}) - \nabla F(x_t)\right)\right\|^p \middle| x_t\right]^{1/p}\end{aligned} \tag{15}$$

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 8 which yields

$$\mathbb{E}\left[\|g_t - \nabla F(x_t)\| \,|\, x_t\right] \overset{(15)}{\leq} \frac{1}{\bar{B}}\mathbb{E}\left[h(x_t,Y) \,|\, x_t\right]^{1/p} \overset{\text{Lem. 8}}{=} \frac{1}{\bar{B}}g(x_t)^{1/p} \overset{(14)}{\leq} 2\frac{\sigma}{\bar{B}^{\frac{p-1}{p}}} \tag{16}$$

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 (3) 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

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

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

$$\mathcal{O}\left(\left(\frac{\Delta_1}{\eta\varepsilon}\right)^{\frac{1}{1-r}} + \left(\frac{\eta L}{\varepsilon}\right)^{\frac{1}{r}} + B\left(\frac{\Delta_1}{\eta\varepsilon}\right)^{\frac{1+q}{1-r}} + B\left(\frac{\eta L}{\varepsilon}\right)^{\frac{1+q}{r}} + \frac{1}{B^{\frac{1}{q}}}\left(\frac{\sigma}{\varepsilon}\right)^{\frac{p(1+q)}{q(p-1)}}\right). \tag{17}$$

This completes the proof. ■

Lastly we will provide the mini-batch version of Theorem 2.

Corollary 15 (Mini-batch Version of Theorem 2) *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 $\mathcal{O}\left(\frac{\Delta_1 L}{\varepsilon^2} + \frac{\Delta_1 L}{\varepsilon^2}\left(\frac{\sigma}{\varepsilon}\right)^{\frac{p}{p-1}}\right)$.

Proof of Corollary 15. Applying Proposition 14 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 $\mathcal{O}\left(\Delta_1 L \varepsilon^{-2}\right)$ and hence a sample complexity of

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

■