OPTIMIZING (L_0, L_1)-SMOOTH FUNCTIONS BY GRADIENT METHODS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study gradient methods for solving an optimization problem with an (L_0, L_1)-smooth objective function. This problem class generalizes that of Lipschitz-smooth problems and has gained interest recently, as it captures a broader range of machine learning applications. We provide novel insights on the properties of this function class and develop a general framework for analyzing optimization methods for (L_0, L_1)-smooth function in a principled manner. While our convergence rate estimates recover existing results for minimizing the gradient norm for nonconvex problems, our approach allows us to significantly improve the current state-of-the-art complexity results in the case of convex problems. We show that both the gradient method with Polyak stepsizes and the normalized gradient method, without any knowledge of the parameters L_0 and L_1, achieve the same complexity bounds as the method with the knowledge of these constants. In addition to that, we show that a carefully chosen accelerated gradient method can be applied to (L_0, L_1)-smooth functions, further improving previously known results. In all cases, the efficiency bounds we establish do not have an exponential dependency on L_0 or L_1, and do not depend on the initial gradient norm.

1 INTRODUCTION

In this paper, we focus on the deterministic unconstrained optimization problem

$$f^* := \min_{x \in \mathbb{R}^d} f(x), \tag{1}$$

where $f\colon \mathbb{R}^d \to \mathbb{R}$ is a smooth function. With the rise of deep learning, ensuring efficient convergence has become increasingly critical. Traditional optimization methods, such as the gradient descent method and its variants, often rely on assumptions like Lipschitz-smoothness to guarantee convergence rates. However, in modern machine learning problems, these assumptions might be too restrictive, especially when optimizing deep neural network models.

In experiments provided in (Zhang et al., 2019), it was shown that the norm of the Hessian correlates with a norm of the gradient of a loss when training neural networks. This observation motivated the authors to introduce a new, more realistic assumption on a function class named (L_0, L_1)-smooth. The class of (L_0, L_1)-smooth functions includes the class of Lipschitz-smooth functions. Also, they provided the convergence rate of the gradient method (GM) with fixed, normalized, and clipped stepsizes for non-convex optimization and showed that normalized and clipped methods are favorable in the new setting. In recent years, many works have studied methods for solving (L_0, L_1)-smooth optimization problems. Despite this interest from the community, the existing convergence results are suboptimal in some important cases, and the analysis of such methods is not satisfactory.

Motivated by this gap, the present work investigates properties of (L_0, L_1)-smooth functions and gradient methods for optimizing these functions. The following subsection discusses existing results for (L_0, L_1)-smooth optimization.

Contributions. Our contributions can be summarized as follows:

- In Section 2, we provide novel results and insights into the (L_0, L_1)-class by (i) providing new examples and operations that preserve the (L_0, L_1)-smoothness of functions and (ii) deriving new properties for this function class, leading to tighter bounds on descent inequalities. In Section 3, we propose new, intuitive step sizes that follow by directly minimizing our tighter upper bounds on the function growth. We also discuss the relation between these stepsizes and those used in the normalized and clipped gradient method.

- For nonconvex functions, we achieve the best-known $\mathcal{O}\left(\frac{L_0 F_0}{\epsilon^2} + \frac{L_1 F_0}{\epsilon}\right)$ (Theorem 3.1) complexity bound for finding an ϵ-stationary point, where $F_0 = f(x_0) - f^*$ is a function residual at the initial point. For convex problems, we significantly improve existing results by achieving the $\mathcal{O}(\frac{L_0 R^2}{\epsilon} + L_1 R \ln \frac{F_0}{\epsilon})$ (Theorem 3.2) complexity bound, where $R = \|x_0 - x^*\|$ is an initial distance to a solution.
- We also study two new methods: normalized gradient method (NGM) and gradient method with Polyak stepsizes (PS-GM), which do not require the knowledge of the parameters L_0, L_1. For both methods, we show that they enjoy the same $\mathcal{O}(\frac{L_0 R^2}{\epsilon} + L_1^2 R^2)$ complexity bound without knowing L_0 and L_1 (see Theorems 4.1 and 5.1).
- Finally, in Section 6, we present a simple procedure achieving the accelerated complexity of $\mathcal{O}(m\sqrt{\frac{L_0 R^2}{\epsilon}} + L_1^2 R^2)$, where $m > 0$ is a number of oracle queries for solving a simple one-dimensional problem. This procedure prescribes running the monotone version of the accelerated gradient method Nesterov et al. (2021) from the initial point constructed after a certain number of iterations of the (nonaccelerated) GM.

In contrast to other results in the literature, all our complexity bounds neither depend on the initial gradient norm, nor have an exponential dependency on L_1.

Related work. Following the introduction of the (L_0, L_1)-class by Zhang et al. (2019), subsequent works have explored other smoothness generalizations and analyzed gradient methods under these new assumptions. Chen et al. (2023) introduced the α-asymmetric class, relaxing the assumption on twice differentiability and allowing a sublinear growth on the norm of a gradient. In (Li et al., 2023), authors went further and proposed the weakest (r, l)-smooth class, which allows even quadratic growth of the norm of the Hessian with respect to the norm of the gradient. Despite the generality of this assumption, there are still some issues and open questions regarding the existing results even for the basic (L_0, L_1)-smooth class.

In (Zhang et al., 2020), the authors analyzed the clipped GM with momentum and improved complexity bound with respect to problem parameters L_0, L_1. Using the right choice of clipping parameters, Koloskova et al. (2023) proved, for nonconvex and convex problems, respectively, the $\mathcal{O}(\frac{L_0 F_0}{\epsilon^2} + \frac{L_1 F_0}{\epsilon})$ and $\mathcal{O}(\frac{L_0 R}{\epsilon} + \sqrt{\frac{L}{\epsilon}} L_1 R^2)$ oracle complexity bounds for obtaining an ϵ-approximate solution, where L is a Lipschitz constant. For convex problems, Li et al. (2023) proposed an (asymptotically) faster accelerated gradient method achieving the complexity of $\mathcal{O}((L_1^2 R^2 + \frac{L_1^2 F_0}{L_0} + 1)\sqrt{\frac{F_0 + L_0 R^2}{\epsilon}})$[1]. Several works have studied adaptive optimization methods that do not require the (L_0, L_1) parameters to be known. Faw et al. (2023); Wang et al. (2023) studied convergence rates for AdaGrad for stochastic nonconvex problems. Hübler et al. (2024) proposed a gradient method with the backtracking line search and showed the $\mathcal{O}(\frac{L_0 F_0}{\epsilon^2} + \frac{L_1^2 F_0^2}{\epsilon^2})$ complexity bound for nonconvex problems. For convex problems, Takezawa et al. (2024) obtained the complexity of $\mathcal{O}(\frac{L_0 R}{\epsilon} + \sqrt{\frac{L}{\epsilon}} L_1 R^2)$, where L is a Lipschitz constant, for the PS-GM, which requires knowing the optimal function value.

One interesting paper that is highly related to our work and independently appeared online during the finalization of our manuscript is Gorbunov et al. (2024). In this paper, the authors propose a new formula (called "smooth clipping") for choosing stepsizes in the GM for convex (L_0, L_1)-smooth functions; up to absolute constants, this formula coincides with one of ours. Their proof techniques differ from ours, which leads to $\mathcal{O}(\frac{L_0 R^2}{\epsilon} + L_1^2 R^2)$ complexity which is slightly worse than our $\mathcal{O}(\frac{L_0 R^2}{\epsilon} + L_1 R \ln(\frac{F_0}{\epsilon}))$, especially when initial function value is reasonably bounded (see Section 3). The authors also show that the PS-GM achieves the same result as in our work. Additionally, they provide an accelerated method with complexity $\mathcal{O}(\sqrt{\frac{L_0 R^2}{\epsilon} L_1 R \exp(L_1 R)})$ and study

[1] See Appendix F.

also the strongly convex and stochastic cases as well as some adaptive methods. In contrast, our work has a slightly different scope and offers deeper insights by providing the intuition on arriving at the "right" stepsize formulas, an analysis for nonconvex functions, the study of normalized gradient methods, and a superior version of the accelerated scheme with significantly better complexity. Importantly, our proof techniques are somewhat different from those in Gorbunov et al. (2024) and are more aligned with classical optimization theory, at least in the specific cases we consider in our paper.

2 Definition and Properties of (L_0, L_1)-Smooth Functions

In this section, we state our assumptions and discuss important properties of generalized smooth functions. We start with defining our main assumption on (L_0, L_1)-smooth functions.

Throughout this paper, unless specified otherwise, we use the standard inner product $\langle \cdot, \cdot \rangle$ and the standard Euclidean norm $\| \cdot \|$ for vectors, and the standard spectral norm $\| \cdot \|$ for matrices. We also assume that there exists a solution for the problem (1).

Definition 2.1. A twice continuously differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is called (L_0, L_1)-*smooth* (for some $L_0, L_1 \geq 0$) if it holds that

$$\|\nabla^2 f(x)\| \leq L_0 + L_1 \|\nabla f(x)\| \qquad \text{for all } x \in \mathbb{R}^d. \tag{2}$$

The class of (L_0, L_1)-smooth function is a wide class of functions, which includes the class of Lipschitz-smooth functions. Definition 2.1 of generalized smooth function was introduced in (Zhang et al., 2019). For twice differentiable function, this definition is equivalent to definition of α-symmetric functions with $\alpha = 1$ provided in (Chen et al., 2023). All our further results also hold for α-symmetric functions, however, we use the stricter assumption for clarity in presentation. Any α-symmetric twice differentiable function is also (L_0, L_1)-smooth function, but with a different choice of parameters; thus, our results also hold for α-symmetric functions. For the purpose of analysis of the methods, we provide an alternative and more useful first-order characterization of the class of (L_0, L_1)-smooth functions.

Lemma 2.2. *Let f be a twice continuously differentiable function, Then, f is (L_0, L_1)-smooth if and only if any of the following inequalities holds for any $x, y \in \mathbb{R}^d$:*[2]

$$\|\nabla f(y) - \nabla f(x)\| \leq (L_0 + L_1 \|\nabla f(x)\|) \frac{e^{L_1 \|y - x\|} - 1}{L_1}, \tag{3}$$

$$|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| \leq (L_0 + L_1 \|\nabla f(x)\|) \frac{\phi(L_1 \|y - x\|)}{L_1^2}, \tag{4}$$

where $\phi(t) := e^t - t - 1$ ($t \geq 0$).

The proof of Lemma 2.2 can be found in Appendix A. It is worth noting that inequality (3) is stronger than that from Corollary A.4 (Zhang et al., 2020). The bound in inequality (4) is tighter than those presented in previous works (see, for example, Lemma A.3 in (Zhang et al., 2020), Lemma 8 in (Hübler et al., 2024)). These tighter estimates allow us to construct gradient methods in the sequel. When the function f is also convex, we have the following useful inequalities.

Lemma 2.3. *Let f be a convex (L_0, L_1)-smooth nonlinear*[3] *function. Then, for any $x, y \in \mathbb{R}^d$,*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L_0 + L_1 \|\nabla f(y)\|}{L_1^2} \phi_* \Big(\frac{L_1 \|\nabla f(y) - \nabla f(x)\|}{L_0 + L_1 \|\nabla f(y)\|} \Big), \tag{5}$$

$$\begin{aligned} \langle \nabla f(x) - \nabla f(y), x - y \rangle \geq &\frac{L_0 + L_1 \|\nabla f(y)\|}{L_1^2} \phi_* \Big(\frac{L_1 \|\nabla f(y) - \nabla f(x)\|}{L_0 + L_1 \|\nabla f(y)\|} \Big) \\ &+ \frac{L_0 + L_1 \|\nabla f(x)\|}{L_1^2} \phi_* \Big(\frac{L_1 \|\nabla f(y) - \nabla f(x)\|}{L_0 + L_1 \|\nabla f(x)\|} \Big), \end{aligned} \tag{6}$$

where ϕ_* is the function from Lemma A.4.

[2] Hereinafter, for $L_1 = 0$ and any $t \geq 0$, we assume that $\frac{e^{L_1 t} - 1}{L_1} \equiv t$, $\frac{\phi(L_1 t)}{L_1^2} \equiv \frac{1}{2} t^2$, etc., which are the limits of these expressions when $L_1 \to 0; L_1 > 0$.

[3] According to Lemma 2.2, this means that $L_0 + L_1 \|\nabla f(x)\| > 0$ for any $x \in \mathbb{R}^d$.

Moreover, using Lemma A.4, we can simplify the lower bound in (5).

Corollary 2.4. *Let f be a convex (L_0, L_1)-smooth nonlinear function. Then, for any $x, y \in \mathbb{R}^d$,*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\|\nabla f(y) - \nabla f(x)\|^2}{2(L_0 + L_1\|\nabla f(y)\|) + L_1\|\nabla f(y) - \nabla f(x)\|}. \tag{7}$$

Lemma 2.3 is a generalization of Theorem 2.1.5(Nesterov, 2018) to (L_0, L_1)-smooth functions, and matches it when $L_1 = 0$ (since $\frac{1}{L_1^2}\phi_*(L_1\alpha) \to \frac{1}{2}\alpha^2$ as $L_1 \to 0$). The proof of Lemma 2.3 is presented in Appendix A.

3 Gradient Method

Having established a few important properties of an (L_0, L_1)-smooth function f, we now turn our attention to the *gradient method* (GM) for minimizing such a function:

$$x_{k+1} = x_k - \eta_k \nabla f(x_k), \qquad k \geq 0, \tag{8}$$

where $x_0 \in \mathbb{R}^d$ is a starting point and $\eta_k \geq 0$ are certain stepsizes.

We start with showing that the gradient update rule (8) and the "right" formula for the stepsize η_k both naturally arise from the classical idea in optimization theory—choosing the next iterate x_{k+1} by minimizing the global upper bound on the objective function value constructed around the current iterate x_k (see (Nesterov, 2018)). Indeed, let $x \in \mathbb{R}^d$ be the current point and let $a := L_0 + L_1\|\nabla f(x)\| > 0$. According to (4), for any $y \in \mathbb{R}^d$,

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{a}{L_1^2}\phi(L_1\|y - x\|).$$

Our goal is to minimize the right-hand of the above inequality in y. Since the last term in this bound depends only on the norm of $y - x$, the optimal point $y^* = T(x)$ is the result of the gradient step $T(x) = x - r^* \frac{\nabla f(x)}{\|\nabla f(x)\|}$ for some $r^* \geq 0$ ensuring the following progress in decreasing the function value:

$$f(x) - f(T(x)) \geq \max_{r \geq 0}\Big\{\|\nabla f(x)\| r - \frac{a}{L_1^2}\phi(L_1 r)\Big\} = \frac{a}{L_1^2}\phi_*\Big(\frac{L_1\|\nabla f(x)\|}{a}\Big),$$

where ϕ_* is the conjugate function to ϕ (see Lemma A.4). Furthermore, r^* is exactly the solution of the above optimization problem, satisfying $L_1\|\nabla f(x)\| = a\phi'(L_1 r^*)$. Solving this equation, using the fact that $(\phi')^{-1}(\gamma) = \phi_*'(\gamma) = \ln(1+\gamma)$, we obtain $r^* = \frac{1}{L_1}\phi_*'(\frac{L_1\|\nabla f(x)\|}{a}) = \frac{1}{L_1}\ln(1 + \frac{L_1\|\nabla f(x)\|}{a})$.

The above considerations lead us to the following *optimal* choice of stepsizes in (8):

$$\boxed{\eta_k^* = \frac{1}{L_1\|\nabla f(x_k)\|}\ln\Big(1 + \frac{L_1\|\nabla f(x_k)\|}{L_0 + L_1\|\nabla f(x_k)\|}\Big),} \qquad k \geq 0, \tag{9}$$

resulting in the following progress in decreasing the objective:

$$f(x_k) - f(x_{k+1}) \geq \frac{L_0 + L_1\|\nabla f(x_k)\|}{L_1^2}\phi_*\Big(\frac{L_1\|\nabla f(x_k)\|}{L_0 + L_1\|\nabla f(x_k)\|}\Big) := \Delta_k. \tag{10}$$

The above expression for Δ_k is quite cumbersome but, in fact, it behaves as the simple fraction $\frac{\|\nabla f(x_k)\|^2}{L_0 + L_1\|\nabla f(x_k)\|}$. More precisely, from Lemma A.4(3), we see that

$$\frac{\|\nabla f(x_k)\|^2}{2L_0 + 3L_1\|\nabla f(x_k)\|} \leq \Delta_k \leq \frac{\|\nabla f(x_k)\|^2}{2(L_0 + L_1\|\nabla f(x_k)\|)}.$$

Thus, there is not much point in keeping the complicated expression (10) and we can safely simplify it as follows:

$$f(x_k) - f(x_{k+1}) \geq \frac{\|\nabla f(x_k)\|^2}{2L_0 + 3L_1\|\nabla f(x_k)\|}. \tag{11}$$

Interestingly, we can also arrive at exactly the same bound (11) by using a simpler choice of stepsizes. Specifically, replacing $\ln(1+\gamma)$ with its lower bound $\frac{2\gamma}{2+\gamma}$ (which is responsible for the inequality in Lemma A.4(3) that we used to simplify (10) into (11)), we obtain the following *simplified stepsizes*:

$$\boxed{\eta_k^{\mathrm{si}} = \frac{1}{L_0 + \frac{3}{2}L_1\|\nabla f(x_k)\|},} \qquad k \geq 0. \tag{12}$$

With this choice, the iterates of method (8) still satisfy (11) (see Lemma B.1).

Further, note that, up to absolute constants, stepsize (12) acts as $\frac{1}{\max\{L_0, L_1\|\nabla f(x_k)\|\}} = \min\{\frac{1}{L_0}, \frac{1}{L_1\|\nabla f(x_k)\|}\}$, which is the so-called clipping stepsize used in many previous works Zhang et al. (2019; 2020); Koloskova et al. (2023). Thus, with the right choice of absolute constants, we can expect the corresponding clipping stepsizes, to satisfy a similar inequality to (11). This is indeed the case, and we can show, in particular, that the *clipping stepsizes*

$$\boxed{\eta_k^{\mathrm{cl}} = \min\Big\{\frac{1}{2L_0}, \frac{1}{3L_1\|\nabla f(x_k)\|}\Big\},} \qquad k \geq 0, \tag{13}$$

do satisfy (11) although with slightly worse absolute constants (see Lemma B.1).

We have thus demonstrated in this section that clipping stepsizes (13) are simply a convenient approximation of the optimal stepsizes (9), ensuring a similar bound on the objective progress. This observation seems to be a new insight into clipping stepsizes which has not been previously explored in the literature.

It is not difficult to see that the three stepsizes we introduced in this section satisfy the following relationships:

$$\eta_k^{\mathrm{cl}} \leq \eta_k^{\mathrm{si}} \leq \eta_k^{*}. \tag{14}$$

3.1 Nonconvex Functions

We are now ready to present a convergence rate result for nonconvex functions.

Theorem 3.1. *Let f be an (L_0, L_1)-smooth function, and let $\{x_k\}$ be iterate sequence of GM* (8) *with one of the stepsize choices given by* (9), (12) *or* (13). *Then, $\min_{0\leq k\leq K}\|\nabla f(x_k)\| \leq \epsilon$ for any given $\epsilon > 0$ whenever*

$$K + 1 \geq \frac{2L_0F_0}{a\epsilon^2} + \frac{3L_1F_0}{a\epsilon},$$

where $a = 1$ for stepsizes (9) *and* (12)*, and $a = \frac{1}{2}$ for stepsize* (13)*.*

The proof of Theorem 3.1 can be found in Appendix B.2. The rate in Theorem 3.1 matches, up to absolute constants, the rate in (Koloskova et al., 2023) for clipped GM with $\eta = \frac{1}{9}(L_0 + cL_1)$ for $c = \frac{L_0}{L_1}$, or equivalently the GM with stepsize $\eta_k = \frac{1}{18L_0}\min\{1, \frac{L_0}{L_1\|\nabla f(x_k)\|}\}$. Furthermore, our rate is significantly better than the rate $\mathcal{O}(\frac{L_0F_0}{\epsilon^2} + \frac{L_1^2F_0}{L_0})$ obtained in (Zhang et al., 2019) for the clipped GM since $\frac{L_1F_0}{\epsilon} \leq \frac{L_0^2F_0}{2\epsilon} + \frac{L_1^2F_0}{2L_0}$, and the latter expression can be arbitrarily far away from the former whenever L_0 is sufficiently small and L_1 is distinct from zero. In addition to that, our convergence rate result does not depend on the gradient norm at the initial point, in contrast to Li et al. (2023) who consider a wider class of generalized-smooth functions but whose rate (polynomially) depends on $\|\nabla f(x_0)\|$. Also, our rate from Theorem 3.1 is better than $\mathcal{O}\big(\frac{L_0F_0}{\epsilon^2} + \frac{L_1^2F_0^2}{\epsilon^2}\big)$ provided in (Hübler et al., 2024) for the GM equipped with a certain backtracking line search.

3.2 Convex Functions

Let us now provide the convergence rate for convex functions.

Theorem 3.2. *Let $\{x_k\}$ be the iterates of GM* (8) *with one of the stepsize choices given in* (9) (12) *or* (13)*, as applied to problem* (1) *with an (L_0, L_1)-smooth convex function f. Let x^* be an arbitrary* *solution to the problem and let* $F_0 := f(x_0) - f^*$*. Then, the sequence* $R_k := \|x_k - x^*\|$, $k \geq 0$, *is nonincreasing, and* $f(x_K) - f^* \leq \epsilon$ *for any given* $0 \leq \epsilon < F_0$ *whenever*

$$K \geq \frac{2}{a}\frac{L_0 R^2}{\epsilon} + \frac{3}{a} L_1 R \ln \frac{F_0}{\epsilon} \qquad \Big(\leq \frac{2 + \frac{3}{e}}{a}\frac{L_0 R^2}{\epsilon} + \frac{3(1+\frac{1}{e})}{a} L_1^2 R^2\Big),$$

where $R := R_0$, *and* $a = 1$ *for stepsizes* (9), (12) *and* $a = \frac{1}{2}$ *for stepsize* (13).

The proof of Theorem 3.2 can be found in Appendix B.3. Notice, that the second estimate $\mathcal{O}(\frac{L_0 R^2}{\epsilon} + L_1^2 R^2)$ in Theorem 3.2 comes from a very pessimistic bound on F_0 with the exponentially large quantity $\exp(L_1 R)\frac{L_0 R^2}{2\epsilon}$ coming from Lemmas 2.2 and A.4. However, in the case when F_0 is reasonably bounded (e.g., we apply "hot-start" or f is a well-behaved function such as the logistic one), the $\mathcal{O}(L_1 R \ln \frac{F_0}{\epsilon})$ term from the main estimate can be much smaller than $\mathcal{O}(L_1^2 R^2)$ from the pessimistic estimate.

In Theorem 3.2, we do not make an assumption on L-smoothness of a function, while this assumption is used in (Koloskova et al., 2023). Moreover, the rate in the theorem is better than $\mathcal{O}(\frac{L_0 R^2}{\epsilon} + \sqrt{\frac{L}{\epsilon}} L_1 R^2)$ rate provided in (Koloskova et al., 2023) for clipped GM, since it does not have $\frac{1}{\sqrt{\epsilon}}$ dependency on L_1 and L. Also, our result does not include the norm $\|\nabla f(x_0)\|$ of the gradient at an initialization point in the estimate, while the rate provided in (Li et al., 2023) does depend on $\|\nabla f(x_0)\|$ which can be large and be an order of L. Consider for example $f(x) = \frac{1}{p}\|x\|^p$ (see Proposition A.1) for $p > 2$ and starting point x_0 sufficiently far from the origin, in this case, the gradient $\|\nabla f(x_0)\| = \|x_0\|^{p-1}$ can be arbitrary large.

4 Normalized Gradient Method

To run GM from Section 3, it is necessary to know the parameters (L_0, L_1) in advance. In many real-life examples, those parameters are unknown, and it might be computationally expensive to estimate them. Furthermore, for any given function f, the pair (L_0, L_1) is generally not unique (see Examples A.1 and A.2), and it is not clear in advance which pair would result in the best possible convergence rate of our optimization method. To address this issue, in this section, we present another version of the gradient method that does not require knowing (L_0, L_1). This is the *normalized gradient method* (NGM):

$$x_{k+1} = x_k - \frac{\beta_k}{\|\nabla f(x_k)\|}\nabla f(x_k), \qquad k \geq 0, \tag{15}$$

where $x_0 \in \mathbb{R}^d$ is a certain starting point, and β_k are positive coefficients. The following result describes the efficiency of the NGM.

Theorem 4.1. *Let* $\{x_k\}$ *be the iterates of NGM* (15)*, as applied to problem* (1) *with an* (L_0, L_1)*-smooth convex function* f*. Consider the constant coefficients* $\beta_k = \frac{\hat{R}}{\sqrt{K+1}}$, $0 \leq k \leq K-1$, *where* $\hat{R} > 0$ *is a parameter and* $K \geq 1$ *is the total number of iterations of the method (fixed in advance). Then,* $\min_{0 \leq k \leq K} f(x_k) - f^* \leq \epsilon$ *for any given* $\epsilon > 0$ *whenever*

$$K + 1 \geq \max\Big\{\frac{L_0 \bar{R}^2}{\epsilon}, \frac{4}{9} L_1^2 \bar{R}^2\Big\},$$

where $\bar{R} := \frac{1}{2}(\frac{R^2}{\hat{R}} + \hat{R})$, $R := \|x_0 - x^*\|$, *and* x^* *is an arbitrary solution of the problem.*

The parameter $\hat{R}$ in the formula for coefficients β_k is an estimation of the initial distance R to a solution, and the best complexity bound of $K^* := \mathcal{O}(\frac{L_0 R^2}{\epsilon} + L_1^2 R^2)$ is achieved whenever $\hat{R} = R$. Note that, even if $\hat{R} \neq R$, the method still converges but with a slightly worse total complexity of $K^* \rho^2$, where $\rho = \max\{\frac{R}{\hat{R}}, \frac{\hat{R}}{R}\}$.

The proof of Theorem 4.1 is based on the following two important facts (Nesterov, 2018, Section 3). First, under the proper choice of coefficients β_k, NGM ensures that the minimal value v_K^* among $v_k := \frac{\langle \nabla f(x_k), x_k - x^* \rangle}{\|\nabla f(x_k)\|}$, $0 \leq k \leq K$, converges to zero at the rate of $\frac{\bar{R}}{\sqrt{K}}$. These quantities v_k have a geometrical meaning—each of them is exactly the distance from the point x^* to the supporting hyperplane to the sublevel set of f at the point x_k. Second, whenever v_K^* converges to zero, so does $\min_{0\le k\le K} f(x_k) - f^*$. Moreover, we can relate the two quantities whenever we can bound, for any given $v \ge 0$, the function residual $f(x) - f^*$ over the ball $\|x - x^*\| \le v$:

Lemma 4.2 ((Nesterov, 2018, Lemma 3.2.1)). *Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be a differentiable convex function. Then, for any $x, y \in \mathbb{R}^d$ and*[4] $v_f(x;y) := \frac{[\langle \nabla f(x), x-y\rangle]_+}{\|\nabla f(x)\|}$, *it holds that*

$$f(x) - f(y) \le \max_{z\in\mathbb{R}^d}\{f(z) - f(y) : \|z - y\| \le v_f(x;y)\}. \tag{16}$$

In our case—when the function f is (L_0, L_1)-smooth—the corresponding bound can be obtained from Lemma 2.2. The complete proof of Theorem 4.1 can be found in Appendix C.

In Theorem 4.1, we fix the number of iterations K before running the method, which is a standard approach for the (normalized)-(sub)gradient methods (Section 3.2 in Nesterov (2018)). However, doing so may be undesirable in practice since it becomes difficult to continue running the method if the time budget was suddenly increased and also prevents the method from using larger stepsizes at the initial iterations. To overcome these drawbacks, one can use time-varying coefficients by setting $\beta_k = \frac{\hat{R}}{\sqrt{k+1}}$, $0 \le k \le K-1$. This results in the same worst-case theoretical complexity as in Theorem 4.1 but with an extra logarithmic factor (see Theorem C.2). Moreover, one can completely eliminate this extra logarithmic factor by switching to an appropriate modification of the standard (sub)gradient method such as Dual Averaging Nesterov (2005).

For $\hat{R} = R$, the complexity of NGM is $\mathcal{O}(\frac{L_0 R^2}{\epsilon} + L_1^2 R^2)$ which is generally worse than that of the previously considered GM (see Theorem 3.2 and the corresponding discussion). However, recall that GM requires knowing (L_0, L_1), and its rate depends on the particular choice of these constants. In contrast, NGM does not require the knowledge of these parameters, and its "real" complexity is

$$\mathcal{O}(1) \min_{L_0, L_1}\Big\{\frac{L_0^2 \bar{R}^2}{\epsilon} + L_1^2 \bar{R}^2 \colon f \text{ is } (L_0, L_1)\text{-smooth}\Big\},$$

where $\mathcal{O}(1)$ is an absolute constant.

5 Gradient Method with Polyak Stepsizes

In the previous sections, the parameters required to run the methods were L_0, L_1 for GM, and the estimation $\hat{R}$ of the initial distance for a solution R for NGM. To achieve good complexity for NGM, the estimate $\hat{R}$ should be close to real R, otherwise the multiplicative factor $\{\frac{R}{\hat{R}}, \frac{\hat{R}}{R}\}$ will lead to an arbitrary large complexity estimate. Sometimes, parameters L_0, L_1 and a good estimate $\hat{R}$ are unknown, while the optimal value of the function is available. One of the examples of such problems is over-parametrized models in machine learning where f^* is usually 0.

In this section, we focus on the case when f^* is known and analyze the gradient method (8) with the *Polyak stepsizes*:

$$\eta_k = \frac{f(x_k) - f^*}{\|\nabla f(x_k)\|^2}, \qquad k \ge 0. \tag{17}$$

Theorem 5.1. *Let $\{x_k\}$ be the iterates of PS-GM* (8), (17), *as applied to problem* (1) *with an (L_0, L_1)-smooth convex function f. Then, it holds that $\min_{0\le k\le K} f(x_k) - f^* \le \epsilon$ for any given $\epsilon > 0$ whenever*

$$K + 1 \ge \max\Big\{\frac{4L_0 R^2}{\epsilon}, 36 L_1^2 R^2\Big\},$$

where $R := \|x_0 - x^*\|$ and x^* is an arbitrary solution of the problem.

We prove the theorem by using a standard inequality for the gradient method with Polyak stepsizes (PS-GM) for convex functions,

$$R_k^2 - R_{k+1}^2 \ge \frac{f_k^2}{g_k^2},$$

[4] Here $[t]_+ := \max\{t, 0\}$ is the nonnegative part of $t \in \mathbb{R}$.

where $R_k = \|x_k - x^*\|$, $f_k = f(x_k) - f^*$, and $g_k = \|\nabla f(x_k)\|$. We then leverage the lower bound (7), and bound gradient norm g_k by $\psi^{-1}(f_k)$, where $\psi(g) := \frac{g^2}{2L_0+3L_1 g}$, obtaining

$$R_k^2 - R_{k+1}^2 \geq \frac{f_k^2}{[\psi^{-1}(f_k)]^2}.$$

Summing up these relations, passing to the minimal value of f_k, and rearranging the resulting inequality, we obtain the desired bound. The complete proof of Theorem 5.1 can be found Appendix D.1.

Notice that the rate $\mathcal{O}(\frac{L_0 R^2}{\epsilon} + L_1^2 R^2)$ in Theorem 5.1 is the same as the rate NGM in Theorem 4.1. Further, our rate is better than $\mathcal{O}(\frac{L_0 R^2}{\epsilon} + \sqrt{\frac{L}{\epsilon}} L_1 R^2)$ provided in Takezawa et al. (2024), since it does not have a dependency on ϵ with respect to L_1 and L parameters. Furthermore, it is more general since we do not assume L-smoothness of the function. Finally, it is worth mentioning that the rate for PS-GM holds for any choice of (L_0, L_1). Thus, the rate holds for the best choice of (L_0, L_1) pair and is the same as for the NGM.

6 Accelerated Gradient Method

This section focuses on developing an accelerated method for minimizing an (L_0, L_1)-smooth function f.

We start with the following observation. Consider a point $x \in \mathbb{R}^d$ with $\|\nabla f(x)\| \leq \frac{L_0}{L_1}$. Then, by the definition of (L_0, L_1)-smoothness, $\|\nabla^2 f(x)\| \leq L_0 + L_1 \|\nabla f(x)\| \leq 2L_0$. Hence, in the region

$$Q := \left\{ x \in \mathbb{R}^d \colon \|\nabla f(x)\| \leq \frac{L_0}{L_1} \right\},$$

the function f behaves like a standard $2L_0$-smooth function. Consequently, we may try to apply the Fast Gradient Method (FGM) to minimize f expecting (after we have found an initial point in Q) the $\mathcal{O}(\sqrt{\frac{L_0 R^2}{\epsilon}})$ oracle complexity for finding an ϵ-approximate solution, where R is the initial distance to the solution. The problem with the above approach is ensuring that the iterates of FGM stay in Q. In general, the set Q may have a complicated structure and might even be nonconvex. One reasonable idea is to find another, better-structured region, contained in Q in which we can keep all the iterates of the method. A good candidate for such a region is the initial feasible set

$$\mathcal{F}_0 := \{x \in \mathbb{R}^d \colon f(x) \leq f(x_0)\}. \tag{18}$$

According to Lemma 2.3, for any $x \in \mathcal{F}_0$, we can upper bound the corresponding gradient norm by the function residual which is, in turn, bounded by the initial function residual:

$$\psi(\|\nabla f(x)\|) \leq f(x) - f^* \leq f(x_0) - f^* =: F_0,$$

where $\psi(\gamma) := \frac{\gamma^2}{2L_0 + 3L_1\gamma}$. Since ψ is increasing, to ensure that $\|\nabla f(x)\| \leq \frac{L_0}{L_1}$, it suffices to require that the initial function residual is sufficiently small:

$$F_0 \leq \psi\Big(\frac{L_0}{L_1}\Big) \equiv \frac{L_0}{5L_1^2} =: \Delta. \tag{19}$$

Thus, whenever $F_0 \leq \Delta$, we have the inclusion $\mathcal{F}_0 \subseteq Q$, meaning that the function f is $2L_0$-smooth over $\mathcal{F}_0$. Note that we can find an initial point x_0 satisfying (19) by using any of our basic (nonaccelerated) methods considered previously. For instance, running GM from Section 3 from a certain initial point x_s, we can ensure (19) in $\mathcal{O}(\frac{L_0 R^2}{\Delta} + L_1^2 R^2) = \mathcal{O}(L_1^2 R^2)$ gradient-oracle queries, where $R = \|x_s - x^*\|$ (see Theorem 3.2); furthermore, the obtained point x_0 does not go far from x^* compared to our initial point x_s, specifically, it holds that $\|x_0 - x^*\| \leq R$ (again, see Theorem 3.2). Thus, if we could guarantee that FGM, when started from x_0, keeps its iterates in the initial sublevel set $\mathcal{F}_0$, we would obtain the total complexity of $\mathcal{O}(\sqrt{\frac{L_0 R^2}{\epsilon}} + L_1^2 R^2)$, which is better than that of the basic methods (at least in the case when ϵ is not too large).

However, for most classical versions of FGM (such as those presented in Nesterov (2018)), the

Algorithm 1 AGMsDR$(x_0, T(\cdot), L, K)$ Nesterov et al. (2021)

1: **Input:** Initial point $x_0 \in \mathbb{R}^d$, update rule $T(\cdot)$, constant $L > 0$, number of iterations $K \geq 1$.
2: $v_0 = x_0$, $A_0 = 0$, $\zeta_0(x) = \frac{1}{2}\|x - x_0\|^2$.
3: **for** $k = 0, 1, \dots, K-1$ **do**
4: $\quad y_k = \arg\min_y \{f(y) \colon y = v_k + \beta(x_k - v_k), \beta \in [0, 1]\}$.
5: $\quad x_{k+1} = T(y_k)$.
6: $\quad$ Find $a_{k+1} > 0$ from the equation $La_{k+1}^2 = A_k + a_{k+1}$. Set $A_{k+1} = A_k + a_{k+1}$.
7: $\quad v_{k+1} = \arg\min_{x\in\mathbb{R}^d}\{\zeta_{k+1}(x) := \zeta_k(x) + a_{k+1}[f(y_k) + \langle\nabla f(y_k), x - y_k\rangle]\}$.
return x_K.

Algorithm 2 Two-Stage Acceleration Procedure

1: **Input:** Initial point $x_s \in \mathbb{R}^d$, constants $\Delta, L > 0$, update rule $T(\cdot)$, number of iterations K.
2: Run GM with stepsize rule (9), (12) or (13) from x_s to get $x_0 \colon f(x_0) - f^* \leq \Delta$.
3: **return** $x_K = \text{AGMsDR}(x_0, T(\cdot), L, K)$.

monotonic decrease of the function value cannot be guaranteed. Nevertheless, one monotone version of FGM does exist, namely, the Accelerated Gradient Method with Small-Dimensional Relaxation (AGMsDR) (Nesterov et al., 2021). We present this method in Algorithm 1, in a slightly more general form compared to the original work. Specifically, instead of computing the point x_{k+1} by the standard gradient step from a point y_k, we allow to use any update rule $T(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$, as long as it ensures a sufficient decrease in the function value:

$$f(x) - f(T(x)) \geq \frac{1}{2L}\|\nabla f(x)\|^2, \qquad \forall x \in \mathcal{F}_0. \tag{20}$$

where $L > 0$ is a certain fixed constant. As long as there exists such a $T(\cdot)$ for our function class, we can prove that the points x_k and y_k constructed by the method remain in $\mathcal{F}_0$, and $f(x_k)$ converges at the $\mathcal{O}(\frac{LR^2}{k^2})$ rate (see Appendix E.1 for the proof):

Theorem 6.1. *Consider problem* (1) *with a differentiable convex objective f. Let AGMsDR (Algorithm 1) be applied to solving this problem, given an update rule $T(\cdot)$ satisfying, for a certain fixed constant $L > 0$, the sufficient decrease property* (20) *over the initial sublevel set* (18)*. Then, for all $k \geq 0$, it holds that*

$$f(x_{k+1}) \leq f(y_k) \leq f(x_k), \qquad f(x_k) - f^* \leq \frac{2LR^2}{k^2},$$

where $R := \|x_0 - x^*\|$ and x^* is an arbitrary solution of the problem.

For our class of (L_0, L_1)-smooth functions, we can choose $T(\cdot)$ as the gradient step $T(x) = x - \eta_x \nabla f(x)$ with any of the stepsize rules (9), (12), or (13) (with x_k replaced by x). Then, according to Lemma B.1 and the fact that, under our assumption (19), for any $x \in \mathcal{F}_0$, it holds that $\|\nabla f(x)\| \leq \frac{L_0}{L_1}$, we obtain

$$f(x) - f(T(x)) \geq \frac{a\|\nabla f(x)\|^2}{2L_0 + 3L_1\|\nabla f(x)\|} \geq \frac{a}{5L_0}\|\nabla f(x)\|^2,$$

where $a = 1$ for stepsize rules (9), (12), and $a = \frac{1}{2}$ for stepsize rule (13). Thus, such a $T(\cdot)$ indeed satisfies (20) with

$$L = \frac{5}{2a}L_0. \tag{21}$$

We are now ready to formally define our two-stage acceleration procedure, see Algorithm 2. Our main result can be summarized as follows (see Appendix E.2 for a formal proof):

Theorem 6.2. *Consider problem* (1) *with an (L_0, L_1)-smooth convex function f. Let x_K be the output of Algorithm 2 as applied to solving this problem with Δ given by* (19)*, $T(\cdot)$ being the gradient update $T(x) = x - \eta_x \nabla f(x)$ with any of the stepsize rules* (9), (12), *or* (13)*, and L given by* (21)*, where $a = 1$ for stepsize rules* (9), (12) *and $a = \frac{1}{2}$ for stepsize rule* (13)*. Then, to ensure that $f(x_K) - f^* \leq \epsilon$ for a given $0 < \epsilon < \Delta$ and an appropriately chosen K, the algorithm requires at most the following number of first-order oracle queries:*

$$(m+1)\left\lceil\sqrt{\frac{5L_0R^2}{a\epsilon}}\right\rceil + \left\lceil\frac{13 + \frac{18}{e}}{a}L_1^2R^2\right\rceil.$$

where $R := \|x_s - x^*\|$, x^* is an arbitrary solution of the problem, and $m \geq 1$ is the number of oracle queries needed to compute y_k at each iteration of AGMsDR *(Algorithm 1).*

Observe that, at every step, AGMsDR requires solving a certain one-dimensional subproblem to find y_k, which we assume can be done in at most m oracle queries. For many practical problems, this subproblem can usually be solved quite efficiently, so the extra factor m in the complexity estimate from Theorem 6.2 is typically insignificant. Nevertheless, from the theoretical point of view, understanding how to to completely remove the one-dimensional search (as in the standard FGM for Lipschitz-smooth functions) is still an important question which we leave for future research.

Let us now compare the obtained complexity with that of other existing accelerated methods. In Li et al. (2023), the authors showed $\mathcal{O}((L_1^2 R^2 + \frac{L_1^2 F_0}{L_0} + 1)\sqrt{\frac{F_0 + L_0 R^2}{\epsilon}})$ complexity for Nesterov Accelerated Gradient (NAG) method, which is significantly worse than ours. Even when F_0 satisfies (19), NAG complexity simplifies to $\mathcal{O}((L_1^2 R^2 + 1)\sqrt{\frac{L_0 R^2}{\epsilon}})$; while our bound have summation instead of a product of two terms which is significantly better. In (Gorbunov et al., 2024), the complexity estimate for a normalized variant of Similar Triangles Method Max (STM-Max) is $\mathcal{O}(1)\exp(\mathcal{O}(1) L_1 R)\sqrt{\frac{L_0 R^2}{\epsilon}}$ which is also worse than our rate provided in Theorem 6.2 and is worse than the complexity of NAG when (19) holds. Both STM and NAG do not guarantee the monotonic decrease of the value function; both methods can escape the initial sublevel set $\mathcal{F}_0$, and consequently, the gradient norm might increase during optimization. While our Algorithm 2 guarantees that after the first stage, the gradient norm becomes smaller than $\frac{L_0}{L_1}$ and all the iterates of the second stage maintain this property and stay in the sublevel set $\mathcal{F}_0$ by construction. However, Algorithm 2 requires additional knowledge of f^* to stop the first stage of the procedure.

7 Conclusion

This work investigates gradient methods for (L_0, L_1)-smooth optimization problems. We have provided new insights into this function class and presented examples along with the operations preserving the (L_0, L_1)-smoothness. Moreover, we have provided new properties of the function class that have rendered tighter bounds on the descent inequalities. Based on these tighter bounds, we derived new stepsizes for the gradient method and connected them with normalized and clipped stepsizes. For such stepsizes, we have shown the best-known $\mathcal{O}(\frac{L_0 F_0}{\epsilon^2} + \frac{L_1 F_0}{\epsilon})$ complexity for finding an ϵ-stationary point in non-convex problems. For convex problems, we have significantly improved the existing results and obtained the $\mathcal{O}(\frac{L_0 R^2}{\epsilon} + L_1 R \ln \frac{F_0}{\epsilon})$ complexity for the gradient method with our stepsizes. We have also analyzed the gradient method with Polyak stepsizes and a normalized gradient method that achieve $\mathcal{O}(\frac{L_0 R^2}{\epsilon} + L_1^2 R^2)$ complexity bound, which is significantly better than previously known complexity bounds. Both of these methods are useful because they automatically adjust to the best possible pair of possible parameters (L_0, L_1). Finally, we have proposed a new procedure by combining our results on the gradient method with the AGMsDR method, and showed fast $\mathcal{O}(m\sqrt{\frac{L_0 R^2}{\epsilon}} + L_1^2 R^2)$ complexity bound which does not have the dependency on the initial gradient norm and does not have an exponential dependency on L_1, in contrast to previous works. Whether it is possible to eliminate the line search for determining y_k in AGMsDR method is an interesting open question for further research. Another interesting question is how to improve the $L_1^2 R^2$ complexity of the first phase.

References

Ziyi Chen, Yi Zhou, Yingbin Liang, and Zhaosong Lu. Generalized-smooth nonconvex optimization is as efficient as smooth nonconvex optimization. In *International Conference on Machine Learning*, 2023.

Matthew Faw, Litu Rout, Constantine Caramanis, and Sanjay Shakkottai. Beyond uniform smoothness: A stopped analysis of adaptive sgd. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 89–160. PMLR, 2023.

Eduard Gorbunov, Nazarii Tupitsa, Sayantan Choudhury, Alen Aliev, Peter Richtárik, Samuel Horváth, and Martin Takáč. Methods for convex (l_0, l_1)-smooth optimization: Clipping, acceleration, and adaptivity. *arXiv preprint arXiv:2409.14989*, 2024.

Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Convex Analysis and Minimization Algorithms II: Advanced Theory and Bundle Methods*. Springer Berlin Heidelberg, 1993. ISBN 9783662064092. doi: 10.1007/978-3-662-06409-2. URL `http://dx.doi.org/10.1007/978-3-662-06409-2`.

Florian Hübler, Junchi Yang, Xiang Li, and Niao He. Parameter-agnostic optimization under relaxed smoothness. In *International Conference on Artificial Intelligence and Statistics*, pp. 4861–4869. PMLR, 2024.

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

Haochuan Li, Jian Qian, Yi Tian, Alexander Rakhlin, and Ali Jadbabaie. Convex and non-convex optimization under generalized smoothness. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120:221–259, 2005. URL `https://api.semanticscholar.org/CorpusID:14935076`.

Yurii Nesterov. *Lectures on Convex Optimization*, volume 137. Springer Optimization and Its Applications Series, 2nd edition, 2018.

Yurii Nesterov, Alexander Gasnikov, Sergey Guminov, and Pavel Dvurechensky. Primal–dual accelerated gradient methods with small-dimensional relaxation oracle. *Optimization Methods and Software*, 36(4):773–810, 2021.

R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

Anton Rodomanov. *Quasi-Newton Methods with Provable Efficiency Guarantees*. PhD thesis, PhD thesis, Université Catholique de Louvain, 2022.

Yuki Takezawa, Han Bao, Ryoma Sato, Kenta Niwa, and Makoto Yamada. Polyak meets parameter-free clipped gradient descent. *arXiv preprint arXiv:2405.15010*, 2024.

Bohan Wang, Huishuai Zhang, Zhiming Ma, and Wei Chen. Convergence of adagrad for non-convex objectives: Simple proofs and relaxed assumptions. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 161–190. PMLR, 2023.

Bohang Zhang, Jikai Jin, Cong Fang, and Liwei Wang. Improved analysis of clipping algorithms for non-convex optimization. *Advances in Neural Information Processing Systems*, 33:15511–15521, 2020.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. *arXiv preprint arXiv:1905.11881*, 2019.

A MISSING PROOFS IN SECTION 2

A.1 PROOF OF LEMMA 2.2

Proof. [(2) $\implies$ (3)] Let $x, y \in \mathbb{R}^d$ be arbitrary and let $h := y - x \neq 0$ (otherwise the claim is trivial). Then, for any $t \in [0, 1]$, using (2), we can estimate

$$\|\nabla f(x+th)-\nabla f(x)\| \leq \|h\| \int_0^t \|\nabla^2 f(x+\tau h)\| d\tau \leq \|h\| \int_0^t (L_0+L_1\|\nabla f(x+\tau h)\|) d\tau =: \nu(t).$$

Our goal is to upper bound $\nu(1)$. We may assume that $L_1 > 0$ since otherwise $\nu(1) = L_0\|h\|$ and the proof is finished. Differentiating, we obtain, for any $t \in [0, 1]$,

$$\nu'(t) = L_0\|h\| + L_1\|h\|\|\nabla f(x + th)\| \leq (L_0 + L_1\|\nabla f(x)\|)\|h\| + L_1\|h\|\nu(t),$$

where the final bound is due to the triangle inequality and the previous display. Hence, for any $t \in [0, 1]$, we have

$$\frac{d}{dt} \ln\big[(L_0 + L_1\|\nabla f(x)\| + \epsilon)\|h\| + L_1\|h\|\nu(t)\big] \leq L_1\|h\|,$$

where $\epsilon > 0$ is arbitrary[5]. Integrating this inequality in $t \in [0, 1]$ and noting that $\nu(0) = 0$, we get

$$\ln \frac{L_0 + L_1\|\nabla f(x)\| + \epsilon + L_1\nu(1)}{L_0 + L_1\|\nabla f(x)\| + \epsilon} \leq L_1\|h\|,$$

or, equivalently,

$$\nu(1) \leq (L_0 + L_1\|\nabla f(x)\| + \epsilon)\frac{e^{L_1\|h\|} - 1}{L_1}.$$

Passing now to the limit as $\epsilon \to 0$, we obtain (3).

[(3) $\implies$ (4)] Let $x, y \in \mathbb{R}^d$ be arbitrary points and let $h := y - x$. Then, using (3), we can estimate

$$\begin{aligned} |f(y) - f(x) - \langle \nabla f(x), y - x\rangle| &\leq \int_0^1 |\langle \nabla f(x + th) - \nabla f(x), h\rangle| dt \\ &\leq (L_0 + L_1\|\nabla f(x)\|)\|h\| \int_0^1 \frac{e^{L_1\|h\|t} - 1}{L_1} dt = (L_0 + L_1\|\nabla f(x)\|)\frac{e^{L_1\|h\|} - L_1\|h\| - 1}{L_1^2}, \end{aligned}$$

which is exactly (4).

[(4) $\implies$ (2)] Let us fix an arbitrary point $x \in \mathbb{R}^d$ and an arbitrary unit vector $h \in \mathbb{R}^d$. Then, for any $t > 0$, it follows from (4) that

$$|f(x + th) - f(x) - t\langle \nabla f(x), h\rangle| \leq (L_0 + L_1\|\nabla f(x)\|)\frac{e^{L_1 t} - L_1 t - 1}{L_1^2}.$$

Dividing both sides by t^2 and passing to the limit as $t \to 0$, we get

$$|\langle \nabla^2 f(x)h, h\rangle| \leq L_0 + L_1\|\nabla f(x)\|.$$

This proves (2) since the unit vector h was allowed to be arbitrary. □

A.2 PROOF OF LEMMA 2.3

Proof. [Proof of (5)] Let $x, y \in \mathbb{R}^d$ be arbitrary points and let us assume w.l.o.g. that $L_1 > 0$. In view of the convexity of f and (4), for any $h \in \mathbb{R}^d$, we can write the following two inequalities:

$$\begin{aligned} 0 &\leq f(y + h) - f(x) - \langle \nabla f(x), y + h - x\rangle \\ &\leq \beta_f(x, y) + \langle \nabla f(y) - \nabla f(x), h\rangle + \frac{L_0 + L_1\|\nabla f(y)\|}{L_1^2}\phi(L_1\|h\|), \end{aligned}$$

[5] This additional term is needed to handle the possibility of $L_0 + L_1\|\nabla f(x)\|$ being zero.

where $\beta_f(x,y) := f(y) - f(x) - \langle \nabla f(x), y-x\rangle$. Denoting $a := L_0 + L_1\|\nabla f(y)\| > 0$ and $s := \nabla f(y) - \nabla f(x)$, we therefore obtain

$$\beta_f(x,y) \geq \max_{h\in\mathbb{R}^d}\Big\{\langle s,h\rangle - \frac{a}{L_1^2}\phi(L_1\|h\|)\Big\} = \max_{r\geq 0}\Big\{\|s\|r - \frac{a}{L_1^2}\phi(L_1 r)\Big\} = \frac{a}{L_1^2}\phi_*\Big(\frac{L_1\|s\|}{a}\Big).$$

[Proof of (6)] Summing up (5) with the same inequality but x and y interchanged, we obtain (6).

[Proof of (7)] By using a lower bound $\phi_*(\gamma) \geq \frac{\gamma^2}{2+\gamma}$ in (5) and denoting $a = \|\nabla f(x) - \nabla f(y)\|$ and $g = \|\nabla f(y)\|$, we obtain

$$\begin{aligned} f(y) &\geq f(x) + \langle \nabla f(x), y-x\rangle + \frac{L_0+L_1 g}{L_1^2}\frac{L_1^2 a^2}{(L_0+L_1 g)^2}\frac{L_0+L_1 g}{2(L_0+L_1 g)+a} \\ &= f(x) + \langle \nabla f(x), y-x\rangle + \frac{a^2}{2(L_0+L_1 g)+a}. \end{aligned}$$

□

A.3 Examples and Properties of (L_0, L_1)-smooth functions

Let us present a few simple examples of (L_0, L_1)-smooth functions.

Example A.1. The function $f(x) = \frac{1}{p}\|x\|^p$, where $p > 2$, is (L_0, L_1)-smooth with arbitrary $L_1 > 0$ and $L_0 = (\frac{p-2}{L_1})^{p-2}$.

Example A.2. The function $f(x) = \ln(1+e^x)$ is (L_0, L_1)-smooth with arbitrary $L_1 \in [0,1]$ and $L_0 = \frac{1}{4}(1-L_1)^2$.

The preceding examples also show that the choice of L_0, L_1 parameters is generally not unique. While we cannot guarantee that the class is closed under all standard operations, such as the summation, affine substitution of the argument, we can still show that some operations do preserve (L_0, L_1)-smoothness under certain additional assumptions.

Proposition A.3. *Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be a twice continuously differentiable (L_0, L_1)-smooth function. Then, the following statements hold:*

1. *Let $g\colon \mathbb{R}^d \to \mathbb{R}$ be an L-smooth and M-Lipschitz twice continuously differentiable function. Then, the sum $f+g$ is (L_0', L_1')-smooth with $L_0' = L_0 + ML_1 + L$ and $L_1' = L_1$.*
2. *Let $f_i\colon \mathbb{R}^{d_i} \to \mathbb{R}$ be an $(L_{0,i}, L_{1,i})$-smooth function for each $i = 1,\ldots,n$. Then, the function $h\colon \mathbb{R}^{d_1}\times\ldots\times\mathbb{R}^{d_n} \to \mathbb{R}$ given by $h(x) = \sum_{i=1}^n f_i(x_i)$, where $x = (x_1,\ldots,x_n)$, is (L_0, L_1)-smooth with $L_0 = \max_{1\leq i\leq n} L_{0,i}$ and $L_1 = \max_{1\leq i\leq n} L_{1,i}$.*
3. *If f is univariate ($d=1$) and $h(x) = f(\langle a,x\rangle + b)$, $x\in\mathbb{R}^d$, where $a \in \mathbb{R}^d$, $b\in\mathbb{R}$, then h is (L_0', L_1')-smooth with parameters $L_0' = \|a\|^2 L_0$ and $L_1' = \|a\|L_1$.*
4. *Let additionally $\nabla^2 f(x) \succ 0$ for all $x\in\mathbb{R}^d$ and f be 1-coercive[6]. Then, f is (L_0, L_1)-smooth iff its conjugate f_* (which is, under our assumptions, defined on the entire space and also twice continuously differentiable) satisfies $\nabla^2 f_*(s) \succeq \frac{1}{L_0+L_1\|s\|}I$ for all $s\in\mathbb{R}^d$, where I is the identity matrix.*

One simple example of the additive term g satisfying the assumptions in the first item of Proposition A.3 is an affine function (for which $L=0$); another interesting example is the soft-max function $g(x) = \mu\ln(\sum_{i=1}^m e^{[(\langle a_i,x\rangle + b_i]/\mu)}$, where $a_i\in\mathbb{R}^d$, $b_i\in\mathbb{R}$, $\mu > 0$. Based on the second statement of Proposition A.3 and Example A.1, the function $f(x) = \frac{1}{p}\|x\|_p^p \equiv \frac{1}{p}\sum_{i=1}^d |x_i|^p$ with $p > 2$ is (L_0, L_1)-smooth with arbitrary $L_1 > 0$ and $L_0 = (\frac{p-2}{L_1})^{p-2}$. Using the third statement, we can generalize Example A.2 and conclude that $f(x) = \ln(1+e^{\langle a,x\rangle})$ is also (L_0, L_1)-smooth with arbitrary $L_1 \in [0, \|a\|]$ and $L_0 = \frac{1}{4}(\|a\| - L_1)^2$. Also, we can use the last statement of the proposition to show that $f(x) = \frac{L_0}{L_1^2}\phi(L_1\|x\|) \equiv \frac{L_0}{L_1^2}(e^{L_1\|x\|} - L_1\|x\| - 1)$ is (L_0, L_1)-smooth since the Hessian of its conjugate $f_*(s) = \frac{L_0}{L_1^2}\phi_*(\frac{L_1\|s\|}{L_0}) \equiv \frac{L_0}{L_1^2}[(1+\frac{L_1\|s\|}{L_0})\ln(1+\frac{L_1\|s\|}{L_0}) - \frac{L_1\|s\|}{L_0}]$ has the

[6]This means that $\frac{f(x)}{\|x\|} \to +\infty$ as $\|x\|\to\infty$.

form $\nabla^2 f_*(s) = \frac{1}{L_0+L_1\|s\|} I$. In particular, we can construct an (L_0, L_1)-smooth function by taking any convex function h_*, adding to it ϕ_* and taking the conjugate (this corresponds to the infimal convolution of h with ϕ).

A.4 Proof of Example A.1

Proof. Differentiating, we obtain, for any $x \in \mathbb{R}^d$,

$$\nabla f(x) = \|x\|^{p-2} x, \qquad \nabla^2 f(x) = \|x\|^{p-2}\Big(I + (p-2)\frac{xx^\top}{\|x\|^2}\Big),$$

where I is the identity matrix. Hence, for any $L_1 > 0$, the minimal value of L_0 satisfying the inequality from Definition 2.1 is given by

$$\begin{aligned} L_0 = \max_{x\in\mathbb{R}^d}\big\{\|\nabla^2 f(x)\| - L_1\|\nabla f(x)\|\big\} &= \max_{x\in\mathbb{R}^d}\big\{(p-1)\|x\|^{p-2} - L_1\|x\|^{p-1}\big\} \\ &= \max_{\tau\geq 0}\big\{(p-1)\tau^{\frac{p-2}{p-1}} - L_1\tau\big\}. \end{aligned}$$

The solution of the latter problem is $\tau^* = (\frac{p-2}{L_1})^{p-1}$. Substituting this value, we obtain

$$L_0 = (p-1)\Big(\frac{p-2}{L_1}\Big)^{p-2} - L_1\Big(\frac{p-2}{L_1}\Big)^{p-1} = \Big(\frac{p-2}{L_1}\Big)^{p-2}. \qquad \square$$

A.5 Proof of Example A.2

Proof. Differentiating, we obtain, for any $x \in \mathbb{R}$,

$$f'(x) = \frac{e^x}{1+e^x} \in (0,1), \quad f''(x) = \frac{e^x}{(1+e^x)^2} = f'(x)(1-f'(x)).$$

Thus, for any $L_1 \in [0,1]$, the minimal value of L_0 satisfying the inequality from Definition 2.1 is

$$\begin{aligned} L_0 = \max_{x\in\mathbb{R}}\{|f''(x)| - L_1|f'(x)|\} &= \max_{\tau\in(0,1)}\{\tau(1-\tau) - L_1\tau\} \\ &= \max_{\tau\in(0,1)}\{(1-L_1)\tau - \tau^2\} = \frac{1}{4}(1-L_1)^2. \qquad \square \end{aligned}$$

A.6 Proof of Proposition A.3

Proof. [Claim 1] Since, g and ∇g are M and L Lipschitz continuous, $\|\nabla g(x)\| \leq M$ and $\|\nabla^2 g(x)\| \leq L$ for all $x \in \mathbb{R}$. Let $F = f + g$, then, for any $x \in \mathbb{R}^d$, we can estimate

$$\begin{aligned} \|\nabla^2 F(x)\| &\leq \|\nabla^2 f(x)\| + \|\nabla^2 g(x)\| \leq L_0 + L + L_1\|\nabla f(x)\| \\ &\leq L_0 + L + L_1\|\nabla g(x)\| + L_1\|\nabla F(x)\| \\ &\leq (L_0 + L_1 M + L) + L_1\|\nabla F(x)\|. \end{aligned}$$

[Claim 2] Notice, that the gradient of f is $\nabla f(x) = (\nabla f_1(x_1)^\top, \ldots, \nabla f_n(x_n)^\top)^\top$ and the Hessian of f is $\nabla^2 f(x)$ is a block-diagonal matrix, with $\nabla^2 f_i(x_i)$ blocks. Thus,

$$\begin{aligned} \|\nabla^2 f(x)\| &= \max_{1\leq i\leq n}\|\nabla^2 f_i(x_i)\| \leq \max_{1\leq i\leq n}\{L_{0,i} + L_{1,i}\|\nabla f_i(x_i)\|\} \\ &\leq \max_{1\leq i\leq n}\{L_{0,i} + L_{1,i}\|\nabla f(x)\|\} \leq \max_{1\leq i\leq n} L_{0,i} + (\max_{1\leq i\leq n} L_{1,i})\|\nabla f(x)\|. \end{aligned}$$

[Claim 3] Observe that the gradient of a function is $\nabla f(x) = g'(\langle a, x\rangle + b)a$, and the Hessian is $\nabla^2 f(x) = g''(\langle a, x\rangle + b)aa^\top$. Hence,

$$\begin{aligned} \|\nabla^2 f(x)\| = |g''(\langle a, x\rangle + b)|\|a\|^2 &\leq (L_0 + L_1|g'(\langle a, x\rangle + b)|)\|a\|^2 \\ &= L_0\|a\|^2 + \|a\|L_1\|\nabla f(x)\|. \end{aligned}$$

[Claim 4] Under our assumptions, $s = \nabla f(x)$ is a one-to-one transformation from $\mathbb{R}^d$ to $\mathbb{R}^d$ (whose inverse transformation is $x = \nabla f_*(s)$); moreover, the Hessians at such a pair of points are inverse to each other: $\nabla^2 f_*(s) = [\nabla^2 f(x)]^{-1}$ (see, e.g., Corollaries 4.1.4 and 4.2.10 in Hiriart-Urruty & Lemaréchal (1993), as well as Example 11.9 from Rockafellar & Wets (2009)). Thus, for any pair of points $x, s \in \mathbb{R}^d$ such that $s = \nabla f(x)$, our assumption $\|\nabla^2 f(x)\| \leq L_0 + L_1\|\nabla f(x)\|$ which, due to the convexity of f, can be equivalently rewritten as $\nabla^2 f(x) \preceq (L_0 + L_1\|\nabla f(x)\|)I$, is equivalent to

$$\nabla^2 f_*(s) \equiv [\nabla^2 f(x)]^{-1} \succeq \frac{1}{L_0 + L_1\|\nabla f(x)\|} I \equiv \frac{1}{L_0 + L_1\|s\|} I.$$

This proves the claim since the transformation $s = \nabla f(x)$ is one-to-one. □

A.7 Proof of Lemma A.4

In our analysis, we often use certain properties of the function ϕ and its conjugate[7] ϕ_*, which we summarize in the following lemma (see Appendix A.7 for the proof).

Lemma A.4. *The following statements for the function $\phi(t) = e^t - t - 1$ hold true:*

1. $\phi(t) \leq \frac{t^2}{2(1-\frac{t}{3})}$ *for all* $t \in [0, 3)$ *and* $\phi(t) \leq \frac{t^2}{2}e^t$ *for all* $t \geq 0$.
2. $\phi_*(\gamma) := \max_{t \geq 0}\{\gamma t - \phi(t)\} = (1+\gamma)\ln(1+\gamma) - \gamma$ *for any* $\gamma \geq 0$.
3. $\frac{\gamma^2}{2+\gamma} \leq \phi_*(\gamma) \leq \frac{\gamma^2}{2}$ *for all* $\gamma \geq 0$.

Proof. [Claim 1] Indeed, for any $t \in [0, 3)$, we have

$$\phi(t) = e^t - t - 1 = \sum_{i=2}^{\infty} \frac{t^i}{i!} = \sum_{i=0}^{\infty} \frac{t^{2+i}}{(2+i)!} = \frac{t^2}{2}\sum_{i=0}^{\infty} \frac{t^i}{\prod_{j=3}^{2+i} j} \leq \frac{t^2}{2}\sum_{i=0}^{\infty}\frac{t^i}{3^i} = \frac{t^2}{2(1-\frac{t}{3})}.$$

Similarly, for any $t \geq 0$,

$$\phi(t) = \frac{t^2}{2}\sum_{i=0}^{\infty}\frac{t^i}{\prod_{j=3}^{2+i} j} \leq \frac{t^2}{2}\sum_{i=0}^{\infty}\frac{t^i}{i!} = \frac{t^2}{2}e^t.$$

[Claim 2] By the definition, for any $\gamma \geq 0$, we have

$$\phi_*(\gamma) = \max_{t\geq 0}\{\gamma t - \phi(t)\} = \max_{t \geq 0}\{(1+\gamma)t - e^t\} + 1.$$

Differentiating, we see that the solution of this optimization problem is $t_* = \ln(1+\gamma)$. Hence,

$$\phi_*(\gamma) = (1+\gamma)\ln(1+\gamma) - (1+\gamma) + 1 = (1+\gamma)\ln(1+\gamma) - \gamma.$$

[Claim 3] We first show that, for any $\gamma \geq 0$,

$$\ln(1+\gamma) \geq \frac{2\gamma}{2+\gamma}.$$

Since both functions coincide at $\gamma = 0$, it suffices to verify the corresponding inequality for the derivatives:

$$\frac{1}{1+\gamma} \geq \frac{4}{(2+\gamma)^2} \equiv \frac{4}{4+4\gamma+\gamma^2} \equiv \frac{1}{1+\gamma+\frac{\gamma^2}{4}}.$$

But this is obviously true. Applying the derived inequality, we get, for any $\gamma \geq 0$,

$$\phi_*(\gamma) \equiv (1+\gamma)\ln(1+\gamma) - \gamma \geq \frac{2\gamma(1+\gamma)}{2+\gamma} - \gamma = \frac{\gamma[2(1+\gamma) - (2+\gamma)]}{2+\gamma} = \frac{\gamma^2}{2+\gamma},$$

which proves the first part of the claim.

For the second part, we note that $\phi_*(\gamma)$ and $\frac{\gamma^2}{2}$ coincide at $\gamma = 0$. Hence, it suffices to check the corresponding inequality for the derivatives, i.e., to verify that, for all $\gamma \geq 0$,

$$\phi_*'(\gamma) \equiv \ln(1+\gamma) \leq \gamma.$$

But this follows from the concavity of the logarithm. □

[7] The conjugate function is defined in the standard way: $\phi_*(\gamma) := \max_{t \geq 0}\{\gamma t - \phi(t)\}$.

B MISSING PROOFS IN SECTION 3

B.1 ONE-STEP PROGRESS

Lemma B.1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be an (L_0, L_1)-smooth function, let $x \in \mathbb{R}^d$, and let $T(x) = x - \eta \nabla f(x)$, where η is given by one of the following formulas:*

$$(1)\ \eta_* = \frac{1}{L_1 \|\nabla f(x)\|} \ln\Big(1 + \frac{L_1 \|\nabla f(x)\|}{L_0 + L_1 \|\nabla f(x)\|}\Big), \quad (2)\ \eta_{\rm si} = \frac{1}{L_0 + \frac{3}{2} L_1 \|\nabla f(x)\|},$$
$$(3)\ \eta_{\rm cl} = \min\Big\{\frac{1}{2L_0}, \frac{1}{3L_1 \|\nabla f(x)\|}\Big\}.$$

Then,

$$f(x) - f(T(x)) \geq \frac{a \|\nabla f(x)\|^2}{2L_0 + 3L_1 \|\nabla f(x)\|},$$

where $a = 1$ in cases (1) and (2), and $a = \frac{1}{2}$ in case (3).

Proof. [Case (1)] The proof of this case was already presented in Section 3.

For the other two cases, we start by applying Lemma 2.2 to get

$$\begin{aligned}\Delta \coloneqq f(x) - f(T(x)) &\geq \langle \nabla f(x), x - T(x) \rangle - \frac{L_0 + L_1 \|\nabla f(x)\|}{L_1^2} \phi(L_1 \|T(x) - x\|) \\ &= \eta_* g^2 - \frac{L_0 + L_1 g}{L_1^2} \phi(\eta_* L_1 g),\end{aligned}$$

where $g \coloneqq \|\nabla f(x)\|$ and $\phi(t) = e^t - t - 1$.

[Case (2)] Estimating $\phi(t) \leq \frac{3t^2}{6-2t} \leq \frac{t^2}{2-t}$ (Lemma A.4) and substituting the definition of $\eta_{\rm si}$, we can continue as follows:

$$\begin{aligned}\Delta &\geq \eta_{\rm si} g^2 - \frac{L_0 + L_1 g}{L_1^2} \frac{\eta_{\rm si}^2 L_1^2 g^2}{2 - \eta_{\rm si} L_1 g} = \Big(1 - \frac{(L_0 + L_1 g)\eta_{\rm si}}{2 - \eta_{\rm si} L_1 g}\Big) \eta_{\rm si} g^2 \\ &= \Big(1 - \frac{L_0 + L_1 g}{(L_0 + \frac{3}{2} L_1 g)(2 - \frac{L_1 g}{L_0 + \frac{3}{2} L_1 g})}\Big) \frac{g^2}{L_0 + \frac{3}{2} L_1 g} = \frac{g^2}{2L_0 + 3L_1 g}.\end{aligned}$$

[Case (3)] Observe that

$$\frac{1}{2L_0 + 3L_1 g} \leq \eta_{\rm cl} \equiv \frac{1}{\max\{2L_0, 3L_1 g\}} \leq \frac{1}{L_0 + \frac{3}{2} L_1 g}.$$

Combining these bounds with $\phi(t) \leq \frac{3t^2}{6-2t}$ (Lemma A.4 (1)), we get

$$\begin{aligned}\Delta &\geq \eta_{\rm cl} g^2 - \frac{L_0 + L_1 g}{L_1^2} \frac{3L_1^2 \eta_{\rm cl}^2 g^2}{6 - 2\eta_{\rm cl} L_1 g} = \Big(1 - \frac{3\eta_{\rm cl}(L_0 + L_1 g)}{6 - 2\eta_{\rm cl} L_1 g}\Big) \eta_{\rm cl} g^2 \\ &\geq \Big(1 - \frac{3(L_0 + L_1 g)}{(L_0 + \frac{3}{2} L_1 g)(6 - \frac{2L_1 g}{L_0 + \frac{3}{2} L_1 g})}\Big) \frac{g^2}{2L_0 + 3L_1 g} \\ &= \Big(1 - \frac{3(L_0 + L_1 g)}{6L_0 + 7L_1 g}\Big) \frac{g^2}{2L_0 + 3L_1 g} \geq \frac{1}{2} \frac{g^2}{2L_0 + 3L_1 g}.\end{aligned}$$

□

B.2 PROOF OF THEOREM 3.1

Proof. According to Lemma B.1, for any $k \geq 0$, we have

$$f(x_k) - f(x_{k+1}) \geq \frac{a \|\nabla f(x_k)\|^2}{2L_0 + 3L_1 \|\nabla f(x_k)\|},$$

where a is an absolute constant defined in the statement depending on the stepsize choice. Denote $f_k = f(x_k) - f^*$ (≥ 0) and $g_k = \|\nabla f(x_k)\|$. In this notation, the above inequality reads

$$f_k - f_{k+1} \geq a\psi(g_k), \qquad \psi(\gamma) := \frac{\gamma^2}{2L_0 + 3L_1\gamma}.$$

Summing up these inequalities for all $0 \leq k \leq K$ and denoting $g_K^* = \min_{0 \leq k \leq K} g_k$, we get

$$F_0 \geq f_0 - f_K \geq a \sum_{k=0}^{K} \psi(g_k) \geq a(K+1)\psi(g_K^*),$$

where the final inequality holds since ψ is an increasing function. Denoting the corresponding inverse function by ψ^{-1}, we come to the conclusion that

$$g_K^* \leq \psi^{-1}\Big(\frac{F_0}{a(K+1)}\Big) \leq \epsilon$$

whenever

$$\frac{F_0}{a(K+1)} \leq \psi(\epsilon),$$

or, equivalently,

$$K + 1 \geq \frac{F_0}{a\psi(\epsilon)} \equiv \frac{2L_0F_0}{a\epsilon^2} + \frac{3L_1F_0}{a\epsilon}.$$

□

B.3 Proof of Theorem 3.2

First, we prove that the distance to the solution is nonincreasing.

Lemma B.2. *Under the conditions of Theorem 3.2, we have $R_{k+1} \leq R_k$ for any $k \geq 0$.*

Proof. Let $k \geq 0$ be arbitrary and denote $\beta_k = \langle \nabla f(x_k), x_k - x^* \rangle$ and $g_k = \|\nabla f(x_k)\|$. According to the update rule of the method, we have

$$R_{k+1}^2 = R_k^2 - 2\eta_k\beta_k + \eta_k^2 g_k^2.$$

Therefore, to prove that $R_{k+1} \leq R_k$, we need to show that

$$\eta_k g_k^2 \leq 2\beta_k.$$

Applying bound (7) twice, we see that

$$\begin{aligned}\beta_k &\equiv [f(x_k) - f^*] + [f^* - f(x_k) - \langle \nabla f(x_k), x^* - x_k \rangle] \\ &\geq \frac{g_k^2}{2L_0 + 3L_1g_k} + \frac{g_k^2}{2L_0 + L_1g_k} \geq \frac{g_k^2}{L_0 + L_1g_k},\end{aligned}$$

where the final inequality follows from the fact that $\frac{1}{a} + \frac{1}{b} \geq \frac{4}{a+b}$ (convexity of $t \mapsto \frac{1}{t}$). Thus, we need to check if

$$\eta_k \leq \frac{2}{L_0 + L_1g_k}. \tag{22}$$

Furthermore, it suffices to check this inequality only for the largest among the three stepsizes we consider. According to (14), this is stepsize (9). Applying $\ln(1+\gamma) \leq \gamma$ (which holds for any $\gamma \geq 0$), we see that

$$\eta_k^* \equiv \frac{1}{L_1g_k} \ln\Big(1 + \frac{L_1g_k}{L_0 + L_1g_k}\Big) \leq \frac{1}{L_0 + L_1g_k},$$

so (22) is indeed satisfied. □

Now we can prove the theorem.

Proof of Theorem 3.2. Let $k \geq 0$ be arbitrary and denote $f_k := f(x_k) - f^*$ and $g_k := \|\nabla f(x_k)\|$. According to Lemma B.1, we have

$$f_k - f_{k+1} \geq a\psi(g_k), \qquad \psi(\gamma) := \frac{\gamma^2}{2L_0 + 3L_1\gamma},$$

where a is an absolute constant defined in the statement depending on the stepsize choice. Further, according to Lemma B.2, the distances $R_k := \|x_k - x^*\|$ are nonincreasing. In particular, $R_k \leq R_0 \equiv R$. Hence, in view of the convexity of f, we can estimate

$$f_k \leq \langle \nabla f(x_k), x_k - x^* \rangle \leq g_k R_k \leq g_k R.$$

Combining the above two displays and using the fact that the function ψ is increasing, we obtain

$$f_k - f_{k+1} \geq a\psi\Big(\frac{f_k}{R}\Big).$$

Consequently,

$$\begin{aligned} a \leq \frac{f_k - f_{k+1}}{\psi(\frac{f_k}{R})} \leq \int_{f_{k+1}}^{f_k} \frac{dt}{\psi(\frac{t}{R})} &= \int_{f_{k+1}}^{f_k} \Big(\frac{2L_0R^2}{t^2} + \frac{3L_1R}{t}\Big) dt \\ &= 2L_0R^2\Big(\frac{1}{f_{k+1}} - \frac{1}{f_k}\Big) + 3L_1R\ln\frac{f_k}{f_{k+1}}. \end{aligned}$$

Summing up these inequalities for all $0 \leq k \leq K-1$ and dropping the negative $\frac{1}{f_0}$ term, we get

$$aK \leq \frac{2L_0R^2}{f_K} + 3L_1R\ln\frac{f_0}{f_K}.$$

Hence, $f_K \leq \epsilon$ whenever

$$K \geq \frac{2L_0R^2}{a\epsilon} + \frac{3}{a}L_1R\ln\frac{f_0}{\epsilon} =: K(\epsilon).$$

To upper bound $K(\epsilon)$, we first estimate f_0 using Lemmas 2.2 and A.4:

$$f_0 \leq \frac{L_0}{L_1^2}\phi(L_1R) \leq \frac{L_0R^2}{2}e^{L_1R}.$$

This gives us

$$aK(\epsilon) \leq \frac{2L_0R^2}{\epsilon} + 3L_1R\Big(L_1R + \ln\frac{L_0R^2}{\epsilon}\Big) = \frac{2L_0R^2}{\epsilon} + 3L_1^2R^2 + 6L_1R\ln\Big(\sqrt{\frac{L_0R^2}{\epsilon}}\Big).$$

Estimating $\ln t \leq \frac{t}{e}$ (holding for any $t > 0$) and applying the AM-GM inequality, we get

$$aK(\epsilon) \leq \frac{2L_0R^2}{\epsilon} + 3L_1^2R^2 + \frac{6}{e}\sqrt{\frac{L_0R^2}{\epsilon}\cdot L_1^2R^2} \leq \frac{(2+\frac{3}{e})L_0R^2}{\epsilon} + \Big(3 + \frac{3}{e}\Big)L_1^2R^2. \qquad \square$$

C Missing Proofs in Section 4

C.1 General Result

Lemma C.1. *Let $\{x_k\}$ be the iterates of NGM* (15) *with arbitrary coefficients $\beta_k > 0$, as applied to problem* (1) *with an (L_0, L_1)-smooth convex function f. Then, $\min_{0\leq k\leq K} f(x_k) - f^* \leq \epsilon$ for any given $K \geq 0$ and $\epsilon > 0$ whenever*

$$\delta_K := \frac{R^2 + \sum_{k=0}^K \beta_k^2}{2\sum_{k=0}^K \beta_k} \leq \delta(\epsilon) := \min\Big\{\frac{3}{2L_1}, \sqrt{\frac{\epsilon}{L_0}}\Big\},$$

where $R := \|x_0 - x^*\|$ is the distance from the initial point to a solution x^* of the problem.

Proof. According to (15), for any $k \geq 0$, we have

$$\|x_{k+1} - x^*\|^2 = \|x_k - x^*\|^2 - 2\eta_k \langle \nabla f(x_k), x_k - x^* \rangle + \eta_k^2 \|\nabla f(x_k)\|^2$$
$$= \|x_k - x^*\|^2 - 2\beta_k v_k + \beta_k^2,$$

where $v_k := \frac{\langle \nabla f(x_k), x_k - x^* \rangle}{\|\nabla f(x_k)\|}$ (≥ 0). Summing up these relations over $k = 0, \ldots, K$ and rearranging the terms, we obtain

$$2 \sum_{k=0}^{K} \beta_k v_k \leq R^2 + \sum_{k=0}^{K} \beta_k^2.$$

Denoting $v_K^* = \min_{0 \leq k \leq K} v_k$, we get

$$v_K^* \leq \frac{R^2 + \sum_{k=0}^{K} \beta_k^2}{2 \sum_{k=0}^{K} \beta_k} =: \delta_K. \tag{23}$$

Let $f_K^* := \min_{0 \leq k \leq K} f(x_k)$. Then, by Lemma 4.2,

$$f_K^* - f^* \leq \max_z \{ f(z) - f^* : \|z - x^*\| \leq v_K^* \}.$$

Applying Lemma 2.2 and the fact that $\phi(t) \leq \frac{3t^2}{6-2t}$ for any $t \in [0, 3)$ (Lemma A.4), we obtain

$$f_K^* - f^* \leq \frac{L_0}{L_1^2} \phi(L_1 v_K^*) \leq \frac{3 L_0 (v_K^*)^2}{6 - 2 L_1 v_K^*}$$

whenever $L_1 v_k^* < 3$. To achieve the desired accuracy ϵ, it thus suffices to ensure that the following two inequalities are satisfied:

$$2 L_1 v_K^* \leq 3, \qquad L_0 (v_K^*)^2 \leq \epsilon.$$

This is equivalent to

$$v_K^* \leq \min\left\{ \frac{3}{2L_1}, \sqrt{\frac{\epsilon}{L_0}} \right\} =: \delta(\epsilon),$$

and follows from $\delta_k \leq \delta(\epsilon)$ in view of (23). $\square$

C.2 Proof of Theorem 4.1

Proof. According to Lemma C.1, we need to ensure that

$$\delta_K := \frac{R^2 + \sum_{k=0}^{K} \beta_k^2}{2 \sum_{k=0}^{K} \beta_k} \leq \delta(\epsilon) := \min\left\{ \frac{3}{2L_1}, \sqrt{\frac{\epsilon}{L_0}} \right\}.$$

In our case,

$$\delta_K = \frac{R^2 + \hat{R}^2}{2\hat{R}\sqrt{K+1}} = \frac{\bar{R}}{\sqrt{K+1}}.$$

Therefore, $\delta_K \leq \delta(\epsilon)$ iff

$$K + 1 \geq \frac{\bar{R}^2}{\delta^2(\epsilon)} \equiv \max\left\{ \frac{4}{9} L_1^2 \bar{R}^2, \frac{L_0 \bar{R}^2}{\epsilon} \right\}. \qquad \square$$

C.3 Analysis for Time-Varying Step Size

Theorem C.2. *Let $\{x_k\}$ be the iterates of NGM (15), as applied to problem (1) with an (L_0, L_1)-smooth nonlinear[8] convex function f. Consider decreasing coefficients $\beta_k = \frac{\hat{R}}{\sqrt{k+1}}$, $k \geq 0$, where $\hat{R} > 0$ is a parameter. Then, $\min_{0 \leq k \leq K} f(x_k) - f^* \leq \epsilon$ for any given $\epsilon > 0$ whenever*

$$K + 1 \geq \max\left\{ 4 N_{\bar{R}}(\epsilon), \left(\frac{e}{e-1} \right)^2 N_{\hat{R}}(\epsilon) [\ln(4 N_{\hat{R}}(\epsilon))]_+^2 \right\},$$

where $\bar{R} := \frac{1}{2}(\frac{R^2}{\hat{R}} + \hat{R})$, $R := \|x_0 - x^*\|$ (x^* is an arbitrary solution of the problem), and

$$N_D(\epsilon) := \max\left\{ \frac{4}{9} L_1^2 D^2, \frac{L_0 D^2}{\epsilon} \right\}.$$

[8]This means that $L_0 + L_1 \|\nabla f(x)\| > 0$ for any $x \in \mathbb{R}^d$, see Lemma 2.2.

Proof. According to Lemma C.1, we need to ensure that

$$\delta_K := \frac{R^2 + \sum_{k=0}^K \beta_k^2}{2\sum_{k=0}^K \beta_k} \leq \delta(\epsilon) := \min\Big\{\frac{3}{2L_1}, \sqrt{\frac{\epsilon}{L_0}}\Big\}.$$

For our choice of β_k, we obtain, by standard results (e.g., Lemma 2.6.3 in Rodomanov (2022)), that

$$\sum_{k=0}^K \beta_k^2 = \hat{R}^2 \sum_{k=1}^{K+1} \frac{1}{k} \leq \hat{R}^2[1 + \ln(K+1)], \qquad \sum_{k=0}^K \beta_k = \hat{R} \sum_{k=1}^{K+1} \frac{1}{\sqrt{k}} \geq \hat{R}\sqrt{K+1}.$$

Hence,

$$\delta_K \leq \frac{R^2 + \hat{R}^2[1 + \ln(K+1)]}{2\hat{R}\sqrt{K+1}} = \frac{\bar{R}}{\sqrt{K+1}} + \frac{\hat{R}\ln(K+1)}{2\sqrt{K+1}}.$$

To ensure that $\delta_K \leq \delta(\epsilon)$, it suffices to ensure that the following two inequalities are satisfied:

$$\frac{\bar{R}}{\sqrt{K+1}} \leq \frac{\delta(\epsilon)}{2}, \qquad \frac{\hat{R}\ln(K+1)}{\sqrt{K+1}} \leq \delta(\epsilon).$$

The first inequality is equivalent to $K+1 \geq \frac{4\bar{R}^2}{\delta^2}$. To get the second one, it suffices to take, according to Lemma C.3 (with $p = \frac{1}{2}$ and $\delta' = \frac{\delta(\epsilon)}{\hat{R}}$),

$$K + 1 \geq \Big(\frac{e}{e-1}\frac{2\hat{R}}{\delta(\epsilon)}\Big[\ln\frac{2\hat{R}}{\delta(\epsilon)}\Big]_+\Big)^2 \equiv \Big(\frac{e}{e-1}\Big)^2 \frac{\hat{R}^2}{\delta^2(\epsilon)}\Big[\ln\frac{4\hat{R}^2}{\delta^2(\epsilon)}\Big]_+^2.$$

Putting these two inequalities together and substituting our formula for $\delta(\epsilon)$, we come to the requirement that

$$\begin{aligned} K + 1 &\geq \max\Big\{\frac{4\bar{R}^2}{\delta^2(\epsilon)}, \Big(\frac{e}{e-1}\Big)^2 \frac{\hat{R}^2}{\delta^2(\epsilon)}\Big[\ln\frac{4\hat{R}^2}{\delta^2(\epsilon)}\Big]_+^2\Big\} \\ &= \max\Big\{4N_{\bar{R}}(\epsilon), \Big(\frac{e}{e-1}\Big)^2 N_{\hat{R}}(\epsilon)[\ln(4N_{\hat{R}}(\epsilon))]_+^2\Big\}, \end{aligned}$$

where

$$N_D(\epsilon) := \frac{D^2}{\delta^2(\epsilon)} = \max\Big\{\frac{4}{9}L_1^2 D^2, \frac{L_0 D^2}{\epsilon}\Big\}.$$ $\square$

Lemma C.3. *For any real $p, \delta > 0$, we have the following implication*[9]*:*

$$t \geq \Big(\frac{e}{e-1}\frac{[\ln\frac{1}{p\delta}]_+}{p\delta}\Big)^{\frac{1}{p}} \qquad \Longrightarrow \qquad \frac{\ln t}{t^p} \leq \delta.$$

Proof. W.l.o.g., we can assume that $p = 1$, and our goal is to prove the implication

$$t \geq \frac{e}{e-1}\frac{[\ln\frac{1}{\delta}]_+}{\delta} =: t(\delta) \qquad \Longrightarrow \qquad \phi(t) := \frac{\ln t}{t} \leq \delta.$$

The general case then follows by the change of variables $t = (t')^p$ and $\delta = p\delta'$.

Further, we can assume that $\delta \leq \frac{1}{e}$ since otherwise $\phi(t) \leq \frac{1}{e} \leq \delta$ for any $t \geq 0$ (since the maximum of ϕ is achieved at $t_* = e$). Under this additional assumption, $[\ln\frac{1}{\delta}]_+ = \ln\frac{1}{\delta}$.

Let us now assume that $t \geq t(\delta)$ ($\geq \frac{e^2}{e-1} \geq e$ since $\delta \leq \frac{1}{e}$). Since the function ϕ is decreasing on the interval $[e, +\infty)$, we have

$$\phi(t) \leq \phi(t(\delta)) = \frac{\ln t(\delta)}{t(\delta)} = \frac{\ln t(\delta)}{\frac{e}{e-1}\ln\frac{1}{\delta}}\delta.$$

[9] For $t = 0$, we define by continuity $\frac{\ln t}{t^p} \equiv -\infty$.

To finish the proof, it remains to show that the final fraction in the above display is ≤ 1, or, equivalently, that

$$t(\delta) \equiv \frac{e}{e-1} \frac{\ln \frac{1}{\delta}}{\delta} \leq \Big(\frac{1}{\delta}\Big)^{\frac{e}{e-1}}.$$

Rearranging and denoting $u := (\frac{1}{\delta})^{\frac{1}{e-1}}$, we see that the above inequality is equivalent to

$$\phi(u) \equiv \frac{\ln u}{u} \leq \frac{1}{e}.$$

But this is indeed true since ϕ attains its maximum value at $u = e$. $\square$

D Missing Proofs in Section 5

D.1 Proof of Theorem 5.1

Proof. Let x^* be an arbitrary solution. By the method's update rule and convexity of $f(\cdot)$, we get, for all $k \geq 0$,

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &= \|x_k - x^*\|^2 - 2\eta_k \langle \nabla f(x_k), x_k - x^* \rangle + \eta_k^2 \|\nabla f(x_k)\|^2 \\ &\leq \|x_k - x^*\|^2 - \frac{(f(x_k) - f^*)^2}{\|\nabla f(x_k)\|^2}.\end{aligned}$$

Denote $R_k = \|x_k - x^*\|$, $g_k = \|\nabla f(x_k)\|$ and $f_k = f(x_k) - f^*$. According to Lemma 2.3, for each $k \geq 0$, it holds that

$$f_k \geq \psi(g_k), \qquad \text{where} \quad \psi(g) := \frac{g^2}{2L_0 + 3L_1 g}, \quad g \geq 0.$$

Observe that the function ψ is increasing, so its inverse ψ^{-1} is well-defined and is increasing as well. In terms of this function, $g_k \leq \psi^{-1}(f_k)$ and hence

$$R_k^2 - R_{k+1}^2 \geq \frac{f_k^2}{g_k^2} \geq \Big(\frac{f_k}{\psi^{-1}(f_k)}\Big)^2.$$

Summing up these inequalities over $0 \leq k \leq K$ and rearranging, we get

$$\sum_{k=0}^{K} \Big(\frac{f_k}{\psi^{-1}(f_k)}\Big)^2 \leq R_0^2 - R_{K+1}^2 \leq R_0^2 \equiv R^2.$$

Note that $\frac{\psi^{-1}(t)}{t}$ is increasing in t (as the composition of increasing in γ function $\frac{\psi(\gamma)}{\gamma} \equiv \frac{\gamma}{2L_0+3L_1\gamma}$ with increasing in t function $\gamma = \psi^{-1}(t)$). Thus, by taking a minimum over the terms on the left-hand side of the above display and denoting $f_K^* := \min_{0 \leq k \leq K} f_k$, we get

$$(K+1)\Big(\frac{f_K^*}{\psi^{-1}(f_K^*)}\Big)^2 \leq R^2.$$

Rearranging, we obtain

$$\psi^{-1}(f_K^*) \geq \frac{\sqrt{K+1} f_K^*}{R},$$

or, equivalently,

$$f_K^* \geq \psi\Big(\frac{\sqrt{K+1} f_K^*}{R}\Big) \equiv \frac{(K+1)(f_K^*)^2}{R^2\big(2L_0 + 3L_1 \frac{\sqrt{K+1} f_K^*}{R}\big)} = \frac{(f_K^*)^2}{\frac{2L_0 R^2}{K+1} + \frac{3L_1 R}{\sqrt{K+1}} f_K^*}.$$

Hence,

$$f_K^* \leq \frac{2L_0 R^2}{(K+1)(1 - 3L_1 R\sqrt{K+1})},$$

whenever $3L_1 R\sqrt{K+1} < 1$. Thus, to achieve desired accuracy $\epsilon > 0$, the number K of iterations should satisfy the following conditions:

$$3L_1 R\sqrt{K+1} \leq \frac{1}{2}, \qquad \frac{4L_0 R^2}{K+1} \leq \epsilon.$$

Thus, the final iteration complexity is $K + 1 \geq \max\{\frac{4L_0R^2}{\epsilon}, 36L_1^2R^2\}$. $\square$

E MISSING PROOFS IN SECTION 6

The proof of Theorem 6.1 is similar to the original proof Theorem 1 in Nesterov et al. (2021), but instead of smoothness of f we use more general property provided in the statement of Theorem 6.1.

E.1 PROOF OF THEOREM 6.1

Proof. Let us prove by induction that, for any $k \geq 0$, we have

$$A_k f(x_k) \leq \zeta_k^* := \zeta_k(v_k). \tag{24}$$

This trivially holds for $k = 0$ since $A_0 = 0$ and $\zeta_0^* = 0$. Now assume that (24) is satisfied for some $k \geq 0$ and let us prove that it is also satisfied for the next index $k' = k + 1$. We start by noting that

$$\begin{aligned}\zeta_{k+1}^* &= \zeta_{k+1}(v_{k+1}) = \zeta_k(v_{k+1}) + a_{k+1}[f(y_k) + \langle \nabla f(y_k), v_{k+1} - y_k\rangle] \\ &\geq \zeta_k^* + \frac{1}{2}\|v_{k+1} - v_k\|^2 + a_{k+1}[f(y_k) + \langle \nabla f(y_k), v_{k+1} - y_k\rangle] \\ &\geq A_k f(x_k) + \frac{1}{2}\|v_{k+1} - v_k\|^2 + a_{k+1}[f(y_k) + \langle \nabla f(y_k), v_{k+1} - y_k\rangle], \end{aligned} \tag{25}$$

where the first inequality holds due to the strong convexity of ζ_k, and the second one is due to the induction hypothesis. Further, note that, by construction, $y_k \in [v_k, x_k]$. Considering separately any of the three possible situations, $y_k = v_k$, $y_k = x_k$ and $y_k \in (v_k, x_k)$, we see that, in all cases,

$$\langle \nabla f(y_k), v_k - y_k\rangle \geq 0.$$

Then, by adding and subtracting $\langle \nabla f(y_k), v_k\rangle$ in (25) and using the estimate in the preceding inequality, as well as $f(y_k) \leq f(x_k)$ (which holds by construction), we obtain

$$\begin{aligned}\zeta_{k+1}^* &\geq A_k f(x_k) + a_{k+1} f(y_k) + \frac{1}{2}\|v_{k+1} - v_k\|^2 + a_{k+1}\langle \nabla f(y_k), v_{k+1} - v_k\rangle \\ &\geq A_{k+1} f(y_k) + \frac{1}{2}\|v_{k+1} - v_k\|^2 + a_{k+1}\langle \nabla f(y_k), v_{k+1} - v_k\rangle \\ &\geq A_{k+1} f(y_k) - \frac{a_{k+1}^2}{2}\|\nabla f(y_k)\|^2 = A_{k+1}\Big[f(y_k) - \frac{1}{2L}\|\nabla f(y_k)\|^2\Big],\end{aligned}$$

where the final identity is due to fact that, by the definition of a_{k+1},

$$L a_{k+1}^2 = A_{k+1}. \tag{26}$$

By the induction hypothesis and by construction, y_k stays in the sublevel set $\mathcal{F}_0$, since $f(y_k) \leq f(x_k) \leq f(x_0)$. Then, by the definition of x_{k+1} and (20), it holds that $f(x_{k+1}) \leq f(y_k) - \frac{1}{2L}\|\nabla f(y_k)\|^2$ $(\leq f(y_k))$. This proves that $\zeta_{k+1}^* \geq A_{k+1} f(x_{k+1})$ completing the induction.

Let $k \geq 1$ be arbitrary. By the convexity of f and the definition of A_k, we have

$$\zeta_k^* \leq \zeta_k(x^*) = \frac{1}{2}R^2 + \sum_{i=0}^{k-1} a_{i+1}[f(y_i) + \langle \nabla f(y_i), x^* - y_i\rangle] \leq \frac{1}{2}R^2 + A_k f^*.$$

where $R = \|x_0 - x^*\|$. Combining this with (24), we conclude that

$$f(x_k) - f^* \leq \frac{LR^2}{2A_k}. \tag{27}$$

It remains to estimate the rate of growth of the coefficients A_k. From (26) and the definition of A_{k+1}, it follows, for any $k \geq 0$, that

$$\begin{aligned}\sqrt{\frac{A_{k+1}}{L}} = a_{k+1} = A_{k+1} - A_k &= (\sqrt{A_{k+1}} + \sqrt{A_k})(\sqrt{A_{k+1}} - \sqrt{A_k}) \\ &\leq 2\sqrt{A_{k+1}}(\sqrt{A_{k+1}} - \sqrt{A_k}).\end{aligned}$$

Dividing both sides of this inequality by $\sqrt{A_{k+1}}$ and summing up the result, we see that, for any $k \geq 1$, it holds that

$$A_k \geq \frac{k^2}{4L}.$$

Substituting this estimate into (27), we obtain the claim. $\square$

E.2 Proof of Theorem 6.2

Proof. Based on the second estimate in Theorem 3.2, GM with stepsizes (9), (9), or (13) finds a point x_0, such that $f(x_0) - f^* \le \Delta$ with Δ defined by (19) in the following number of iterations / oracle queries (note that one iteration needs precisely one gradient computation):

$$N_0 := \left\lceil \frac{2+\frac{3}{e}}{a} \frac{L_0 R^2}{\frac{L_0}{5L_1^2}} + \frac{3(1+\frac{1}{e})}{a} L_1^2 R^2 \right\rceil = \left\lceil \frac{13+\frac{18}{e}}{a} L_1^2 R^2 \right\rceil,$$

where $R := \|x_s - x^*\|$, and a is an absolute constant from the statement depending on the stepsize rule. Moreover, for the obtained point x_0, it holds that $R_0 := \|x_0 - x^*\| \le R$.

Further, as discussed in the paragraph after Theorem 6.1, for the specific value of Δ we have chosen, our update rule $T(\cdot)$ satisfies the sufficient decrease property (20) with L given by (21). Hence, by Theorem 6.1, after $k \ge 1$ iterations of AGMsDR, we have

$$f(x_k) - f^* \le \frac{2LR_0^2}{k^2} = \frac{5L_0 R_0^2}{ak^2} \le \frac{5L_0 R^2}{ak^2}.$$

To obtain $f(x_k) - f^* \le \epsilon$, it therefore suffices to perform the following number of iterations:

$$K(\epsilon) := \left\lceil \sqrt{\frac{5L_0R^2}{a\epsilon}} \right\rceil.$$

Each iteration of AGMsDR requires one computation of the gradient plus at most m oracle queries for the line search. Hence, the oracle complexity of AGMsDR is at most

$$N(\epsilon) := (m+1)K(\epsilon).$$

Summing up N_0 and $N(\epsilon)$, we obtained the claimed complexity. $\square$

F Complexity of NAG

Unfortunately, the NAG algorithm presented in Li et al. (2023) is not scale-invariant and its complexity reported in (Li et al., 2023, Theorem 4.4) is not written explicitly. To streamline the comparison of the complexity bound for NAG with those for other methods for minimizing an (L_0, L_1)-smooth function, we provide a simple fix making the algorithm scale-invariant and also rewrite the result of (Li et al., 2023, Theorem 4.4) (assuming it is true) in an explicit form.

Theorem F.1. *Consider problem* (1) *with an (L_0, L_1)-smooth convex function f assuming $L_0 > 0$. Let NAG Li et al. (2023) be applied to solving the rescaled version of this problem:*

$$\tilde{f}^* := \min_{x \in \mathbb{R}^d} \Big\{ \tilde{f}(x) := \frac{1}{L_0} f(x) \Big\},$$

starting from a certain point $x_0 \in \mathbb{R}^d$. Then, for an appropriate choice of parameters, NAG finds a point $\bar{x} \in \mathbb{R}^d$ such that $f(\bar{x}) - f^* \le \epsilon$ for a given $\epsilon > 0$ after at most the following number of iterations / gradient-oracle queries:

$$16\Big(128 L_1^2 R^2 + \frac{128 L_1^2 F_0}{L_0} + 1\Big) \sqrt{\frac{F_0 + L_0 R^2}{\epsilon}},$$

where $F_0 := f(x_0) - f^*$, $R := \|x_0 - x^*\|$ and x^* is an arbitrary solution of our problem.

Proof. By construction, $\tilde{f}$ is an $(\tilde{L}_0, \tilde{L}_1)$-smooth with $\tilde{L}_0 = 1$ and $\tilde{L}_1 = L_1$. In the terminology of Li et al. (2023), this means that $\tilde{f}$ is ℓ-smooth w.r.t. the function

$$\ell(G) := \tilde{L}_0 + \tilde{L}_1 G \equiv 1 + L_1 G.$$

Theorem 4.4 from Li et al. (2023) then tells us that the sequence of the iterates $\{x_t\}$ constructed by NAG satisfies

$$\tilde{f}(x_t) - \tilde{f}^* \le \frac{4(\tilde{F}_0 + R^2)}{\eta t^2 + 4}, \tag{28}$$

where $\tilde{F}_0 := \tilde{f}(x_0) - \tilde{f}^*$, $R := \|x_0 - x^*\|$, and $\eta > 0$ is the stepsize parameter required to satisfy

$$\eta \leq \min\Big\{\frac{1}{16[\ell(2G)]^2}, \frac{1}{2\ell(2G)}\Big\} \equiv \frac{1}{16[\ell(2G)]^2} \equiv \frac{1}{16(1+2L_1G)^2}, \tag{29}$$

where G is an arbitrary constant such that

$$G \geq \max\{8\sqrt{\ell(2G)(\tilde{F}_0 + R^2)}, \tilde{g}_0\} \equiv \max\{8\sqrt{(1+2L_1G)(\tilde{F}_0 + R^2)}, \tilde{g}_0\}. \tag{30}$$

where $\tilde{g}_0 := \|\nabla \tilde{f}(x_0)\|$. In terms of our original function f, the guarantee (28) reads

$$f_t := f(x_t) - f^* \leq \frac{4(F_0 + L_0R^2)}{\eta t^2 + 4}.$$

To achieve the fastest possible convergence, we select the largest possible stepsize η which is, according to (29),

$$\eta = \frac{1}{16(1+2L_1G)^2}.$$

Substituting this formula into the previous display and dropping the (useless for improving the convergence rate) constant 4 from the denominator, we obtain

$$f_t \leq \frac{64(1+2L_1G)^2(F_0 + L_0R^2)}{t^2} \leq \epsilon$$

whenever

$$t \geq 8(1+2L_1G)\sqrt{\frac{F_0 + L_0R^2}{\epsilon}} =: t(\epsilon). \tag{31}$$

The obtained $t(\epsilon)$ is exactly the iteration complexity of the algorithm for obtaining an ϵ-approximate solution for the original problem, and is also its gradient oracle complexity since the method makes precisely one gradient-oracle query at each iteration.

It remains to choose the smallest possible parameter G satisfying (30). We start with rewriting this inequality in terms of the original function:

$$G \geq \max\Big\{8\sqrt{(1+2L_1G)\Big(\frac{F_0}{L_0} + R^2\Big)}, \frac{g_0}{L_0}\Big\} \equiv \max\Big\{\sqrt{(1+2L_1G)\Delta}, \frac{g_0}{L_0}\Big\}$$

where $g_0 := \|\nabla f(x_0)\|$ and $\Delta := 64(\frac{F_0}{L_0} + R^2)$. This inequality is equivalent to the system of two inequalities:

$$G^2 \geq (1+2L_1G)\Delta, \qquad G \geq \frac{g_0}{L_0}.$$

Rearranging, we see that the first inequality is equivalent to

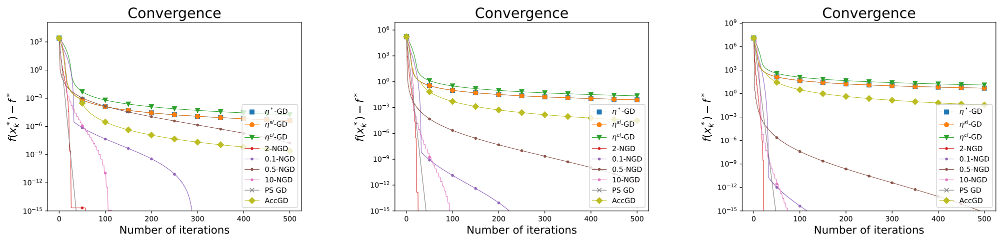
$$G \geq \sqrt{\Delta + L_1^2\Delta^2} + L_1\Delta =: G_*$$

Further, it turns out that $G_* \geq \frac{g_0}{L_0}$. Indeed, according to (7), we have $F_0 \geq \frac{g_0^2}{2L_0 + 3L_1g_0}$, meaning that $g_0 \leq \sqrt{2L_0F_0 + \frac{9}{4}L_1^2F_0^2} + \frac{3}{2}L_1F_0 \leq \sqrt{2L_0F_0} + 3L_1F_0$; on the other hand, estimating $\Delta \geq \frac{64F_0}{L_0}$, we see that $L_0(\sqrt{\Delta} + L_1\Delta) \geq 8\sqrt{L_0F_0} + 64L_1F_0$. Thus, the smallest possible value of G satisfying the original requirement (30) is in fact $G = G_*$.

Choosing now $G = G_*$ and substituting the definition of Δ, we obtain

$$1 + 2L_1G = \frac{G_*^2}{\Delta} \leq \frac{2(\Delta + L_1^2\Delta^2) + 2L_1^2\Delta^2}{\Delta} = 2(1 + 2L_1^2\Delta) = 2\Big(1 + \frac{128L_1^2F_0}{L_0} + 128L_1^2R^2\Big).$$

Substituting this bound into (31), we obtain the claimed bound on $t(\epsilon)$. $\square$



(a) $p = 4$ (b) $p = 6$ (c) $p = 8$

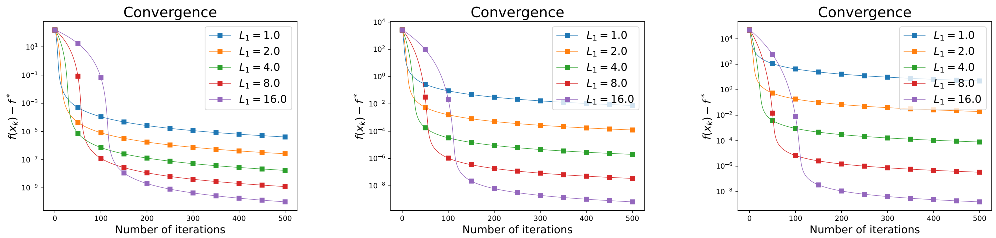
Figure 1: Comparison of gradient methods for $f(x) = \frac{1}{p}\|x\|^p$. $\frac{\hat{R}}{R}$-NGD stands for Normalized Gradient Method, where $\hat{R}$ is an estimation of the true initial distance to a solution R. η_*-GD, η^{si}-GD, η^{cl}-GD stand for gradient method with stepsizes (9), (12) and (13) respectively, PS GD stands for Polyak stepsizes gradient method, and AccGD stands for two-stage accelerated procedure (Algorithm 2).

G Numerical Results

In Fig. 1, we compare the performance of the analyzed methods for solving optimization problem (1) with a function $f(x) = \frac{1}{p}\|x\|^p$. To generate a starting point we We fix $L_1 = 1$ and choose $L_0 = (\frac{p-2}{L_1})^{p-2}$ according to Example A.1. For GM, we choose stepsizes according to (9), (12) and (13). For NGM, we use time-varying coefficients $\beta_k = \frac{\hat{R}}{k+1}$ with different values of $\hat{R} \in \{\frac{1}{2}R, 2R, 10R\}$, which allows us to study the robustness of this method to our initial guess of the unknown initial distance to the solution. Note that, for this particular problem, the choice of $\hat{R} = R$ is rather special and allows the method to find the exact solution after one iteration, so we are not considering it. We observe that, NGM and PS-GM outperform GM with stepsizes from (9), (12) and (13). This can be explained by the fact that the complexity of GM depends on the particular choice of (L_0, L_1), while complexity of NGM and PS-GM involves the optimal parameters L_0, L_1 as discussed in Section 4. Moreover, closer initial distance estimation $\hat{R}$ to a true value R leads to a faster convergence of NGM to a solution.

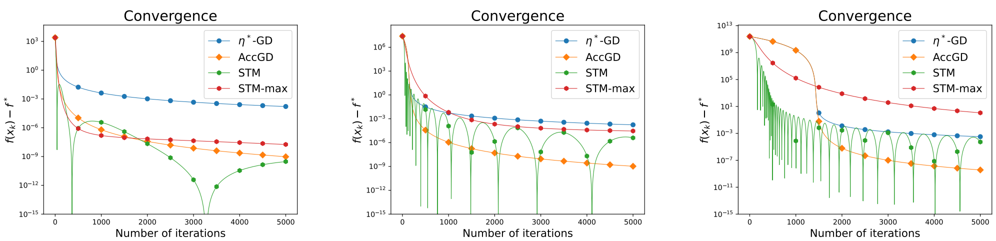
In Fig. 2, we present an experiment studying the performance of the GM with the stepsize rule (9) based on the choice of (L_0, L_1). For each choice of $L_1 \in \{1, 2, 4, 8, 16\}$ we set $L_0 = (\frac{p-2}{L_1})^{p-2}$, according to Example A.1. As expected from the theory (see the corresponding discussion at the end of Section 4), the choice of (L_0, L_1) pair is crucial in practice for the performance of GM and depends on a target accuracy ϵ.

In Fig. 3, we conduct an experiment on the performance of accelerated methods and consider the GM 9, the proposed two-stage procedure from Section 6 with $T_L(x)$ being the gradient update with stepsizes (9) and $L = 3L_0$, and two variants of normalized Similar Triangles Methods (STM, and STM-Max) from Gorbunov et al. (2024). STM uses normalization by the norm of the gradient at the current point in a gradient step, while STM-Max normalizes by the largest norm of the gradient over the optimization trajectory. It is worth noticing that only STM-Max has theoretical convergence guarantees. We set $p = 6$, $L_1 = 1$, $L_0 = (\frac{p-2}{L_1})^{p-2}$ (see Example A.1) and vary the initial distance to the solution $R = \|x_0 - x^*\|$. We observe that for a large initial distance to the solution, our proposed procedure outperforms STM-max. This fact supports our theoretical founding and reasoning on accelerated methods provided in Section 6. We also notice that the second variant STM outperforms all considered methods in terms of the best iterate convergence. However, there is no theoretical analysis for it. Additionally, we compare the performance of Algorithm 2, STM and STM-Max with good starting point x_0, such that $f(x_0) - f^* \leq \frac{L_0}{5L_1^2}$. In Fig. 4, we plot the function values residual for different values of $p \in \{4.0, 6.0, 8.0\}$. Since $f(x_0) - f^* \leq \frac{L_0}{5L_1^2}$, Algorithm 2 will run only the second stage, which is AGMsDR. We observe that for smaller values of $p = 4.0$, all three accelerated methods are comparable, while for larger values of $p \in \{6.0, 8.0\}$ STM and STM-Max outperform Algorithm 2.



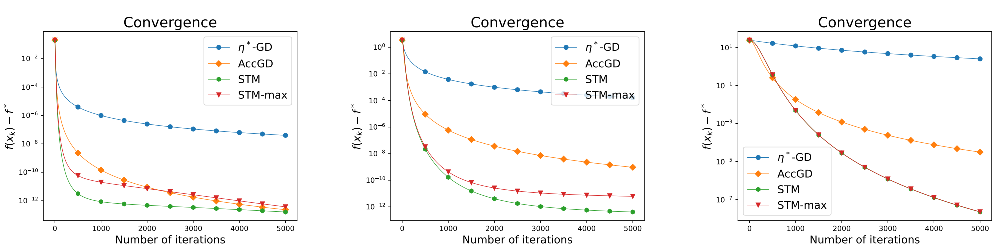
(a) $p = 4$ (b) $p = 6$ (c) $p = 8$

Figure 2: Convergence of the gradient method on the same function but with different choices of (L_0, L_1).



(a) $R = 10$ (b) $R = 100$ (c) $R = 1000$

Figure 3: Comparison of two-stage procedure (Algorithm 2) denoted by AccGD with Similar Triangles Method (SMT) and Similar Triangles Method Max (STM-max) for $f(x) = \frac{1}{6}\|x\|^6$, with different initial distance R.



(a) $p = 4$ (b) $p = 6$ (c) $p = 8$

Figure 4: Comparison of two-stage procedure (Algorithm 2) denoted by AccGD with Similar Triangles Method (SMT) and Similar Triangles Method Max (STM-max) for $f(x) = \frac{1}{p}\|x\|^p$, with different values p and a good starting point.