Parameter-Agnostic Optimization under Relaxed Smoothness

Florian Hübler ETH Zurich
Junchi Yang ETH Zurich
Xiang Li ETH Zurich
Niao He ETH Zurich

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 exponential term is inevitable to such schemes by introducing a theoretical framework of lower bounds tailored explicitly for parameter-agnostic algorithms. 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.

Proceedings of the 27[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

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 (Bottou et al., 2018).

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, notably by Arjevani et al. (2022), 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 but with stepsizes depending on problem parameters like L (Ghadimi and Lan, 2013). 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 (Faw et al., 2022; Yang et al., 2022). We call algorithms with this characteristic *parameter-agnostic*, and *parameter-dependent* otherwise.

However, Zhang et al. (2020b) highlight that not all machine learning applications adhere to the L-smoothness assumption. 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 further validated in various machine learning tasks (Zhang et al., 2020a; Crawshaw et al., 2022).

In light of this more realistic smoothness assumption, a substantial body of literature has emerged. The nearly rate-optimal complexity $\widetilde{\mathcal{O}}(\varepsilon^{-4})$ has been established for various algorithms, including SGD (Li et al., 2023b), Clipped SGD (Zhang et al., 2020b,a), Normalized SGD (Zhao et al., 2021), AdaGrad-Norm (Faw et al., 2023; Wang et al., 2023) and ADAM (Li et al., 2023a). 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 (Wang et al., 2023), 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 (Ward et al., 2019). 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?*

With the growing interest in the development of parameter-agnostic algorithms, a fundamental trade-off becomes evident: while these algorithms demand less prior knowledge about the problem, they may also offer weaker convergence guarantees. For instance, under L-smoothness, SGD with decaying stepsizes $\eta_t = \eta/\sqrt{t}$ achieves the near-optimal complexity $\widetilde{\mathcal{O}}(L\Delta_1\sigma^2\varepsilon^4)$ when η is selected based on knowledge of problem parameters (Ghadimi and Lan, 2013). Without this information, however, using the same stepsizes has been shown to suffer from a lower bound of $\Omega\Big(\eta^{-4}L^{-2}e^{(\eta L)^2/8}\varepsilon^{-4}\Big)$, even in the deterministic setting (Yang et al., 2022).

This underscores the need to differentiate between parameter-agnostic and parameter-dependent algorithms when establishing lower bounds to truly grasp the potential of parameter-agnostic algorithms. However, the existing lower bound framework is constructed in a way that implicitly allows algorithms to have access to problem-specific parameters. This work addresses another pivotal question:

> *Can we develop a lower bound framework that distinguishes between parameter-agnostic and parameter-dependent algorithms?*

1.1 Our Contributions

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

Firstly, we show that under the general (L_0, L_1)-smoothness condition, Normalized Stochastic Gradient Descent with Momentum (NSGD-M), as introduced by (Cutkosky and Mehta, 2020), converges with a nearly rate-optimal complexity of $\widetilde{\mathcal{O}}(\varepsilon^{-4})$ without any prior knowledge of the problem parameters L_0, L_1. 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 $\mathcal{O}\big((L_0\Delta_1 + L_1^2\Delta_1^2)\varepsilon^{-2}\big)$. To the best of our knowledge, these are the first parameter-agnostic convergence results in the (L_0, L_1)-smoothness setting.

Secondly, we provide a novel framework for lower bound analysis tailored to parameter-agnostic algorithms. Within this framework, we show 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. This framework distinctly delineates the parameter-agnostic setting from the parameter-dependent setting, in which NSGD-M does not suffer from the exponential term. Additionally, it suggests that the (L_0, L_1)-smoothness setting may be more challenging than the L-smoothness setting.

1.2 Related Work

Parameter-Agnostic Algorithms. If the objective function is L-smooth, convergence results are typically contingent upon stepsizes being less than $2/L$ (Bottou et al., 2018). In the deterministic context, GD with a constant stepsize that does not satisfy this threshold may diverge (Nesterov, 2018). However, this can be rectified using a Backtracking Line Search to determine the stepsize, which does not rely on knowing problem parameters, and achieves an optimal complexity of $\mathcal{O}(\epsilon^{-2})$ (Armijo, 1966). Conversely, in stochastic environments, Vaswani et al. (2022) 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 (Yang et al., 2023). Various adaptive methods, such as AdaGrad (Duchi et al., 2011; McMahan and Streeter, 2010), its variants AdaGrad-Norm (Streeter and McMahan, 2010) and NSGD-M (Cutkosky and Mehta, 2020), bypass this exponential term, even without knowledge of the problem parameters, as recently shown in (Faw et al., 2022; Yang et al., 2023). These adaptive methods are typically considered more robust to different problem parameters (Ward et al., 2019; Kavis et al., 2019), given their ability to tune algorithm hyperparameters dynamically during training. In the convex setting, the issue can furthermore be

rectified using Polyak stepsizes, which only requires knowledge of Δ_1 (Polyak, 1987). While this stepsize schedule achieves an $\mathcal{O}(\varepsilon^{-2})$ oracle complexity in deterministic environments (Hazan and Kakade, 2019), it again falls short in stochastic environments, where convergence is only guaranteed to a neighbourhood (Loizou et al., 2021; Orvieto et al., 2022). Only under stronger assumptions, modifications of the Polyak stepsize guarantee convergence to a minimum (Orvieto et al., 2022; Jiang and Stich, 2023). There is another line of research dedicated to "parameter-free" algorithms for online convex optimization (Orabona and Pal, 2016; Cutkosky and Orabona, 2018). 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. (2020b) introduced the concept of (L_0, L_1)-smoothness, defined by the following affine bound on the Hessian-norm: $\|\nabla^2 F(x)\| \leq L_0 + L_1 \|\nabla F(x)\|$. The convergence of both `GD` and `SGD` was only recently established in this setting (Li et al., 2023b). 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` (Zhang et al., 2020b), and its momentum-augmented counterpart (Zhang et al., 2020a), 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. (2020a) also provided a convergence result for Finite Horizon `NSGD-M` 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` (Zhao et al., 2021), signed `SGD` (Crawshaw et al., 2022), `AdaGrad-Norm` (Faw et al., 2023; Wang et al., 2023), and `ADAM` (Li et al., 2023a; Wang et al., 2022). 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 (Wang et al., 2023).

Lower Bound Theory. Lower bounds for seeking near-stationary points have been extensively studied within the L-smoothness setting. Nesterov (2012) first addressed constrained optimization under box constraints. Subsequently, a seminal study by Carmon et al. (2020) established a tight lower bound of $\Omega(\Delta_1 L \varepsilon^{-2})$ for the deterministic setting. Arjevani et al. (2022) extended the results to the stochastic setting, introducing the $\Omega(\Delta_1 L \sigma^2 \varepsilon^{-4})$ lower bound. Specific algorithms, such as `SGD` (Drori and Shamir, 2020) and `Newton's method` (Cartis et al., 2010), also have associated lower bounds. However, the algorithm classes considered by these lower bounds include algorithms with stepsizes that can depend on problem parameters, so they might not be tight in the parameter-agnostic setting. Vaswani et al. (2022) discovered that parameter-agnostic SGD with a specific exponentially decreasing stepsize suffers from an exponential dependence during its initial phase when minimizing strongly convex functions. Later, Yang et al. (2023) also derived a lower bound for `SGD` under a polynomially decreasing stepsize in the nonconvex setting. Yet the implications of the parameter-agnostic lower bound for a class of algorithms remain ambiguous. The aforementioned studies consider the function class of L-smooth functions, so they are also applicable to (L_0, L_1)-smooth functions. In the realm of online convex optimization, Cutkosky and Boahen (2016, 2017) have introduced a lower bound featuring an exponential term when the norm of the predictor and the Lipschitz constant are allowed to scale with the total number of iterations.

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 dimension 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 $x_1, \ldots, x_n \in \mathbb{R}^d$ is denoted by $\operatorname{cone}(x_1, \ldots, x_n) := \{\sum_{i=1}^n \lambda_i x_i \colon \lambda_1, \ldots, \lambda_n \geq 0\}$.

Problem Setup Since finding a solution to (1) is computationally intractable (Nemirovskij and Yudin, 1983), we aim to find an ε-stationary point. Furthermore we only allow access to a (possibly noisy) gradient oracle $\nabla f(\cdot, \xi)$ of ∇F, where ξ is a random vector. Due to this randomness, our specific goal is finding an approximate solution $x \in \mathbb{R}^d$ with $\mathbb{E}\left[\|\nabla F(x)\|\right] \leq \varepsilon$.

Assumptions Building on established work in stochastic optimization (Ghadimi and Lan, 2013; Arjevani et al., 2022), we employ the following two de-facto standard assumptions in various results of this study.

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 ex-*

ists $\sigma \geq 0$ such that

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

Instead of the traditional L-smoothness assumption, we adopt the weaker concept of (L_0, L_1)-smoothness, as proposed by Zhang et al. (2020b). Following the work of Zhang et al. (2020a), 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. (2020b, Definition 1).

Definition 3. *Let $L_0, L_1 \geq 0$ and $f\colon \mathbb{R}^d \to \mathbb{R}$ be a differentiable function. Then f is called (L_0, L_1)-smooth if for all $x, y \in \mathbb{R}^d$ and all $c > 0$ with $L_1 \left\|x - y\right\| \leq c$ it holds that*

$$\begin{aligned}&\left\|\nabla f(x) - \nabla f(y)\right\| \\ &\leq \left(A_0(c)L_0 + A_1(c)L_1 \left\|\nabla f(x)\right\|\right)\left\|x - y\right\|,\end{aligned}$$

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

Assumption 4 ((L_0, L_1)-smoothness)**.** *The objective function F is (L_0, L_1)-smooth.*

Notably, the two definitions are equivalent if the objective function is twice differentiable as the following Lemma shows.

Lemma 1. *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 3.*

3 PARAMETER-AGNOSTIC UPPER BOUNDS

In this section, we present the first parameter-agnostic convergence results for (L_0, L_1)-smooth functions. In Section 3.1, 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 3.2 we show that in the deterministic setting, GD with Backtracking Line Search can avoid this exponential dependence, while still being parameter-agnostic.

3.1 Stochastic Setting

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 grows exponentially with L_1. This behaviour is intrinsic to NSGD-M and cannot be eliminated, as we will show in Section 4. Once the stepsize decreases below a threshold (which is polynomial in L_1), the algorithm transitions into the *convergence phase*. In this latter phase, the error decays at a rate of $T^{-1/4}\log(T)$. The following Theorem 2 formalizes this behaviour. Its more verbose version (Theorem 14), and proof can be found in Appendix C.1.1.

Algorithm 1: Normalized SGD with Momentum (NSGD-M) (Cutkosky and Mehta, 2020)

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**
  Indep. 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

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

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

where $\Delta_1 \coloneqq F(x_1) - F^*$ 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 an 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 proposition shows that this cost arises from the parameter-agnostic stepsize; that is, the exponential term disappears when the stepsize is determined based on the parameters.

Proposition 3. *Under the assumptions of Theorem 2, define the parameters $\beta_t \coloneqq 1 - t^{-1/2}$ and $\eta_t \coloneqq \frac{t^{-3/4}}{12L_1}$. Then NSGD-M with starting point $x_1 \in \mathbb{R}^d$ satisfies*

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

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

These results indicate that NSGD-M is potentially more robust to hyper-parameter selection than other existing algorithms. In comparison, SGD necessitates knowledge of both L_0 and L_1, as well as the exact value of $\|\nabla f(x_1)\|$ (Li et al., 2023b). Clipped SGD requires

Algorithm 2: GD with Backtracking Line Search

Input: Starting point $x_1 \in \mathbb{R}^d$, Armijo Parameters $\beta \in (0,1)$ and $\gamma \in (0,1)$

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

  Choose k minmal such that $\beta^k \leq \eta_{t-1}$ and $F\big(x_t - \beta^k \nabla F(x_t)\big) \leq F(x_t) - \beta^k \gamma \|\nabla F(x_t)\|^2$

  $\eta_t \leftarrow \beta^k$

  $x_{t+1} \leftarrow x_t - \eta_t \nabla F(x_t)$

end

to know L_0 and L_1 (Zhang et al., 2020a), and even AdaGrad-Norm demands knowledge of L_1 (Faw et al., 2023; Wang et al., 2023). It is important to note that our analysis is significantly different from the previous analysis for NSGD-M in (Zhang et al., 2020a). The latter focused on constant stepsizes and momentum parameters determined by L_0, L_1, target accuracy ε, and variance σ.

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. This is achieved by using Gradient Descent with a Backtracking Line-search (see Algorithm 2).

Theorem 4. *Assume (Lower Boundedness) and ((L_0, L_1)-smoothness) in the deterministic setting. Then GD with Backtracking Line Search (see Algorithm 2) with any parameters $\beta, \gamma \in (0,1)$ satisfies*

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

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 if it is addressing $(L_0 + L_1 C)$-smooth functions, where C represents the gradient norm's upper bound. The formal proof can be found in Appendix C.1.2. We have not extended our considerations to the stochastic setting for this algorithm, as a stochastic line search can potentially fail even under the stricter L-smoothness assumption (Vaswani et al., 2022).

4 PARAMETER-AGNOSTIC LOWER BOUNDS

In the previous section, we highlighted that the first provable parameter-agnostic algorithm, NSGD-M, comes at the cost of an exponential L_1-dependence. This naturally raises the question: Is such an undesirable term unavoidable for this class of algorithms? Since most existing lower bounds focus on the parameter-dependent setting — where hyper-parameters of algorithms can be set based on problem parameters — we begin by introducing the concept of lower bounds specifically designed for parameter-agnostic setting in Section 4.1. Subsequently, in Section 4.2, we utilize this concept to show that NSGD-M indeed suffers from an exponential dependence on L_1.

4.1 A Lower Bound Framework

To motivate the need for specific lower bounds for parameter-agnostic algorithms, let us consider the algorithm class $\mathcal{A}$ consisting of GD with all constant stepsizes $\{\eta : \eta > 0\}$. Furthermore, we consider the well-studied function class comprising of L-smooth functions with initialization gap $F(x_1) - F^* \leq \Delta_1$, denoted as $\mathcal{F}_{L,\Delta_1}$. A well-established lower bound for $\mathcal{A}$ to find an ϵ-stationary point in this setting is $\Omega\big(L\Delta_1\varepsilon^{-2}\big)$, as demonstrated by the seminal work of Carmon et al. (2020). This lower bound is tightly matched by GD with a parameter-dependent stepsize smaller than $2/L$, and hence cannot be improved. However, without knowledge of L, GD with constant stepsize generally fails to converge (Nesterov, 2018). Thus, a parameter-agnostic notion of lower bounds would be more informative under this setting.

For simplicity, our discussion focuses on deterministic algorithms. However, this can be readily generalized to the stochastic setting by incorporating a stochastic oracle into the algorithm's definition, as detailed in (Arjevani et al., 2022). Additionally, algorithms with a deterministic gradient oracle can be viewed as specific instances of their stochastic equivalents when there is no gradient noise. Consequently, the lower bounds established for deterministic algorithms are generally stronger and are also applicable to their stochastic counterparts.

Definition 5 (Deterministic Algorithm (Carmon et al., 2020))**.** *We say that A is a* first order deterministic algorithm *if it, given a differentiable function f, produces iterates of the form*

$$x_t = A_t(f(x_1), \nabla f(x_1), \ldots, f(x_{t-1}), \nabla f(x_{t-1})),$$

where A_t is a (Lebesgue-) measurable mapping. We denote the set of all such algorithms as $\mathcal{A}_{\text{det}}$.

It is important to note that an algorithm $A \in \mathcal{A}_{\text{det}}$ is *a function that takes a differentiable function as its argument and outputs a sequence* in $\mathbb{R}^d$. When we mention an "algorithm with a specific stepsize scheme", we are technically referring to a set of algorithms $\{A_\eta\}_{\eta>0}$, where η serves as the hyperparameter of this stepsize scheme, and each distinct η defines a unique algorithm.

To begin with, we consider general parameterized function spaces, denoted by $\mathcal{F}_\theta$, where the parameter θ lives in a parameter space Θ. Specifically, in our application, it is given by $\theta = (\Delta_1, L_0, L_1)$. We use $\mathscr{F}_\Theta := \{\mathcal{F}_\theta \colon \theta \in \Theta\}$ to denote a parameterized family of function spaces. For an algorithm class $\mathcal{A} \subseteq \mathcal{A}_{\text{det}}$, existing lower bound literature usually considers the following challenge (Carmon et al., 2020): for any problem parameter $\theta \in \Theta$ and target accuracy $\epsilon > 0$, find a lower bound for

$$\inf_{A\in\mathcal{A}} \sup_{f\in\mathcal{F}_\theta} T_\varepsilon(A, f), \tag{2}$$

where $T_\varepsilon(A, f) := \inf\{t \in \mathbb{N}_{\geq 1} \mid \|\nabla f(x_t)\| \leq \varepsilon\}$ denotes the number of iterations required for A to reach an ε-stationary point of f. Importantly, the parameter θ is fixed for the function class $\mathcal{F}_\theta$ before taking the infimum over the algorithm class and supremum over the function class. This implicitly assumes that algorithms have the ability to adjust their hyperparameters based on θ.

In contrast to the framework above, we propose the concept of *parameter-agnostic lower bounds*.

Definition 6 (Parameter-Agnostic Lower Bound)**.** *Let $\mathcal{A} \subseteq \mathcal{A}_{\text{det}}$ be an algorithm class and $\mathscr{F}_\Theta := \{\mathcal{F}_\theta \colon \theta \in \Theta\}$ a parameterized family of function spaces. A function $g\colon (0,\infty) \times \Theta \to [0,\infty]$ is called a* parameter-agnostic lower bound of $\mathcal{A}$ on $\mathscr{F}_\Theta$ *if there does not exist an algorithm $A \in \mathcal{A}$ such that for all $\kappa > 0$, there exists $\varepsilon_0, \theta_0 > 0$ such that for all $\varepsilon \leq \varepsilon_0, \theta \geq \theta_0$, $\sup_{f\in\mathcal{F}_\theta} T_\varepsilon(A, f) \leq \kappa g(\varepsilon, \theta)$. The comparisons of $\theta \in \Theta$ and scalars are to be understood component-wise.*

In other words, this definition states that $g(\varepsilon, \theta)$ serves as a parameter-agnostic lower bound if no algorithm has a complexity that is better than g asymptotically. Here, ε_0, θ_0 and κ can depend on A, which excludes the possibility that A can pick its hyperparameters according to the parameter θ. The performance of A is therefore evaluated across all function spaces with θ large enough. On the other hand, the conventional definition in Equation (2) states that for any $\theta \in \Theta$ — which means the parameter is determined first — there does not exist an algorithm $A \in \mathcal{A}$ such that for all $\kappa > 0$, $\sup_{f\in\mathcal{F}_\theta} T_\varepsilon(A, f) \leq \kappa g(\varepsilon, \theta)$.

Note that Definition 6 also outlines a way to compare complexities for different algorithms within the parameter-agnostic framework by assessing their performance with asymptotically large θ. The earlier parameter-agnostic lower bounds presented in (Yang et al., 2023; Vaswani et al., 2022) apply solely to a particular algorithm with a specified stepsize. Consequently, it is ambiguous how one might define a lower bound across a class of algorithms.

To establish that g serves as an parameter-agnostic lower bound of $\mathcal{A}$ on $\mathscr{F}_\Theta$, one could instead prove the following stronger result.

Proposition 5. *If for any algorithm $A \in \mathcal{A}$ there exist constants $\varepsilon_0, \theta_0, K > 0$ such that*

$$\forall\, \varepsilon \in (0, \varepsilon_0], \theta \geq \theta_0 \colon \sup_{f\in\mathcal{F}_\theta} T_\varepsilon(A, f) \geq K g(\varepsilon, \theta),$$

where $\theta \in \Theta$, then g is a parameter-agnostic lower bound of $\mathcal{A}$ on $\mathscr{F}_\Theta$.

The condition presented in Proposition 5 is more handy for use in proofs, and will be our primary tool for deriving lower bounds in the subsequent subsection. However, Definition 6 offers a more precise depiction of the lower bounds' asymptotic behaviors compared to the condition in Proposition 5. We will delve deeper into this distinction in Appendix D.

The upcoming example demonstrates how the notion of parameter-agnostic lower bounds is able to close the gap described in the beginning of this section.

Example (Parameter-Agnostic Lower Bound for Constant Stepsize `GD`)**.** *In the parameter-dependent regime, `GD` with properly tuned constant stepsize converges for L-smooth functions. However, it is well-known that `GD` with stepsize $\eta > 2/L$ does not converge in general. We now show how this is reflected by our framework of parameter-agnostic lower bounds. Let A_η denote `GD` with constant stepsize $\eta > 0$ and $\mathcal{F}_L$ the set of L-smooth functions. It is well-known that A_η diverges on the function $F(x) = \frac{L}{2}x^2$ if $L > 2/\eta$. In particular, we have for all $L \geq L_0 := 3/\eta$ that $\sup_{f\in\mathcal{F}_L} T_\varepsilon(A_\eta, f) = \infty$. By choosing $\varepsilon_0 = 1, K = 1$ and L_0 as above we obtain*

$$\forall\, \varepsilon \in (0, \varepsilon_0], L \geq L_0 \colon \sup_{f\in\mathcal{F}_L} T_\varepsilon(A_\eta, f) \geq 1 \cdot \infty.$$

Since we chose η arbitrary in the start, Proposition 5 implies that $g \equiv \infty$ is a parameter-agnostic lower bound for the family of `GD` with all constant stepsizes.

4.2 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:

$$\begin{aligned} &g_t \leftarrow \nabla f(x_t, \xi_t) \\ &\text{Choose } m_t \in \operatorname{cone}(g_1, \dots, g_t) \\ &x_{t+1} \leftarrow x_t - \frac{\eta}{t^\alpha} \frac{m_t}{\|m_t\|} \end{aligned} \tag{3}$$

We call algorithms that follow this procedure *General Normalized Momentum Methods* (see also Algorithm 3 in Appendix C.2). It is clear that NSGD-M from Theorem 2 is a member of this family of algorithms.

Theorem 6 (Parameter Agnostic Lower Bound for General Normalized Momentum Methods)**.** *Let $\mathcal{A}$ be the class of algorithms defined by (3) with $\alpha \geq 1/2$ and $\mathcal{F}_{L_0,L_1,\Delta_1}$ the set of (L_0, L_1)-smooth functions with $F(x_1) - \inf_{x\in\mathbb{R}^d} F(x) \leq \Delta_1$. Furthermore assume the deterministic setting, i.e. the gradient oracle returns the true gradient. Then, for any $\delta \in (0, 1)$,*

$$g(\varepsilon, \theta) \coloneqq \left(\frac{\Delta_1 + \frac{1}{L_1}\left(e^{L_1^{1-\delta}} - 1\right)}{\varepsilon} \right)^2$$

is a parameter-agnostic lower bound for $\mathcal{A}$ on $\{\mathcal{F}_{L_0,L_1,\Delta_1} : L_0, L_1, \Delta_1 \geq 0\}$. The subset $\tilde{\mathcal{A}} \subseteq \mathcal{A}$ that satisfies $\alpha \geq 3/4$ furthermore has the parameter-agnostic lower bound

$$\tilde{g}(\varepsilon, \theta) \coloneqq \left(\frac{\Delta_1 + \frac{1}{L_1}\left(e^{L_1^{1-\delta}} - 1\right)}{\varepsilon} \right)^4.$$

In particular this lower bound applies to NSGD-M in Theorem 2.

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*. Consequently, this finding does not contradict Proposition 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, aligning with previous upper bounds (Yang et al., 2023; Cutkosky and Mehta, 2020).

Proof Sketch As shown in Proposition 5, to establish such a lower bound, we need a set of hard functions for each algorithm and large enough parameters. The subsequent Lemma accomplishes this.

Lemma 7. *Consider a General Normalized Momentum Method 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 and initialization x_1 with $F(x_1) - \inf_{x\in\mathbb{R}^d} F(x) \leq \Delta_1$ for which 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.

To prove the lemma, we consider the following function constructed by its derivatives: $F(x) \coloneqq \Delta_1 + \int_0^x F' d\lambda$, with

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

$$\text{where} \quad z_1 \coloneqq \frac{2}{L_0}, \quad z_2 \coloneqq \eta - \frac{1 - 2\epsilon}{L_0}, \quad z_3 \coloneqq \eta + \frac{M}{2\varepsilon},$$
$$z_4 \coloneqq z_3 + \frac{2\varepsilon}{L_0}, \quad \text{and} \quad M \coloneqq \Delta_1 + \frac{2}{L_1}\left(e^{\frac{\eta L_1}{4}} - 1\right).$$

In the appendix, we provide a plot of the function (see Figure 3). Notably, within the range $[z_1, \eta/2]$, the gradient increases exponentially with x. This steep gradient change is permissible due to the relaxed smoothness assumption. Initiating from $x_1 = 0$ and taking a step, the iterate arrives at $x_2 = \eta$. At this point, we can demonstrate that $F(x_2) \geq M$, signifying the emergence of an exponential dependency on L_1.

Subsequently, along the x-axis, the function's value descends with a gradient of -2ε. Iterations will consistently shift to the right due to these negative gradients. Given the algorithm's intrinsic normalization, the shift in x is limited to η/t^α during the t-th iteration. To move beyond the interval $[z_2, z_3]$ (where gradients remain large at -2ε), the condition $\sum_{t=1}^T \eta/t^\alpha \geq z_3$ must be satisfied, which gives us the lower bound for T. This completes the proof of Lemma 7.

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

Now we are ready to use Proposition 5 to finish the proof of Theorem 6. Therefore choose $\delta \in (0, 1)$ and let $A \in \mathcal{A}$ be specified by $\eta > 0$, $\alpha \in (1/2, 1)$ and any momentum generating rule. Define $\varepsilon_0 \coloneqq 1/2, C_1 \coloneqq \max\{1/2, 8/\eta\}$ and $K \coloneqq ((1-\alpha)/2\eta)^{\frac{1}{1-\alpha}}$. By Lemma 7

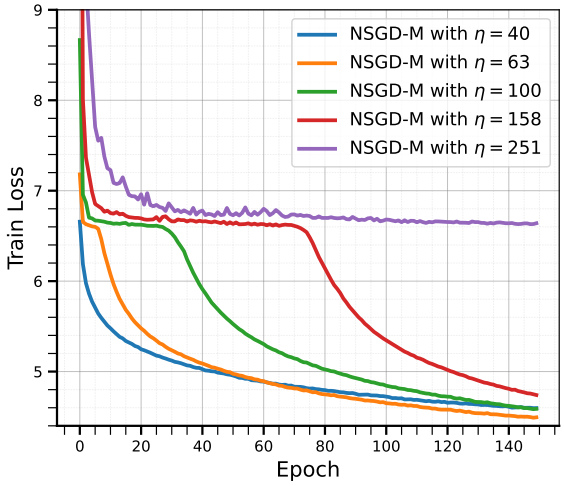


Figure 1: Training curve of NSGD-M for different stepsizes.

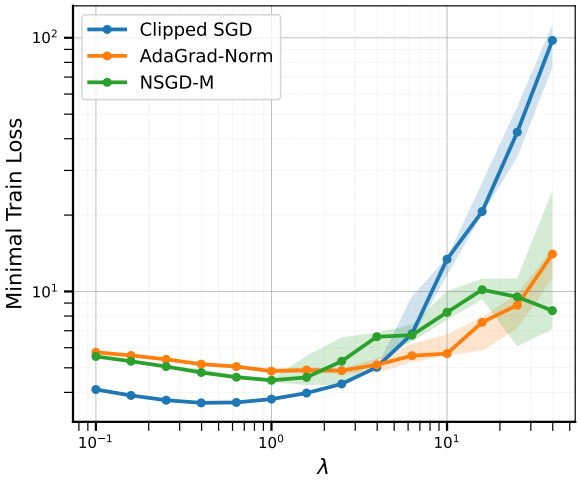


Figure 2: Minimal training loss within 150 epochs of different algorithms with stepsize $\lambda \cdot \eta_{\texttt{opt}}$. Shaded areas represent the minimal and maximal value within 5 seeds, the line the median.

we have that

$$T_\varepsilon(A, \mathcal{F}_{L_0,L_1,\Delta_1}) \geq K\left(\frac{\Delta_1 + \frac{1}{L_1}\left(e^{\frac{\eta L_1}{4}} - 1\right)}{\varepsilon}\right)^{\frac{1}{1-\alpha}}$$

for all $\Delta_1, L_0, L_1 \geq C_1$. Finally we define $\theta_0 := \max\left\{C_1, (4/\eta)^{\frac{1}{\delta}}\right\}$ to obtain $\frac{\eta L_1}{4} \geq L_1^{1-\delta}$ for all $L_1 \geq \theta_0$. This completes the proof.

5 EXPERIMENTS

In this section, we present experiments designed to empirically validate the theoretical findings of this paper. In concordance with our theory, the primary focus is to demonstrate the robustness of NSGD-M to hyperparameter selection in the context of (L_0, L_1)-smoothness. Language modeling tasks with LSTM and Transformer architectures are well-known settings for which (L_0, L_1)-smoothness was empirically confirmed to be necessary (Crawshaw et al., 2022; Zhang et al., 2020b). We therefore focus on these tasks.

Experimental Setup. To match the assumptions of our theory, we conduct training on the Penn Treebank (PTB) dataset (Mikolov et al., 2010) using the AWD-LSTM architecture (Merity et al., 2018). Additional experiments can be found in Appendix E. Besides NSGD-M, we also include AdaGrad-Norm (Faw et al., 2023) and Clipped SGD (Zhang et al., 2020b). For each algorithm we first chose the optimal stepsize $\eta_{\texttt{opt}}$ based on a course grid search in a 50 epoch training. The clipping threshold was fixed to be 0.25 in concordance to previous work (Zhang et al., 2020b) and the decay-rates of NSGD-M were chosen according to Theorem 2. For each algorithm, the final training was then carried out with stepsizes $\eta = \lambda \cdot \eta_{\texttt{opt}}$, where $\lambda = 10^{k/5}, k \in \{-5, -4, \ldots, 8\}$, for 150 epochs. This procedure is replicated with five different seeds to get more reliable results. The code is based on the experiments by Zhang et al. (2020a).

Discussion. Figure 1 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. Only after reaching a threshold, NSGD-M starts to decrease the loss. Figure 2 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.

6 CONCLUSION

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. Further, we introduce a lower bound framework specifically for the parameter-agnostic context, revealing that this exponential term is inescapable for a family of General Normalized Momentum Methods. 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 fully parameter-agnostic algorithm in the stochastic setting without an exponential term. Another interesting topic is the derivation of lower bounds for all first-order parameter-agnostic methods.

Acknowledgments

The work is supported by ETH research grant and Swiss National Science Foundation (SNSF) Project Funding No. 200021-207343.

References

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

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

Bottou, L., Curtis, F. E., and Nocedal, J. (2018). Optimization Methods for Large-Scale Machine Learning. *SIAM Review*, 60(2):223–311.

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

Cartis, C., Gould, N. I. M., and Toint, P. L. (2010). On the Complexity of Steepest Descent, Newton's and Regularized Newton's Methods for Nonconvex Unconstrained Optimization Problems. *SIAM Journal on Optimization*, 20(6):2833–2852.

Crawshaw, M., Liu, M., Orabona, F., Zhang, W., and Zhuang, Z. (2022). Robustness to Unbounded Smoothness of Generalized SignSGD. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems*, volume 35, pages 9955–9968. Curran Associates, Inc.

Cutkosky, A. and Boahen, K. (2017). Online Learning Without Prior Information. In Kale, S. and Shamir, O., editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 643–677. PMLR, PMLR.

Cutkosky, A. and Boahen, K. A. (2016). Online Convex Optimization with Unconstrained Domains and Losses. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

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

Cutkosky, A. and Orabona, F. (2018). Black-Box Reductions for Parameter-free Online Learning in Banach Spaces. In Bubeck, S., Perchet, V., and Rigollet, P., editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1493–1529. PMLR, PMLR.

Drori, Y. and Shamir, O. (2020). The Complexity of Finding Stationary Points with Stochastic Gradient Descent. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2658–2667. PMLR.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research*, 12(61):2121–2159.

Fatkhullin, I., Barakat, A., Kireeva, A., and He, N. (2023). Stochastic Policy Gradient Methods: Improved Sample Complexity for Fisher-non-degenerate Policies. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J., editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 9827–9869. PMLR.

Faw, M., Rout, L., Caramanis, C., and Shakkottai, S. (2023). Beyond Uniform Smoothness: A Stopped Analysis of Adaptive SGD. In Neu, G. and Rosasco, L., editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 89–160. PMLR.

Faw, M., Tziotis, I., Caramanis, C., Mokhtari, A., Shakkottai, S., and Ward, R. (2022). The Power of Adaptivity in SGD: Self-Tuning Step Sizes with Unbounded Gradients and Affine Variance. In Loh, P.-L. and Raginsky, M., editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178

of *Proceedings of Machine Learning Research*, pages 313–355. PMLR.

Ghadimi, S. and Lan, G. (2013). Stochastic First- and Zeroth-Order Methods for Nonconvex Stochastic Programming. *SIAM Journal on Optimization*, 23(4):2341–2368.

Hazan, E. and Kakade, S. (2019). Revisiting the Polyak step size. *arXiv preprint arXiv:1905.00313*.

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

Jiang, X. and Stich, S. U. (2023). Adaptive SGD with Polyak stepsize and Line-search: Robust Convergence and Variance Reduction. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Kavis, A., Levy, K. Y., Bach, F., and Cevher, V. (2019). UniXGrad: A Universal, Adaptive Algorithm with Optimal Guarantees for Constrained Optimization. In Wallach, H., Larochelle, H., Beygelzimer, A., d' Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Li, H., Jadbabaie, A., and Rakhlin, A. (2023a). Convergence of Adam under Relaxed Assumptions. *arXiv preprint arXiv:2304.13972*.

Li, H., Qian, J., Tian, Y., Rakhlin, A., and Jadbabaie, A. (2023b). Convex and Non-Convex Optimization under Generalized Smoothness. *arXiv preprint arXiv:2306.01264*.

Loizou, N., Vaswani, S., Hadj Laradji, I., and Lacoste-Julien, S. (2021). Stochastic Polyak Step-size for SGD: An Adaptive Learning Rate for Fast Convergence. In Banerjee, A. and Fukumizu, K., editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1306–1314. PMLR.

McMahan, H. B. and Streeter, M. J. (2010). Adaptive Bound Optimization for Online Convex Optimization. In *Annual Conference Computational Learning Theory*.

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

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

Mikolov, T., Karafiát, M., Burget, L., Černocký, J., and Khudanpur, S. (2010). Recurrent neural network based language model. In *Proc. Interspeech 2010*, pages 1045–1048.

Nemirovskij, A. S. and Yudin, D. B. (1983). Problem Complexity and Method Efficiency in Optimization.

Nesterov, Y. (2012). How to make the gradients small. *Optima. Mathematical Optimization Society Newsletter*, (88):10–11.

Nesterov, Y. (2018). *Lectures on Convex Optimization*, volume 137 of *Springer Optimization and Its Applications*. Springer, Cham. Second edition of [MR2142598].

Orabona, F. and Pal, D. (2016). Coin Betting and Parameter-Free Online Learning. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Orvieto, A., Lacoste-Julien, S., and Loizou, N. (2022). Dynamics of SGD with Stochastic Polyak Stepsizes: Truly Adaptive Variants and Convergence to Exact Solution. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K., editors, *Advances in Neural Information Processing Systems*.

Polyak, B. T. (1987). *Introduction to Optimization*. Translations Series in Mathematics and Engineering. Optimization Software, Inc., Publications Division, New York. Translated from the Russian, With a foreword by Dimitri P. Bertsekas.

Streeter, M. and McMahan, H. B. (2010). Less Regret via Online Conditioning. *arXiv preprint arXiv:1002.4862*.

Vaswani, S., Dubois-Taine, B., and Babanezhad, R. (2022). Towards Noise-adaptive, Problem-adaptive (Accelerated) Stochastic Gradient Descent. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S., editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 22015–22059. PMLR.

Wang, B., Zhang, H., Ma, Z., and Chen, W. (2023). Convergence of AdaGrad for Non-convex Objectives: Simple Proofs and Relaxed Assumptions. In Neu, G. and Rosasco, L., editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 161–190. PMLR.

Wang, B., Zhang, Y., Zhang, H., Meng, Q., Ma, Z.-M., Liu, T.-Y., and Chen, W. (2022). Provable adaptivity in adam. *arXiv preprint arXiv:2208.09900*.

Ward, R., Wu, X., and Bottou, L. (2019). AdaGrad Stepsizes: Sharp Convergence Over Nonconvex Landscapes. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97

of *Proceedings of Machine Learning Research*, pages 6677–6686. PMLR.

Yang, J., Li, X., Fatkhullin, I., and He, N. (2023). Two Sides of One Coin: the Limits of Untuned SGD and the Power of Adaptive Methods. *arXiv preprint arXiv:2305.12475*.

Yang, J., Li, X., and He, N. (2022). Nest Your Adaptive Algorithm for Parameter-Agnostic Nonconvex Minimax Optimization. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems*, volume 35, pages 11202–11216. Curran Associates, Inc.

Zhang, B., Jin, J., Fang, C., and Wang, L. (2020a). Improved Analysis of Clipping Algorithms for Nonconvex Optimization. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15511–15521. Curran Associates, Inc.

Zhang, J., He, T., Sra, S., and Jadbabaie, A. (2020b). Why Gradient Clipping Accelerates Training: A Theoretical Justification for Adaptivity. In *International Conference on Learning Representations*.

Zhao, S.-Y., Xie, Y.-P., and Li, W.-J. (2021). On the convergence and improvement of stochastic normalized gradient descent. *Science China Information Sciences*, 64(3):132103.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Parameter-Agnostic Optimization under Relaxed Smoothness Supplementary Materials

A BASIC PROPERTIES OF (L_0, L_1)-SMOOTHNESS

In this section, we prove basic properties of (L_0, L_1)-Smoothness. We start with the proof of the relation to the original definition by Zhang et al. (2020b).

Proof of Lemma 1. "⇒": This implication was already shown by Zhang et al. (2020a, Corollary A.4).

"⇐": We slightly adapt the proof by Faw et al. (2023, Proposition 1). Assume F is (L_0, L_1)-smooth according to Definition 3. 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 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 8 (c.f. (Zhang et al., 2020a, 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

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

tend to 1 as c tends towards 0.

Proof. This proof closely follows the arguments from Zhang et al. (2020a). We include the proof for completeness. Let $x, y \in \mathbb{R}^d$ and calculate

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

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

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

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 9 (Gradient Bound, c.f. (Zhang et al., 2020a, Lemma A.5))**.** *Let $L_0, L_1 > 0$ and assume that $f \colon \mathbb{R}^d \to \mathbb{R}$ is (L_0, L_1)-smooth. Further assume that f is lower bounded by f^*. Then all $x \in \mathbb{R}^d$ satisfy*

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

Proof. This proof is again based on Zhang et al. (2020a). We include it for the sake of completeness. Let $x \in \mathbb{R}^d$. Firstly note that, for A_1 from Definition 3, 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 $\lambda \coloneqq \frac{1}{2A_1(c)(L_0 + L_1 \|\nabla f(x)\|)}$ and $y \coloneqq x - \lambda \nabla f(x)$. Then Lemma 8 yields

$$f^* \le f(y) \le f(x) - \lambda \|\nabla f(x)\|^2 + A_1(c)(L_0 + L_1 \|\nabla f(x)\|) \lambda^2 \|\nabla f(x)\|^2 = f(x) - \frac{\lambda}{2} \|\nabla f(x)\|^2.$$

We now differentiate between the two cases $\|\nabla f(x)\| \le \frac{L_0}{L_1}$ and $\|\nabla f(x)\| > \frac{L_0}{L_1}$. Therefore,

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

This shows the claim. □

B 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 10 (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. (Fatkhullin et al., 2023, 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

$$\prod_{t=a}^{b}\left(1-t^{-q}\right) \leq \exp\left(-\sum_{\tau=a}^{b} t^{-q}\right) \leq \exp\left(-\int_{a}^{b+1} t^{-q}dt\right) = \exp\left(\frac{1}{1-q}\left(a^{1-q}-(b+1)^{1-q}\right)\right), \tag{4}$$

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

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

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

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

[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 $(1-q)^{-1}$.

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 (Fatkhullin et al., 2023). 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 Lemma 10.

Lemma 11 (Technical Lemma)**.** *Let $\eta > 0$ and for $t \in \mathbb{N}_{\geq 1}$ we set*

$$\beta_t \coloneqq 1 - t^{-1/2},$$
$$\eta_t \coloneqq \eta t^{-3/4}.$$

Then, for $E_t \coloneqq e^{L_1 \eta_t}$, the following statements hold.

a) For all $T \in \mathbb{N}_{\geq 1}$ we have

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

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

b) For all $T \in \mathbb{N}_{\geq 1}$ we have

i) $\sum_{t=1}^{T} \eta_t^2 E_t \leq 3\eta^2 e^{\eta L_1}$;

ii) $\sum_{t=1}^{T} \eta_t \sum_{\tau=2}^{t} \eta_{\tau-1} E_{\tau-1} \prod_{\kappa=\tau}^{t} \beta_\kappa \leq \frac{5}{2}\eta^2\big(3e^{\eta L_1} + \log(T)\big)$.

c) For $t \in \mathbb{N}_{\geq 1}$ define $\delta_t \coloneqq 4\eta(t-1)^{\frac{1}{4}} - 3\eta$. Then, for all $b \in \mathbb{N}_{\geq 2}$, we have

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

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

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

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

a) i) The inequality follows from

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

where we used Lemma 10 *ii)* in the first inequality.

a) ii) For ease of notation let $\alpha_t \coloneqq 1 - \beta_t$. We start by regrouping

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

Applying Lemma 10 *i)*, *iii)* and *a) i)* now yields the statement:

$$\sum_{t=1}^{T} \eta_t \sqrt{\sum_{\tau=1}^{t} \alpha_\tau^2 \prod_{\kappa=\tau+1}^{t} \beta_\kappa^2} \overset{i),10}{\leq} \frac{7}{2}\eta + \sum_{t=2}^{T} \eta_t \sqrt{2e^2 (t+1)^{-q}} \leq \eta\Big(\frac{7}{2} + \sqrt{2e^2}\log(T)\Big).$$

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) We start by calculating

$$\sum_{t=1}^{T} \eta_t^2 E_t = \eta^2 e^{\eta L_1} + \sum_{t=2}^{T} \eta_t^2 E_t \leq \eta^2 e^{\eta L_1} + \eta^2 \int_1^T t^{-2p} e^{\eta L_1 t^{-p}}.$$

Next, the local uniform convergence of the exponential series yields

$$\int_1^T t^{-2p} e^{\eta L_1 t^{-p}} = \int_1^T t^{-2p} \sum_{k=0}^{\infty} \frac{(\eta L_1 t^{-p})^k}{k!} dt = \sum_{k=0}^{\infty} \frac{(\eta L_1)^k}{k!} \int_1^T t^{-p(k+2)} dt \leq \sum_{k=0}^{\infty} \frac{(\eta L_1)^k}{k!} \frac{4}{3k+2} \leq 2e^{\eta L_1},$$

and combining these results yields the claim.

b) *ii*) We first use Lemma 10 *ii*) to derive

$$\begin{aligned}\sum_{t=1}^{T}\eta_t\sum_{\tau=2}^{t}\eta_{\tau-1}E_{\tau-1}\prod_{\kappa=\tau}^{t}\beta_\kappa &= \sum_{\tau=2}^{T}\eta_{\tau-1}E_{\tau-1}\sum_{t=\tau}^{T}\eta_t\prod_{\kappa=\tau}^{t}\beta_\kappa\\ &\le \sum_{\tau=2}^{T}\eta_{\tau-1}E_{\tau-1}\eta\exp\left(2\left(\sqrt{\tau}-\sqrt{\tau-1}\right)\right)(\tau-1)^{-1/4}\\ &\le \frac{5}{2}\eta^2\sum_{\tau=2}^{T}(\tau-1)^{-1}E_{\tau-1}\\ &\le \frac{5}{2}\eta^2\left(e^{\eta L_1}+\int_1^T\tau^{-1}e^{\eta L_1\tau^{-p}}d\tau\right).\end{aligned}$$

As in *b*) *i*) we now calculate

$$\int_1^T\tau^{-1}e^{\eta L_1\tau^{-p}}d\tau=\sum_{k=0}^{\infty}\frac{(\eta L_1)^k}{k!}\int_1^T\tau^{-pk-1}d\tau\le\log(T)+\sum_{k=1}^{\infty}\frac{(\eta L_1)^k}{k!}\frac{4}{3k}\le\log(T)+2e^{\eta L_1}.$$

Combining these two results yields the claim.

c) *i*) We start off by calculating

$$\begin{aligned}\sum_{t=2}^{b}L_1\eta_tE_tt^{-\frac{1}{4}}\delta_te^{L_1\delta_t} &\le L_1\eta_2E_22^{-\frac{1}{4}}\eta e^{\eta L_1}+\sum_{t=3}^{b}L_1\eta_tE_tt^{-\frac{1}{4}}\delta_te^{L_1\delta_t}\\ &\le \frac{1}{2}\eta^2L_1e^{2\eta L_1}+4\eta\sum_{t=3}^{b}L_1\eta_tE_te^{L_1\delta_t}\end{aligned}$$

and further

$$\begin{aligned}\sum_{t=3}^{b}L_1\eta_tE_te^{L_1\delta_t} &\le \sum_{t=3}^{b}L_1\eta_t\exp\left(L_1\left(4\eta(t-1)^{\frac{1}{4}}-3\eta+\eta_t\right)\right)\\ &\le e^{-\frac{5}{2}\eta L_1}\sum_{t=3}^{b}L_1\eta_{t-1}e^{4\eta L_1(t-1)^{\frac{1}{4}}}\\ &\le e^{-\frac{5}{2}\eta L_1}\int_2^{b+1}\eta L_1(t-1)^{-p}e^{4\eta L_1(t-1)^{1-p}}dt\\ &= e^{-\frac{5}{2}\eta L_1}\left(e^{4\eta L_1b^{\frac{1}{4}}}-e^{4\eta L_1}\right).\end{aligned}\tag{5}$$

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 (5) also holds for $b=2$ yields the claim.

ii) We have

$$\begin{aligned}\sum_{t=1}^{b}L_1\eta_tE_tt^{-\frac{1}{4}}e^{L_1\delta_t} &= \eta L_1E_1+\frac{1}{2}\eta L_1E_2e^{\eta L_1}+\sum_{t=3}^{b}L_1\eta_tE_tt^{-\frac{1}{4}}e^{L_1\delta_t}\\ &\le \eta L_1e^{\eta L_1}+\frac{1}{2}\eta L_1e^{\left(1+2^{-\frac{3}{4}}\right)\eta L_1}+\sum_{t=3}^{b}L_1\eta_tE_te^{L_1\delta_t}\end{aligned}\tag{6}$$

and using (5) yields

$$\sum_{t=1}^{b}L_1\eta_tE_tt^{-\frac{1}{4}}e^{L_1\delta_t}\le\frac{3}{2}\eta L_1e^{\frac{5}{3}\eta L_1}+e^{-\frac{5}{2}\eta L_1}\left(e^{4\eta L_1b^{\frac{1}{4}}}-e^{4\eta L_1}\right).$$

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 Lemma 10 *iii*), using partial integration to derive

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

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 (6) yields the claim.

□

C MISSING PROOFS

This section contains the proofs for Section 3 and Section 4.

C.1 Proofs for Parameter-Agnostic Upper Bounds

C.1.1 Stochastic Setting

We start with the proof of Theorem 2, which has the same structure as in the L-smooth setting (Cutkosky and Mehta, 2020): 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 11 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 (Yang et al., 2022). In particular, this implies that $p = 3/4$ is the only possible choice to achieve the optimal complexity (Cutkosky and Mehta, 2020; Zhang et al., 2020a).

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 $\mathcal{F}_t := \sigma(\xi_1, \xi_2, \ldots, \xi_t)$.

Lemma 12 (Descent Lemma)**.** *Assume ((L_0, L_1)-smoothness) and let $t \in \mathbb{N}_{\geq 2}$. Then*

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

where $E_t := e^{L_1 \eta_t}$. If we further assume (Lower Boundedness) we also get

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

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

Proof. The proof follows the arguments by Zhao et al. (2021). Using Lemma 8 we get

$$\begin{aligned} F(x_{t+1}) - F(x_t) &\leq \nabla F(x_t)^\top (x_{t+1} - x_t) + \frac{\eta_t^2}{2}(L_0 B_0(L_1 \eta_t) + L_1 B_1(L_1 \eta_t) \|\nabla F(x_t)\|) \\ &= -\frac{\eta_t}{\|m_t\|} \nabla F(x)^\top m_t + \frac{\eta_t^2 E_t}{2}(L_0 + L_1 \|\nabla F(x_t)\|) \\ &= -\frac{\eta_t}{\|m_t\|}(\nabla F(x_t) - m_t)^\top m_t - \eta_t \|m_t\| + \frac{\eta_t^2 E_t}{2}(L_0 + L_1 \|\nabla F(x_t)\|), \end{aligned}$$

where we used that $B_0(c), B_1(c) \leq e^c$. 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

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

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

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

□

Lemma 13 (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*

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

where $\beta_{a:b}$ denotes $\prod_{t=a}^{b} \beta_t$ and $E_\tau = e^{L_1 \eta_\tau}$.

Proof. This proof is motivated by Cutkosky and Mehta (2020), and similar arguments are carried by Zhang et al. (2020a) and Yang et al. (2022). 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

$$\begin{aligned} \mathbb{E}\left[\gamma_j^\top \gamma_i\right] &= \mathbb{E}\left[\mathbb{E}\left[\gamma_j^\top \gamma_i \,\middle|\, \mathcal{F}_{j-1}\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\gamma_j \,|\, \mathcal{F}_{j-1}\right]^\top \gamma_i\right] \\ &= 0, \end{aligned} \tag{7}$$

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 equality. 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, (7) 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

$$\|S_t\| \leq \eta_{t-1}(A_0(L_1\eta_{t-1})L_0 + A_1(L_1\eta_{t-1})L_1 \|\nabla F(x_t)\|) \leq \eta_{t-1} E_{t-1}(L_0 + L_1 \|\nabla F(x_t)\|)$$

and hence

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

Putting these results together we get the claim. □

Now we are ready for the main result.

Theorem 14 (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*

$$\beta_t \coloneqq 1 - t^{-1/2}$$
$$\eta_t \coloneqq \eta t^{-3/4}.$$

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

$$\sum_{t=1}^{T} \frac{\eta_t}{2} \mathbb{E}\left[\|\nabla F(x_t)\|\right] \leq \Delta_1 + \eta\sigma\left(7 + 2\sqrt{2e^2}\log(T)\right) + \eta^2 L_0\left(15e^{\eta L_1} + 5\log(T)\right) + 21\eta^2 L_0 e^{48(\eta L_1)^2} + 6\eta e^{48(\eta L_1)^2}\|\nabla F(x_1)\|,$$

where $\Delta_1 \coloneqq F(x_1) - F^*$. 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 11. 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 12 and Lemma 13 to obtain

$$\begin{aligned}
\sum_{t=1}^{T} \eta_t \mathbb{E}\left[\|\nabla F(x_t)\|\right] &\overset{12}{\leq} \Delta_1 + \frac{L_0}{2}\sum_{t=1}^{T}\eta_t^2 E_t + \frac{L_1}{2}\sum_{t=1}^{T}\eta_t^2 E_t \mathbb{E}\left[\|\nabla F(x_t)\|\right] + 2\sum_{t=1}^{T}\eta_t \mathbb{E}\left[\|\nabla F(x_t) - m_t\|\right] \\
&\overset{13}{\leq} \Delta_1 + \frac{L_0}{2}\sum_{t=1}^{T}\eta_t^2 E_t + \frac{L_1}{2}\sum_{t=1}^{T}\eta_t^2 E_t \mathbb{E}\left[\|\nabla F(x_t)\|\right] + 2\sigma\sum_{t=1}^{T}\eta_t\sqrt{\sum_{\tau=1}^{t}\alpha_\tau^2\left(\beta_{(\tau+1):t}\right)^2} \\
&\quad + 2L_0\sum_{t=1}^{T}\eta_t\sum_{\tau=2}^{t}\eta_{\tau-1}E_{\tau-1}\beta_{\tau:t} + 2L_1\sum_{t=1}^{T}\eta_t\sum_{\tau=2}^{t}\eta_{\tau-1}E_{\tau-1}\beta_{\tau:t}\mathbb{E}\left[\|\nabla F(x_\tau)\|\right].
\end{aligned}$$

Next, we use Lemma 11 *a*) and *b*) to bound all terms that are independent of the iterates x_t. This leaves us with

$$\begin{aligned}
\sum_{t=1}^{T} \eta_t \mathbb{E}\left[\|\nabla F(x_t)\|\right] &\leq \Delta_1 + \eta\sigma\left(7 + 2\sqrt{2e^2}\log(T)\right) + \eta^2 L_0\left(17e^{\eta L_1} + 5\log(T)\right) \\
&\quad + \underbrace{\frac{L_1}{2}\sum_{t=1}^{T}\eta_t^2 E_t \mathbb{E}\left[\|\nabla F(x_t)\|\right] + 2L_1\sum_{\tau=2}^{T}\eta_{\tau-1}E_{\tau-1}\left(\sum_{t=\tau}^{T}\eta_t\beta_{\tau:t}\right)\mathbb{E}\left[\|\nabla F(x_\tau)\|\right]}_{=:(A)},
\end{aligned} \tag{8}$$

where we rearranged the sums of the last term. We then focus on upper bounding (A). Therefore we use Lemma 10 *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

$$(A) \leq \underbrace{\sum_{t=1}^{t_0-1} \eta_t \Big(M L_1 \eta t^{-1/4} E_t \Big) \mathbb{E}\left[\|\nabla F(x_t)\|\right]}_{(B)} + \sum_{t=t_0}^{T} \frac{\eta_t}{2} \mathbb{E}\left[\|\nabla F(x_t)\|\right]. \tag{9}$$

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 11 $c)$ $i)$ allows us to upper bound $(B1)$ via

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

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 11 $c)$ $ii)$ 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 (9) 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 (8) yields

$$\begin{aligned}
\frac{1}{2} \sum_{t=1}^{T} \eta_t \mathbb{E}\left[\|\nabla F(x_t)\|\right] \leq \Delta_1 + \eta\sigma \Big(7 + 2\sqrt{2e^2} \log(T) \Big) + \eta^2 L_0 \big(17 e^{\eta L_1} + 5 \log(T) \big) \\
+ 21 \eta^2 L_0 e^{48(\eta L_1)^2} L_0 + 6\eta e^{48(\eta L_1)^2} \|\nabla F(x_1)\|.
\end{aligned}$$

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 11 *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 15. *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 $x_1 \in \mathbb{R}^d$ and $T \in \mathbb{N}_{\geq 3}$ satisfies*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla F(x_t)\|\right] \leq \frac{\left(14 + 96L_1e^{L_1^2}\right)\Delta_1 + \left(6e^{L_1/7} + 2\log(T) + 6e^{L_1^2}\right)L_0}{T^{\frac{1}{4}}} + \frac{12e\log(T)\sigma + 12e^{L_1^2}\min\left\{\frac{L_0}{L_1}, \sqrt{8L_0\Delta_1}\right\}}{T^{\frac{1}{4}}},$$

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 14 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 9, 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

$$\frac{1}{T^{3/4}}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla F(x_t)\|\right] \leq \left(14 + 96L_1e^{L_1^2}\right)\Delta_1 + 12e\log(T)\sigma + \left(6e^{L_1/7} + 2\log(T) + 6e^{L_1^2}\right)L_0 + 12e^{L_1^2}\min\left\{\frac{L_0}{L_1}, \sqrt{8L_0\Delta_1}\right\}$$

and hence proved the first claim.

For the second claim assume $L_1 \geq 7/2$. We now can use the second statement in Theorem 14 to get

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

where we used that $14e^{L_1/7} + 2L_1^{-2}e^{L_1^2} \leq 2e^{L_1^2}$ for $L_1 \geq 7/2$. □

Finally we provide the formal statement of Proposition 3.

Corollary 16 (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*

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

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

Proof. Denote $\eta := 1/12$. By plugging our choice of η_t into (8) we obtain

$$\sum_{t=1}^{T}\frac{1}{2}\eta_t\mathbb{E}\left[\|\nabla F(x_t)\|\right] \leq \Delta_1 + \eta\sigma\left(7 + 2\sqrt{2e^2}\log(T)\right) + \eta^2 L_0(19 + 5\log(T))$$

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

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla F(x_t)\|\right] \leq \frac{\frac{2\Delta_1}{\eta} + \left(14 + 4\sqrt{2e^2}\log(T)\right)\sigma + \eta(38 + 10\log(T))L_0}{T^{1/4}}.$$

□

C.1.2 Deterministic Setting

In this subsection, we provide the result for GD with Backtracking Line Search.

Proof of Theorem 4. By Lemma 9 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 $\|\nabla F(x_t)\| \leq \max\left\{8L_1\Delta_1, \frac{L_0}{L_1}\right\} =: u(L_0, L_1, \Delta_1)$ 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 8 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 $L := 2L_0 + (e-1)L_1 u(L_0, L_1, \Delta_1)$. 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

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

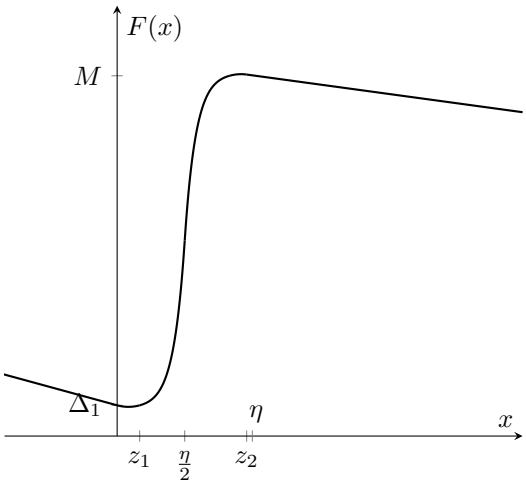
and thus

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

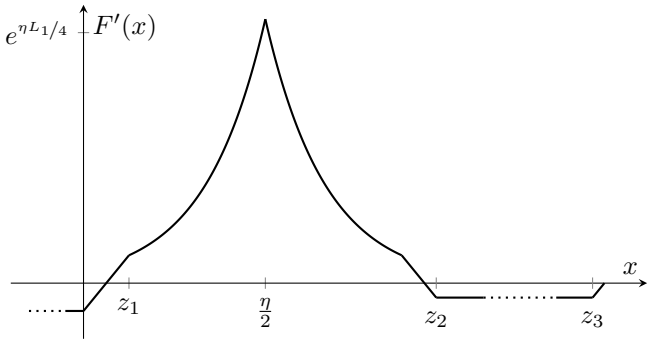
This finishes the proof. □

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 $m_t \in \operatorname{cone}(g_1, \ldots, g_t) \setminus \{0\}$
  $x_{t+1} \leftarrow x_t - \frac{\eta}{t^\alpha} \frac{m_t}{\|m_t\|}$
end



(a) Plot of $F(x)$



(b) Plot of $F'(x)$

Figure 3: Plot of the hard function used in the proof of Lemma 7.

C.2 Proofs for Parameter-Agnostic Lower Bounds

In this section, we provide the formal proofs for Section 4.

Proof of Lemma 7. Define $z_1 \coloneqq \frac{2}{L_0}$ and the derivatives

$$\begin{aligned} p'(x) &\coloneqq L_0 x & p_1'(x) &\coloneqq p'(x) - 1, & p_2'(x) &\coloneqq p'(\eta - x) - 1, \\ q'(x) &\coloneqq e^{L_1 x} & q_1'(x) &\coloneqq q'(x - z_1), & q_2'(x) &\coloneqq q'(\eta - z_1 - x). \end{aligned}$$

We now define the function F via its derivative

$$F' \coloneqq -\mathbb{1}_{(-\infty,0)} + \mathbb{1}_{[0,z_1)} p_1' + \mathbb{1}_{[z_1,\eta/2)} q_1' + \mathbb{1}_{[\eta/2,\eta-z_1)} q_2' + \mathbb{1}_{[\eta-z_1,z_2)} p_2' - 2\varepsilon \mathbb{1}_{[z_2,z_3)} + \mathbb{1}_{[z_3,z_4)} h$$

where $z_2 \coloneqq \eta - z_1 + \frac{1+2\varepsilon}{L_0} \leq \eta$ and z_3, z_4 and h will be determined later. Then $F(x) \coloneqq \Delta_1 + \int_0^x F' d\lambda$ (see Figure 3a) satisfies

$$\begin{aligned} F(x) = \Delta_1 &+ \frac{2}{L_1}\Big(e^{L_1(\eta/2 - z_1)} - 1\Big)\mathbb{1}_{[\eta/2,\infty)}(x) - \mathbb{1}_{(-\infty,0)}(x) \cdot x \\ &+ \mathbb{1}_{[0,z_1)}(x) \cdot \left(\frac{L_0}{2}x^2 - x\right) + \mathbb{1}_{[z_1,\eta/2)}(x) \cdot \frac{1}{L_1}\Big(e^{L_1(x - z_1)} - 1\Big) \\ &- \mathbb{1}_{[\eta/2,\eta-z_1)}(x)\frac{1}{L_1}\Big(e^{L_1(\eta - z_1 - x)} - 1\Big) + \mathbb{1}_{[\eta-z_1,z_2)}(x)\left(x - \eta - z_1 - \frac{L_0(x - 1 - z_1)^2}{2}\right) \\ &- 2\varepsilon \mathbb{1}_{[z_2,z_3)}(x) + \mathbb{1}_{[z_3,z_4)}(x)h(x) \end{aligned}$$

and in particular

$$F(\eta) \geq \Delta_1 + \frac{2}{L_1}\Big(e^{L_1(\eta/2 - z_1)} - 1\Big).$$

By our choice of L_0 we get that $\frac{\eta}{2} - z_1 \geq \frac{\eta}{4}$ which implies $F(\eta) \geq \Delta_1 + \frac{2}{L_1}\Big(e^{\frac{\eta L_1}{4}} - 1\Big) =: C$. We have

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

and hence

$$F(x_T) \geq C - 2\varepsilon(x_T - \eta) \geq 2\eta\varepsilon + C - \frac{2\eta\varepsilon}{1-\alpha} T^{1-\alpha}.$$

Since

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

now implies that $F(x_T) \geq 2\eta\varepsilon$, the gradient of F at x_T is still $2\varepsilon > \varepsilon$ and we have not yet reached an ε-stationary point. Finally we are left with the task of flattening F out while making sure it never attains negative values and is still (L_0, L_1)-smooth. Therefore set $z_3 \coloneqq \eta + \frac{C}{2\varepsilon}$ and $z_4 \coloneqq z_3 + \frac{2\varepsilon}{L_0}$. Now let $h(x) \coloneqq p'(x - z_3) - 2\varepsilon$ and note that this achieves the exact goal we were aiming for.

The only thing left to do, is to show that F is indeed (L_0, L_1)-smooth. It is clear that F is (L_0, L_1)-smooth on each of the subintervals $(-\infty, 0), ..., [z_3, z_4), [z_4, \infty)$. The claim hence follows from the upcoming Lemma 17. □

Lemma 17. *Let $I \subseteq \mathbb{R}$ be an interval, $a \in I$ and set $I_- \coloneqq \{x \in I \mid x \leq a\}$, $I_+ \coloneqq \{x \in I \mid x \geq a\}$. Further Let $f\colon \mathbb{R} \to \mathbb{R}$ be continuously differentiable and suppose that f satisfied the inequality from Definition 3 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

$$\begin{aligned}\|\nabla f(x) - \nabla f(y)\| &= \|\nabla f(x) - \nabla f(a) + \nabla f(a) - \nabla f(y)\| \\ &\le L_0(\|x-a\| A_0(c_1) + \|a-y\| A_0(c_2)) \\ &\quad + L_1 \|x-a\| A_1(c_1) \|\nabla f(x)\| + L_1 \|a-y\| A_1(c_2) \|\nabla f(a)\|. \end{aligned} \tag{10}$$

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

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

and hence

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

We now plug this result into (10) and rearrange to obtain

$$\begin{aligned}\|\nabla f(x) - \nabla f(y)\| \le L_0(e^{c_2} \|x-a\| A_0(c_1) + \|a-x\| A_0(c_2)) \\ + L_1 \|\nabla f(x)\| (\|x-a\| A_1(c_1) + \|a-y\| A_1(c_2) e^{c_1}).\end{aligned} \tag{11}$$

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 (11), which corresponds to the L_0-dependence. Calculating yields

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

In the last inequality we used that for all $a, b, L_1 \ge 0$ the following inequality holds: $b + (a+b)e^{L_1 b} + b e^{L_1(a+b)} \le a + b + (a+b)^{L_1(a+b)}$. This follows by taking partial derivatives with respect to L_1. Finally we plug everything into (11) and obtain

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

This finishes the proof. □

D ADDITIONAL DISCUSSION ON PARAMETER-AGNOSTIC LOWER BOUNDS

In this section, we provide further discussion on the notion of parameter-agnostic lower bounds. Additionally, we highlight the difference between Definition 6 and the condition in Proposition 5.

The section is organised as follows: We start by introducing the necessary notation, assumptions, and definitions in Appendix D.1. Subsequently, in Appendix D.2, we present an alternative way to motivate our definition of parameter-agnostic lower bounds. This alternative perspective allows for a more intuitive distinctions between Definition 6 and the condition in Proposition 5, as discussed in Remark 1.

D.1 Preliminaries

Notation. *Throughout this section, let $\Theta \subseteq \mathbb{R}^k$ denote a parameter space that is unbounded in each dimension, i.e. there exists a sequence $\theta^{(n)} \in \Theta$ such that $\theta_i^{(n)} \to \infty$ for all $i \in [k]$. Additionally, let $\mathscr{F}_\Theta = \{\mathcal{F}_\theta \colon \theta \in \Theta\}$ be a parameterized family of function spaces.*

For simplicity, we furthermore assume that all algorithms satisfy $x_1 = 0$. If this is not the case, we can apply A to the shifted function $\tilde{f}(x) = f(x - x_1)$. For the scope of this section, we therefore restrict $\mathcal{A}_{\text{det}}$ to deterministic algorithms that use $x_1 = 0$.

Lastly, we introduce a multivariate $\mathcal{O}$-notation. While the extension of the $\mathcal{O}$-notation to a multivariate setting comes with technical complexities, as noted by Howell (2008), the straightforward extension is sufficient for our purposes.

Definition 7 (Multivariate $\mathcal{O}$-Notation)**.** *Consider a function $h \colon (0,\infty) \times \Theta \to [0,\infty]$. We employ the following definitions:*

i) The multivariate $\mathcal{O}$ is given by the set

$$\mathcal{O}(h) \coloneqq \{f \colon (0,\infty) \times \Theta \to [0,\infty] \mid \exists\, \varepsilon_0, \theta_0, K > 0\, \forall\, \varepsilon \in (0,\varepsilon_0], \theta \geq \theta_0 \colon f(\varepsilon,\theta) \leq K h(\varepsilon,\theta)\}.$$

ii) Analogously, the multivariate o is defined as the set

$$o(h) \coloneqq \{f \colon (0,\infty) \times \Theta \to [0,\infty] \mid \forall\, \kappa > 0\, \exists\, \varepsilon_0, \theta_0 > 0\, \forall\, \varepsilon \in (0,\varepsilon_0], \theta \geq \theta_0 \colon f(\varepsilon,\theta) \leq \kappa h(\varepsilon,\theta)\}.$$

Here $\theta \geq C$ is to be understood component-wise. We also adopt standard $\mathcal{O}$-notation $f(\varepsilon,\theta) = \mathcal{O}(h(\varepsilon,\theta))$, $(\varepsilon \to 0, \theta \to \infty)$ to indicate $f \in \mathcal{O}(h)$. Analogously, we use $f(\varepsilon,\theta) = o(h(\varepsilon,\theta))$, $(\varepsilon \to 0, \theta \to \infty)$ to signify $f \in \mathcal{O}(h)$.

D.2 Another Point of View

In this section, we re-examine the definition of parameter-agnostic lower bounds through the lens of order theory. This perspective serves two purposes. Firstly, it enables us to formally compare the performance of two parameter-agnostic algorithms. Secondly, it better highlights the differences between Definition 6 and Proposition 5.

To start off, we address the question of how to compare different parameter-agnostic algorithms to determine which one is "better". To this end, we first introduce the concept of parameter-agnostic complexity of an algorithm, which maps each combination of θ and ε to the corresponding worst-case performance.

Definition 8 (Parameter-Agnostic Complexity of an Algorithm)**.** *For any $A \in \mathcal{A}_{\text{det}}$ we call $h_A \colon (0,\infty) \times \Theta \to [1,\infty]$,*

$$h_A(\varepsilon,\theta) \coloneqq \sup_{f \in \mathcal{F}_\theta} T_\varepsilon(A,f)$$

the parameter-agnostic complexity for A on $\mathscr{F}_\Theta$. *Here $T_\varepsilon(A,f) = \inf\{t \in \mathbb{N}_{\geq 1} \mid \|\nabla f(x_t)\| \leq \varepsilon\}$ denotes the number of iterations required for A to reach an ε-stationary point of f.*

To illustrate this notion, let us consider the example of Gradient Descent with constant stepsizes applied to L-smooth functions.

Example 1. *Let $\{A_\eta : \eta > 0\} = \mathcal{A} \subseteq \mathcal{A}_{\text{det}}$ be the set of Gradient Descent algorithms with constant stepsizes $\eta > 0$ and $x_1 = 0$. Furthermore, for each $L \geq 0$ and $\Delta_1 \geq 0$, let $\mathcal{F}_{L,\Delta_1}$ denote the set of all L-smooth functions with initialization gap $F(0) - \inf_{x\in\mathbb{R}^d} F(x) \leq \Delta_1$, and $\Theta = [0,\infty)^2$. For each $A_\eta \in \mathcal{A}$ we will now calculate the parameter-agnostic complexity on $\mathscr{F}_\Theta$. Firstly, for $\eta < 2/L$ it is well known that*

$$h_{A_\eta}(\varepsilon, L, \Delta_1) \leq \left\lceil \frac{\Delta_1}{\eta\left(\frac{L\eta}{2} - 1\right)} \varepsilon^{-2} \right\rceil.$$

On the other hand, if $\eta \geq 2/L$, we can construct the function $F(x) = \frac{L}{2}\left(x + \sqrt{2}\frac{\varepsilon}{L}\right)^2$ that is L-smooth and on which A_η will not converge. Hence we get that

$$T_\varepsilon(A_\eta, F) = \infty. \tag{12}$$

Now note that F belongs to $\mathcal{F}_{L,\Delta_1}$ for all $\Delta_1 \geq \varepsilon^2/L$. Therefore (12) implies that for all such Δ_1 and $\eta \geq 2/L$ we have $h_{A_\eta}(\varepsilon, L, \Delta_1) = \infty$. In particular, as $L, \Delta_1 \to \infty$ and $\varepsilon \to 0$ we get that $h_{A_\eta}(\varepsilon, L, \Delta_1) = \infty$.

Now that we have established a measure for the parameter-agnostic complexity of an individual algorithm, the next logical step is to consider how to compare two algorithms to determine which one is "better". We argue that in general algorithms are considered better than others, if they have a preferable behaviour as problems get harder. We therefore introduce the following (pre-)order for parameter-agnostic complexities.

Definition 9 (Ordering Complexities)**.** *Let $\mathscr{C} = \{f : (0,\infty) \times \Theta \to [1,\infty]\}$ denote the set of all possible complexities. Then we define the relation $\preceq$ on $\mathscr{C}$ as*

$$h_1 \preceq h_2 \Leftrightarrow h_1(\varepsilon, \theta) = \mathcal{O}(h_2(\varepsilon, \theta)), \qquad (\varepsilon \to 0, \theta \to \infty)$$

where $\mathcal{O}$ denotes the multivariate $\mathcal{O}$-notation (see Definition 7).

This definition paves the way for comparing the parameter-agnostic complexities of different algorithms. We say that a (parameter-agnostic) algorithm A is at least as good as algorithm B, if $h_A \preceq h_B$. This observation naturally leads to the following definition.

Definition 10 (Naïve Parameter-Agnostic Lower Bound)**.** *Let $\mathcal{A} \subseteq \mathcal{A}_{\text{det}}$ be an algorithm class and $g : (0,\infty) \times \Theta \to [1,\infty]$. Then we call g* weak parameter-agnostic lower bound for $\mathcal{A}$ on $\mathscr{F}_\Theta$, *if*

$$\forall\, A \in \mathcal{A} : g \preceq h_A. \tag{13}$$

When comparing the definition of $\preceq$ with the assumption in Proposition 5, we can observe that (13) is equivalent to the assumption stated in the proposition. Therefore, discussing the difference between Proposition 5 and Definition 6 boils down to understanding how Definition 6 and Definition 10 differ.

Though the concept of a weak parameter-agnostic lower bound is intuitive and straightforward, its limitations become evident when examined more closely. The following example highlights this issue.

Example 2. *Consider $\mathcal{A} = \{A_1, A_2\} \subseteq \mathcal{A}_{\text{det}}$ and let $\mathcal{F}_{L,\Delta_1}, \mathscr{F}_\Theta$ be defined as in Example 1. Suppose the parameter-agnostic complexities of A_1 and A_2 are given by*

$$h_{A_1}(\varepsilon, L, \Delta_1) = \frac{\Delta_1 + e^L}{\varepsilon^2},$$
$$h_{A_2}(\varepsilon, L, \Delta_1) = \frac{e^{\Delta_1} + L}{\varepsilon^2}.$$

The best possible weak parameter-agnostic lower bound for $\mathcal{A}$ on $\mathscr{F}_\Theta$ is then given by $g(\varepsilon, L, \Delta_1) = \frac{\Delta_1 + L}{\varepsilon^2}$. However, this lower bound fails to capture the fact that all algorithms in $\mathcal{A}$ suffer from an exponential dependence on at least one parameter.

Motivated by this shortcoming of weak parameter-agnostic lower bounds, we instead chose Definition 6 for our notion of parameter-agnostic lower bounds. In our current setting, Definition 6 can be rephrased as follows.

Proposition 18. *Let $\mathcal{A} \subseteq \mathcal{A}_{\text{det}}$ be an algorithm class and $g\colon (0,\infty)\times\Theta \to [1,\infty]$. Then g is a parameter-agnostic lower bound of $\mathcal{A}$ on $\mathscr{F}_\Theta$ as defined in Definition 6 if and only if*

$$\nexists A \in \mathcal{A}\colon h_A \prec g. \tag{14}$$

Here we define $h_1 \prec h_2$ if $h_1(\varepsilon,\theta) = o(h_2(\varepsilon,\theta))$ for $\varepsilon \to 0, \theta \to \infty$ (see Definition 7).

Specifically, (14) ensures that no algorithm A in the class $\mathcal{A}$ can have a parameter-agnostic complexity h_A that is "better" (in the little-o sense) than the proposed lower bound g.

Let us revisit Example 2 to see how this definition fixes the previously discussed issue.

Example 3. *Consider the same setting as in Example 2 and define $g_1(\varepsilon, L, \Delta_1) \coloneqq \frac{\Delta_1 + e^L}{\varepsilon^2}$, $g_2(\varepsilon, L, \Delta_1) \coloneqq \frac{e^{\Delta_1} + L}{\varepsilon^2}$. Then both, g_1 and g_2 are parameter-agnostic lower bounds of $\mathcal{A}$ on $\mathscr{F}_\Theta$, while neither of them is a weak parameter-agnostic lower bound. This notion of lower bound does hence capture the fact, that there is exponential dependence in at least one variable.*

This demonstrates the utility of employing the definition in Proposition 18 over weak parameter-agnostic lower bounds. The more nuanced criterion allows for a better representation of the complexities from algorithms in $\mathcal{A}$. The following remark delves deeper into this distinction.

Remark 1. The main difference between Definition 6 and Definition 10 (and therefore Proposition 5) is how they handle incomparable algorithms, i.e. algorithms for which neither $h_A \preceq h_B$ nor $h_B \preceq h_A$. Definition 10 enforces *a*) that g is comparable with all complexities and *b*) that g must be at least as good as all complexities. Definition 6 on the other hand only requires that complexities which are comparable with g must not be strictly better than g.

When focusing on parameters, the difference can be characterized as follows: A weak parameter-agnostic lower bound guarantees that there does not exist an algorithm in $\mathcal{A}$, that has a better dependence in *any single parameter*. Parameter-agnostic lower bounds on the other hand guarantee, that there does not exist an algorithm A which has better dependencies *in all parameters*.

From an order-theoretic standpoint, the difference is nearly the same as the difference between lower bounds and minimal elements. The only small difference is that we do not force g to be in the set of complexities $\{h_A\colon A \in \mathcal{A}\}$.

Finally we show that every weak parameter-agnostic lower bound is also a parameter-agnostic lower bound, as claimed by Proposition 5.

Lemma 19 (Rephrased Proposition 5)**.** *Let $\mathcal{A} \subseteq \mathcal{A}_{\text{det}}$ be an algorithm class and $g\colon (0,\infty)\times\Theta \to [1,\infty]$. If g is a weak parameter-agnostic lower bound of $\mathcal{A}$ on $\mathscr{F}_\Theta$, then g is also a parameter-agnostic lower bound of $\mathcal{A}$ on $\mathscr{F}_\Theta$.*

Proof. Let us first first recall the logical statements behind the two version of lower bounds. Firstly, (13) can be rewritten to

$$\forall\, A \in \mathcal{A}\, \exists\, \varepsilon_0, \theta_0, K > 0\, \forall\, \varepsilon \in (0, \varepsilon_0], \theta \geq \theta_0\colon g(\varepsilon,\theta) \leq K h_A(\varepsilon,\theta). \tag{15}$$

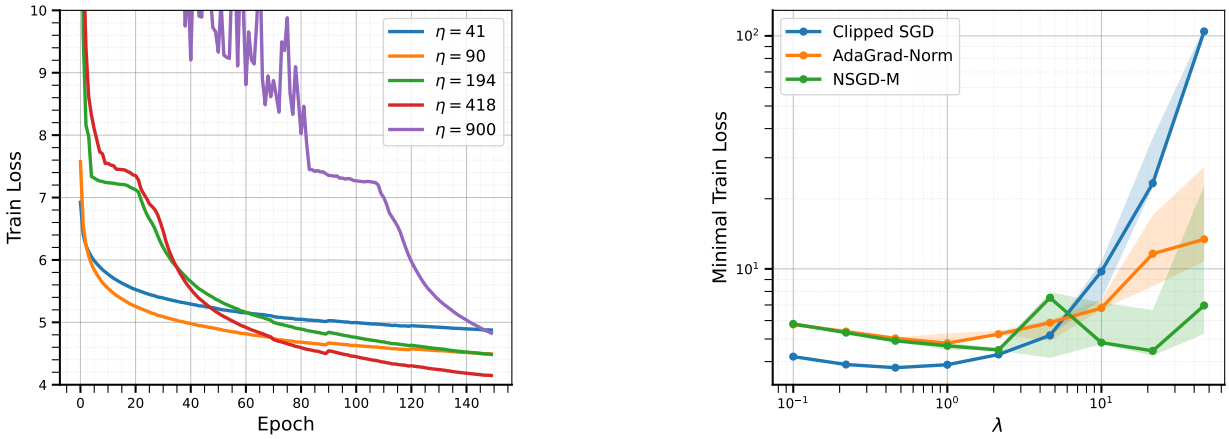
Secondly, (14) corresponds to

$$\forall\, A \in \mathcal{A}\, \exists\, \kappa > 0\, \forall\, \varepsilon_0', \theta_0' > 0\, \exists\, \varepsilon \in (0, \varepsilon_0'], \theta \geq \theta_0'\colon h_A(\varepsilon,\theta) \geq \kappa g(\varepsilon,\theta). \tag{16}$$

Now the proof is straightforward. Suppose g satisfies (15) and let $A \in \mathcal{A}$. Choose $\kappa \coloneqq 1/K$ and let $\varepsilon_0', \theta_0' > 0$ be arbitrary. Lastly define $\varepsilon \coloneqq \min\{\varepsilon_0, \varepsilon_0'\}$ and $\theta \coloneqq \max\{\theta_0, \theta_0'\}$, where the max for θ is to be understood component-wise. Since $\varepsilon \in (0, \varepsilon_0]$ and $\theta \geq \theta_0$ we get that

$$g(\varepsilon,\theta) \leq K h_A(\varepsilon,\theta) = \frac{1}{\kappa} h_A(\varepsilon,\theta).$$

This completes the proof. □



(a) Training curves of NSGD-M.

(b) Minimal training loss.

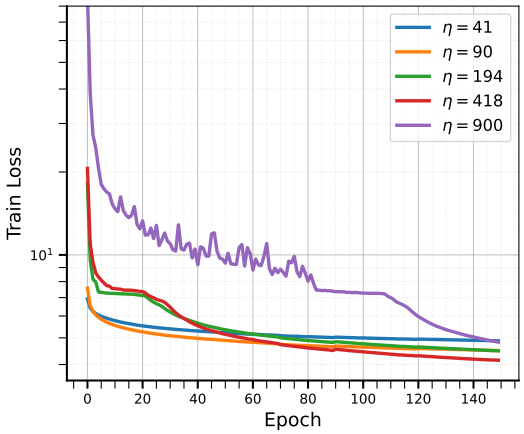
Figure 4: Results on the WikiText-2 dataset. Figure 4a represents the training curves of NSGD-M for stepsizes $\eta = 10^{k/3} \cdot \eta_{\texttt{opt}}$, where $\eta_{\texttt{opt}} = 90$ and $k \in \{-1, 0, 1, 2, 3\}$. Figure 4b shows the best train loss within 150 epochs of different algorithms with stepsizes $\lambda \cdot \eta_{\texttt{opt}}$. Shaded areas represent the minimal and maximal value within 3 seeds, the line the median.

E ADDITIONAL EXPERIMENTS

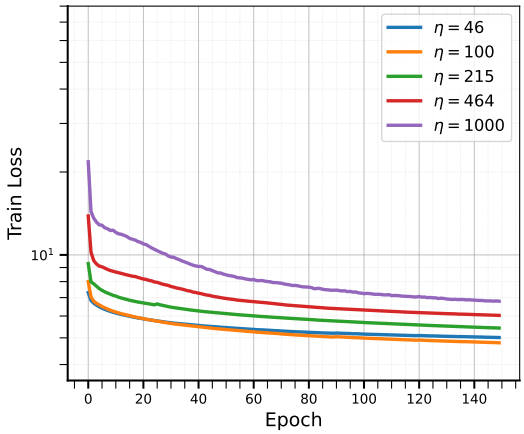
In this section, we provide additional experiments on the WikiText-2 dataset (Merity et al., 2017). Crawshaw et al. (2022) empirically motivated, that this task also requires the weaker notion of (L_0, L_1)-smoothness.

Experimental Setup. We conduct training on the WikiText-2 dataset (Merity et al., 2017) using the AWD-LSTM architecture (Merity et al., 2018). As in the previous experiment, we compare NSGD-M to AdaGrad-Norm (Faw et al., 2023) and Clipped SGD (Zhang et al., 2020b). For each algorithm we first chose the optimal stepsize $\eta_{\texttt{opt}}$ based on a course grid search in a 20 epoch training. The clipping threshold for Clipped SGD was fixed to be 0.25 in concordance to previous work (Zhang et al., 2020b), the decay-rates of NSGD-M were chosen according to Theorem 2 and b_0 of AdaGrad-Norm was set to be $b_0 = 10^{-6}$. For each algorithm, the final training was then carried out with stepsizes $\eta = \lambda \cdot \eta_{\texttt{opt}}$, where $\lambda = 10^{k/3}, k \in \{-3, -2, \dots, 6\}$, for 150 epochs. This procedure is replicated with three different seeds to get more reliable results. The code is based on the experiments by Zhang et al. (2020a).

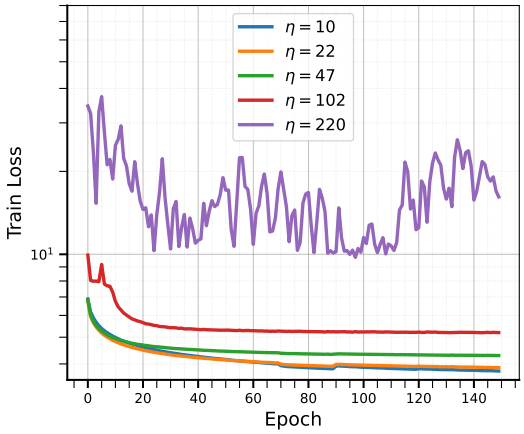
Discussion. In Figure 4a we can again notice the same threshold behaviours for NSGD-M as experienced on the PTB dataset. Instead of a plateau we do however observe higher trainings losses before the fast decrease. Training curves of Clipped SGD and AdaGrad-Norm can be found in Figure 5. Figure 4b showcases the robustness of NSGD-M to hyperparmeter-tuning to an greater extend than Figure 2. We can see that NSGD-M outperforms AdaGrad-Norm for nearly all stepsizes, with the gap increasing as stepsizes increase relative to the optimal stepsize. While Clipped SGD outperforms the adaptive methods when using the optimally-tuned stepsize or less, it suffers from an order of magnitude higher training loss as stepsizes increase relative to the optimally tuned stepsize. When compared to Figure 2, a large improvement in performance can be noticed for NSGD-M. We offer the following explanation: While, in both cases, we trained for 150 epochs, the training on the smaller PTB dataset consisted of roughly 680 batches per epoch. On the larger WikiText-2 dataset, epochs consisted of roughly 1500 batches, increasing the total number of iterations from roughly 100000 to roughly 230000. When assuming similar values of L_0, L_1, NSGD-M hence more likely reached the threshold needed, entering the fast convergence phase, while AdaGrad-Norm behaves more steadily, as can be seen in Figure 5b.



(a) NSGD-M with $\eta_{\texttt{opt}} = 90$.



(b) AdaGrad-Norm with $\eta_{\texttt{opt}} = 100$.



(c) Clipped SGD with $\eta_{\texttt{opt}} = 22$.

Figure 5: Logarithmic training curves of NSGD-M, AdaGrad-Norm and Clipped SGD on WikiText-2 for stepsizes $\eta = 10^{k/3} \cdot \eta_{\texttt{opt}}$ with $k \in \{-1, 0, 1, 2, 3\}$.