OPT2023: 15th Annual Workshop on Optimization for Machine Learning

Parameter-Agnostic Optimization under Relaxed Smoothness

Florian Hübler FHUEBLER@STUDENT.ETHZ.CH
Junchi Yang JUNCHI.YANG@INF.ETHZ.CH
Xiang Li XIANG.LI@INF.ETHZ.CH
Niao He NIAO.HE@INF.ETHZ.CH
ETH Zurich, Switzerland

Abstract

Tuning hyperparameters, such as the stepsize, presents a major challenge of training machine learning models. To address this challenge, numerous adaptive optimization algorithms have been developed that achieve near-optimal complexities, even when stepsizes are independent of problem-specific parameters, provided that the loss function is L-smooth. However, as the assumption is relaxed to the more realistic (L_0, L_1)-smoothness, all existing convergence results still necessitate tuning of the stepsize. In this study, we demonstrate that Normalized Stochastic Gradient Descent with Momentum (NSGD-M) can achieve a (nearly) rate-optimal complexity without prior knowledge of any problem parameter, though this comes at the cost of introducing an exponential term dependent on L_1 in the complexity. We further establish that this term is inescapable to such schemes. Interestingly, in deterministic settings, the exponential factor can be neutralized by employing Gradient Descent with a Backtracking Line Search. To the best of our knowledge, these findings represent the first parameter-agnostic convergence results under the generalized smoothness condition. Our empirical experiments further confirm our theoretical insights.

1. Introduction

We consider the unconstrained optimization problem

$$\min_{x\in\mathbb{R}^d} F(x), \tag{1}$$

where $F\colon \mathbb{R}^d \to \mathbb{R}$ may be non-convex and admits access to unbiased stochastic gradients. This setting has been extensively studied due to its prevalence in modern machine learning and data-driven optimization [3].

When the objective function F is L-smooth, i.e., F has L-Lipschitz gradients, the problem is well-explored. For the goal of finding an ε-stationary point, lower bounds have been established by Arjevani et al. [1], setting a limit of $\Omega\big(L\Delta_1\sigma^2\varepsilon^{-4}\big)$ for stochastic first-order methods. Here σ denotes the variance of the stochastic gradient and Δ_1 the initialization gap. Stochastic Gradient Descent (`SGD`) achieves this complexity with stepsizes depending on problem parameters like L [11]. Remarkably, several algorithms such as `AdaGrad-Norm`, oblivious to problem parameters, are recently proven to achieve a nearly rate-optimal complexity $\widetilde{\mathcal{O}}\big(\varepsilon^{-4}\big)$, up to the dependency on problem parameters and logarithmic factors [9, 25]. We call algorithms with this characteristic *parameter-agnostic*, and *parameter-dependent* otherwise.

However, Zhang et al. [28] highlighted that not all machine learning applications satisfy L-smoothness. Their experiments in language modeling tasks revealed that the norm of the Hessian

is not uniformly upper-bounded as required by L-smoothness. Rather, it may increase affinely with the gradient norm. To bridge the gap between theory and this observation, they introduced a more general smoothness condition termed (L_0, L_1)-smoothness: $\left\|\nabla^2 F(x)\right\| \leq L_0 + L_1 \left\|\nabla F(x)\right\|$. This condition has since been validated in various machine learning tasks [4, 27].

In light of this more realistic smoothness assumption, a substantial body of literature has emerged. The nearly rate-optimal complexity $\widetilde{\mathcal{O}}\left(\varepsilon^{-4}\right)$ has been established for various algorithms, including `SGD` [14], Clipped `SGD` [27, 28], Normalized `SGD` [29], `AdaGrad-Norm` [10, 23] and `ADAM` [13]. Yet, all of these algorithms require prior information of the problem, such as the values of L_0 and L_1. Notably, unlike the L-smooth setting, `AdaGrad-Norm` may diverge without access to L_1 [23], shedding its fully parameter-agnostic nature. This dependence on problem parameters poses a significant challenge as these parameters are usually unknown in practical applications, necessitating resource-intensive tuning [24]. These observations culminate in the pressing question:

> *Is there an algorithm that converges with near-optimal complexity, without having access to any problem parameters in the (L_0, L_1)-smoothness setting?*

To tackle these challenges, this work makes the following contributions.

We show that, under the relaxed (L_0, L_1)-smoothness assumption, Normalized Stochastic Gradient Descent with Momentum (`NSGD-M`), as introduced by [5], converges with a nearly rate-optimal complexity of $\widetilde{\mathcal{O}}\left(\varepsilon^{-4}\right)$ without any prior knowledge of the problem parameters. However, it results in an exponential dependency on L_1, which vanishes when the stepsize is informed by L_1. Furthermore, we prove that this exponential dependency can also be avoided in the deterministic setting using Gradient Descent (`GD`) with Backtracking Line Search, resulting in a complexity of $\mathcal{O}\left(\left(L_0\Delta_1 + L_1^2\Delta_1^2\right)\varepsilon^{-2}\right)$. To the best of our knowledge, these are the first parameter-agnostic convergence results in the (L_0, L_1)-smoothness setting.

We furthermore demonstrate that the exponential term in L_1 is indispensable for a class of Normalized Momentum Methods, including `NSGD-M`, when the problem parameters are unknown.

1.1. Related Work

Parameter-Agnostic Algorithms. If the objective function is L-smooth, convergence results are typically contingent upon stepsizes being less than $2/L$ [3]. In the deterministic context, `GD` with a constant stepsize that does not satisfy this threshold may diverge [18]. However, this can be rectified using a Backtracking Line Search, which does not rely on knowing problem parameters, and achieves an optimal complexity of $\mathcal{O}(\epsilon^{-2})$ [2]. Conversely, in stochastic environments, Vaswani et al. [21] highlighted that line search techniques might not always converge. `SGD` with a parameter-agnostic diminishing stepsize of $1/\sqrt{t}$ still reaches a near-optimal complexity of $\widetilde{\mathcal{O}}(\epsilon^{-4})$, though it introduces an inescapable exponential term in L [26]. Various adaptive methods, such as `AdaGrad` [7, 15], its variants `AdaGrad-Norm` [20] and `NSGD-M` [5], bypass this exponential term, even without knowledge of the problem parameters, as recently shown in [9, 26]. These adaptive methods are typically considered more robust to different problem parameters [12, 24], given their ability to tune algorithm hyperparameters dynamically during training. There is another line of research dedicated to "parameter-free" algorithms for online convex optimization [6, 19]. However, this research emphasizes the optimal dependence on $\|x^* - x_0\|$, where x^* is the predictor in the regret bound.

(L_0, L_1)-Smoothness. Zhang et al. [28] introduced the concept of (L_0, L_1)-smoothness, defined by the following affine bound on the Hessian-norm: $\left\|\nabla^2 F(x)\right\| \leq L_0 + L_1 \|\nabla F(x)\|$. The convergence of both GD and SGD was only recently established in this setting [14]. However, their stepsizes require prior knowledge of L_0, L_1, and also the exact gradient norm of the initial point, which can be unavailable in stochastic settings. Clipped SGD [28], and its momentum-augmented counterpart [27], both demand knowledge of L_0 and L_1 for convergence. They attain an optimal complexity of $\mathcal{O}(\varepsilon^{-4})$ and are believed to improve over SGD in constants. Additionally, Zhang et al. [27] also provided a convergence result for NSGD-M with constant stepsizes in the appendix. Their analysis does however make use of a stronger noise assumption and requires access to all parameters. Similar complexities have been established for Normalized SGD [29], signed SGD [4], AdaGrad-Norm [10, 23], and ADAM [13, 22]. However, each of these methods requires prior knowledge of problem-specific parameters. Notably, in stark contrast to the L-smooth setting, even AdaGrad-Norm is not wholly parameter-agnostic. It risks divergence if the stepsize is not informed by L_1, despite the method generally demanding knowledge of fewer problem parameters than other algorithms [23].

2. Preliminaries

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

Notation. Throughout the paper, $d \in \mathbb{N}_{\geq 1}$ denotes the dimensionality of the variable to be optimized, $F\colon \mathbb{R}^d \to \mathbb{R}$ the objective and $\nabla f(\cdot, \cdot)$ the gradient oracle. We use the common convention that empty sums and products are given by their corresponding neutral element. The conic combination of $x_1, \ldots, x_n \in \mathbb{R}^d$ will be denoted by $\operatorname{cone}(x_1, \ldots, x_n) \coloneqq \{\sum_{i=1}^n \lambda_i x_i \colon \lambda_1, \ldots, \lambda_n \geq 0\}$.

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

Assumption 2 (Bounded Variance) *The gradient oracle is unbiased and has finite variance, i.e. there exists $\sigma \geq 0$ such that*

$$\mathbb{E}\left[\nabla f(x, \xi)\right] = \nabla F(x) \qquad \text{and} \qquad \mathbb{E}\left[\left\|\nabla f(x, \xi) - \nabla F(x)\right\|^2\right] \leq \sigma^2.$$

Instead of the traditional L-smoothness assumption, we adopt the weaker concept of (L_0, L_1)-smoothness, as proposed by Zhang et al. [28]. Following the work of Zhang et al. [27], we choose a definition that does not require the Hessian. This definition is therefore weaker than the original (L_0, L_1)-smoothness assumption by Zhang et al. [28, Definition 1], but equivalent if the objective function is twice differentiable as shown in Appendix B, Lemma 6.

Definition 1 *Let $L_0, L_1 \geq 0$ and $f\colon \mathbb{R}^d \to \mathbb{R}$ be a differentiable function for some $d \in \mathbb{N}_{\geq 1}$. Then f is called* (L_0, L_1)-smooth *if for all $x, y \in \mathbb{R}^d$ and all $c > 0$ with $L_1 \|x - y\| \leq c$ it holds that*

$$\|\nabla f(x) - \nabla f(y)\| \leq \left(A_0(c) L_0 + A_1(c) L_1 \|\nabla f(x)\|\right) \|x - y\|,$$

where $A_0(c) \coloneqq 1 + e^c - \frac{e^c - 1}{c}$ and $A_1(c) \coloneqq \frac{e^c - 1}{c}$.

Assumption 3 ((L_0, L_1)-smoothness) *The objective function F is (L_0, L_1)-smooth.*

3. Parameter-Agnostic Convergence under (L_0, L_1)-Smoothness

In this section, we present the first parameter-agnostic convergence results on (L_0, L_1)-smooth functions. We show that in the stochastic setting, NSGD-M (see Algorithm 1) achieves the nearly rate-optimal complexity of $\widetilde{\mathcal{O}}\left(\varepsilon^{-4}\right)$, even without access to problem-dependent parameters. However, this is accompanied by an undesirable exponential dependence on L_1. In Section 3.1 we will then show that the undesirable dependence is unavoidable for NSGD-M, even in the deterministic setting. However, in Section 3.2 we will show that this exponential dependence can be avoided in the deterministic setting, by using GD with Backtracking Line Search which is parameter-agnostic. Experiments in Appendix E empirically confirm our theoretical insights.

Algorithm 1: Normalized SGD with Momentum (NSGD-M) [5]

Input: Starting point $x_1 \in \mathbb{R}^d$, stepsizes $\eta_t > 0$, moving average parameters $\beta_t \in [0, 1)$
$m_0 \leftarrow 0$
for $t = 1, 2, \ldots$ **do**
  Independently sample ξ_t from the distribution of ξ.
  $g_t \leftarrow \nabla f(x_t, \xi_t)$
  $m_t \leftarrow \beta_t m_{t-1} + (1 - \beta_t) g_t$
  $x_{t+1} \leftarrow x_t - \frac{\eta_t}{\|m_t\|} m_t$
end

The convergence of NSGD-M occurs in two phases. In the initial *adaptation phase*, the algorithm accumulates error due to a large stepsize. Unfortunately, this error may grow exponentially in L_1. Once the stepsize decreases below a threshold (polynomial in L_1), the algorithm transitions into the *convergence phase*. In this latter phase, the error decays proportionally to $T^{-1/4} \log(T)$.

Theorem 2 (Convergence of NSGD-M under (L_0, L_1)-smoothness) *Assume (Lower Boundedness), ((L_0, L_1)-smoothness) and (Bounded Variance). Furthermore define the parameters $\beta_t := 1 - t^{-1/2}$ and $\eta_t := \frac{t^{-3/4}}{7}$. Then NSGD-M with starting point $x_1 \in \mathbb{R}^d$ satisfies*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\|\nabla F(x_t)\|\right] \leq \frac{\left(14 + 168e^{L_1^2}\right)\Delta_1 + 15\sigma\log(T) + \left(4\log(T) + 6e^{L_1} + 3e^{L_1^2}\right)L_0}{T^{\frac{1}{4}}},$$

where $\Delta_1 := F(x_1) - F^*$ is the initialization gap.

Since (L_0, L_1)-smoothness includes L-smoothness as a special case, the lower bound of $\mathcal{O}(\varepsilon^{-4})$ to find an ε-stationary point is still applicable here. Theorem 2 implies a near-optimal complexity in ε up to the logarithmic factor without any prior knowledge of the problem parameters, but it comes with the cost of an exponential term in L_1. The following remark shows that this cost indeed comes from the parameter-agnostic stepsize.

Remark 3 *In Corollary 15, we show that tuning the stepsize using L_1 only is sufficient to shave off the exponential term and achieve a rate of*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\|\nabla F(x_t)\|\right] = \widetilde{\mathcal{O}}\left(\frac{\Delta_1 L_1 + \sigma + \frac{L_0}{L_1}}{T^{\frac{1}{4}}}\right).$$

These results have indicated that NSGD-M is potentially more robust to hyper-parameters selection than other existing algorithms. In comparison, SGD necessitates knowledge of both L_0 and L_1, as well as the exact value of $\|\nabla f(x_1)\|$ [14]. Clipped SGD requires to know L_0 and L_1 [27], and even AdaGrad-Norm demands knowledge of L_1 [10, 23].

3.1. Lower Bound for A Family of Normalized Momentum Methods

In this subsection we establish a parameter-agnostic lower bound for a generalized version of NSGD-M. More specifically, for $\eta > 0$ and $\alpha \in (0, 1)$, we consider the following iteration rule:

$$\begin{aligned} &g_t \leftarrow \nabla f(x_t, \xi_t) \\ &\text{Choose } m_t \in \operatorname{cone}(g_1, \dots, g_t) \\ &x_{t+1} \leftarrow x_t - \frac{\eta}{t^\alpha} \frac{m_t}{\|m_t\|} \end{aligned} \tag{2}$$

We call algorithms following this procedure *General Normalized Momentum Methods* (see Algorithm 3 in Appendix D.2). Clearly, NSGD-M from Theorem 2 belongs to this family of algorithms.

The lower bound is based on two rationales. Firstly, the relaxed smoothness assumption allows for rapid changes in the gradient. This enables the construction of a function on which the optimality gap increases to about $F(x_2) - F^* \geq \Delta_1 + e^{\eta L_1}$ after just one iteration. Secondly, due to the normalization in stepsizes, we can bound the distance $\|x_t - x_1\|$ by $\eta \sum_{\tau=1}^{t-1} \tau^{-\alpha} \leq \frac{\eta}{1-\alpha} t^{1-\alpha}$, regardless of the momentum norm. This behaviour allows us to continue the construction of F in such way, that reaching an ε-stationary point takes $\Omega\left(\varepsilon^{-1/(1-\alpha)}\right)$ iterations.

It is worth noting that these key ideas are also applicable to other algorithms and settings, such as SGD with diminishing stepsizes under L-smoothness.

Theorem 4 (Lower Bound for General Normalized Momentum Methods) *Consider a General Normalized Momentum Method $\mathcal{A}$ with parameters $\eta > 0$ and $\alpha \in (0, 1)$. Let $0 < \varepsilon < 1/2, \Delta_1 \geq 1/4, L_0 \geq 8/\eta, L_1 > 0$. Then there exists an (L_0, L_1)-smooth function F with $F(x_1) - F^* \leq \Delta_1$ for which $\mathcal{A}$ requires at least*

$$T \geq \left(\frac{1-\alpha}{2}\right)^{\frac{1}{1-\alpha}} \left(\frac{\Delta_1}{\eta} + \frac{2}{\eta L_1}\left(e^{\frac{\eta L_1}{4}} - 1\right)\right)^{\frac{1}{1-\alpha}} \varepsilon^{-\frac{1}{1-\alpha}}$$

iterations to find an ε-stationary point in the deterministic setting. In particular, NSGD-M specified in Theorem 2 has a lower complexity bound of

$$\Omega\left(\left(\Delta_1 + \frac{e^{L_1/28} - 1}{L_1}\right)\varepsilon^{-4}\right).$$

This lower bound reveals that one cannot achieve a parameter-agnostic convergence result for NSGD-M without an exponential dependence on L_1. It is important to note that the above is a *parameter-agnostic lower bound*: We first fix an algorithm $\mathcal{A}$ before adversarially choosing a hard function. Consequently, this finding does not contradict Remark 3. Moreover, it also suggests that finding an ε-stationary point in a parameter-agnostic fashion is strictly harder in this relaxed smoothness setting: in the L-smooth setting, equivalent to $(L, 0)$-smoothness, the exponential term in Theorem 2 vanishes, which also aligns with previous upper bounds [5, 26].

3.2. Deterministic Setting

Given the prior results, one might naturally wonder if there exists any algorithm that can attain parameter-agnostic convergence without exponential dependence on L_1. The subsequent theorem confirms that this is indeed possible, at least in the deterministic setting, by Gradient Descent with a Backtracking Line-search (see Algorithm 2).

Theorem 5 *Assume (Lower Boundedness) and ((L_0, L_1)-smoothness) in the deterministic setting. Then GD with Backtracking Line Search (see Algorithm 2) with parameters $\beta, \gamma \in (0, 1)$ and starting point $x_1 \in \mathbb{R}^d$ satisfies*

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla F(x_t)\|^2 \leq \frac{4L_0\Delta_1 + 14L_1^2\Delta_1^2}{\beta\gamma(1-\gamma)T} = \mathcal{O}\left(\frac{L_0\Delta_1 + L_1^2\Delta_1^2}{T}\right),$$

where $\Delta_1 \coloneqq F(x_1) - F^*$.

This implies a complexity of $\mathcal{O}\big(\big(L_0\Delta_1 + L_1^2\Delta_1^2\big)\varepsilon^{-2}\big)$, which is optimal in the dependence of ε and L_0 in the deterministic setting. The proof rests on the observation that GD with Backtracking Line Search is a descent algorithm and hence both the function value and gradient norm remain upper bounded along the trajectory. Consequently, the algorithm behaves as though it is addressing $(L_0 + L_1 C)$-smooth functions, where C represents the gradient norm's upper bound. We have not extended our considerations to the stochastic setting for this algorithm, as a stochastic line search could potentially fail even under the stricter L-smoothness assumption [21].

4. Conclusion and Future Work

In this work, we conduct a theoretical investigation into parameter-agnostic algorithms under the (L_0, L_1)-smoothness assumption. In the stochastic setting, we show that without requiring any knowledge about problem parameters, Normalized Stochastic Gradient Descent with Momentum (NSGD-M) converges at an order-optimal rate, albeit with an exponential term in L_1. In the deterministic setting, we show the exponential dependency can be circumvented using GD with Backtracking Line Search while being parameter-agnostic.

This work motivates several questions for future research. The most pressing one is whether there exists a parameter-agnostic algorithm in the stochastic setting without an exponential term. A further interesting topic is the derivation of lower bounds for first-order parameter-agnostic methods.

References

[1] Yossi Arjevani, Yair Carmon, John C. Duchi, Dylan J. Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, Jun 2022. ISSN 1436-4646. doi: 10.1007/s10107-022-01822-7. URL https://doi.org/10.1007/s10107-022-01822-7.

[2] Larry Armijo. Minimization of Functions having Lipschitz continuous first partial Derivatives. *Pacific Journal of Mathematics*, 16(1):1 – 3, 1966.

[3] Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization Methods for Large-Scale Machine Learning. *SIAM Review*, 60(2):223–311, 2018. doi: 10.1137/16M1080173. URL https://doi.org/10.1137/16M1080173.

[4] Michael Crawshaw, Mingrui Liu, Francesco Orabona, Wei Zhang, and Zhenxun Zhuang. Robustness to Unbounded Smoothness of Generalized SignSGD. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 9955–9968. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/40924475a9bf768bdac3725e67745283-Paper-Conference.pdf.

[5] Ashok Cutkosky and Harsh Mehta. Momentum Improves Normalized SGD. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2260–2268. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/cutkosky20b.html.

[6] Ashok Cutkosky and Francesco Orabona. Black-Box Reductions for Parameter-free Online Learning in Banach Spaces. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1493–1529. PMLR, PMLR, 06–09 Jul 2018. URL https://proceedings.mlr.press/v75/cutkosky18a.html.

[7] John Duchi, Elad Hazan, and Yoram Singer. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research*, 12(61):2121–2159, 2011. URL http://jmlr.org/papers/v12/duchi11a.html.

[8] Ilyas Fatkhullin, Anas Barakat, Anastasia Kireeva, and Niao He. Stochastic Policy Gradient Methods: Improved Sample Complexity for Fisher-non-degenerate Policies. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 9827–9869. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/fatkhullin23a.html.

[9] Matthew Faw, Isidoros Tziotis, Constantine Caramanis, Aryan Mokhtari, Sanjay Shakkottai, and Rachel Ward. The Power of Adaptivity in SGD: Self-Tuning Step Sizes with Unbounded Gradients and Affine Variance. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 313–355. PMLR, 02–05 Jul 2022. URL https://proceedings.mlr.press/v178/faw22a.html.

[10] Matthew Faw, Litu Rout, Constantine Caramanis, and Sanjay Shakkottai. Beyond Uniform Smoothness: A Stopped Analysis of Adaptive SGD. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 89–160. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/faw23a.html.

[11] Saeed Ghadimi and Guanghui Lan. Stochastic First- and Zeroth-Order Methods for Nonconvex Stochastic Programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013. doi: 10.1137/120880811. URL https://doi.org/10.1137/120880811.

[12] Ali Kavis, Kfir Y. Levy, Francis Bach, and Volkan Cevher. UniXGrad: A Universal, Adaptive Algorithm with Optimal Guarantees for Constrained Optimization. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/88855547570f7ff053fff7c54e5148cc-Paper.pdf.

[13] Haochuan Li, Ali Jadbabaie, and Alexander Rakhlin. Convergence of Adam under Relaxed Assumptions. *arXiv preprint arXiv:2304.13972*, 2023.

[14] Haochuan Li, Jian Qian, Yi Tian, Alexander Rakhlin, and Ali Jadbabaie. Convex and Non-Convex Optimization under Generalized Smoothness. *arXiv preprint arXiv:2306.01264*, 2023.

[15] H. B. McMahan and Matthew J. Streeter. Adaptive Bound Optimization for Online Convex Optimization. In *Annual Conference Computational Learning Theory*, 2010.

[16] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer Sentinel Mixture Models. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Byj72udxe.

[17] Stephen Merity, Nitish Shirish Keskar, and Richard Socher. Regularizing and Optimizing LSTM Language Models. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=SyyGPP0TZ.

[18] Yurii Nesterov. *Lectures on Convex Optimization*, volume 137 of *Springer Optimization and Its Applications*. Springer, Cham, 2018. ISBN 978-3-319-91578-4. doi: 10.1007/978-3-319-91578-4. Second edition of [MR2142598].

[19] Francesco Orabona and David Pal. Coin Betting and Parameter-Free Online Learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/320722549d1751cf3f247855f937b982-Paper.pdf.

[20] Matthew Streeter and H Brendan McMahan. Less Regret via Online Conditioning. *arXiv preprint arXiv:1002.4862*, 2010.

[21] Sharan Vaswani, Benjamin Dubois-Taine, and Reza Babanezhad. Towards Noise-adaptive, Problem-adaptive (Accelerated) Stochastic Gradient Descent. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 22015–22059. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/vaswani22a.html.

[22] Bohan Wang, Yushun Zhang, Huishuai Zhang, Qi Meng, Zhi-Ming Ma, Tie-Yan Liu, and Wei Chen. Provable adaptivity in adam. *arXiv preprint arXiv:2208.09900*, 2022.

[23] Bohan Wang, Huishuai Zhang, Zhiming Ma, and Wei Chen. Convergence of AdaGrad for Non-convex Objectives: Simple Proofs and Relaxed Assumptions. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 161–190. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/wang23a.html.

[24] Rachel Ward, Xiaoxia Wu, and Leon Bottou. AdaGrad Stepsizes: Sharp Convergence Over Nonconvex Landscapes. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6677–6686. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/ward19a.html.

[25] Junchi Yang, Xiang Li, and Niao He. Nest Your Adaptive Algorithm for Parameter-Agnostic Nonconvex Minimax Optimization. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 11202–11216. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/488b8db9ec118c3d750c34d1812a5a3a-Paper-Conference.pdf.

[26] 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. *arXiv preprint arXiv:2305.12475*, 2023.

[27] Bohang Zhang, Jikai Jin, Cong Fang, and Liwei Wang. Improved Analysis of Clipping Algorithms for Non-convex Optimization. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15511–15521. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/b282d1735283e8eea45bce393cefe265-Paper.pdf.

[28] Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why Gradient Clipping Accelerates Training: A Theoretical Justification for Adaptivity. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BJgnXpVYwS.

[29] Shen-Yi Zhao, Yin-Peng Xie, and Wu-Jun Li. On the convergence and improvement of stochastic normalized gradient descent. *Science China Information Sciences*, 64(3):132103, Feb 2021. ISSN 1869-1919. doi: 10.1007/s11432-020-3023-7. URL https://doi.org/10.1007/s11432-020-3023-7.

Appendix A. Organisation of the Appendix

The appendix is organised as follows.

Appendix B contains basic properties of (L_0, L_1)-smooth functions.
Appendix C lists the technical Lemmas needed for the main result.
Appendix D contains the proofs of Section 3.
Appendix E contains experiments on language modelling tasks.

Appendix B. Basic Properties of (L_0, L_1)-Smoothness

In this section we will prove basic properties of (L_0, L_1)-Smoothness. We will start with the proof of the relation to the original definition by [28].

Lemma 6 *Let $F\colon \mathbb{R}^d \to \mathbb{R}$ be twice continuously differentiable and $L_0, L_1 \geq 0$. Then F satisfies $\left\|\nabla^2 F(x)\right\| \leq L_0 + L_1 \left\|\nabla F(x)\right\|$ if and only if F is (L_0, L_1)-smooth according to Definition 1.*

Proof "⇒": This implication was already shown by [27, Corollary A.4].

"⇐": We slightly adapt the proof of [10, Proposition 1]. Assume F is (L_0, L_1)-smooth according to Definition 1. Let $x, s \in \mathbb{R}^d$ with $\|s\| = 1$. For $\alpha > 0$ our assumption gives

$$\|\nabla F(x + \alpha s) - \nabla F(x)\| \leq (A_0(\alpha L_1)L_0 + A_1(\alpha L_1)L_1 \|\nabla F(x)\|)\alpha$$

and hence

$$\left\|\frac{\nabla F(x + \alpha s) - \nabla F(x)}{\alpha}\right\| \leq A_0(\alpha L_1)L_0 + A_1(\alpha L_1)L_1 \|\nabla F(x)\| .$$

Using the continuity of norms and the assumption that F is twice continously differentiable, we hence get

$$\begin{aligned} L_0 + L_1 \|\nabla F(x)\| &= \lim_{\alpha \to 0} A_0(\alpha L_1)L_0 + A_1(\alpha L_1)L_1 \|\nabla F(x)\| \\ &\geq \lim_{\alpha \to 0} \left\|\frac{\nabla F(x + \alpha s) - \nabla F(x)}{\alpha}\right\| \\ &= \left\|\lim_{\alpha \to 0} \frac{\nabla F(x + \alpha s) - \nabla F(x)}{\alpha}\right\| \\ &= \left\|\nabla^2 F(x) s\right\| . \end{aligned}$$

Taking the sup over all such s yields the claim. ■

The following Lemma serves as the (L_0, L_1)-smooth counterpart to the well-known quadratic upper bound on the function value change in the L-smooth setting.

Lemma 7 (c.f. [27, Lemma A.3]) *Let $d \in \mathbb{N}_{\geq 1}$ and $L_0, L_1 \geq 0$. Assume that $f\colon \mathbb{R}^d \to \mathbb{R}$ is (L_0, L_1)-smooth. Then all $x, y \in \mathbb{R}^d$ satisfy*

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{1}{2}(B_0(L_1 \|x - y\|)L_0 + B_1(L_1 \|x - y\|)L_1 \|\nabla f(x)\|)\|x - y\|^2,$$

where

$$B_0(c) = 1 + 2\frac{e^c - 1}{c} - 4\frac{e^c - 1 - c}{c^2},$$
$$B_1(c) = 2\frac{e^c - 1 - c}{c^2}$$

tend to 1 as c tends towards 0.

Proof *This proof closely follows the arguments from [27]. We include the proof for completeness. Let $x, y \in \mathbb{R}^d$ and calculate*

$$\begin{aligned} f(y) - f(x) - \nabla f(x)^\top (y - x) &= \int_0^1 \nabla f(x + t(y - x))^\top (y - x) dt - \nabla f(x)^\top (y - x) \\ &\leq \int_0^1 \|\nabla f(x + t(y - x)) - \nabla f(x)\| \, \|x - y\| \, dt \\ &\leq \|x - y\|^2 \left(L_0 \int_0^1 t A_0(tc) dt + L_1 \|\nabla f(x)\| \int_0^1 t A_1(tc) dt \right) \end{aligned}$$

where $c := L_1 \|x - y\|$. We now calculate

$$\int_0^1 t A_0(tc) = \frac{1}{2} + \frac{e^c - 1}{c} - 2\frac{e^c - 1}{c^2} =: \frac{1}{2} B_0(c)$$

and

$$\int_0^1 t A_1(tc) dt = \frac{e^c - 1 - c}{c^2} =: \frac{1}{2} B_1(c).$$

This shows the claim. ■

Analogous to the L-smooth setting, we can also derive an upper bound for the gradient norm based on the suboptimality gap.

Lemma 8 (Gradient Bound, c.f. [27, Lemma A.5]) *Let $L_0, L_1 > 0$ and assume that $f : \mathbb{R}^d \to \mathbb{R}$ is (L_0, L_1)-smooth. Further assume that f is lower bounded by f^*. Then all $x \in \mathbb{R}^d$ satisfy*

$$\min \left\{ \frac{\|\nabla f(x)\|}{L_1}, \frac{\|\nabla f(x)\|^2}{L_0} \right\} \leq 8(f(x) - f^*).$$

Proof This proof is again based on [27]. We include it since we require parts of the proof later. Let $x \in \mathbb{R}^d$. Firstly note that, for A_1 from Definition 1, the equation

$$c = \frac{L_1 \|\nabla F(x)\|}{A_1(c) L_0 + L_1 A_1(c) \|\nabla F(x)\|}$$

has a solution $c \in (0, 1)$. Now we set $\lambda := \frac{1}{2A_1(c)(L_0 + L_1 \|\nabla F(x)\|)}$ and $y := x - \lambda \nabla F(x)$. Then Lemma 7 yields

$$\begin{aligned} F^* \leq F(y) &\leq F(x) - \lambda \|\nabla F(x)\|^2 + A_1(c)(L_0 + L_1 \|\nabla F(x)\|) \lambda^2 \|\nabla F(x)\|^2 \\ &= F(x) - \frac{\lambda}{2} \|\nabla F(x)\|^2. \end{aligned}$$

We now differentiate between the two cases $\|\nabla F(x)\| \leq \frac{L_0}{L_1}$ and $\|\nabla F(x)\| > \frac{L_0}{L_1}$. Therefore calculate

$$2(F(x) - F^*) \geq \frac{\|\nabla F(x)\|^2}{A_1(c)(L_0 + L_1 \|\nabla F(x)\|)} \geq \begin{cases} \frac{\|\nabla F(x)\|^2}{4L_0}, & \text{if } \|\nabla F(x)\| \leq \frac{L_0}{L_1} \\ \frac{\|\nabla F(x)\|}{4L_1}, & \text{otherwise.} \end{cases}$$

This shows the claim. ■

Appendix C. Technical Lemmas

This section presents crucial technical lemmas and their proofs. These results may be of interest on their own as they can potentially be applied in the analysis of other momentum-based algorithms.

Lemma 9 (Technical Lemma) *Let $q \in (0,1), p \geq 0$ and $t > 0$. Further let $a, b \in \mathbb{N}_{\geq 2}$ with $a \leq b$. Then the following statements are true.*

i) We have

$$\prod_{t=a}^{b}\left(1-t^{-q}\right) \leq \exp\left(\frac{1}{1-q}\left(a^{1-q}-b^{1-q}\right)\right).$$

ii) If $p \geq q$, then

$$\sum_{t=a}^{b} t^{-p} \prod_{\tau=a}^{t}\left(1-\tau^{-q}\right) \leq \frac{(a-1)^{q-p}\exp\left(\frac{a^{1-q}-(a-1)^{1-q}}{1-q}\right) - b^{q-p}\exp\left(\frac{a^{1-q}-b^{1-q}}{1-q}\right)}{1+(p-q)b^{q-1}},$$

and in particular,

$$\sum_{t=a}^{b} t^{-p} \prod_{\tau=a}^{t}\left(1-\tau^{-q}\right) \leq (a-1)^{q-p}\exp\left(\frac{a^{1-q}-(a-1)^{1-q}}{1-q}\right) = \mathcal{O}\left(a^{q-p}\right).$$

iii) (c.f. [8, Lemma 15][1]) If $a \geq p^{\frac{1}{1-q}}$ and $a \geq \left(\frac{p-q}{2}\right)^{\frac{1}{1-q}}$, then

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

Note that these requirements are always fulfilled for $p \leq 1$.

Proof $i)$ The first claim follows from the calculation

$$\prod_{t=a}^{b}\left(1-t^{-q}\right) \leq \exp\left(-\sum_{\tau=a}^{b} t^{-q}\right) \leq \exp\left(-\int_{a}^{b+1} t^{-q}dt\right) = \exp\left(\frac{1}{1-q}\left(a^{1-q}-(b+1)^{1-q}\right)\right), \tag{3}$$

where we used $1-x \leq e^{-x}$ in the first, and the monotonicity of t^{-q} in the second inequality. Weakening the inequality by replacing $(b+1)$ with b finishes the proof.
$ii)$ For the second inequality we use $i)$ to derive

$$\sum_{t=a}^{b} t^{-p} \prod_{\tau=a}^{t}\left(1-t^{-q}\right) \leq \exp\left(\frac{a^{1-q}}{1-q}\right)\sum_{t=a}^{b} t^{-p}\exp\left(-\frac{t^{1-q}}{1-q}\right).$$

1. Note that the proof in the paper has a typo in the last line of page 42. Instead of $(1-q)$ the authors meant $(1-q)^{-1}$.

Using the monotonicity of $t^{-p}\exp\left(-t^{1-q}\right)$ we obtain

$$\sum_{t=a}^{b} t^{-p}\exp\left(-\frac{t^{1-q}}{1-q}\right) \leq \int_{a-1}^{b} t^{-p}\exp\left(-\frac{t^{1-q}}{1-q}\right)dt = \int_{a-1}^{b} t^{q-p}t^{-q}\exp\left(-\frac{t^{1-q}}{1-q}\right)dt.$$

Partial integration now yields

$$\begin{aligned}
&\int_{a-1}^{b} t^{q-p}t^{-q}\exp\left(-\frac{t^{1-q}}{1-q}\right)dt \\
&= \left[-t^{q-p}\exp\left(-\frac{t^{1-q}}{1-q}\right)\right]_{t=a-1}^{t=b} - (p-q)\int_{a-1}^{b} t^{q-p-1}\exp\left(-\frac{t^{1-q}}{1-q}\right)dt \\
&= (a-1)^{q-p}\exp\left(-\frac{(a-1)^{1-q}}{1-q}\right) - b^{q-p}\exp\left(-\frac{b^{1-q}}{1-q}\right) + (q-p)\int_{a-1}^{b} t^{q-p-1}\exp\left(-\frac{t^{1-q}}{1-q}\right)dt.
\end{aligned}$$

Finally, we use that $t^{q-p-1}\exp\left(-\frac{t^{1-q}}{1-q}\right)$ is monotonically decreasing and $p \geq q$ to derive

$$(q-p)\int_{a-1}^{b} t^{q-p-1}\exp\left(-\frac{t^{1-q}}{1-q}\right)dt \leq (q-p)b^{q-1}\int_{a-1}^{b} t^{-p}\exp\left(-\frac{t^{1-q}}{1-q}\right)dt.$$

Noting that this is the integral we started with and rearranging yields the claim.
iii) The proof of the last claim uses the same arguments as in [8]. First we use *i*) to obtain

$$\sum_{t=a}^{b} t^{-p}\prod_{\tau=t+1}^{b}\left(1-\tau^{-q}\right) \leq \sum_{t=a}^{b} t^{-p}\exp\left(-\sum_{\tau=t+1}^{b}\tau^{-q}\right) = \exp\left(-\sum_{\tau=1}^{b}\tau^{-q}\right)\sum_{t=a}^{b} t^{-p}\exp\left(\sum_{\tau=1}^{t}\tau^{-q}\right).$$

Using the monotonicity of τ^{-q}, we get

$$\exp\left(-\sum_{\tau=1}^{b}\tau^{-q}\right) \leq \exp\left(-\int_{1}^{b+1}\tau^{-q}d\tau\right) = \exp\left(\frac{1-(b+1)^{1-q}}{1-q}\right)$$

and

$$\exp\left(\sum_{\tau=1}^{t}\tau^{-q}\right) \leq \exp\left(\int_{0}^{t}\tau^{-q}d\tau\right) = \exp\left(\frac{t^{1-q}}{1-q}\right).$$

We now proceed to bound

$$\sum_{t=a}^{b} t^{-p}\exp\left(\sum_{\tau=1}^{t}\tau^{-q}\right) \leq \sum_{t=a}^{b} t^{-p}\exp\left(\frac{t^{1-q}}{1-q}\right).$$

Therefore, note that $f(t) \coloneqq t^{-p}\exp\left(\frac{t^{1-q}}{1-q}\right)$ is monotonically increasing for $t \geq a$ by our assumption on a. This implies

$$\sum_{t=a}^{b} t^{-p}\exp\left(\frac{t^{1-q}}{1-q}\right) \leq \int_{a}^{b+1} t^{-p}\exp\left(\frac{t^{1-q}}{1-q}\right)dt =: I.$$

Integration by party now yields

$$\begin{aligned} I &= \int_a^{b+1} t^{q-p} t^{-q} \exp\left(\frac{t^{1-q}}{1-q}\right) dt \\ &= \left[t^{q-p} \exp\left(\frac{t^{1-q}}{1-q}\right)\right]_{t=a}^{t=b+1} - (q-p)\int_a^{b+1} t^{q-p-1} \exp\left(\frac{t^{1-q}}{1-q}\right) dt \\ &\leq (b+1)^{q-p} \exp\left(\frac{(b+1)^{1-q}}{1-q}\right) - a^{q-p} \exp\left(\frac{a^{1-q}}{1-q}\right) + (p-q)a^{q-1} I, \end{aligned}$$

where we used $p \geq q$ in the last inequality. By our second assumption on a we now get that $(p-q)a^{q-1} \leq 1/2$ and hence

$$I \leq 2(b+1)^{q-p} \exp\left(\frac{(b+1)^{1-q}}{1-q}\right) - 2a^{q-p} \exp\left(\frac{a^{1-q}}{1-q}\right).$$

Putting together the pieces yields

$$\begin{aligned} \sum_{t=a}^{b} t^{-p} \prod_{\tau=t+1}^{b} \left(1-\tau^{-q}\right) &\leq 2\exp\left(\frac{1-(b+1)^{1-q}}{1-q}\right)\left((b+1)^{q-p}\exp\left(\frac{(b+1)^{1-q}}{1-q}\right) - a^{q-p}\exp\left(\frac{a^{1-q}}{1-q}\right)\right) \\ &= 2\exp\left(\frac{1}{1-q}\right)(b+1)^{q-p} - a^{q-p}\exp\left(\frac{1-(b+1)^{1-q}+a^{1-q}}{1-q}\right), \end{aligned}$$

thus proving the last claim. ■

The following lemma applies the specific values of p and q to Theorem 9.

Lemma 10 (Technical Lemma) *Let $\eta > 0$ and for $t \in \mathbb{N}_{\geq 1}$ we set*

$$\begin{aligned} \beta_t &:= 1 - t^{-1/2}, \\ \eta_t &:= \eta t^{-3/4}. \end{aligned}$$

Then, for $\alpha_t := 1 - \beta_t$, we have

a) For all $T \in \mathbb{N}_{\geq 1}$ the following inequalities hold:

i) $\sum_{t=1}^{T} \eta_t \prod_{\tau=2}^{t} \beta_\tau \leq \frac{7}{2}\eta$;

ii) $\sum_{t=1}^{T} \eta_t \sqrt{\sum_{\tau=1}^{t} \alpha_\tau^2 \prod_{\kappa=\tau+1}^{t} \beta_\kappa^2} \leq \eta\left(\frac{7}{2} + \sqrt{2e^2}\log(T)\right)$.

b) Let $T \in \mathbb{N}_{\geq 1}$ and define $C_t := 1 + 2\frac{e^{L_1\eta_t}-1}{L_1\eta_t} - 4\frac{e^{L_1\eta_t}-1-L_1\eta_t}{(L_1\eta_t)^2}$, $E_t := e^{L_1\eta_t}$. Then the following inequalities hold:

i) $\sum_{t=1}^{T} \eta_t^2 C_t \leq 6\eta^2 \frac{e^{L_1\eta}-1}{L_1\eta}$;

ii) $\sum_{t=1}^{T} \eta_t \sum_{\tau=2}^{t} \eta_{\tau-1} E_{\tau-1} \prod_{\kappa=\tau}^{t} \beta_\kappa \leq \frac{7}{2}\eta^2\left(3e^{\eta L_1} + \log(T)\right)$.

c) For $t \in \mathbb{N}_{\geq 1}$ we define $E_t := e^{L_1 \eta_t}$ and $\delta_t := 4\eta(t-1)^{\frac{1}{4}} - 3\eta$. Then for all $b \in \mathbb{N}_{\geq 2}$, the following inequalities hold:

i) $\sum_{t=2}^{b} L_1 \eta_t E_t t^{-\frac{1}{4}} \delta_t e^{L_1 \delta_t} \leq \frac{1}{2}\eta^2 L_1 e^{2\eta L_1} + 4\eta e^{-\frac{5}{2}\eta L_1}\left(e^{4\eta L_1 b^{\frac{1}{4}}} - e^{4\eta L_1}\right)$;

ii) $\sum_{t=1}^{b} L_1 \eta_t E_t t^{-\frac{1}{4}} e^{L_1 \delta_t} \leq \frac{3}{2}\eta L_1 e^{\frac{5}{3}\eta L_1} + e^{-\frac{5}{2}\eta L_1}\left(e^{4\eta L_1 b^{\frac{1}{4}}} - e^{4\eta L_1}\right)$;

iii) If additionally $\eta L_1 \geq \frac{1}{2}$, we have
$\sum_{t=1}^{b} L_1 \eta_t E_t t^{-\frac{1}{4}} e^{L_1 \delta_t} \leq \frac{3}{2}\eta L_1 e^{\frac{5}{3}\eta L_1} + e^{-\frac{5}{2}\eta L_1}\left(2b^{-\frac{1}{4}} e^{4\eta L_1 b^{\frac{1}{4}}} - e^{4\eta L_1}\right)$.

Proof Let $T \in \mathbb{N}_{\geq 1}$ and denote $p := 3/4, q := 1/2, \alpha_t := 1 - \beta_t$ for simplicity.

a) i) The inequality follows from

$$\sum_{t=1}^{T} \eta_t \prod_{\tau=2}^{t} \beta_\tau = \eta + \sum_{t=2}^{T} \eta_t \prod_{\tau=2}^{t} \beta_\tau \leq \eta + \eta \exp\left(2\sqrt{2} - 2\right) \leq \frac{7}{2}\eta,$$

where we used Theorem 9 *ii)* in the first inequality.

a) ii) We start by regrouping

$$\sum_{t=1}^{T} \eta_t \sqrt{\sum_{\tau=1}^{t} \alpha_\tau^2 \prod_{\kappa=\tau+1}^{t} \beta_\kappa^2} < \sum_{t=1}^{T} \eta_t \left(\prod_{\kappa=2}^{t}\left(1 - \kappa^{-q}\right) + \sqrt{\sum_{\tau=2}^{t} \tau^{-2q} \prod_{\kappa=\tau+1}^{t}\left(1 - \kappa^{-q}\right)}\right).$$

Applying Theorem 9 *i)*, *iii)* and *a) i)* now yields the statement:

$$\sum_{t=1}^{T} \eta_t \sqrt{\sum_{\tau=1}^{t} \alpha_\tau^2 \prod_{\kappa=\tau+1}^{t} \beta_\kappa^2} \overset{i),9}{\leq} \frac{7}{2}\eta + \sum_{t=2}^{T} \eta_t \sqrt{2e^2(t+1)^{-q}} \leq \eta\left(\frac{7}{2} + \sqrt{2e^2}\log(T)\right).$$

Note that the first inequality is rather loose, a more precise analysis might yield a better result. The above result does however suffice for our use-case.

b) i) First note that

$$C_t \leq 2\frac{e^{L_1\eta_t} - 1}{L_1 \eta_t} \tag{4}$$

and hence

$$\sum_{t=1}^{T} \eta_t^2 C_t \leq \frac{2}{L_1} \sum_{t=1}^{T} \eta_t \left(e^{L_1 \eta_t} - 1\right). \tag{5}$$

Now we calculate

$$\sum_{t=1}^{T} \eta_t \left(e^{L_1\eta_t} - 1\right) \leq \eta^2 \frac{e^{\eta L_1} - 1}{\eta} + \int_1^T \eta_t \left(e^{L_1 \eta_t} - 1\right) dt = \eta^2 \left(\frac{e^{\eta L_1} - 1}{\eta} + \frac{1}{\eta}\int_1^T t^{-p}\left(e^{L_1\eta_t} - 1\right)\right) \tag{6}$$

and further

$$\int_1^T t^{-p}\big(e^{L_1\eta_t}-1\big) = \int_1^T t^{-p}\sum_{k=1}^{\infty}\frac{(\eta L_1 t^{-p})^k}{k!}dt = \sum_{k=1}^{\infty}\int_1^T \frac{(\eta L_1)^k}{k!}t^{-p(k+1)}dt, \tag{7}$$

where we used that the exponential series converges locally uniformly in the second equality. Finally we calculate for $k \geq 2$

$$\int_1^T t^{-p(k+1)}dt = \frac{4}{-3k+1}\Big(T^{-p(k+1)+1}-1\Big) \leq \frac{4}{3k-1} \leq \frac{4}{k+1}. \tag{8}$$

Combining (7) and (8) now yields

$$\int_1^T t^{-p}\big(e^{L_1\eta_t}-1\big) \leq 4\sum_{k=1}^{\infty}\frac{(\eta L_1)^k}{(k+1)!} = 4\frac{e^{\eta L_1}-1-\eta L_1}{\eta L_1} \tag{9}$$

and hence

$$\begin{aligned}\sum_{t=1}^T \eta_t^2 C_t &\overset{(5)}{\leq} \frac{2}{L_1}\sum_{t=1}^T \eta_t\big(e^{L_1\eta_t}-1\big)\\ &\overset{(6)}{\leq} \eta^2\left(2\frac{e^{\eta L_1}-1}{\eta L_1} + \frac{2}{\eta L_1}\int_1^T t^{-p}\big(e^{L_1\eta_t}-1\big)dt\right)\\ &\overset{(9)}{\leq} \eta^2\left(2\frac{e^{\eta L_1}-1}{\eta L_1} + 8\frac{e^{\eta L_1}-1-\eta L_1}{(\eta L_1)^2}\right).\end{aligned}$$

The claim now follows by noting that for all $x \geq 0$ the inequality $2\frac{e^x-1-x}{x^2} \leq \frac{e^x-1}{x}$ is satisfied.

b) ii) Firstly, *a) i)* yields

$$\begin{aligned}\sum_{t=1}^T\sum_{\tau=2}^t \eta_t\eta_{\tau-1}\left(\prod_{\kappa=\tau}^t\beta_\kappa\right)E_{\tau-1} &= \sum_{\tau=2}^T \eta_{\tau-1}E_{\tau-1}\sum_{t=\tau}^T\eta_t\prod_{\kappa=\tau}^t\beta_\kappa\\ &\overset{a)i)}{\leq} \frac{7}{2}\eta\sum_{\tau=2}^T\eta_{\tau-1}E_{\tau-1}(\tau-1)^{q-p}\\ &\leq \frac{7}{2}\eta^2\sum_{\tau=1}^T\tau^{-1}E_\tau.\end{aligned} \tag{10}$$

To upper bound $\sum_{t=1}^T t^{-1}E_t = e^{\eta L_1} + \sum_{t=2}^T t^{-1}E_t$ we again use the locally uniform convergence of the exponential series to get

$$\begin{aligned}\sum_{t=2}^{T} t^{-1} E_t &\leq \int_1^T t^{-1} \sum_{k=0}^{\infty} \frac{(L_1\eta_t)^k}{k!} dt \\ &= \log(T) + \sum_{k=1}^{\infty} \frac{(\eta L_1)^k}{k!} \int_1^T t^{-pk-1} dt \\ &= \log(T) + \sum_{k=1}^{\infty} \frac{(\eta L_1)^k}{k!} \frac{1 - T^{-pk}}{pk} \\ &\leq \log(T) + \frac{4}{3}\big(e^{\eta L_1} - 1\big).\end{aligned}$$

Combining these results yields

$$\begin{aligned}\sum_{t=1}^{T} \eta_t \sum_{\tau=2}^{t} \eta_{\tau-1} E_{\tau-1} \prod_{\kappa=\tau}^{t} \beta_\kappa &\leq \frac{7}{2}\eta^2 \left(e^{\eta L_1} + \log(T) + \frac{4}{3}\big(e^{\eta L_1} - 1\big) \right) \\ &\leq \frac{7}{2}\eta^2 \big(\log(T) + 3e^{\eta L_1} \big)\end{aligned}$$

and hence proves the claim.

c) i) We start off by calculating

$$\begin{aligned}\sum_{t=2}^{b} L_1 \eta_t E_t t^{-\frac{1}{4}} \delta_t e^{L_1 \delta_t} &\leq L_1 \eta_2 E_2 2^{-\frac{1}{4}} \eta e^{\eta L_1} + \sum_{t=3}^{b} L_1 \eta_t E_t t^{-\frac{1}{4}} \delta_t e^{L_1 \delta_t} \\ &\leq \frac{1}{2}\eta^2 L_1 e^{2\eta L_1} + 4\eta \sum_{t=3}^{b} L_1 \eta_t E_t e^{L_1 \delta_t}\end{aligned}$$

and further

$$\begin{aligned}\sum_{t=3}^{b} L_1 \eta_t E_t e^{L_1 \delta_t} &\leq \sum_{t=3}^{b} L_1 \eta_t \exp\left(L_1 \left(4\eta (t-1)^{\frac{1}{4}} - 3\eta + \eta_t \right) \right) \\ &\leq e^{-\frac{5}{2}\eta L_1} \sum_{t=3}^{b} L_1 \eta_{t-1} e^{4\eta L_1 (t-1)^{\frac{1}{4}}} \\ &\leq e^{-\frac{5}{2}\eta L_1} \int_2^{b+1} \eta L_1 (t-1)^{-p} e^{4\eta L_1 (t-1)^{1-p}} dt \\ &= e^{-\frac{5}{2}\eta L_1} \left(e^{4\eta L_1 b^{\frac{1}{4}}} - e^{4\eta L_1} \right).\end{aligned} \tag{11}$$

Here we used that $g(t) \coloneqq L_1 \eta_{t-1} e^{4\eta L_1 (t-1)^{\frac{1}{4}}}$ is non-negative and monotonically decreasing before turning monotonically increasing in the third inequality. Noting that (11) also holds for $b = 2$ yields the claim.

ii) We have

$$\begin{aligned}\sum_{t=1}^{b} L_1\eta_t E_t t^{-\frac{1}{4}} e^{L_1\delta_t} &= \eta L_1 E_1 + \frac{1}{2}\eta L_1 E_2 e^{\eta L_1} + \sum_{t=3}^{b} L_1\eta_t E_t t^{-\frac{1}{4}} e^{L_1\delta_t} \\ &\leq \eta L_1 e^{\eta L_1} + \frac{1}{2}\eta L_1 e^{\left(1+2^{-\frac{3}{4}}\right)\eta L_1} + \sum_{t=3}^{b} L_1\eta_t E_t e^{L_1\delta_t}\end{aligned} \tag{12}$$

and using (11) yields

$$\sum_{t=1}^{b} L_1\eta_t E_t t^{-\frac{1}{4}} e^{L_1\delta_t} \leq \frac{3}{2}\eta L_1 e^{\frac{5}{3}\eta L_1} + e^{-\frac{5}{2}\eta L_1}\left(e^{4\eta L_1 b^{\frac{1}{4}}} - e^{4\eta L_1}\right).$$

iii) We first again calculate

$$\sum_{t=3}^{b} L_1\eta_t E_t t^{-\frac{1}{4}} e^{L_1\delta_t} \leq e^{-\frac{5}{2}\eta L_1}\int_2^{b+1} t^{-\frac{1}{4}} L_1\eta(t-1)^{-\frac{3}{4}} e^{4\eta L_1(t-1)^{\frac{1}{4}}} dt$$

before, similar to the proof of Theorem 9 *iii*), using partial integration to derive

$$\begin{aligned} I &:= \int_2^{b+1} t^{-\frac{1}{4}} L_1\eta(t-1)^{-\frac{3}{4}} e^{4\eta L_1(t-1)^{\frac{1}{4}}} dt \\ &= \left[t^{-\frac{1}{4}} e^{4\eta L_1(b-1)^{\frac{1}{4}}}\right]_{t=2}^{t=b+1} + \frac{1}{4}\int_2^{b+1} t^{-\frac{5}{4}} e^{4\eta L_1(t-1)^{\frac{1}{4}}} dt \\ &\leq b^{-\frac{1}{4}} e^{4\eta L_1 b^{\frac{1}{4}}} - \frac{1}{2}e^{4\eta L_1} + \frac{1}{2^{\frac{1}{4}} 4\eta L_1}\int_2^{b+1} \eta L_1(t-1)^{-1} e^{4\eta L_1(t-1)^{\frac{1}{4}}} \\ &\leq b^{-\frac{1}{4}} e^{4\eta L_1 b^{\frac{1}{4}}} - \frac{1}{2}e^{4\eta L_1} + \frac{1}{4\eta L_1} I. \end{aligned}$$

By our assumption we have $\frac{1}{4\eta L_1} \leq \frac{1}{2}$ and hence

$$I \leq 2b^{-\frac{1}{4}} e^{4\eta L_1 b^{\frac{1}{4}}} - e^{4\eta L_1}.$$

Finally (12) yields the claim.

■

Appendix D. Missing Proofs

This section contains the missing proofs and mentioned results from Section 3.

D.1. Upper Bounds

D.1.1. PARAMETER-AGNOSTIC

We start with the proof of Theorem 2, which has the same structure as in the L-smooth setting [5]: We first derive a Descent Lemma, second bound the momentum deviation $\|m_t - \nabla F(x_t)\|$ and third combine these two to show the result. The last step is however more intricate, as large stepsizes in the beginning can lead to an exponential increase in the gradient norm. The main intuitions behind the third step are the following:

Due to potentially too large stepsizes, we cannot use the descent lemma to control the expected gradient norm in the beginning. Only after reaching a threshold $t_0 \propto (\eta L_1)^4$ the gradient norms can be controlled in this fashion. Before this threshold, in the *adaption phase*, we instead use (L_0, L_1)-smoothness to control the gradient norms based on $\|\nabla F(x_1)\|$. After this threshold, in the *convergence phase*, Lemma 10 essentially establishes that the diminishing step-size rule $\eta_t = t^{-p}$ exhibits the same asymptotically behaviour as if the stepsizes were chosen constantly as $\eta_t \equiv T^{-p}$, where T denotes the iteration horizon. This aligns with the behaviour of NSGD-M in the L-smooth setting [25]. In particular, this implies that $p = 3/4$ is the only possible choice to achieve the optimal complexity [5, 27].

Unless stated otherwise, the notations $\{\xi_1, \xi_2, \ldots\}, \{g_1, g_2, \ldots\}, \{m_1, m_2, \ldots\}$ and $\{x_1, x_2, \ldots\}$ correspond to the iterations generated by NSGD-M throughout this section. We denote the natural filtration of $\xi_1, \ldots, \xi_t$ with respect to the underlying probability space by $\mathcal{F}_t \coloneqq \sigma(\xi_1, \xi_2, \ldots, \xi_t)$.

Lemma 11 (Descent Lemma) *Assume ((L0, L1)-smoothness) and let $t \in \mathbb{N}_{\geq 2}$. Then*

$$F(x_{t+1}) - F(x_t) \leq -\eta_t \|\nabla F(x_t)\| + 2\eta_t \|\nabla F(x_t) - m_t\| + \frac{\eta_t^2}{2}(L_0 C_t + L_1 D_t \|\nabla F(x_t)\|),$$

where $C_t \coloneqq B_0(L_1\eta_t)$ and $D_t \coloneqq B_1(L_1\eta_t)$, where B_0, B_1 are as defined in Theorem 7. If we further assume (Lower Boundedness) we also get

$$\sum_{t=1}^{T} \left(\eta_t - \frac{L_1 \eta_t^2 D_t}{2}\right) \|\nabla F(x_t)\| \leq \Delta_1 + \frac{L_0}{2} \sum_{t=1}^{T} \eta_t^2 C_t + 2 \sum_{t=1}^{T} \eta_t \|\nabla F(x_t) - m_t\|,$$

where $\Delta_1 \coloneqq F(x_1) - F^*$.

Proof The proof follows the arguments by Zhao et al. [29]. Using Lemma 7 we get

$$\begin{aligned} F(x_{t+1}) - F(x_t) &\leq \nabla F(x_t)^\top (x_{t+1} - x_t) + \frac{\eta_t^2}{2}(L_0 C_t + L_1 D_t \|\nabla F(x_t)\|) \\ &= -\frac{\eta_t}{\|m_t\|} \nabla F(x)^\top m_t + \frac{\eta_t^2}{2}(L_0 C_t + L_1 D_t \|\nabla F(x_t)\|) \\ &= -\frac{\eta_t}{\|m_t\|} (\nabla F(x_t) - m_t)^\top m_t - \eta_t \|m_t\| + \frac{\eta_t^2}{2}(L_0 C_t + L_1 D_t \|\nabla F(x_t)\|). \end{aligned}$$

Utilizing Cauchy-Schwarz and $\eta_t \|\nabla F(x_t)\| \leq \eta_t \|\nabla F(x_t) - m_t\| + \eta_t \|m_t\|$ now yields

$$F(x_{t+1}) - F(x_t) \leq -\eta_t \|\nabla F(x_t)\| + 2\eta_t \|\nabla F(x_t) - m_t\| + \frac{\eta_t^2}{2}(L_0 C_t + L_1 D_t \|\nabla F(x_t)\|)$$

and hence the first claim. For the second statement we sum up to get

$$\sum_{t=1}^{T} \left(\eta_t - \frac{L_1 \eta_t^2 D_t}{2}\right) \|\nabla F(x)\| \leq \Delta_1 + \frac{1}{2}\sum_{t=1}^{T} L_0 \eta_t^2 C_t + 2\sum_{t=1}^{T} \eta_t \|\nabla F(x_t) - m_t\| .$$

■

Lemma 12 (General Momentum Deviation Bound) *Assume ((L0, L1)-smoothness), (Bounded Variance) and let $t \in \mathbb{N}_{\geq 1}$. Suppose $\beta_1 = 0$. Then we have*

$$\mathbb{E}\left[\|m_t - \nabla F(x_t)\|\right] \leq \sigma \sqrt{\sum_{\tau=1}^{t} \beta_{(\tau+1):t}^2 (1-\beta_\tau)^2} + L_0 \sum_{\tau=2}^{t} \eta_{\tau-1} E_{\tau-1} \beta_{\tau:t} + L_1 \sum_{\tau=2}^{t} \eta_{\tau-1} E_{\tau-1} \beta_{\tau:t} \mathbb{E}\left[\|\nabla F(x_{\tau-1})\|\right],$$

where $\beta_{a:b}$ denotes $\prod_{t=a}^{b} \beta_t$ and $E_t \coloneqq e^{L_1 \eta_t}$.

Proof This proof is motivated by Cutkosky and Mehta [5], and similar arguments are carried by Zhang et al. [27] and Yang et al. [25]. To simplify notation we first define

$$\begin{aligned} \mu_t &\coloneqq m_t - \nabla F(x_t), \\ \gamma_t &\coloneqq g_t - \nabla F(x_t), \\ \alpha_t &\coloneqq 1 - \beta_t, \\ \beta_{a:b} &\coloneqq \prod_{t=a}^{b} \beta_t. \end{aligned}$$

Now let $i, j \in \mathbb{N}, i < j$ and calculate

$$\begin{aligned} \mathbb{E}\left[\gamma_j^\top \gamma_i\right] &= \mathbb{E}\left[\mathbb{E}\left[\gamma_j^\top \gamma_i \,\middle|\, \mathcal{F}_{j-1}\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\gamma_j \,|\, \mathcal{F}_{j-1}\right]^\top \gamma_i\right] \\ &= 0, \end{aligned} \tag{13}$$

where we used that $\mathbb{E}\left[\gamma_j \,|\, \mathcal{F}_{j-1}\right] = 0$ in the last equality. Next we define $S_t \coloneqq \nabla F(x_{t-1}) - \nabla F(x_t)$ and calculate

$$\begin{aligned} m_t &= \beta_t m_{t-1} + (1-\beta_t) g_t \\ &= \beta_t(\nabla F(x_{t-1}) + \mu_{t-1}) + (1-\beta_t)(\gamma_t + \nabla F(x_t)) \\ &= \nabla F(x_t) + (1-\beta_t)\gamma_t + \beta_t S_t + \beta_t \mu_{t-1}. \end{aligned}$$

This yields

$$\mu_t = \beta_{2:t}\mu_1 + \sum_{\tau=2}^{t} \beta_{(\tau+1):t}\alpha_\tau\gamma_\tau + \sum_{\tau=2}^{t} \beta_{\tau:t}S_\tau = \sum_{\tau=1}^{t} \beta_{(\tau+1):t}\alpha_\tau\gamma_\tau + \sum_{\tau=2}^{t} \beta_{\tau:t}S_\tau,$$

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

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

To further concretize this upper bound, (13) firstly yields

$$\mathbb{E}\left[\left\|\sum_{\tau=1}^{t} \beta_{(\tau+1):t}\alpha_\tau\gamma_\tau\right\|\right] \leq \sqrt{\sum_{\tau=1}^{t} \beta_{(\tau+1):t}^2\alpha_t^2\sigma^2}.$$

Secondly, ((L_0, L_1)-smoothness) implies

$$\begin{aligned}\|S_t\| &\leq \eta_{t-1}(A_0(L_1\eta_{t-1})L_0 + A_1(L_1\eta_{t-1})L_1\|\nabla F(x_{t-1})\|) \\ &\leq \eta_{t-1}(E_{t-1}L_0 + E_{t-1}L_1\|\nabla F(x_{t-1})\|)\end{aligned}$$

and hence

$$\sum_{\tau=2}^{t} \beta_{\tau:t}\mathbb{E}\left[\|S_\tau\|\right] \leq L_0\sum_{\tau=2}^{t} \eta_{\tau-1}E_{\tau-1}\beta_{\tau:t} + L_1\sum_{\tau=2}^{t} \eta_{\tau-1}E_{\tau-1}\beta_{\tau:t}\mathbb{E}\left[\|\nabla F(x_{\tau-1})\|\right].$$

Putting these results together we get the claim. ■

Now we are ready for the main result.

Theorem 13 (NSGD-M for (L_0, L_1)-smoothness) *Assume (Lower Boundedness), ((L_0, L_1)-smoothness) and (Bounded Variance). Let $\eta > 0$ and define the parameters*

$$\begin{aligned}\beta_t &:= 1 - t^{-1/2} \\ \eta_t &:= \eta t^{-3/4}.\end{aligned}$$

Then NSGD-M with starting point $x_1 \in \mathbb{R}^d$ satisfies

$$\begin{aligned}\sum_{t=1}^{T} \frac{\eta_t}{2}\mathbb{E}\left[\|\nabla F(x_t)\|\right] \leq \Delta_1 + \eta\sigma\left(7 + 2\sqrt{2e^2}\log(T)\right) + \eta^2 L_0\left(21e^{\eta L_1} + 7\log(T)\right) \\ + 21\eta^2 L_0 e^{48(\eta L_1)^2} + 6\eta e^{48(\eta L_1)^2}\|\nabla F(x_1)\|,\end{aligned}$$

where $\Delta_1 := F(x_1) - F^*$. Furthermore, if $L_1 \geq 1/2\eta$, the statement also holds when replacing $6\eta e^{48(\eta L_1)^2}\|\nabla F(x_1)\|$ with $\frac{e^{48(\eta L_1)^2}}{L_1}\|\nabla F(x_1)\|$.

The main workhorse behind the following proof is Lemma 10. It intuitively states that the quantities which emerge due to the nonconstant parameters behave (nearly) *asymptotically the same* as constant stepsizes would.

Proof To simplify notation we define

$$\beta_{a:b} \coloneqq \prod_{\tau=a}^{b} \beta_\tau.$$

We start the proof by combining Lemma 11 and Lemma 12 to obtain

$$\begin{aligned}
\sum_{t=1}^{T} \eta_t \mathbb{E}\left[\|\nabla F(x_t)\|\right] &\overset{11}{\leq} \Delta_1 + \frac{L_0}{2}\sum_{t=1}^{T}\eta_t^2 C_t + \frac{L_1}{2}\sum_{t=1}^{T}\eta_t^2 D_t \mathbb{E}\left[\|\nabla F(x_t)\|\right] + 2\sum_{t=1}^{T}\eta_t \mathbb{E}\left[\|\nabla F(x_t) - m_t\|\right] \\
&\overset{12}{\leq} \Delta_1 + \frac{L_0}{2}\sum_{t=1}^{T}\eta_t^2 C_t + \frac{L_1}{2}\sum_{t=1}^{T}\eta_t^2 D_t \mathbb{E}\left[\|\nabla F(x_t)\|\right] + 2\sigma\sum_{t=1}^{T}\eta_t\sqrt{\sum_{\tau=1}^{t}\alpha_\tau^2\left(\beta_{(\tau+1):t}\right)^2} \\
&\quad + 2L_0\sum_{t=1}^{T}\eta_t\sum_{\tau=2}^{t}\eta_{\tau-1}E_{\tau-1}\beta_{\tau:t} + 2L_1\sum_{t=1}^{T}\eta_t\sum_{\tau=2}^{t}\eta_{\tau-1}E_{\tau-1}\beta_{\tau:t}\mathbb{E}\left[\|\nabla F(x_{\tau-1})\|\right].
\end{aligned}$$

Next, we use Lemma 10 a) and b) to bound all terms that are independent of the iterates x_t. This leaves us with

$$\begin{aligned}
\sum_{t=1}^{T} \eta_t \mathbb{E}\left[\|\nabla F(x_t)\|\right] &\leq \Delta_1 + \eta\sigma\left(7 + 2\sqrt{2e^2}\log\left(T\right)\right) + \eta^2 L_0\left(21e^{\eta L_1} + 7\log\left(T\right)\right) \\
&\quad + \underbrace{\frac{L_1}{2}\sum_{t=1}^{T}\eta_t^2 D_t \mathbb{E}\left[\|\nabla F(x_t)\|\right] + 2L_1\sum_{\tau=2}^{T}\left(\sum_{t=\tau}^{T}\eta_t\beta_{\tau:t}\right)\eta_{\tau-1}E_{\tau-1}\mathbb{E}\left[\|\nabla F(x_{\tau-1})\|\right]}_{=:(A)},
\end{aligned} \tag{14}$$

where we rearranged the sums of the last term. We then focus on upper bounding (A). Therefore we use Lemma 9 ii) which yields

$$(A) \leq \sum_{t=1}^{T}\eta_t E_t\left(\frac{L_1}{2}\eta_t + 2e^{2\left(\sqrt{2}-1\right)}L_1\eta t^{-\frac{1}{4}}\right)\mathbb{E}\left[\|\nabla F(x_t)\|\right] \leq \sum_{t=1}^{T}\eta_t E_t\left(ML_1\eta t^{-1/4}\right)\mathbb{E}\left[\|\nabla F(x_t)\|\right],$$

where $M \coloneqq \frac{1}{2} + 2\exp\left(2\sqrt{2}-2\right) \leq 5.1$. In a setting with access to problem parameters, we could now set $\eta \coloneqq \frac{1}{12L_1}$ and hence guarantee that $M\eta L_1 t^{-\frac{1}{4}}E_t \leq \frac{1}{2}$, which would complete the proof. In the parameter agnostic setting we have to wait until the stepsize decreased below this threshold. We therefore define the threshold $t_0 \coloneqq \lceil (12\eta L_1)^4 \rceil$ after which we again have $M\eta L_1 t^{-\frac{1}{4}}E_t \leq \frac{1}{2}$. This is due to $E_t \leq E_{t_0} \leq \frac{12}{2M}$ for $t \geq t_0$. We are therefore left with the task of controlling the sum in (A) up to t_0, i.e. (B) in

$$(A) \leq \underbrace{\sum_{t=1}^{t_0-1}\eta_t\left(ML_1\eta t^{-1/4}E_t\right)\mathbb{E}\left[\|\nabla F(x_t)\|\right]}_{(B)} + \sum_{t=t_0}^{T}\frac{\eta_t}{2}\mathbb{E}\left[\|\nabla F(x_t)\|\right]. \tag{15}$$

We start by upper bounding $\|\nabla F(x_t)\|$ using ((L_0, L_1)-smoothness). For $\delta_t \coloneqq \|x_t - x_1\| \leq 4\eta t^{\frac{1}{4}} - 3\eta$ our smoothness assumption implies

$$\|\nabla F(x_t)\| \leq \|\nabla F(x_1)\| + \|\nabla F(x_t) - \nabla F(x_1)\| \leq e^{L_1\delta_t} L_0 \delta_t + e^{L_1\delta_t} \|\nabla F(x_1)\|$$

and plugging into (B) yields

$$(B) \leq \underbrace{\left(\eta M \sum_{t=2}^{t_0-1} L_1 \eta_t t^{-\frac{1}{4}} \delta_t E_t e^{L_1\delta_t}\right)}_{=:(B1)} L_0 + \underbrace{\left(\eta M \sum_{t=1}^{t_0-1} L_1 \eta t^{-1} E_t e^{L_1\delta_t}\right)}_{=:(B2)} \|\nabla F(x_1)\|.$$

Now Lemma 10 $c)$ $i)$ allows us to upper bound $(B1)$ via

$$\begin{aligned}
(B1) &\leq \eta^2 M L_0 \left(\frac{\eta L_1}{2} e^{2\eta L_1} + 4e^{-\frac{5}{2}\eta L_1}\left(e^{4\eta L_1 (t_0-1)^{\frac{1}{4}}} - e^{4\eta L_1}\right)\right) \\
&\leq \eta^2 M L_0 \left(\left(\frac{\eta L_1}{2} - 4\right) e^{2\eta L_1} + 4e^{4\eta L_1 (t_0-1)^{\frac{1}{4}}}\right) \\
&\leq \eta^2 M L_0 \left(\left(\frac{\eta L_1}{2} - 4\right) e^{2\eta L_1} + 4e^{48(\eta L_1)^2}\right),
\end{aligned}$$

where we used the definition of t_0 in the last inequality. Next we use that, for all $x \geq 0$, we have $(x/2 - 4)e^{4x} + e^{48x^2} \leq \frac{21}{4M} e^{48x^2}$ and hence

$$(B1) \leq 21\eta^2 L_0 e^{48(\eta L_1)^2}.$$

Using Lemma 10 $c)$ $ii)$ and the same technique as for $(B1)$ we obtain

$$\begin{aligned}
(B2) &\leq \eta M \left(\frac{3}{2}\eta L_1 e^{5/3\eta L_1} + e^{-5/2\eta L_1}\left(e^{4\eta L_1 (t_0-1)^{\frac{1}{4}}} - e^{4\eta L_1}\right)\right) \\
&\leq \eta M \left(\left(\frac{3}{2}\eta L_1 - 1\right) e^{2\eta L_1} + e^{48(\eta L_1)^2}\right) \\
&\leq 6\eta e^{48(\eta L_1)^2} < \frac{3}{L_1} e^{48(\eta L_1)^2}.
\end{aligned}$$

We plug these results into (15) to obtain

$$(A) \leq 21\eta^2 L_0 e^{48(\eta L_1)^2} L_0 + 6\eta e^{48(\eta L_1)^2} \|\nabla F(x_1)\| + \sum_{t=t_0}^{T} \frac{\eta_t}{2} \mathbb{E}\left[\|\nabla F(x_t)\|\right]$$

and combing with (14) yields

$$\begin{aligned}
\frac{1}{2}\sum_{t=1}^{T} \eta_t \mathbb{E}\left[\|\nabla F(x_t)\|\right] \leq \Delta_1 + \eta\sigma\left(7 + 2\sqrt{2e^2}\log(T)\right) + \eta^2 L_0 \left(45\frac{e^{\eta L_1} - 1}{\eta L_1} + 14\log(T)\right) \\
+ 21\eta^2 L_0 e^{48(\eta L_1)^2} L_0 + 6\eta e^{48(\eta L_1)^2} \|\nabla F(x_1)\|.
\end{aligned}$$

This finishes the proof of the first statement.

For the second statement assume $\eta L_1 \geq 1/2$. In this case we apply Lemma 10 $c)$ $iii)$ and get

$$\begin{aligned}(B2) &\leq \eta M\left(\frac{3}{2}\eta L_1 e^{5/3\eta L_1} + e^{-5/2\eta L_1}\left(2(t_0-1)^{-1/4}e^{4\eta L_1 (t_0-1)^{1/4}} - e^{4\eta L_1}\right)\right)\\ &\leq \eta M\left(\left(\frac{3}{2}\eta L_1 - 1\right)e^{2\eta L_1} + \frac{1}{6\eta L_1}e^{48(\eta L_1)^2}\right)\\ &\leq \frac{1}{L_1}e^{48(\eta L_1)^2}\end{aligned}$$

Proceeding as before yields the second claim. ■

By plugging in $\eta = 1/7$ we now get the formal result of Theorem 2.

Corollary 14 *Assume (Lower Boundedness), ((L_0, L_1)-smoothness), (Bounded Variance) and $T \geq 3$. Furthermore define the parameters $\beta_t := 1 - t^{-1/2}$ and $\eta_t := \frac{t^{-3/4}}{7}$. Then NSGD-M with starting point $x_1 \in \mathbb{R}^d$ satisfies*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla F(x_t)\|\right] \leq \frac{\left(14 + 96L_1 e^{L_1^2}\right)\Delta_1 + \left(6e^{L_1/7} + 2\log(T) + 6e^{L_1^2}\right)L_0}{T^{\frac{1}{4}}} + \frac{12e\log(T)\sigma + 12e^{L_1^2}\min\left\{\frac{L_0}{L_1}, \sqrt{8L_0\Delta_1}\right\}}{T^{\frac{1}{4}}},$$

where $\Delta_1 := F(x_1) - F^*$ is the initialization gap. Furthermore, if $L_1 \geq 7/2$, we get the following improved dependence on L_1:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla F(x_t)\|\right] \leq \frac{126e^{L_1^2}\Delta_1 + 12e\log(T)\sigma + \left(8e^{L_1^2} + 2\log(T)\right)L_0}{T^{\frac{1}{4}}}.$$

Proof Plugging the choice of $\eta = \frac{1}{7}$ into Theorem 13 and using that $\log(T) \geq 1$ yields

$$\frac{\eta}{2}\sum_{t=1}^{T} t^{-3/4}\mathbb{E}\left[\|\nabla F(x_t)\|\right] \leq \Delta_1 + 6e\eta\log(T)\sigma + 6\eta e^{L_1^2}\|\nabla F(x_1)\| + \eta L_0\left(3e^{L_1/7} + \log(T) + 3e^{L_1^2}\right).$$

Next, from the proof of Lemma 8, we get that

$$\|\nabla F(x_1)\| \leq 8L_1\Delta_1 + \min\left\{\frac{L_0}{L_1}, \sqrt{8L_0\Delta_1}\right\}$$

and hence, by noting that $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla F(x_t)\|\right] \leq T^{-\frac{1}{4}}\sum_{t=1}^{T} t^{-3/4}\mathbb{E}\left[\|\nabla F(x_t)\|\right]$ we obtain

$$\frac{1}{T^{3/4}}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla F(x_t)\|\right] \leq \left(14 + 96L_1 e^{L_1^2}\right)\Delta_1 + \left(6e^{L_1/7} + 2\log(T) + 6e^{L_1^2}\right)L_0 + 12e\log(T)\sigma + 12e^{L_1^2}\min\left\{\frac{L_0}{L_1}, \sqrt{8L_0\Delta_1}\right\}$$

and hence proved the first claim.

For the second claim assume $L_1 \geq 7/2$. We now can use the second statement in Lemma 13 to get

$$\begin{aligned}\frac{1}{T^{3/4}}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla F(x_t)\|\right] &\leq \left(14+112e^{L_1^2}\right)\Delta_1 + 12e\log(T)\sigma + \left(6e^{L_1/7}+2\log(T)+6e^{L_1^2}\right)L_0 \\ &\quad + 2\frac{e^{L_1^2}}{L_1}\min\left\{\frac{L_0}{L_1}, \sqrt{8L_0\Delta_1}\right\} \\ &\leq 126e^{L_1^2}\Delta_1 + 12e\log(T)\sigma + \left(8e^{L_1^2}+2\log(T)\right)L_0,\end{aligned}$$

where we used that $6e^{L_1/7} + 2L_1^{-2}e^{L_1^2} \leq 2e^{L_1^2}$ for $L_1 \geq 7/2$. ■

D.1.2. PARAMETER-DEPENDENT

This section contains the proof of Remark 3.

Corollary 15 (Non parameter-agnostic NSGD-M) *Assume (Lower Boundedness), ((L_0, L_1)-smoothness) and (Bounded Variance). Furthermore define the parameters $\beta_t := 1 - t^{-1/2}$ and $\eta_t := \frac{t^{-3/4}}{12L_1}$. Then NSGD-M with starting point $x_1 \in \mathbb{R}^d$ satisfies*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla F(x_t)\|\right] \leq \frac{24L_1\Delta_1 + \left(14+4\sqrt{2e^2}\log(T)\right)\sigma + (4+2\log(T))\frac{L_0}{L_1}}{T^{1/4}}.$$

where $\Delta_1 := F(x_1) - F^*$ is the initialization gap.

Proof Denote $\eta := 1/12$. By plugging our choice of η_t into (14) we obtain

$$\sum_{t=1}^{T}\frac{1}{2}\eta_t\mathbb{E}\left[\|\nabla F(x_t)\|\right] \leq \Delta_1 + \eta\sigma\left(7+2\sqrt{2e^2}\log(T)\right) + \eta^2 L_0(21+7\log(T))$$

and by using the same arguments as in the proof of Theorem 2 we get

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla F(x_t)\|\right] \leq \frac{\frac{2\Delta_1}{\eta} + \left(14+4\sqrt{2e^2}\log(T)\right)\sigma + \eta(42+14\log(T))L_0}{T^{1/4}}.$$

■

D.1.3. DETERMINISTIC SETTING

This section contains the proof of Theorem 5.

Proof of Theorem 5. By Lemma 8 we have that $\|\nabla F(x)\| \leq \max\left\{8L_1(F(x)-F^*), \frac{L_0}{L_1}\right\}$. Since GD with Backtracking Line Search is a descent algorithm, we get that $\|\nabla F(x_t)\| \leq \max\left\{8L_1\Delta_1, \frac{L_0}{L_1}\right\} =:$

Algorithm 2: GD with Backtracking Line Search

Input: Starting point $x_1 \in \mathbb{R}^d$, Armijo Parameters $\beta \in (0,1)$ and $\gamma \in (0,1)$
$k \leftarrow 1$
for $t = 1, 2, \ldots$ **do**
  $g_t \leftarrow \nabla F(x_t)$
  while $\left(F\left(x_t - \beta^k g_t\right) > F(x_t) - \beta^k \gamma \|\nabla F(x_t)\|^2\right)$ **do**
    $k \leftarrow k + 1$
  end
  $\eta_t \leftarrow \beta^k$
  $x_{t+1} \leftarrow x_t - \eta_t g_t$
end

$u(L_0, L_1, \Delta_1)$ for all $t \in \mathbb{N}$. Now let $x \in \mathbb{R}^d$ be an iterate of GD with Backtracking Line Search and $\eta \leq \frac{1}{L_1}$. Then Lemma 7 implies

$$\begin{aligned} F(x - \eta \nabla F(x)) &\leq F(x) - \eta \|\nabla F(x)\|^2 + \eta^2 (2L_0 + (e-1)L_1 \|\nabla F(x)\|) \|\nabla F(x)\|^2 \\ &\leq F(x) - \eta \|\nabla F(x)\|^2 + \eta^2 (2L_0 + (e-1)u(L_0, L_1, \Delta_1)) \|\nabla F(x)\|^2 \\ &= F(x) - \eta(1 - \eta L) \|\nabla F(x)\|^2, \end{aligned}$$

where $L \coloneqq 2L_0 + (e-1)L_1 u(L_0, L_1, \Delta_1)$. In particular we have that $F(x - \eta \nabla F(x)) \leq F(x) - \eta\beta \|\nabla F(x)\|^2$ whenever $\eta \leq \frac{1-\beta}{L}$. This allows us to lower bound our stepsizes by $\eta_t > \frac{\gamma(1-\beta)}{L}$. As in the L-smooth setting, the definition of x_{t+1} now yields

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

and thus

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

This finishes the proof. ■

D.2. Lower Bound

This section contains the proof for Theorem 4.

Algorithm 3: General Normalized Momentum Method

Input: Starting point $x_1 \in \mathbb{R}^d$, stepsize $\eta > 0$, power $\alpha > 0$

$m_0 \leftarrow 0$

for $t = 1, 2, \ldots$ **do**

  Independently sample ξ_t from the distribution of ξ.

  $g_t \leftarrow \nabla f(x_t, \xi_t)$

  Choose $m_t \in \operatorname{cone}(g_1, \ldots, g_t) \setminus \prime$

  $x_{t+1} \leftarrow x_t - \frac{\eta}{t^\alpha} \frac{m_t}{\|m_t\|}$

end

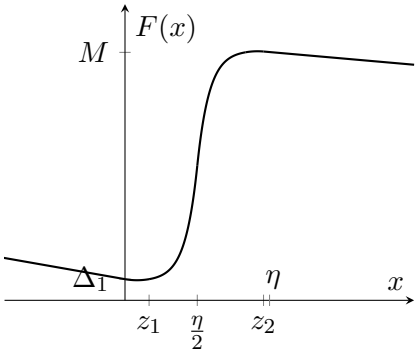

Figure 1: Plot of $F(x)$

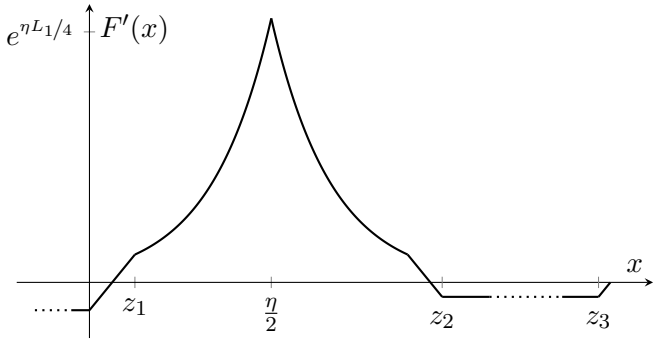

Figure 2: Plot of $F'(x)$

Plots of the hard function used in the proof of Theorem 4.

Proof of Theorem 4. We start by constructing F via its derivative F'. Therefore let $z_1 := \frac{2}{L_0}$, $z_2 := \eta - z_1 + \frac{1+2\varepsilon}{L_0} \leq \eta$ and set

$$F'(x) = \begin{cases} -1, & \text{if } x \leq 0 \\ L_0 x - 1, & \text{if } 0 < x \leq z_1 \\ e^{L_1(x - z_1)}, & \text{if } z_1 < x \leq \frac{\eta}{2} \\ F'(\eta - x), & \text{if } \frac{\eta}{2} < x \leq z_2 \\ -2\varepsilon, & \text{if } z_2 < x \leq z_3, \\ L_0(x - z_3) - 2\varepsilon, & \text{if } z_3 < x \leq z_4, \\ 0, & \text{if } x > z_4 \end{cases},$$

where z_3 and z_4 will be determined later. Note that z_2 is chosen in such way, that F' is continuous. A plot of F' can be seen in Figure 2. Then $F(x) := \Delta_1 + \int_0^x F' d\lambda$ (see Figure 1) satisfies

$$F(\eta) \geq \Delta_1 + \frac{2}{L_1}\left(e^{L_1(\eta/2 - z_1)} - 1\right) \geq \Delta_1 + \frac{2}{L_1}\left(e^{\frac{\eta L_1}{4}} - 1\right) =: M,$$

where the second inequality follows by our choice of $L_0 \geq 8/\eta$, which implies $\frac{\eta}{2} - z_1 \geq \frac{\eta}{4}$. We now choose z_3, z_4 in such a way, that a) F never attains negative values and b) F' is continuous. By doing

so, we guarantee that a) $F^* \geq 0$ and hence $F(x_1) - F^* \leq \Delta_1$, and b) that F is (L_0, L_1)-smooth. Therefore set $z_3 \coloneqq \frac{M}{2\varepsilon}$ and calculate

$$F(z_3) = F(\eta) + \left(\frac{M}{2\varepsilon} - \eta\right)(-2\varepsilon) \geq 2\eta\varepsilon.$$

Finally we choose $z_4 = z_3 + \frac{2\varepsilon}{L_0}$ and get

$$F(z_4) = F(z_3) + \frac{L_0}{2}\left(\frac{2\varepsilon}{L_0}\right)^2 - 2\varepsilon\frac{2\varepsilon}{L_0} \geq 2\eta\varepsilon - \frac{2\varepsilon^2}{L_0} > 0,$$

where we again used $L_0 \geq 8/\eta$ in the last inequality. To show that F is indeed (L_0, L_1)-smooth, first notice that F is (L_0, L_1)-smooth on each of the subintervals $(-\infty, 0), ..., [z_3, z_4), [z_4, \infty)$. The claim now follows from the upcoming Lemma 16. F therefore satisfies all the properties required from the statement and we can turn out attention to the behaviour of the iterates.

Induction yields that each gradient points in the same direction $F'(x_t) \leq 0$ and hence so does the momentum. Therefore, by the normalizing nature of the algorithm, we get

$$x_T = \eta \sum_{t=1}^{T-1} t^{-\alpha} \leq \eta\left(1 + \frac{1}{1-\alpha}\left((T-1)^{1-\alpha} - 1\right)\right) \leq \frac{\eta}{1-\alpha}T^{1-\alpha}. \tag{16}$$

Now suppose the inequality in Theorem 4 is violated, i.e.

$$T < \left(\frac{(1-\alpha)M}{2\eta\varepsilon}\right)^{\frac{1}{1-\alpha}}.$$

Plugging into (16) yields $x_T < \frac{M}{2\varepsilon} = z_3$ and therefore, by construction, $|F'(x_t)| \geq 2\varepsilon$ for all $t \in [T]$. This completes the proof.

■

Lemma 16 *Let $I \subseteq \mathbb{R}$ be an interval, $a \in I$ and set $I_- \coloneqq \{\S \in \mathcal{I} \mid \S \leq \dashv\}$, $I_+ \coloneqq \{\S \in \mathcal{I} \mid \S \geq \dashv\}$. Further Let $f\colon \mathbb{R} \to \mathbb{R}$ be continuously differentiable and suppose that f satisfied the inequality from Theorem 1 on I_+ and I_-. Then the inequality is also satisfied on I, i.e. it also holds for $x \in I_-, y \in I_+$.*

Proof W.l.o.g. let $x \in I_-, y \in I_+$ and set $c \coloneqq L_1\|x - y\|$. Furthermore set $c_1 \coloneqq L_1\|x - a\|, c_2 \coloneqq L_1\|a - y\|$ and calculate

$$\begin{aligned}\|\nabla f(x) - \nabla f(y)\| &= \|\nabla f(x) - \nabla f(a) + \nabla f(a) - \nabla f(y)\| \\ &\leq L_0(\|x - a\| A_0(c_1) + \|a - y\| A_0(c_2)) \\ &\quad + L_1\|x - a\| A_1(c_1)\|\nabla f(x)\| + L_1\|a - y\| A_1(c_2)\|\nabla f(a)\|.\end{aligned} \tag{17}$$

Next, since $a \in I_-$, we get that

$$\|\nabla f(a)\| \leq L_0\|x - a\| A_0(c_1) + e^{c_1}\|\nabla f(x)\|$$

and hence

$$\begin{aligned} L_1 \|a-y\| A_1(c_2) \|\nabla F(a)\| &\leq L_0 L_1 \|a-y\| A_1(c_2) \|x-a\| A_0(c_1) + L_1 \|a-y\| A_1(c_2) e^{c_1} \|\nabla f(x)\| \\ &= L_0(e^{c_2}-1) \|x-a\| A_0(c_1) + L_1 \|a-y\| A_1(c_2) e^{c_1} \|\nabla f(x)\| \end{aligned}$$

We now plug this result into (17) and rearrange to obtain

$$\begin{aligned} \|\nabla f(x) - \nabla f(y)\| \leq\ & L_0(e^{c_2} \|x-a\| A_0(c_1) + \|a-x\| A_0(c_2)) \\ & + L_1 \|\nabla f(x)\| (\|x-a\| A_1(c_1) + \|a-y\| A_1(c_2) e^{c_1}). \end{aligned} \tag{18}$$

Now we focus on the second term, involving $L_1 \|\nabla f(x)\|$. Therefore we calculate

$$\begin{aligned} &\|x-a\| A_1(c_1) + \|a-y\| A_1(c_2) e^{c_1} \\ &= \frac{e^{L_1\|x-a\|}-1}{L_1} + \frac{e^{L_1\|x-y\|} - e^{L_1\|x-a\|}}{L_1} \\ &= A_1(c) \|x-y\| . \end{aligned}$$

Next we focus on the first term in (18), which corresponds to the L_0-dependence. Calculating yields

$$\begin{aligned} &e^{c_2} \|x-a\| A_0(c_1) + \|a-y\| A_0(c_2) \\ &= \|x-a\| e^{L_1\|a-y\|} + \|x-a\| e^{L_1\|x-y\|} - \frac{e^{L_1\|x-y\|} - e^{L_1\|a-y\|}}{L_1} \\ &\quad + \|a-y\| + \|a-y\| e^{L_1\|a-y\|} - \frac{e^{L_1\|a-y\|}-1}{L_1} \\ &= \|a-y\| + \|x-y\| e^{L_1\|a-y\|} + \|x-a\| e^{L_1\|x-y\|} - \frac{e^{L_1\|x-y\|}-1}{L_1} \\ &\leq \|x-y\| + \|x-y\| e^{L_1\|x-y\|} - \frac{e^{L_1\|x-y\|}-1}{L_1} = A_0(L_1\|x-y\|) \|x-y\| . \end{aligned}$$

In the last inequality we used that for all $a, b, L_1 \geq 0$ the following inequality holds: $b + (a+b)e^{L_1 b} + b e^{L_1(a+b)} \leq a + b + (a+b)^{L_1(a+b)}$. This follows by taking partial derivatives with respect to L_1. Finally we plug everything into (18) and obtain

$$\|\nabla f(x) - \nabla f(y)\| \leq (A_0(c) L_0 + A_1(c) L_1 \|\nabla f(x)\|) \|x-a\| .$$

This finishes the proof. ■

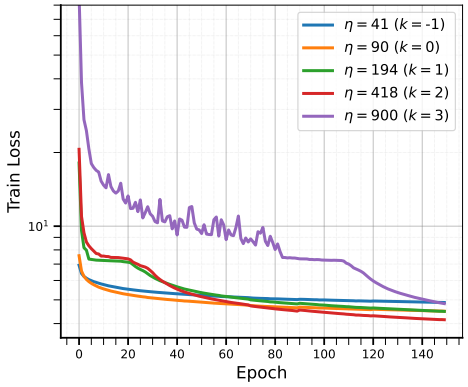

Figure 3: Training curves of NSGD-M.

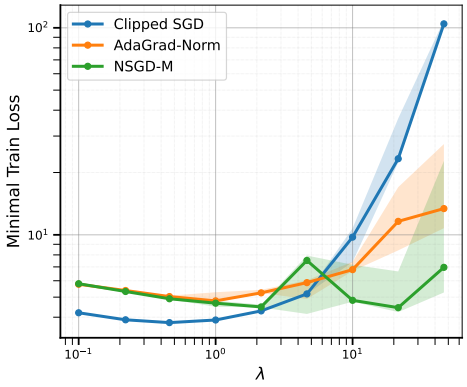

Figure 4: Minimal training loss.

Figure 3 shows training curves of NSGD-M for stepsizes $\eta = 10^{k/3} \cdot \eta_{\texttt{opt}}$, where $\eta_{\texttt{opt}} = 90$. Figure 4 shows the smallest train loss within 150 epochs of different algorithms with stepsizes $\lambda \cdot \eta_{\texttt{opt}}$. Shaded areas represent the minimal and maximal value within 3 seeds, the line the median.

Appendix E. Experiments

In concordance with our theory, the experiment's primary focus is to demonstrate the robustness of NSGD-M to stepsize selection in the context of (L_0, L_1)-smoothness. Based on the empirical findings in [4, 28], confirming the necessity of (L_0, L_1)-smoothness in this setting, we train the AWD-LSTM architecture [17] on the WikiText-2 dataset [16]. We first conduct a 50 epoch coarse grid search to tune the stepsize of NSGD-M, AdaGrad-Norm [10] and Clipped SGD [28]. The clipping threshold for Clipped SGD was fixed to be 0.25 in concordance to previous work [28], the decay-rates of NSGD-M were chosen according to Theorem 2 and b_0 of AdaGrad-Norm was set to be $b_0 = 10^{-6}$ [24]. For each algorithm, the final training was then carried out with stepsizes $\eta = \lambda \cdot \eta_{\texttt{opt}}$, where $\lambda = 10^{k/3}, k \in \{-3, -2, \ldots, 5\}$, for 150 epochs. This procedure is replicated with three different seeds to get more reliable results. In order to observe the actual algorithm behaviour, we disabled the averaging mechanism of the model. The code is based on [27].

Discussion. Figure 3 shows the behaviour of NSGD-M with different stepsizes. The result supports the narrative behind Theorem 2 that NSGD-M needs an adaption phase before transitioning to a convergence phase. During the adaption phase, NSGD-M plateaus instead of accumulating an exponential error. Figure 4 focuses on the robustness to hyperparameter selection. It compares the smallest training loss across 150 epochs of different algorithms on scaled versions of their optimally tuned stepsize. As expected, well-tuned Clipped SGD with constant stepsize outperforms all decaying algorithms, while decaying algorithms are more robust to untuned stepsizes. Between NSGD-M and AdaGrad-Norm we notice that NSGD-M has slightly preferable behaviour for small stepsizes. Furthermore the trend for large stepsizes points towards a more robust behaviour of NSGD-M.