

Linear Convergence Rate in Convex Setup is Possible! Gradient Descent Method Variants under (L_0, L_1)-Smoothness

Aleksandr Lobanov*
MIPT, Skoltech, ISP RAS
lobbsasha@mail.ru

Alexander Gasnikov
Innopolis University, MIPT, ISP RAS
gasnikov@yandex.ru

Eduard Gorbunov
MBZUAI
eduard.gorbunov@mbzuai.ac.ae

Martin Takáč
MBZUAI
martin.takac@mbzuai.ac.ae

February 20, 2025[†]

Abstract

The gradient descent (GD) method – is a fundamental and likely the most popular optimization algorithm in machine learning (ML), with a history traced back to a paper in 1847 Cauchy (1847). It was studied under various assumptions, including so-called (L_0, L_1) -smoothness, which received noticeable attention in the ML community recently. In this paper, we provide a refined convergence analysis of gradient descent and its variants, assuming generalized smoothness. In particular, we show that (L_0, L_1) -GD has the following behavior in the *convex setup*: as long as $\|\nabla f(x^k)\| \geq \frac{L_0}{L_1}$ the algorithm has *linear convergence* in function suboptimality, and when $\|\nabla f(x^k)\| < \frac{L_0}{L_1}$ is satisfied, (L_0, L_1) -GD has standard sublinear rate. Moreover, we also show that this behavior is common for its variants with different types of oracle: *Normalized Gradient Descent* as well as *Clipped Gradient Descent* (the case when the full gradient $\nabla f(x)$ is available); *Random Coordinate Descent* (when the gradient component $\nabla_i f(x)$ is available); *Random Coordinate Descent with Order Oracle* (when only $\text{sign}[f(y) - f(x)]$ is available). In addition, we also extend our analysis of (L_0, L_1) -GD to the strongly convex case.

Contents

1	Introduction	2
1.1	Notations and main assumptions	4
1.2	Paper structure	5

*Part of the work was done while A. Lobanov was an intern at MBZUAI.

[†]The first version was submitted to arXiv on December 22, 2024. The second version contains the following changes: 1) minor inaccuracies in the proofs from Section 3 were fixed, the results remained the same; 2) major inaccuracies in the proofs from Section 4 were fixed, and the results were simplified to the uniform sampling case; 3) major inaccuracies in the results from Section 5 were fixed; 4) writing was improved, missing references were added (including the updated version of (Vankov et al., 2024a)).

2	Related Works	5
3	Full-Gradient Methods	6
3.1	Gradient descent method	7
3.2	Normalized GD method	8
3.3	Clipped GD method	9
4	Coordinate Descent Type Methods	10
4.1	Random coordinate descent	10
4.2	Random coordinate descent with Order Oracle	11
5	Extension to Strongly Convex Setup	12
6	Discussion and Future Work	13
7	Conclusion	13
A	Auxiliary Results	17
B	Monotonicity of Gradient Norms	17
C	Missing Proofs for Full-Gradient Algorithms	22
C.1	Proof of Theorem 3.1	23
C.2	Proof of Theorem 3.3	24
C.3	Proof of Theorem 3.5	25
D	Missing Proofs for Coordinate Descent Type Methods	30
D.1	Proof of Theorem 4.1	30
D.2	Proof of Theorem 4.3	33
E	Missing Proof for GD in the Strongly Convex Setup	34
F	Motivation Strong Growth Conditions on the Example of Logistic Regression	36

1 Introduction

We consider the standard unconstrained minimization

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

where $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is a convex differentiable function. This problem configuration is quite general and encompasses a broad range of applications in ML scenarios. For such problems, the traditional optimization algorithm is the *gradient descent method* (GD) Cauchy (1847), which has a sublinear convergence rate in the convex setting under the Lipschitz smoothness assumption (see, e.g., Nesterov, 2013). In particular, GD is the core of optimization for machine learning, and various modifications of this method have been studied in different assumptions suited to ML applications.

In this paper, we consider one of such assumptions called (L_0, L_1) -smoothness (Zhang et al., 2020b,a; Chen et al., 2023), which in the case of twice differentiable functions, states that $\|\nabla^2 f(x)\| \leq L_0 + L_1 \|\nabla f(x)\|$, i.e., the smoothness constant can grow as a linear function of the gradient norm. Under this assumption, different variants of GD are analyzed, including GD with clipping (Clip-GD) (Zhang et al., 2020b,a; Koloskova et al., 2023; Vankov et al., 2024b), (L_0, L_1) -GD (Gorbunov et al., 2024; Vankov et al., 2024b), Normalized GD (NGD) (Zhao et al., 2021; Chen et al., 2023; Vankov et al., 2024b), and other variants (Crawshaw et al., 2022; Wang et al., 2022; Faw et al., 2023; Wang et al., 2023; Hübner et al., 2024; Li et al., 2024b). More precisely, in the deterministic convex case, the state-of-the-art results for Clip-GD, (L_0, L_1) -GD, and NGD are obtained by Gorbunov et al. (2024); Vankov et al. (2024b) showing the $\mathcal{O}\left(\frac{L_0 R^2}{N}\right)$ rates for function suboptimality when $N = \Omega(L_1^2 R^2)$ ¹ leaving open questions about the refined methods behavior characterization for $N = \mathcal{O}(L_1^2 R^2)$.

However, beyond the first-order methods, the algorithms for (L_0, L_1) -smooth optimization are weakly studied. In particular, *random coordinate descent* (RCD) (Nesterov (2012); Shalev-Shwartz and Tewari (2009); Richtárik and Takáč (2016)), which is useful in the situations when the computation of the full gradient is prohibitively expensive, is not analyzed in the context of (L_0, L_1) -smooth optimization. Moreover, in some cases, e.g., in the reinforcement learning with human feedback (Tang et al., 2024), even objective values are available, and for given points $x, y \in \mathbb{R}^d$ one can only evaluate $\text{sign}[f(y) - f(x)]$. To the best of our knowledge, there are no theoretical convergence results for such methods under (L_0, L_1) -smoothness, and, in particular, the convergence of *random coordinate descent with order oracle* (OrderRCD) (Lobanov et al., 2024) is not studied in this setup.

In this paper, we address this gap in the literature and provide the first analysis of RCD and OrderRCD for convex (L_0, L_1) -smooth optimization. Moreover, we improve the existing results for (L_0, L_1) -GD, NGD, and Clip-GD: we prove that these methods enjoy *linear convergence rates without any additional assumptions* for the initial optimization phase when $\|\nabla f(x^k)\| \geq \frac{L_0}{L_1}$.

Our contributions can be summarized as follows.

- We show under what conditions variants of the gradient descent achieve linear convergence in the convex setup.
- We prove better complexity bounds for (L_0, L_1) -GD, NGD, and Clip-GD than previously known ones, assuming convexity and (L_0, L_1) -smoothness of the objective function. We show that these algorithms converge linearly at first, and slow down as they approach the solution, converging sublinearly. We also show that for the phase of convergence of NGD, when the iterates satisfy $\|\nabla f(x^k)\| \geq c$, the method converges linearly. Table 1 demonstrates the conditions under which Clip-GD converges linearly. In particular, the case of $\lambda_k = 1$ corresponds to the convergence of GD, and the case of $\lambda_k = \frac{c}{\|\nabla f(x^k)\|}$ corresponds to the convergence of NGD.
- We provide the first convergence results for the RCD and OrderRCD algorithms under the convexity and (L_0, L_1) -coordinate smoothness assumptions. We demonstrate that the linear convergence phenomenon of the full-gradient methods exists for both of the mentioned methods.

¹After the first version of our work appeared on arXiv, Vankov et al. (2024b) let us know that they also independently derived $\mathcal{O}\left(\frac{L_0 R^2}{N} + \left(1 - \frac{1}{L_1 R}\right)^N F_0\right)$ rate for (L_0, L_1) -GD, where $F_0 = f(x^0) - f(x^*)$ during the discussion with reviewers of their work to a ML conference. Then, the authors updated their paper on arXiv (Vankov et al., 2024a). At the moment of writing our paper, we were unaware of the updated version of (Vankov et al., 2024a).

Table 1: Comparison of the convergence rates for Clip-GD in the convex case. Clip-GD update scheme: $x^{k+1} = x^k - \eta_k \cdot \text{clip}_c(\nabla f(x^k))$. Notation: $\text{clip}_c(\nabla f(x^k)) = \lambda_k \cdot \nabla f(x^k)$; $\lambda_k = \min\{1, c/\|\nabla f(x^k)\|\}$; $c > 0$ – clipping radius; $\eta_k > 0$ – step size; N = number of iterations; $F_0 = f(x^0) - f^*$; $R = \|x^0 - x^*\|$; $T = \min\{k \in \{0, 1, \dots, N-1\} \mid \|\nabla f(x^k)\| < L_0/L_1\}$; LCR = linear convergence rate.

Reference	Clipping threshold	λ_k	Smoothness case: L_0 (?) cL_1	Convergence rate $f(x^N) - f^* \lesssim$	LCR?
Koloskova et al. (2023)	arbitrary	1	larger	$\mathcal{O}\left(\frac{L_0 R^2}{N}\right)$	✗
			less or equal	$\mathcal{O}\left(\frac{c L_1 R^2}{N}\right)$	✗
		$\frac{c}{\ \nabla f(x^k)\ }$	larger	$\mathcal{O}\left(\frac{L_0^2 L R^4}{c^2 N^2}\right)$	✗
			less or equal	$\mathcal{O}\left(\frac{L_1^2 L R^4}{N^2}\right)$	✗
Gorbunov et al. (2024) & Vankov et al. (2024b)	$c = \frac{L_0}{L_1}$	1	equal	$\mathcal{O}\left(\frac{L_0 R^2}{N}\right)$	✗
		$\frac{c}{\ \nabla f(x^k)\ }$	equal	✗	✗
Theorem 3.5 (Our work)	arbitrary	1	larger	$\mathcal{O}\left(\frac{L_0 R^2}{N}\right)$	✗
			less or equal	$\mathcal{O}\left(\min\left\{\frac{L_0 R^2}{N-T}, \left(1 - \frac{1}{L_1 R}\right)^T F_0\right\}\right)$	✓
		$\frac{c}{\ \nabla f(x^k)\ }$	larger	$\mathcal{O}\left(\left(1 - \frac{c}{L_0 R}\right)^N F_0\right)$	✓
			less or equal	$\mathcal{O}\left(\left(1 - \frac{1}{L_1 R}\right)^N F_0\right)$	✓

- We extend our analysis of (L_0, L_1) -GD to the case when the function is μ -strongly convex.

1.1 Notations and main assumptions

Before discussing related work, we first introduce the notations and assumptions that are used in this paper.

Notations. We use $\langle x, y \rangle := \sum_{i=1}^d x_i y_i$ to denote standard inner product of $x, y \in \mathbb{R}^d$. We denote Euclidean norm in \mathbb{R}^d as $\|x\| := \sqrt{\sum_{i=1}^d x_i^2} = \sqrt{\langle x, x \rangle}$. We use $\mathbf{e}_i \in \mathbb{R}^d$ to denote the i -th unit vector. For $\mathbf{L} = (L^{(1)}, \dots, L^{(d)})^\top \in \mathbb{R}^d$ and $\alpha \in \mathbb{R}$, we define the norms $\|x\|_{[\mathbf{L}, \alpha]} := \sqrt{\sum_{i=1}^d (L^{(i)})^\alpha x_i^2}$ and $\|x\|_{[L_p, \alpha]}^* := \sqrt{\sum_{i=1}^d \frac{1}{(L_p^{(i)})^\alpha} x_i^2}$. We denote by $\nabla f(x)$ the full gradient of function f at point $x \in \mathbb{R}^d$, and by $\nabla_i f(x)$ the i -th coordinate gradient of function f at point $x \in \mathbb{R}^d$. We also introduce $S_\alpha^\mathbf{L} := \sum_{i=1}^d (L^{(i)})^\alpha$. We use $\tilde{O}(\cdot)$ to hide the logarithmic coefficients. We denote $f^* := f(x^*)$ and $x^* \in X^* := \arg \min_{x \in \mathbb{R}^d} f(x)$ to be any solution of (1). We also use $R := \|x^0 - x^*\|$ and $F_0 := f(x^0) - f^*$.

The most common assumption about smoothness in the literature (see, e.g., Nesterov, 2013) is L -smoothness.

Assumption 1.1 (L -smoothness). Function f is L -smooth if the following inequality is satisfied for any $x, y \in \mathbb{R}^d$:

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

However, instead of standard L -smoothness, we focus on the so-called (L_0, L_1) -smoothness (Zhang et al., 2020b,a).

Assumption 1.2 ((L_0, L_1) -smoothness). Function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is (L_0, L_1) -smooth if the following inequality is satisfied for any $x, y \in \mathbb{R}^d$ with $\|y - x\| \leq \frac{1}{L_1}$:

$$\|\nabla f(y) - \nabla f(x)\| \leq (L_0 + L_1 \|\nabla f(x)\|) \|y - x\|. \quad (2)$$

If $L_1 = 0$, the above assumption recovers Assumption 1.1 with $L = L_0$. Moreover, (L_0, L_1) -smoothness is strictly more general than L -smoothness, see the examples in Zhang et al. (2020b); Chen et al. (2023); Koloskova et al. (2023); Gorbunov et al. (2024).

Next, we also use a coordinate-wise version of Assumption 1.2 introduced by Crawshaw et al. (2022).

Assumption 1.3 ((L_0, L_1) -coordinate-smoothness). A function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is (L_0, L_1) -coordinate-smooth for $L_0^{(1)}, L_0^{(2)}, \dots, L_0^{(d)}, L_1^{(1)}, L_1^{(2)}, \dots, L_1^{(d)} \geq 0$ if for any $i \in [d]$, $x \in \mathbb{R}^d$ and $h \in \mathbb{R}$, $|h| \leq \frac{1}{\max_{i \in [d]} L_1^{(i)}}$ the following inequality holds:

$$|\nabla_i f(x + h \mathbf{e}_i) - \nabla_i f(x)| \leq \left(L_0^{(i)} + L_1^{(i)} |\nabla_i f(x)| \right) |h|.$$

The above assumption generalizes the standard coordinate L -smoothness (Lin et al., 2014; Allen-Zhu et al., 2016; Zhang and Xiao, 2017) similarly to how (L_0, L_1) -smoothness generalizes L -smoothness.

We also assume that the function f is $(\mu$ -strongly) convex.

Assumption 1.4. Function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is $\mu \geq 0$ strongly convex if for any $x, y \in \mathbb{R}^d$ the following inequality holds:

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|^2. \quad (3)$$

Assumption 1.4 is classical and widely used in the literature (see, e.g., Boyd and Vandenberghe, 2004; Nesterov, 2018).

1.2 Paper structure

Further, our paper has the following structure. In Section 2, we discuss the related work. In Section 3, we provide the results for full-gradient methods. The case where the oracle only has access to the gradient coordinate or the comparison of the values of two functions is considered in Section 4. In Section 5, we generalize our results for (L_0, L_1) -GD to the strongly convex case. Discussions of this work and future work plans are given in Section 6. Section 7 concludes the paper. All missing proofs of the theoretical results are provided in the Appendix.

2 Related Works

The literature on the analysis of GD-type methods is very rich. Below, we discuss only closely related works.

Full-gradient methods for the (L_0, L_1) -smooth convex optimization. Although most of the existing works on (L_0, L_1) -smoothness focus on the non-convex case, there are several papers considering the (strongly) convex problems as well. Koloskova et al. (2023) gives the first analysis Clip-GD (Pascanu et al., 2013) under (L_0, L_1) -smoothness and L -smoothness and proves $\mathcal{O}\left(\max\left\{\frac{(L_0+cL_1)R^2}{N}, \frac{R^4L(L_0+cL_1)^2}{c^2N^2}\right\}\right)$ rate (see Table 1 for the details). This bound is derived under the additional L -smoothness assumption, which is not always satisfied for (L_0, L_1) -smooth problems. Moreover, when $\lambda_k = 1$ and $L_0 \leq cL_1$, the derived rate is proportional to c . In addition, the analysis from (Koloskova et al., 2023) implies the sublinear convergence rate for NGD (see the case $\lambda_k = \frac{c}{\|\nabla f(x^k)\|}$ in Table 1). Takezawa et al. (2024) derive similar results for GD with Polyak Stepsizes (GD-PS), i.e., they show $\mathcal{O}\left(\max\left\{\frac{L_0R^2}{N}, \frac{R^4LL_1^2}{c^2N^2}\right\}\right)$ convergence rate. Next, Li et al. (2024a) derive convergence rates for GD and its accelerated version under (r, ℓ) -smoothness assumption, which generalizes (L_0, L_1) -smoothness. In particular, for GD Li et al. (2024a) prove $\mathcal{O}(\frac{\ell R^2}{N})$ convergence rate, where $\ell = \mathcal{O}(L_0 + L_1G)$ and constant G depends in $L_0, L_1, R, \|\nabla f(x^0)\|$, and $f(x^0) - f^*$, meaning that it can be exponentially large in terms of L_1 and R . Finally, Gorbunov et al. (2024); Vankov et al. (2024b) independently improve the convergence rates of (L_0, L_1) -GD/Clip-GD by considering the special case of clipping radius $c = \frac{L_0}{L_1}$. More precisely, they prove $\mathcal{O}\left(\frac{L_0R^2}{N}\right)$ convergence rate if $N \geq L_1^2R^2$ and extend this result to GD-PS (Vankov et al. (2024b) also show a similar result for NGD). However, the results from Gorbunov et al. (2024); Vankov et al. (2024b) do not provide convergence rates in terms of $f(x^N) - f^*$ for the stage when $\|\nabla f(x^k)\| > L_0/L_1$, which can be noticeable when L_0 is small and L_1 is large. *In our work, we propose the analysis that addresses this limitation (see Table 1).*

Coordinate descent type methods. Convergence of coordinate methods is also relatively well-studied. For example, under the standard L -smoothness assumption $((\nabla^2 f(x))_{i,i} \leq L)$, the coordinate descent (CD) method has the following convergence rate $\mathcal{O}\left(\frac{dLR^2}{N}\right)$ (see, e.g., Bubeck et al., 2015). Using the fact that $\frac{1}{d} \sum_{i=1}^d L^{(i)} \leq L$ and assuming L -coordinate-smoothness $((\nabla^2 f(x))_{i,i} \leq L^{(i)})$, the previous result can be improved to $\mathcal{O}\left(\frac{\sum_{i=1}^d L^{(i)}R^2}{N}\right)$ rate. Next, assuming that the active coordinate i_k can be obtained (independently) from the distribution $p_\alpha(i) = (L^{(i)})^\alpha/s_\alpha$, then RCD converges at $\mathcal{O}\left(\frac{S_\alpha R_{[\mathbf{L}, 1-\alpha]}^2}{N}\right)$ rate Nesterov (2012), where

$R_{[\mathbf{L}, 1-\alpha]} := \max_{x \in \mathbb{R}^d} \left\{ \max_{x^* \in X^*} \|x - x^*\|_{[\mathbf{L}, 1-\alpha]} : f(x) \leq f(x^0) \right\}$. Moreover, Lobanov et al. (2024) show that it is possible to create an OrderRCD algorithm based on RCD, whose oracle has access only to function comparisons (this oracle can be motivated by, e.g., RLHF (Ouyang et al., 2022; Bai et al., 2022)). More precisely, Lobanov et al. (2024) prove that the iteration complexity of OrderRCD is the same as for RCD, and the oracle complexity is inferior only in $\log(1/\epsilon)$ factor, where ϵ is the accuracy of the solution of the linear search problem. *In our paper, we extend these results to the more general case of (L_0, L_1) -smoothness.*

3 Full-Gradient Methods

In this section, we present our result for full-gradient algorithms (see GD in Subsection 3.1, NGD in Subsection 3.2, and Clip-GD in Subsection 3.3).

3.1 Gradient descent method

The first algorithm we consider has the following algorithm:

Algorithm 1 Gradient Descent Method (GD)

Input: $x_0 \in \mathbb{R}^d$, iterations number N , step size $\eta_k > 0$
for $k = 0$ **to** $N - 1$ **do**
 $x^{k+1} \leftarrow x^k - \eta_k \nabla f(x^k)$
end for
Return: x^N

We prove the following result for Algorithm 1 with stepsize $\eta_k = (L_0 + L_1 \|\nabla f(x^k)\|)^{-1}$ (to emphasize the specificity of the step size we call it (L_0, L_1) -GD for brevity).

Theorem 3.1. *Let function f satisfy Assumption 1.2 $((L_0, L_1)$ -smoothness) and Assumption 1.4 (convexity, $\mu = 0$), then GD (Algorithm 1) with step size $\eta_k = (L_0 + L_1 \|\nabla f(x^k)\|)^{-1}$ guarantees*

- *linear convergence, if $\|\nabla f(x^k)\| \geq \frac{L_0}{L_1}$ for $k \in [N - 1]$*

$$f(x^N) - f^* \leq \left(1 - \frac{1}{4L_1R}\right)^N F_0;$$

- *sublinear convergence, if $\|\nabla f(x^{N-1})\| < \frac{L_0}{L_1}$:*

$$f(x^N) - f^* < \frac{4L_0R^2}{N}.$$

In the general case, the convergence rate is

$$f(x^N) - f^* \leq \min \left\{ \frac{4L_0R^2}{N - T}, \left(1 - \frac{1}{4L_1R}\right)^T F_0 \right\},$$

where $T \geq 0$ is the smallest index such as $\|\nabla f(x^T)\| < \frac{L_0}{L_1}$.

Given the monotonicity of the gradient norm (see Appendix B), Theorem 3.1 characterizes in details the convergence behavior of GD for convex (L_0, L_1) -smooth problems. More precisely, as long as the gradient norm is larger than L_0/L_1 , GD converges with linear rate, but when the method approaches the solution ($\|\nabla f(x^k)\| < L_0/L_1$) the convergence slows down to the standard sublinear rate. That is, $\mathcal{O}(\frac{L_0R^2}{N})$ rate is common to the previous works (Gorbunov et al., 2024; Vankov et al., 2024b) (see Table 1). However, in contrast to (Gorbunov et al., 2024; Vankov et al., 2024b), our analysis shows $\mathcal{O}(1 - 1/L_1R)^N F_0$ rate when $\|\nabla f(x^k)\| \geq L_0/L_1$ (see the case of $\lambda_k = c/\|\nabla f(x^k)\|$ in Table 1). Moreover, our result significantly improves the one from (Koloskova et al., 2023) (see smoothness case “less or equal” with $\lambda_k = c/\|\nabla f(x^k)\|$ in Table 1) from sublinear to linear and

gets rid of potentially large parameter $c \geq L_0/L_1$. The proof of the Theorem 3.1 is provided in Appendix C.1.

The significance of the improved estimate can be observed under the assumption of strong growth condition².

Remark 3.2. Theorem 3.1 implies that under Assumption 1.2 with $L_0 = 0$, Algorithm 1 converges to the desired accuracy ε ($f(x^N) - f^* \leq \varepsilon$) after $N = \mathcal{O}\left(L_1 R \log \frac{F_0}{\varepsilon}\right)$ iterations.

The result of Remark 3.2 significantly outperforms all known results in this regime. In particular, Koloskova et al. (2023) show $\mathcal{O}(L_1 c R^2/\varepsilon)$ complexity bound for Clip-GD, and Gorbunov et al. (2024); Vankov et al. (2024b) do not provide explicit rates in this case.

3.2 Normalized GD method

From the previous section, we see that GD with step size with step size $\eta_k = (L_0 + L_1 \|\nabla f(x^k)\|)^{-1}$ enjoys linear convergence in the convex setting, when $\|\nabla f(x^k)\| \geq L_0/L_1$. However, in this regime, we have $L_1 \|\nabla f(x^k)\| \geq L_0$, meaning that $(2L_1 \|\nabla f(x^k)\|)^{-1} \leq \eta_k \leq (L_1 \|\nabla f(x^k)\|)^{-1}$, i.e., the method is very close to NGD (Algorithm 2). Therefore, it is natural to expect similar behavior from NGD as for GD.

Algorithm 2 Normalized Gradient Descent Method (NGD)

Input: $x_0 \in \mathbb{R}^d$, iterations number N , step size $\eta_k > 0$
for $k = 0$ **to** $N - 1$ **do**
 if $\|\nabla f(x^k)\| = 0$ **then**
 Return: x^k
 end if
 $x^{k+1} \leftarrow x^k - \eta_k \frac{\nabla f(x^k)}{\|\nabla f(x^k)\|}$
end for
Return: x^N

The following result formalizes this observation.

Theorem 3.3. *Let function f satisfy Assumption 1.2 ((L_0, L_1) -smoothness) and Assumption 1.4 (convexity, $\mu = 0$), then Algorithm 2 with step size $\eta_k = \eta \leq c/(L_0 + L_1 c)$, where constant $c > 0$ is such that $\|\nabla f(x^k)\| \geq c$ for all $k = 0, 1, \dots, N - 1$, has linear convergence:*

$$f(x^N) - f^* \leq \left(1 - \frac{\eta}{2R}\right)^N F_0.$$

Theorem 3.3 shows that in the case of $c \geq L_0/L_1$, NGD has $\mathcal{O}\left((1 - \frac{1}{L_1 R})^N F_0\right)$ convergence rate similarly to GD, which is natural to expect due to $\|\nabla f(x^k)\| \geq c$ and the discussion given in the beginning of this subsection. However, if we select large enough N , one has to select c small enough

²We refer to Assumption 1.2 with $L_0 = 0$ as strong growth condition for smoothness assumption by analogy with Vaswani et al. (2019) for variance.

such that $\|\nabla f(x^k)\| \geq c$ holds for all $k = 0, 1, \dots, N-1$. If $c < L_0/L_1$, then the rate reduces to $\mathcal{O}((1 - c/(L_0 R))^N F_0)$ and the method is guaranteed to converge only to the error $\varepsilon \sim cR$. Therefore, to guarantee the convergence to ε -accuracy, one has to take $c \sim \varepsilon/R$ in the worst case. In this case, our result implies $\mathcal{O}(L_0 R^2 \log(F_0/\varepsilon)/\varepsilon)$ complexity for NGD. However, hyperparameter c depends only on the gradient norm, so in problems where the high accuracy on gradient norm is not required, Algorithm 2 is efficient and shows linear convergence. The proof of Theorem 3.3 see Appendix C.2.

Remark 3.4. Theorem 3.3 implies that under Assumption 1.2 with $L_0 = 0$, Algorithm 2 converges to the desired accuracy ε ($f(x^N) - f^* \leq \varepsilon$) after $N = \mathcal{O}(L_1 R \log \frac{F_0}{\varepsilon})$ iterations.

As previously noted, when $\|\nabla f(x^k)\| \geq \frac{L_0}{L_1}$ GD and NGD with $c \geq \frac{L_0}{L_1}$ are almost the same. Therefore, the result of Remark 3.4 is expected given Remark 3.2. Moreover, our results imply that NGD has $\mathcal{O}(\max\{L_0 R^2 \log(F_0/\varepsilon)/\varepsilon, L_1 R \log(F_0/\varepsilon)\})$ complexity. Compared to $\mathcal{O}(\max\{L_0 \bar{R}^2/\varepsilon, L_1^2 \bar{R}^2\})$ complexity bound derived for NGD with $\eta_k = \hat{R}/\sqrt{N+1}$, $\bar{R} := \hat{R} + R^2/\hat{R}$ by Vankov et al. (2024b), our bound has an additional logarithmic factor in the first term but has much better second term when $L_1 R$ is large and $\log(F_0/\varepsilon)$ is much smaller than $L_1 R$.

3.3 Clipped GD method

In this section, we consider Clip-GD (Algorithm 3), which applies the clipping operator to the gradient:

$$\text{clip}_c(\nabla f(x)) = \min \left\{ 1, \frac{c}{\|\nabla f(x)\|} \right\} \nabla f(x), \quad (4)$$

where $c > 0$ is the clipping radius. Clip-GD can also be seen as a combination of GD (when $\|\nabla f(x^k)\| \leq c$) and NGD (when $\|\nabla f(x^k)\| > c$).

Algorithm 3 Clipped Gradient Descent Method (Clip-GD)

Input: initial point $x_0 \in \mathbb{R}^d$, iterations number N , step size $\eta_k > 0$ and clipping radius $c > 0$
for $k = 0$ **to** $N - 1$ **do**
 $x^{k+1} \leftarrow x^k - \eta_k \cdot \text{clip}_c(\nabla f(x^k))$ according to (4)
end for
Return: x^N

Then, following similar reasoning as in the previous sections, we obtain the next convergence result for Clip-GD method.

Theorem 3.5. *Let function f satisfy Assumption 1.2 ((L_0, L_1) -smoothness) and Assumption 1.4 (convexity, $\mu = 0$), then Algorithm 3 with step size $\eta_k = (L_0 + L_1 \min\{\|\nabla f(x^k)\|, c\})^{-1}$ guarantees the following error:*

$$f(x^N) - f^* = \mathcal{O} \left(\min \left\{ \frac{L_0 R^2}{N - T}, \left(1 - \frac{\rho}{R} \right)^T F_0 \right\} \right),$$

where $\rho := c/\max\{L_0, L_1 c\}$ and $T \geq 0$ is the smallest index such as $\|\nabla f(x^T)\| < \min\{c, L_0/L_1\}$

Since NGD and GD are monotonically decreasing in terms of the gradient norm, it follows that Algorithm 3 is also monotonically decreasing in terms of the gradient norm (see Appendix B for details). Given this fact, Theorem 3.5 shows that Algorithm 3 has two convergence regimes depending on the ratio of c and L_0/L_1 . If $c \geq L_0/L_1$, then Clip-GD starts its convergence with a linear rate $\mathcal{O}\left((1 - (1/L_1 R))^N F_0\right)$, and as soon as it approaches the solution, i.e., when $\|\nabla f(x^k)\| < L_0/L_1$, it slows down to a sublinear $\mathcal{O}(L_0 R^2/N)$ rate. If $c < L_0/L_1$, then Clip-GD has inferior linear convergence rate $\mathcal{O}\left((1 - (c/L_0 R))^N F_0\right)$ at the beginning, and approaching the solution, i.e., when $\|\nabla f(x^k)\| < c$, it slows down to the same sublinear rate. The cases in Appendix C.3 are discussed in more detail. Table 1 summarizes the derived results and compares them with the closely related works analyzing Clip-GD. It is worth noting that Theorem 3.5 shows when Algorithm 3 has linear convergence and gets rid of standard smoothness constant L (in contrast to (Koloskova et al., 2023)). Moreover, Theorem 3.5 is valid for an arbitrary clipping threshold c (in contrast to (Gorbunov et al., 2024; Vankov et al., 2024b)).

Remark 3.6 (Strong growth condition). Theorem 3.5 implies that under Assumption 1.2 with $L_0 = 0$, Algorithm 3 converges to the desired accuracy ε ($f(x^N) - f^* \leq \varepsilon$) after $N = \mathcal{O}(L_1 R \log \frac{F_0}{\varepsilon})$.

4 Coordinate Descent Type Methods

In this section, we present our main results for the algorithms that does not use access to the full gradient (see RCD in Subsection 4.1, and OrderRCD see Subsection 4.2).

4.1 Random coordinate descent

RCD is formalized as Algorithm 4. At each iteration, the method computes the gradient coordinate $\nabla_{i_k} f(x^k)$, where active coordinate i_k is selected uniformly at random from $[d]$ independently from previous steps.

Algorithm 4 Random Coordinate Descent Method (RCD)

Input: initial point $x_0 \in \mathbb{R}^d$, iterations number N , step size $\eta_k > 0$
for $k = 0$ **to** $N - 1$ **do**
 1. sample i_k uniformly at random from $[d]$
 2. $x^{k+1} \leftarrow x^k - \eta_k \nabla_{i_k} f(x^k) \mathbf{e}_{i_k}$
end for
Return: x^N

Our main results for RCD are given below.

Theorem 4.1. *Let function f satisfy Assumption 1.3 $((L_0, L_1)$ -coordinate-smoothness) and Assumption 1.4 (convexity, $\mu = 0$), then RCD (Algorithm 4) with step size $\eta_k \leq (L_0 + L_1 |\nabla_{i_k} f(x^k)|)^{-1}$, where $L_0 = \max_{i \in [d]} L_0^{(i)}$ and $L_1 = \max_{i \in [d]} L_1^{(i)}$, guarantees the following error:*

$$\mathbb{E} [f(x^N)] - f^* = \mathcal{O} \left(\max \left\{ \left(1 - \frac{\rho}{dR}\right)^N F_0, \frac{dL_0 R^2}{N} \right\} \right),$$

where $\rho := 1/(4\sqrt{2}L_1)$.

Theorem 4.1 provides a generalization of the results of Nesterov (2012) to the case of (L_0, L_1) -coordinate-smoothness (see Assumption 1.3). In particular, following Section 3, we separated L_0 and L_1 in the convergence results and also show that there is no need to assume standard L -smoothness since the case $L_1 = 0$ covers it. Moreover, in the case of L_0 being much smaller than L_1 , the results of Theorem 4.1 are strictly better than previously known ones. Furthermore, in the case of $L_0 = 0$, RCD converges linearly to any accuracy.

Remark 4.2 (Strong growth condition). Theorem 4.1 implies that under Assumption 1.3 with $L_0 = 0$, Algorithm 4 converges to the desired accuracy ε ($\mathbb{E}[f(x^N)] - f^* \leq \varepsilon$) after $N = \mathcal{O}(dL_1 R \log \frac{F_0}{\varepsilon})$ iterations.

For a detailed proof of Theorem 4.1, see Appendix D.1.

4.2 Random coordinate descent with Order Oracle

In this section, we consider the OrderRCD (Algorithm 5). In contrast to all previously considered methods in this paper, OrderRCD does not have access to a first-order oracle. Instead, the algorithm uses so-called Order Oracle: for any $x, y \in \mathbb{R}^d$, one can compute

$$\psi(x, y) = \text{sign}[f(y) - f(x)] \quad (5)$$

Algorithm 5 RCD with Order Oracle (OrderRCD)

Input: initial point $x_0 \in \mathbb{R}^d$, iterations number N , random generator $\mathcal{R}_\alpha(L_0, L_1)$
for $k = 0$ **to** $N - 1$ **do**
 1. sample i_k uniformly at random from $[d]$
 2. compute $\zeta_k = \text{argmin}_\zeta \{f(x^k + \zeta \mathbf{e}_{i_k})\}$ via (GRM)
 3. $x^{k+1} \leftarrow x^k + \zeta_k \mathbf{e}_{i_k}$
end for
Return: x^N

Algorithm 5 is similar to Algorithm 4, but it does not have access to the gradient coordinate $\nabla_{i_k} f(x^k)$. Following Lobanov et al. (2024), we address this challenge using the standard steepest descent trick, namely, we solve at each iteration the auxiliary linear search problem using the golden ratio method (GRM, see Algorithm 6 in Appendix D.2) with ϵ accuracy allowing to match RCD with step size η_k .

Below, we present the convergence result for Algorithm 5.

Theorem 4.3. *Let function f satisfy Assumption 1.3 $((L_0, L_1)$ -coordinate-smoothness) and Assumption 1.4 (convexity, $\mu = 0$), then Algorithm 5 (OrderRCD) with oracle (5) guarantees the following error:*

$$\mathbb{E} [f(x^N)] - f^* = \mathcal{O} \left(\max \left\{ \left(1 - \frac{\rho}{dR}\right)^N F_0, \frac{dL_0 R^2}{N} \right\} \right),$$

where $\rho := 1/(4\sqrt{2}L_1)$, $L_0 = \max_{i \in [d]} L_0^{(i)}$, and $L_1 = \max_{i \in [d]} L_1^{(i)}$.

That is, Theorem 4.3 gives exactly the same rate as Theorem 4.1 with one exception. However, it is important to note that Algorithm 5 requires $\log(1/\epsilon)$ oracle calls per iteration to solve the linear search problem at each iteration using GRM, where Order Oracle (5) is directly used. In the special case of $L_1 = 0$, Theorem 4.3 recovers known results, e.g., Gorbunov et al. (2019); Saha et al. (2021). However, when L_0 is much smaller than L_1 , Theorem 4.3 shows better results, i.e., linear convergence.

Remark 4.4 (Strong growth condition). Theorem 4.3 implies that under Assumption 1.3 with $L_0 = 0$, Algorithm 5 converges to the desired accuracy ε ($\mathbb{E}[f(x^N)] - f^* \leq \varepsilon$) after $N = \mathcal{O}(dL_1 R \log \frac{F_0}{\varepsilon})$ iterations and $T = \mathcal{O}(N \log \frac{1}{\varepsilon})$ oracle calls.

For a detailed proof of Theorem 4.1, see Appendix D.2.

5 Extension to Strongly Convex Setup

In this section, we answer the question:

“Are there convergence improvements of algorithms under the (L_0, L_1) -smoothness assumption compared to standard smoothness in a strongly convex setup?”

In particular, we consider GD (Algorithm 1) and derive the following convergence result.

Theorem 5.1. *Let function f satisfy Assumption 1.2 $((L_0, L_1)$ -smoothness) and Assumption 1.4 (strongly convexity, $\mu > 0$), then gradient descent method (Algorithm 1) with step size $\eta_k = (L_0 + L_1 \|\nabla f(x^k)\|)^{-1}$ guarantees:*

$$F_N \leq (1 - \rho_3)^{N - \mathcal{T}_2} (1 - \rho_2)^{\mathcal{T}_2 - \mathcal{T}_1} (1 - \rho_1)^{\mathcal{T}_1 + 1} F_0.$$

where $F_k = f(x^k) - f^*$ for $k \in [N]$, $\rho_3 = \frac{\mu}{2L_0}$, $\rho_2 = \max \left\{ \frac{\sqrt{\mu}}{2\sqrt{2}L_1}, \frac{1}{4L_1 R} \right\}$, $\rho_1 = \frac{1}{4L_1 R}$, $N_3 = N - \mathcal{T}_2$, $\mathcal{T}_2 = \max\{k \in [N - 1] \mid \|\nabla f(x^k)\| \geq \frac{L_0}{L_1}\}$ (if there are no such k , we let $\mathcal{T}_2 = -1$), and $\mathcal{T}_1 = \max\{k \in [N - 1] \mid \|\nabla f(x^k)\| \geq \frac{L_0}{L_1} \text{ and } F_k > 1\}$ (if there are no such k , we let $\mathcal{T}_1 = -1$). In particular,

- if $\mathcal{T}_1 = -1$ and $\mathcal{T}_2 = N - 1$, then

$$F_N \lesssim \left(1 - \max \left\{ \frac{\sqrt{\mu}}{2\sqrt{2}L_1}, \frac{1}{4L_1R} \right\}\right)^N F_0;$$

- if $\mathcal{T}_1 = \mathcal{T}_2 = -1$, then

$$F_N \lesssim \left(1 - \frac{\mu}{2L_0}\right)^N F_0.$$

The above theorem improves the result from (Gorbunov et al., 2024) that show $\|x^N - x^*\|^2 = \mathcal{O}((1 - \rho_3)^{N-\mathcal{T}_2}R)$ rate for GD, and, in contrast to the result from (Koloskova et al., 2023), Theorem 5.1 does not require L -smoothness. Moreover, the derived bound contains factor $(1 - \rho_2)^{\mathcal{T}_2-\mathcal{T}_1}(1 - \rho_1)^{\mathcal{T}_1+1}$, which might be better than $(1 - \rho_1)^{\mathcal{T}_2}$ when $R > \sqrt{2/\mu}$. Moreover, if $\mu/2L_0 < \sqrt{\mu}/(2\sqrt{2}L_1)$, then the derived result is strictly better than the known ones for GD under the standard smoothness. The proof of the Theorem 5.1 is provided in Appendix E.

6 Discussion and Future Work

In this paper (see Sections 3 and 4) we have shown that linear convergence in a convex setup is possible in the case of (L_0, L_1) -smooth problems with small enough L_0 . However, looking at the convergence of Algorithms 1-5, in particular Theorem 3.1-4.3, we see that the dominant part is sublinear $\mathcal{O}(1/N)$ and might be further improved. Nevertheless, as Remarks 3.2-4.4 demonstrate, in the case of the strong growth condition ($L_0 = 0$), we can observe significant improvements compared to previous works by (see e.g., Koloskova et al., 2023; Gorbunov et al., 2024; Vankov et al., 2024b). A prime example of a function that satisfies the strong growth condition is logistic function (see Example 1.6, Gorbunov et al., 2024), which is classical in the field of machine learning.

As future work, we see the following directions: generalizing Algorithms 2-5 to the strongly convex case; investigating whether the proposed technique can be used in the analysis of stochastic methods; investigating the convergence advantages of assuming generalized smoothness (Assumption 1.2) in accelerated optimization algorithms and many other directions. We believe this work opens up a number of research directions, including answering the question of whether it is possible to create adaptive methods, as well as methods in different settings (such as federated learning, overparameterization, etc.) that will exhibit similar advantages.

7 Conclusion

This paper demonstrates that generalized smoothness allows us to achieve linear convergence rates in convex setups. We explained the convergence behavior of gradient descent theoretically and showed that the advantages of generalized smoothness extend to gradient descent method variants, in particular, we significantly improved convergence estimates for GD, NGD, Clip-GD, and demonstrated novel convergence results for algorithms that do not have access to the full gradient as well as to the function values themselves (RCD and OrderRCD). We have demonstrated that this work opens up a number of directions for future research (see Section 6).

Acknowledgments

The authors would like to thank Anastasia Koloskova and Nazarii Tupitsa for useful discussions.

References

- Allen-Zhu, Z., Qu, Z., Richtárik, P., and Yuan, Y. (2016). Even faster accelerated coordinate descent using non-uniform sampling. In *International Conference on Machine Learning*, pages 1110–1119. PMLR.
- Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.
- Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.
- Bubeck, S. et al. (2015). Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357.
- Cauchy, A. (1847). Méthode générale pour la résolution des systemes d’équations simultanées. *Comp. Rend. Sci. Paris*, 25(1847):536–538.
- Chen, Z., Zhou, Y., Liang, Y., and Lu, Z. (2023). Generalized-smooth nonconvex optimization is as efficient as smooth nonconvex optimization. In *International Conference on Machine Learning*, pages 5396–5427. PMLR.
- Crawshaw, M., Liu, M., Orabona, F., Zhang, W., and Zhuang, Z. (2022). Robustness to unbounded smoothness of generalized signsgd. *Advances in neural information processing systems*, 35:9955–9968.
- Faw, M., Rout, L., Caramanis, C., and Shakkottai, S. (2023). Beyond uniform smoothness: A stopped analysis of adaptive sgd. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 89–160. PMLR.
- Gorbunov, E., Bibi, A., Sener, O., Bergou, E. H., and Richtarik, P. (2019). A stochastic derivative free optimization method with momentum. In *International Conference on Learning Representations*.
- Gorbunov, E., Tupitsa, N., Choudhury, S., Aliev, A., Richtárik, P., Horváth, S., and Takáč, M. (2024). Methods for convex (l_0, l_1) -smooth optimization: Clipping, acceleration, and adaptivity. *arXiv preprint arXiv:2409.14989*.
- Hübler, F., Yang, J., Li, X., and He, N. (2024). Parameter-agnostic optimization under relaxed smoothness. In *International Conference on Artificial Intelligence and Statistics*, pages 4861–4869. PMLR.
- Koloskova, A., Hendrikx, H., and Stich, S. U. (2023). Revisiting gradient clipping: Stochastic bias and tight convergence guarantees. In *International Conference on Machine Learning*, pages 17343–17363. PMLR.

- Li, H., Qian, J., Tian, Y., Rakhlin, A., and Jadbabaie, A. (2024a). Convex and non-convex optimization under generalized smoothness. *Advances in Neural Information Processing Systems*, 36.
- Li, H., Rakhlin, A., and Jadbabaie, A. (2024b). Convergence of adam under relaxed assumptions. *Advances in Neural Information Processing Systems*, 36.
- Lin, Q., Lu, Z., and Xiao, L. (2014). An accelerated proximal coordinate gradient method. *Advances in Neural Information Processing Systems*, 27.
- Lobanov, A., Gasnikov, A., and Krasnov, A. (2024). Acceleration exists! optimization problems when oracle can only compare objective function values. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Lojasiewicz, S. (1963). A topological property of real analytic subsets. *Coll. du CNRS, Les équations aux dérivées partielles*, 117:87–89.
- Nesterov, Y. (2012). Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362.
- Nesterov, Y. (2013). *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media.
- Nesterov, Y. (2018). *Lectures on convex optimization*, volume 137. Springer.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- Pascanu, R., Mikolov, T., and Bengio, Y. (2013). On the difficulty of training recurrent neural networks. In Dasgupta, S. and McAllester, D., editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1310–1318, Atlanta, Georgia, USA. PMLR.
- Polyak, B. T. (1963). Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878.
- Richtárik, P. and Takáč, M. (2016). Distributed coordinate descent method for learning with big data. *Journal of Machine Learning Research*, 17(75):1–25.
- Saha, A., Koren, T., and Mansour, Y. (2021). Dueling convex optimization. In *International Conference on Machine Learning*, pages 9245–9254. PMLR.
- Shalev-Shwartz, S. and Tewari, A. (2009). Stochastic methods for l_1 regularized loss minimization. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 929–936.
- Takezawa, Y., Bao, H., Sato, R., Niwa, K., and Yamada, M. (2024). Polyak meets parameter-free clipped gradient descent. *arXiv preprint arXiv:2405.15010*.
- Tang, Z., Rybin, D., and Chang, T.-H. (2024). Zeroth-order optimization meets human feedback: Provable learning via ranking oracles. In *The Twelfth International Conference on Learning Representations*.

- Vankov, D., Rodomanov, A., Nedich, A., Sankar, L., and Stich, S. U. (2024a). Optimizing (L_0, L_1) -smooth functions by gradient methods. *arXiv preprint arXiv:2410.10800, version 2*.
- Vankov, D., Rodomanov, A., Nedich, A., Sankar, L., and Stich, S. U. (2024b). Optimizing (L_0, L_1) -smooth functions by gradient methods. *arXiv preprint arXiv:2410.10800, version 1*.
- Vaswani, S., Bach, F., and Schmidt, M. (2019). Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd international conference on artificial intelligence and statistics*, pages 1195–1204. PMLR.
- Wang, B., Zhang, H., Ma, Z., and Chen, W. (2023). Convergence of adagrad for non-convex objectives: Simple proofs and relaxed assumptions. In *The Thirty Sixth Annual Conference on Learning Theory*, 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*.
- Zhang, B., Jin, J., Fang, C., and Wang, L. (2020a). Improved analysis of clipping algorithms for non-convex optimization. *Advances in Neural Information Processing Systems*, 33:15511–15521.
- 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*.
- Zhang, Y. and Xiao, L. (2017). Stochastic primal-dual coordinate method for regularized empirical risk minimization. *Journal of Machine Learning Research*, 18(84):1–42.
- 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*, 64:1–13.

A Auxiliary Results

In this section, we provide auxiliary technical results that are used in our analysis.

Basic inequalities. For all $a, b \in \mathbb{R}^d$ ($d \geq 1$), the following inequalities hold:

$$2\langle a, b \rangle - \|b\|^2 = \|a\|^2 - \|a - b\|^2, \quad (6)$$

$$\langle a, b \rangle \leq \|a\| \cdot \|b\|. \quad (7)$$

Generalized-Lipschitz-smoothness. In the analysis of full-gradient methods, we assume that the (L_0, L_1) -smoothness condition (Assumption 1.2) is satisfied. This inequality can be represented in the equivalent form for any $x, y \in \mathbb{R}^d$ (Zhang et al., 2020a):

$$f(y) - f(x) \leq \langle \nabla f(x), y - x \rangle + \frac{L_0 + L_1 \|\nabla f(x)\|}{2} \|y - x\|^2, \quad (8)$$

where $L_0, L_1 \geq 0$ for any $x \in \mathbb{R}^d$ and $\|y - x\| \leq \frac{1}{L_1}$.

Generalized-coordinate-Lipschitz-smoothness. In the analysis of coordinate-wise methods, we assume that the smoothness condition (Assumption 1.3) is satisfied. This inequality can be represented in the equivalent form (Crawshaw et al., 2022, Lemma 1):

$$f(x + h\mathbf{e}_i) \leq f(x) + h\nabla_i f(x) + \frac{(L_0^{(i)} + L_1^{(i)} |\nabla_i f(x)|) h^2}{2}, \quad (9)$$

where $L_0^{(1)}, L_0^{(2)}, \dots, L_0^{(d)}, L_1^{(1)}, L_1^{(2)}, \dots, L_1^{(d)} \geq 0$ for any $i \in [d], x \in \mathbb{R}^d$ and $|h| \leq \frac{1}{L_1^{(i)}}$.

B Monotonicity of Gradient Norms

In this section, we give a proof of monotonicity of convergence of algorithms by gradient norm. In particular, see Lemma B.2 for the proof for Algorithm 1, see Lemma B.3 for Algorithm 2, and see Lemma B.4 for Algorithm 3.

First of all, we start with the auxiliary result.

Lemma B.1. *Let function f satisfy Assumption 1.2 ((L_0, L_1) -smoothness) and Assumption 1.4 (convexity, $\mu = 0$), then for $x, y \in \mathbb{R}^d$ such that $\|y - x\| \leq \frac{1}{L_1}$ we have:*

$$\frac{\|\nabla f(y) - \nabla f(x)\|^2}{2(L_0 + L_1 \|\nabla f(x)\|)} \leq f(x) - f(y) - \langle \nabla f(y), x - y \rangle. \quad (10)$$

Proof. The proof of this statement is based on the results from Nesterov (2018); Gorbunov et al. (2024).

Let us define the following function $\varphi_a(b)$ for a given $a \in \mathbb{R}^d$:

$$\varphi_a(b) = f(b) - \langle \nabla f(a), b \rangle.$$

Then this function is differentiable and $\nabla\varphi_a(b) = \nabla f(b) - \nabla f(a)$. Moreover, for any $b, c \in \mathbb{R}^d$ such that $\|b - c\| \leq \frac{1}{L_1}$ we have:

$$\|\nabla\varphi_a(c) - \nabla\varphi_a(b)\| = \|\nabla f(b) - \nabla f(c)\| \stackrel{\textcircled{1}}{\leq} (L_0 + L_1 \|\nabla f(c)\|) \|b - c\|, \quad (11)$$

where in $\textcircled{1}$ we applied Assumption 1.2. Next, for given a and for any $b, c \in \mathbb{R}^d$ such that $\|b - c\| \leq \frac{1}{L_1}$ we define function $\psi_{abc}(t) : \mathbb{R} \rightarrow \mathbb{R}$ as

$$\psi_{abc}(t) = \varphi_a(c + t(b - c)).$$

Then, by definition of ψ_{abc} , we have $\varphi_a(c) = \psi_{abc}(0)$, $\varphi_a(b) = \psi_{abc}(1)$ and $\psi'_{abc} = \langle \nabla\varphi_a(c + t(b - c)), b - c \rangle$. Therefore, using the Newton-Leibniz formula, we have:

$$\begin{aligned} \varphi_a(b) - \varphi_a(c) &= \psi_{abc}(1) - \psi_{abc}(0) = \int_0^1 \psi'_{abc} dt = \int_0^1 \langle \nabla\varphi_a(c + t(b - c)), b - c \rangle dt \\ &= \langle \nabla\varphi_a(c), b - c \rangle + \int_0^1 \langle \nabla\varphi_a(c + t(b - c)) - \nabla\varphi_a(c), b - c \rangle dt \\ &\stackrel{(7)}{\leq} \langle \nabla\varphi_a(c), b - c \rangle + \int_0^1 \|\nabla\varphi_a(c + t(b - c)) - \nabla\varphi_a(c)\| \|b - c\| dt \\ &\stackrel{(11)}{\leq} \langle \nabla\varphi_a(c), b - c \rangle + \int_0^1 (L_0 + L_1 \|\nabla f(c)\|) \|b - c\|^2 \cdot t \cdot dt \\ &= \langle \nabla\varphi_a(c), b - c \rangle + \frac{(L_0 + L_1 \|\nabla f(c)\|)}{2} \|b - c\|^2. \end{aligned} \quad (12)$$

Let $b = c - \frac{1}{L_0 + L_1 \|\nabla f(c)\|} \nabla\varphi_a(c)$ and assume that $\|a - c\| \leq \frac{1}{L_1}$, then we have

$$\|b - c\| = \frac{\|\nabla\varphi_a(c)\|}{L_0 + L_1 \|\nabla f(c)\|} = \frac{\|\nabla f(c) - \nabla f(a)\|}{L_0 + L_1 \|\nabla f(c)\|} \stackrel{(2)}{\leq} \|c - a\| \leq \frac{1}{L_1},$$

meaning that for this choice of c and b we can apply (12) and get:

$$\varphi_a(b) - \varphi_a(c) \leq -\frac{\|\nabla\varphi_a(c)\|^2}{L_0 + L_1 \|\nabla f(c)\|} + \frac{\|\nabla\varphi_a(c)\|^2}{2(L_0 + L_1 \|\nabla f(c)\|)} = -\frac{\|\nabla\varphi_a(c)\|^2}{2(L_0 + L_1 \|\nabla f(c)\|)}.$$

Using the fact that a is an optimum for $\varphi_a(c)$ (since $\nabla\varphi_a(a) = 0$) and by definition of $\varphi_a(c)$ we obtain the following inequality:

$$f(a) - \langle \nabla f(a), a \rangle \leq f(c) - \langle \nabla f(a), c \rangle - \frac{\|\nabla f(c) - \nabla f(a)\|^2}{2(L_0 + L_1 \|\nabla f(c)\|)}.$$

Using the fact that this inequality is satisfied for any $a, c \in \mathbb{R}^d$ such that $\|a - c\| \leq \frac{1}{L_1}$, we take $a = y$ and $c = x$ and we get the original statement of the Lemma:

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

□

We are now ready to present the proofs of the gradient norm monotonicity along the trajectories of the considered first-order methods.

Lemma B.2. *Let function f satisfy Assumption 1.2 ((L_0, L_1) -smoothness) and Assumption 1.4 (convexity, $\mu = 0$), then for all $k \geq 0$ Algorithm 1 with $\eta_k = (L_0 + L_1 \|\nabla f(x^k)\|)^{-1}$ satisfies*

$$\|\nabla f(x^{k+1})\| \leq \|\nabla f(x^k)\|.$$

Proof. We note that for GD with $\eta_k = (L_0 + L_1 \|\nabla f(x^k)\|)^{-1}$ iterates x^k and x^{k+1} satisfy

$$\|x^k - x^{k+1}\| = \frac{\|\nabla f(x^k)\|}{L_0 + L_1 \|\nabla f(x^k)\|} \leq \frac{1}{L_1},$$

meaning that one can apply Lemma B.1 for these points. Introducing for convenience the new notation $\omega_k = L_0 + L_1 \|\nabla f(x^k)\|$ and summing (10) with $x = x^k, y = x^{k+1}$ and $x = x^{k+1}, y = x^k$, we get the following inequality:

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

Multiplying both sides by $2\omega_k$, we obtain

$$\begin{aligned} \left(1 + \frac{\omega_k}{\omega_{k+1}}\right) \left(\|\nabla f(x^{k+1})\|^2 - 2\langle \nabla f(x^{k+1}), \nabla f(x^k) \rangle + \|\nabla f(x^k)\|^2\right) \\ \leq -2\omega_k \eta_k \langle \nabla f(x^{k+1}) - \nabla f(x^k), \nabla f(x^k) \rangle, \end{aligned}$$

which is equivalent to

$$\begin{aligned} \left(1 + \frac{\omega_k}{\omega_{k+1}}\right) \|\nabla f(x^{k+1})\|^2 &\leq \left(1 + \frac{\omega_k}{\omega_{k+1}}\right) \|\nabla f(x^k)\|^2 \\ &\quad + 2\left(1 + \frac{\omega_k}{\omega_{k+1}} - \omega_k \eta_k\right) \langle \nabla f(x^{k+1}) - \nabla f(x^k), \nabla f(x^k) \rangle \\ &= \left(1 + \frac{\omega_k}{\omega_{k+1}}\right) \|\nabla f(x^k)\|^2 \\ &\quad - \frac{2}{\eta_k} \left(1 + \frac{\omega_k}{\omega_{k+1}} - \omega_k \eta_k\right) \langle \nabla f(x^{k+1}) - \nabla f(x^k), x^{k+1} - x^k \rangle \\ &\stackrel{\textcircled{1}}{=} \left(1 + \frac{\omega_k}{\omega_{k+1}}\right) \|\nabla f(x^k)\|^2 - \frac{2\omega_k}{\omega_{k+1}\eta_k} \langle \nabla f(x^{k+1}) - \nabla f(x^k), x^{k+1} - x^k \rangle \\ &\stackrel{\textcircled{2}}{\leq} \left(1 + \frac{\omega_k}{\omega_{k+1}}\right) \|\nabla f(x^k)\|^2, \end{aligned}$$

where in $\textcircled{1}$ we used $\eta_k = \frac{1}{\omega_k}$; and in $\textcircled{2}$ we used $\eta_k, \omega_k, \omega_{k+1} \geq 0$ and convexity of function f . Hence, we obtain the original statement of the Lemma:

$$\|\nabla f(x^{k+1})\| \leq \|\nabla f(x^k)\|.$$

□

Next, we provide a similar result for Algorithm 2.

Lemma B.3. *Let function f satisfy Assumption 1.2 ((L_0, L_1) -smoothness) and Assumption 1.4 (convexity, $\mu = 0$), then for all $k \geq 0$ Algorithm 2 with $\eta_k = \eta \leq \frac{c}{L_0 + cL_1}$, where $\|\nabla f(x^k)\| \geq c$, satisfies*

$$\|\nabla f(x^{k+1})\| \leq \|\nabla f(x^k)\|.$$

Proof. We note that for NGD with $\eta_k = \eta \leq \frac{c}{L_0 + cL_1}$ iterates x^k and x^{k+1} satisfy

$$\|x^k - x^{k+1}\| = \eta \leq \frac{1}{L_1},$$

meaning that one can apply Lemma B.1 for these points. Introducing for convenience the new notation $\omega_k = L_0 + L_1 \|\nabla f(x^k)\|$ and summing (10) with $x = x^k, y = x^{k+1}$ and $x = x^{k+1}, y = x^k$, we get the following inequality:

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

Multiplying both sides by $2\omega_k$, we obtain

$$\begin{aligned} \left(1 + \frac{\omega_k}{\omega_{k+1}} \right) \left(\|\nabla f(x^{k+1})\|^2 - 2 \langle \nabla f(x^{k+1}), \nabla f(x^k) \rangle + \|\nabla f(x^k)\|^2 \right) \\ \leq -\frac{2\omega_k \eta_k}{\|\nabla f(x^k)\|} \langle \nabla f(x^{k+1}) - \nabla f(x^k), \nabla f(x^k) \rangle, \end{aligned}$$

which is equivalent to

$$\begin{aligned} \left(1 + \frac{\omega_k}{\omega_{k+1}} \right) \|\nabla f(x^{k+1})\|^2 &\leq \left(1 + \frac{\omega_k}{\omega_{k+1}} \right) \|\nabla f(x^k)\|^2 \\ &\quad + 2 \left(1 + \frac{\omega_k}{\omega_{k+1}} - \frac{\omega_k \eta_k}{\|\nabla f(x^k)\|} \right) \langle \nabla f(x^{k+1}) - \nabla f(x^k), \nabla f(x^k) \rangle \\ &= \left(1 + \frac{\omega_k}{\omega_{k+1}} \right) \|\nabla f(x^k)\|^2 \\ &\quad - \frac{2\|\nabla f(x^k)\|}{\eta_k} \left(1 + \frac{\omega_k}{\omega_{k+1}} - \frac{\omega_k \eta_k}{\|\nabla f(x^k)\|} \right) \langle \nabla f(x^{k+1}) - \nabla f(x^k), x^{k+1} - x^k \rangle \\ &\stackrel{\textcircled{1}}{\leq} \left(1 + \frac{\omega_k}{\omega_{k+1}} \right) \|\nabla f(x^k)\|^2 \\ &\quad - \frac{2\|\nabla f(x^k)\|}{\eta_k} \left(1 + \frac{\omega_k}{\omega_{k+1}} - \frac{\omega_k c}{\|\nabla f(x^k)\| (L_0 + L_1 c)} \right) \langle \nabla f(x^{k+1}) - \nabla f(x^k), x^{k+1} - x^k \rangle \\ &\stackrel{\textcircled{2}}{\leq} \left(1 + \frac{\omega_k}{\omega_{k+1}} \right) \|\nabla f(x^k)\|^2 - \frac{2\omega_k \|\nabla f(x^k)\|}{\omega_{k+1} \eta_k} \langle \nabla f(x^{k+1}) - \nabla f(x^k), x^{k+1} - x^k \rangle \\ &\stackrel{\textcircled{3}}{\leq} \left(1 + \frac{\omega_k}{\omega_{k+1}} \right) \|\nabla f(x^k)\|^2, \end{aligned}$$

where in ① we used $\eta_k \leq \frac{c}{L_0 + L_1 c}$, in ② we used $\|\nabla f(x^k)\| \geq c$ implying $\frac{c}{L_0 + L_1 c} \leq \frac{\|\nabla f(x^k)\|}{\omega_k}$, and in ③ we used $\|\nabla f(x^k)\|, \eta_k, \omega_k, \omega_{k+1} \geq 0$ and convexity of function f . Hence, we obtain the original statement of the Lemma:

$$\|\nabla f(x^{k+1})\| \leq \|\nabla f(x^k)\|.$$

□

Finally, we present a similar result for Algorithm 3 that can be viewed as a combination of the previous two.

Lemma B.4. *Let function f satisfy Assumption 1.2 ((L_0, L_1) -smoothness) and Assumption 1.4 (convexity, $\mu = 0$), then for all $k \geq 0$ Algorithm 3 with step size $\eta_k = (L_0 + L_1 \max\{\|\nabla f(x^k)\|, c\})^{-1}$ satisfies*

$$\|\nabla f(x^{k+1})\| \leq \|\nabla f(x^k)\|.$$

Proof. We note that for Clip-GD with $\eta_k = (L_0 + L_1 \max\{\|\nabla f(x^k)\|, c\})^{-1}$ iterates x^k and x^{k+1} satisfy

$$\|x^k - x^{k+1}\| = \frac{\max\{\|\nabla f(x^k)\|, c\}}{L_0 + L_1 \max\{\|\nabla f(x^k)\|, c\}} \leq \frac{1}{L_1},$$

meaning that one can apply Lemma B.1 for these points. Introducing for convenience the new notation $\omega_k = L_0 + L_1 \|\nabla f(x^k)\|$ and summing (10) with $x = x^k, y = x^{k+1}$ and $x = x^{k+1}, y = x^k$, we get the following inequality:

$$\begin{aligned} \left(\frac{1}{2\omega_k} + \frac{1}{2\omega_{k+1}}\right) \|\nabla f(x^{k+1}) - \nabla f(x^k)\|^2 &\leq \langle \nabla f(x^{k+1}) - \nabla f(x^k), x^{k+1} - x^k \rangle \\ &= -\eta_k \cdot \underbrace{\min\left\{1, \frac{c}{\|\nabla f(x^k)\|}\right\}}_{\lambda_k} \langle \nabla f(x^{k+1}) - \nabla f(x^k), \nabla f(x^k) \rangle. \end{aligned}$$

Multiplying both sides by $2\omega_k$, we obtain

$$\begin{aligned} \left(1 + \frac{\omega_k}{\omega_{k+1}}\right) \left(\|\nabla f(x^{k+1})\|^2 - 2\langle \nabla f(x^{k+1}), \nabla f(x^k) \rangle + \|\nabla f(x^k)\|^2\right) \\ \leq -2\omega_k \eta_k \lambda_k \langle \nabla f(x^{k+1}) - \nabla f(x^k), \nabla f(x^k) \rangle. \end{aligned} \quad (13)$$

Consider two cases: $\lambda_k = 1$ or $\lambda_k = \frac{c}{\|\nabla f(x^k)\|}$. If $\lambda_k = 1$, then $c \geq \|\nabla f(x^k)\|$. Then (13) is equivalent to the following:

$$\begin{aligned} \left(1 + \frac{\omega_k}{\omega_{k+1}}\right) \|\nabla f(x^{k+1})\|^2 &\leq \left(1 + \frac{\omega_k}{\omega_{k+1}}\right) \|\nabla f(x^k)\|^2 \\ &\quad + 2\left(1 + \frac{\omega_k}{\omega_{k+1}} - \omega_k \eta_k\right) \langle \nabla f(x^{k+1}) - \nabla f(x^k), \nabla f(x^k) \rangle \\ &= \left(1 + \frac{\omega_k}{\omega_{k+1}}\right) \|\nabla f(x^k)\|^2 - \frac{2}{\eta_k} \left(1 + \frac{\omega_k}{\omega_{k+1}} - \omega_k \eta_k\right) \langle \nabla f(x^{k+1}) - \nabla f(x^k), x^{k+1} - x^k \rangle \\ &\stackrel{\text{①}}{\leq} \left(1 + \frac{\omega_k}{\omega_{k+1}}\right) \|\nabla f(x^k)\|^2 \end{aligned}$$

$$\begin{aligned}
& -\frac{2}{\eta_k} \left(1 + \frac{\omega_k}{\omega_{k+1}} - \frac{\omega_k}{L_0 + L_1 c}\right) \langle \nabla f(x^{k+1}) - \nabla f(x^k), x^{k+1} - x^k \rangle \\
& \stackrel{\textcircled{2}}{\leq} \left(1 + \frac{\omega_k}{\omega_{k+1}}\right) \|\nabla f(x^k)\|^2 - \frac{2\omega_k}{\omega_{k+1}\eta_k} \langle \nabla f(x^{k+1}) - \nabla f(x^k), x^{k+1} - x^k \rangle \\
& \stackrel{\textcircled{3}}{\leq} \left(1 + \frac{\omega_k}{\omega_{k+1}}\right) \|\nabla f(x^k)\|^2,
\end{aligned}$$

where in ① we used $\eta_k \leq \frac{1}{L_0 + L_1 c}$, in ② we used $\frac{1}{L_0 + L_1 c} \leq \frac{1}{\omega_k}$, and in ③ we used $\eta_k, \omega_k, \omega_{k+1} \geq 0$ and convexity of function f .

Next, we consider the case when $\lambda_k = \frac{c}{\|\nabla f(x^k)\|}$, implying $c \leq \|\nabla f(x^k)\|$. Then, (13) is equivalent to the following:

$$\begin{aligned}
& \left(1 + \frac{\omega_k}{\omega_{k+1}}\right) \|\nabla f(x^{k+1})\|^2 \leq \left(1 + \frac{\omega_k}{\omega_{k+1}}\right) \|\nabla f(x^k)\|^2 \\
& \quad + 2 \left(1 + \frac{\omega_k}{\omega_{k+1}} - \frac{\omega_k \eta_k c}{\|\nabla f(x^k)\|}\right) \langle \nabla f(x^{k+1}) - \nabla f(x^k), \nabla f(x^k) \rangle \\
& = \left(1 + \frac{\omega_k}{\omega_{k+1}}\right) \|\nabla f(x^k)\|^2 \\
& \quad - \frac{2\|\nabla f(x^k)\|}{\eta_k c} \left(1 + \frac{\omega_k}{\omega_{k+1}} - \frac{\omega_k \eta_k c}{\|\nabla f(x^k)\|}\right) \langle \nabla f(x^{k+1}) - \nabla f(x^k), x^{k+1} - x^k \rangle \\
& \stackrel{\textcircled{1}}{\leq} \left(1 + \frac{\omega_k}{\omega_{k+1}}\right) \|\nabla f(x^k)\|^2 \\
& \quad - \frac{2\|\nabla f(x^k)\|}{\eta_k c} \left(1 + \frac{\omega_k}{\omega_{k+1}} - \frac{\omega_k}{(L_0 + L_1 c)}\right) \langle \nabla f(x^{k+1}) - \nabla f(x^k), x^{k+1} - x^k \rangle \\
& \stackrel{\textcircled{2}}{\leq} \left(1 + \frac{\omega_k}{\omega_{k+1}}\right) \|\nabla f(x^k)\|^2 - \frac{2\omega_k \|\nabla f(x^k)\|}{\omega_{k+1} \eta_k c} \langle \nabla f(x^{k+1}) - \nabla f(x^k), x^{k+1} - x^k \rangle \\
& \stackrel{\textcircled{3}}{\leq} \left(1 + \frac{\omega_k}{\omega_{k+1}}\right) \|\nabla f(x^k)\|^2,
\end{aligned}$$

where in ① we used $\eta_k \leq \frac{1}{L_0 + L_1 c}$ and $c \leq \|\nabla f(x^k)\|$, in ② we used $\frac{c}{L_0 + L_1 c} \leq \frac{1}{\omega_k}$, and in ③ we used $\|\nabla f(x^k)\|, \eta_k, \omega_k, \omega_{k+1} \geq 0$ and convexity of function f .

That is, in both cases, we obtain the original statement of the Lemma:

$$\|\nabla f(x^{k+1})\| \leq \|\nabla f(x^k)\|.$$

□

C Missing Proofs for Full-Gradient Algorithms

In this section, we give missing proofs from the main part of the paper. In particular, see Subsection C.1 for the proof of convergence results for Algorithm 1, see Subsection C.2 for Algorithm 2, and see Subsection C.3 for Algorithm 3.

C.1 Proof of Theorem 3.1

Using Assumption 1.2, we derive

$$\begin{aligned}
f(x^{k+1}) - f(x^k) &= f(x^k - \eta_k \nabla f(x^k)) - f(x^k) \\
&\stackrel{(8)}{\leq} -\eta_k \langle \nabla f(x^k), \nabla f(x^k) \rangle + \eta_k^2 \frac{L_0 + L_1 \|\nabla f(x^k)\|}{2} \|\nabla f(x^k)\|^2 \\
&\stackrel{\textcircled{1}}{\leq} -\eta_k \|\nabla f(x^k)\|^2 + \frac{\eta_k}{2} \|\nabla f(x^k)\|^2 \\
&= -\frac{\eta_k}{2} \|\nabla f(x^k)\|^2,
\end{aligned} \tag{14}$$

where in $\textcircled{1}$ we used $\eta_k \leq \frac{1}{L_0 + L_1 \|\nabla f(x^k)\|}$. Next, let us consider two cases.

The case of $\|\nabla f(x^k)\| \geq \frac{L_0}{L_1}$. Taking $\eta_k = \frac{1}{L_0 + L_1 \|\nabla f(x^k)\|}$ and using the convexity assumption of the function (see Assumption 1.4, $\mu = 0$), we have the following:

$$\begin{aligned}
f(x^k) - f^* &\leq \langle \nabla f(x^k), x^k - x^* \rangle \stackrel{(7)}{\leq} \|\nabla f(x^k)\| \|x^k - x^*\| \stackrel{\textcircled{1}}{\leq} \|\nabla f(x^k)\| \underbrace{\|x^0 - x^*\|}_R = \frac{\eta_k}{\eta_k} \|\nabla f(x^k)\| R \\
&= \eta_k (L_0 + L_1 \|\nabla f(x^k)\|) \|\nabla f(x^k)\| R \leq 2\eta_k L_1 \|\nabla f(x^k)\|^2 R,
\end{aligned}$$

where $\textcircled{1}$ follows from $\|x^k - x^*\| \leq \|x^0 - x^*\|$ (Gorbunov et al., 2024, proof of Theorem 3.3). The above inequality implies

$$\eta_k \geq \frac{f(x^k) - f^*}{2L_1 R \|\nabla f(x^k)\|^2}. \tag{15}$$

Plugging (15) into (14), we obtain

$$f(x^{k+1}) - f(x^k) \leq -\eta_k \|\nabla f(x^k)\|^2 \leq \frac{1}{4L_1 R} (f(x^k) - f^*),$$

which is equivalent to

$$f(x^{k+1}) - f^* \leq \left(1 - \frac{1}{4L_1 R}\right) (f(x^k) - f^*). \tag{16}$$

Moreover, Lemma B.2 implies that for all $t = 0, \dots, k$ a similar inequality holds. We denote $T := \min \left\{ k \in \{0, 1, 2, \dots, N-1\} \mid \|\nabla f(x^k)\| < \frac{L_0}{L_1} \text{ and } \|\nabla f(x^{k-1})\| \geq \frac{L_0}{L_1} \right\}$ as the first index k such that $\|\nabla f(x^k)\| < \frac{L_0}{L_1}$ (note that $T = 0$ is possible). Then, for the first T iterations, we have linear convergence:

$$f(x^T) - f^* \leq \left(1 - \frac{1}{4L_1 R}\right)^T (f(x^0) - f^*), \tag{17}$$

which follows from unrolling (16).

The case of $\|\nabla f(x^k)\| < \frac{L_0}{L_1}$. Taking $\eta_k = \frac{1}{L_0 + L_1 \|\nabla f(x^k)\|}$ and using the convexity assumption of the function (see Assumption 1.4, $\mu = 0$), we have the following:

$$f(x^k) - f^* \leq \langle \nabla f(x^k), x^k - x^* \rangle \stackrel{(7)}{\leq} \|\nabla f(x^k)\| \|x^k - x^*\| \leq \|\nabla f(x^k)\| \underbrace{\|x^0 - x^*\|}_R \tag{18}$$

$$= \frac{\eta_k}{\eta_k} \left\| \nabla f(x^k) \right\| R = \eta_k (L_0 + L_1 \left\| \nabla f(x^k) \right\|) \left\| \nabla f(x^k) \right\| R < 2\eta_k L_0 \left\| \nabla f(x^k) \right\| R.$$

The above inequality implies

$$\eta_k > \frac{f(x^k) - f^*}{2L_0 R \left\| \nabla f(x^k) \right\|}. \quad (19)$$

Then, plugging (19) into (14) and using the notation $F_k = f(x^k) - f^*$, we obtain:

$$F_{k+1} < F_k - \frac{\left\| \nabla f(x^k) \right\|}{4L_0 R} F_k \stackrel{(18)}{\leq} F_k - \frac{1}{4L_0 R^2} F_k^2,$$

which is equivalent to

$$\frac{1}{4L_0 R^2} F_k^2 < F_k - F_{k+1}.$$

Next, we divide both sides by $F_{k+1} F_k$

$$\frac{1}{4L_0 R^2} \cdot \frac{F_k}{F_{k+1}} < \frac{1}{F_{k+1}} - \frac{1}{F_k}$$

and use that $F_{k+1} \leq F_k$ due to (14):

$$\frac{1}{4L_0 R^2} < \frac{1}{F_{k+1}} - \frac{1}{F_k}.$$

Summing up the above inequality for $k = T, T+1, \dots, N$, we get

$$\frac{N-T}{4L_0 R^2} = \sum_{k=T}^{N-1} \frac{1}{4L_0 R^2} < \sum_{k=T}^{N-1} \left(\frac{1}{F_{k+1}} - \frac{1}{F_k} \right) = \frac{1}{F_N} - \frac{1}{F_T} < \frac{1}{F_N},$$

which is equivalent to

$$f(x^N) - f^* < \frac{4L_0 R^2}{N-T}. \quad (20)$$

Finally, combining inequalities (17) and (20) and taking into account that $F_N \leq F_T$, we obtain the convergence rate of Algorithm 1 in the convex case:

$$f(x^N) - f^* = \mathcal{O} \left(\min \left\{ \frac{L_0 R^2}{N-T}, \left(1 - \frac{1}{L_1 R} \right)^T F_0 \right\} \right),$$

where $T := \min \left\{ k \in \{0, 1, 2, \dots, N-1\} \mid \left\| \nabla f(x^k) \right\| < \frac{L_0}{L_1} \text{ and } \left\| \nabla f(x^{k-1}) \right\| \geq \frac{L_0}{L_1} \right\}$.

C.2 Proof of Theorem 3.3

Using Assumption 1.2, we derive

$$\begin{aligned} f(x^{k+1}) - f(x^k) &= f \left(x^k - \eta_k \frac{\nabla f(x^k)}{\left\| \nabla f(x^k) \right\|} \right) - f(x^k) \\ &\stackrel{(8)}{\leq} -\frac{\eta_k}{\left\| \nabla f(x^k) \right\|} \left\langle \nabla f(x^k), \nabla f(x^k) \right\rangle + \eta_k^2 \frac{L_0 + L_1 \left\| \nabla f(x^k) \right\|}{2 \left\| \nabla f(x^k) \right\|^2} \left\| \nabla f(x^k) \right\|^2 \end{aligned}$$

$$\begin{aligned}
&\stackrel{\textcircled{1}}{\leq} -\eta_k \left\| \nabla f(x^k) \right\| + \frac{\eta_k}{2} \left\| \nabla f(x^k) \right\| \\
&= -\frac{\eta_k}{2} \left\| \nabla f(x^k) \right\|,
\end{aligned} \tag{21}$$

where in ① we used $\eta_k = \eta \leq \frac{c}{L_0+L_1c} \leq \frac{\|\nabla f(x^k)\|}{L_0+L_1\|\nabla f(x^k)\|}$ since $\|\nabla f(x^k)\| \geq c$ for all $k = 0, 1, \dots, N-1$ and function $\varphi(u) = \frac{u}{L_0+L_1u}$ is increasing function in $u \geq 0$.

Next, we use the convexity assumption of the function (see Assumption 1.4, $\mu = 0$):

$$f(x^k) - f^* \leq \left\langle \nabla f(x^k), x^k - x^* \right\rangle \stackrel{(7)}{\leq} \left\| \nabla f(x^k) \right\| \left\| x^k - x^* \right\| \stackrel{\textcircled{1}}{\leq} \left\| \nabla f(x^k) \right\| \underbrace{\left\| x^0 - x^* \right\|}_R, \tag{22}$$

where ① follows from $\|x^k - x^*\| \leq \|x^0 - x^*\|$:

$$\begin{aligned}
\|x^k - x^*\|^2 &= \|x^{k-1} - x^*\|^2 - \frac{2\eta_k}{\|\nabla f(x^k)\|} \langle \nabla f(x^k), x^k - x^* \rangle + \eta_k^2 \\
&\stackrel{(3)}{\leq} \|x^{k-1} - x^*\|^2 - \frac{2\eta(f(x^k) - f^*)}{\|\nabla f(x^k)\|} + \eta^2 \\
&= \|x^{k-1} - x^*\|^2 - \eta \left(\frac{2(f(x^k) - f^*)}{\|\nabla f(x^k)\|} - \eta \right) \\
&\leq \|x^{k-1} - x^*\|^2,
\end{aligned}$$

where in the last step, we use

$$\frac{\eta \|\nabla f(x^k)\|}{2} \leq \frac{c \|\nabla f(x^k)\|}{2(L_0 + L_1c)} \leq \frac{\|\nabla f(x^k)\|^2}{2(L_0 + L_1 \|\nabla f(x^k)\|)} \stackrel{(10)}{\leq} f(x^k) - f^*.$$

Next, inequality (22) gives

$$\left\| \nabla f(x^k) \right\| \geq \frac{f(x^k) - f^*}{R}. \tag{23}$$

Then, plugging (23) into (21), we obtain:

$$f(x^{k+1}) - f(x^k) \leq -\frac{\eta}{2} \left\| \nabla f(x^k) \right\| \leq -\frac{\eta}{2R} (f(x^k) - f^*),$$

which is equivalent to

$$f(x^{k+1}) - f^* \leq \left(1 - \frac{\eta}{2R}\right) (f(x^k) - f^*).$$

Unrolling the above recurrence, we derive the linear convergence for NGD with step size $\eta_k = \eta \leq \frac{c}{L_0+L_1c}$:

$$f(x^N) - f^* \leq \left(1 - \frac{\eta}{2R}\right)^N (f(x^0) - f^*).$$

C.3 Proof of Theorem 3.5

Since $\lambda_k = \min \left\{ 1, \frac{c}{\|\nabla f(x^k)\|} \right\}$, we there are only two possible cases for λ_k : either $\lambda_k = 1$ or $\lambda_k = \frac{c}{\|\nabla f(x^k)\|}$.

i) Consider the case of $\lambda_k = \frac{c}{\|\nabla f(x^k)\|}$, i.e., $c \leq \|\nabla f(x^k)\|$. Using Assumption 1.2, we derive

$$\begin{aligned}
f(x^{k+1}) - f(x^k) &= f(x^k - \eta_k \lambda_k \nabla f(x^k)) - f(x^k) \\
&= f\left(x^k - \eta_k \frac{c}{\|\nabla f(x^k)\|} \nabla f(x^k)\right) - f(x^k) \\
&\stackrel{(8)}{\leq} -\eta_k \frac{c}{\|\nabla f(x^k)\|} \langle \nabla f(x^k), \nabla f(x^k) \rangle \\
&\quad + \eta_k^2 \frac{c^2}{\|\nabla f(x^k)\|^2} \frac{L_0 + L_1 \|\nabla f(x^k)\|}{2} \|\nabla f(x^k)\|^2 \\
&= -\eta_k c \|\nabla f(x^k)\| + \eta_k^2 c^2 \frac{L_0 + L_1 \|\nabla f(x^k)\|}{2} \\
&\stackrel{\textcircled{1}}{\leq} -\eta_k c \|\nabla f(x^k)\| + \frac{\eta_k c}{2} \|\nabla f(x^k)\| \\
&= -\frac{\eta_k c}{2} \|\nabla f(x^k)\|, \tag{24}
\end{aligned}$$

where in $\textcircled{1}$ we used $\eta_k \leq \frac{\|\nabla f(x^k)\|}{c(L_0 + L_1 \|\nabla f(x^k)\|)}$, which follows from $c \leq \|\nabla f(x^k)\|$:

$$\frac{\|\nabla f(x^k)\|}{c(L_0 + L_1 \|\nabla f(x^k)\|)} = \frac{1}{L_0 \frac{c}{\|\nabla f(x^k)\|} + L_1 c} \geq \frac{1}{L_0 + L_1 c} = \eta = \eta_k.$$

Next, using the convexity assumption of the function (see Assumption 1.4, $\mu = 0$), we get

$$f(x^k) - f^* \leq \langle \nabla f(x^k), x^k - x^* \rangle \stackrel{(7)}{\leq} \|\nabla f(x^k)\| \|x^k - x^*\| \stackrel{\textcircled{1}}{\leq} \|\nabla f(x^k)\| \underbrace{\|x^0 - x^*\|}_R, \tag{25}$$

where $\textcircled{1}$ follows from $\|x^k - x^*\| \leq \|x^0 - x^*\|$:

$$\begin{aligned}
\|x^k - x^*\|^2 &= \|x^{k-1} - x^*\|^2 - \frac{2c\eta_k}{\|\nabla f(x^k)\|} \langle \nabla f(x^k), x^k - x^* \rangle + c^2 \eta_k^2 \\
&\stackrel{(3)}{\leq} \|x^{k-1} - x^*\|^2 - \frac{2c\eta(f(x^k) - f^*)}{\|\nabla f(x^k)\|} + c^2 \eta^2 \\
&= \|x^{k-1} - x^*\|^2 - c\eta \left(\frac{2(f(x^k) - f^*)}{\|\nabla f(x^k)\|} - c\eta \right) \\
&\leq \|x^{k-1} - x^*\|^2,
\end{aligned}$$

where in the last step, we use

$$\frac{c\eta \|\nabla f(x^k)\|}{2} \leq \frac{c \|\nabla f(x^k)\|}{2(L_0 + L_1 c)} \leq \frac{\|\nabla f(x^k)\|^2}{2(L_0 + L_1 \|\nabla f(x^k)\|)} \stackrel{(10)}{\leq} f(x^k) - f^*.$$

Inequality (25) gives

$$\|\nabla f(x^k)\| \geq \frac{f(x^k) - f^*}{R}. \tag{26}$$

Then, plugging (26) into (24), we obtain

$$f(x^{k+1}) - f(x^k) \stackrel{(24)}{\leq} -\frac{\eta c}{2} \|\nabla f(x^k)\| \leq \frac{\eta c}{2R} (f(x^k) - f^*),$$

which is equivalent to

$$f(x^{k+1}) - f^* \leq \left(1 - \frac{\eta c}{2R}\right) (f(x^k) - f^*).$$

Next, we consider two possible scenarios for the convergence of the algorithm depending on the relation between $\|\nabla f(x^k)\|, c$ and $\frac{L_0}{L_1}$ (note that $\|\nabla f(x^k)\| \geq c$ in this case), given the monotonicity of the gradient norm (Lemma B.4).

(\mathcal{T}) If for $k = 0, 1, 2, \dots, \mathcal{T}_1 - 1$, the iterates of Clip-GD satisfy $\|\nabla f(x^k)\| \geq c \geq \frac{L_0}{L_1}$, then $\eta \geq \frac{1}{2L_1 c}$ and we have linear convergence for the first \mathcal{T}_1 iterations:

$$f(x^{\mathcal{T}_1}) - f^* \leq \left(1 - \frac{1}{4L_1 R}\right)^{\mathcal{T}_1} (f(x^0) - f^*). \quad (27)$$

(\mathcal{K}) If for $k = 0, 1, 2, \dots, \mathcal{K}_1 - 1$, the iterates of Clip-GD satisfy $\|\nabla f(x^k)\| \geq \frac{L_0}{L_1} \geq c$ or $\frac{L_0}{L_1} \geq \|\nabla f(x^k)\| \geq c$, then $\eta \geq \frac{1}{2L_0}$ and we have linear convergence of the first \mathcal{K}_1 iterations:

$$f(x^{\mathcal{K}_1}) - f^* \leq \left(1 - \frac{c}{4L_0 R}\right)^{\mathcal{K}_1} (f(x^0) - f^*). \quad (28)$$

ii) Consider the case of $\lambda_k = 1$, i.e., $c \geq \|\nabla f(x^k)\|$. Using Assumption 1.2, we derive

$$\begin{aligned} f(x^{k+1}) - f(x^k) &= f(x^k - \eta_k \lambda_k \nabla f(x^k)) - f(x^k) \\ &= f(x^k - \eta_k \nabla f(x^k)) - f(x^k) \\ &\stackrel{(8)}{\leq} -\eta_k \langle \nabla f(x^k), \nabla f(x^k) \rangle + \eta_k^2 \frac{L_0 + L_1 \|\nabla f(x^k)\|}{2} \|\nabla f(x^k)\|^2 \\ &= -\eta_k \|\nabla f(x^k)\|^2 + \eta_k^2 \frac{L_0 + L_1 \|\nabla f(x^k)\|}{2} \|\nabla f(x^k)\|^2 \\ &\stackrel{\textcircled{1}}{=} -\frac{1}{2(L_0 + L_1 \|\nabla f(x^k)\|)} \|\nabla f(x^k)\|^2, \end{aligned} \quad (29)$$

where in $\textcircled{1}$ we used $\eta_k = \frac{1}{L_0 + L_1 \|\nabla f(x^k)\|}$. Using the convexity assumption of the function (see Assumption 1.4, $\mu = 0$), we get

$$f(x^k) - f^* \leq \langle \nabla f(x^k), x^k - x^* \rangle \stackrel{(7)}{\leq} \|\nabla f(x^k)\| \|x^k - x^*\| \stackrel{\textcircled{1}}{\leq} \|\nabla f(x^k)\| \underbrace{\|x^0 - x^*\|}_R, \quad (30)$$

where $\textcircled{1}$ follows from $\|x^k - x^*\| \leq \|x^0 - x^*\|$:

$$\begin{aligned} \|x^k - x^*\|^2 &= \|x^{k-1} - x^*\|^2 - 2\eta_k \langle \nabla f(x^k), x^k - x^* \rangle + \eta_k^2 \|\nabla f(x^k)\|^2 \\ &\stackrel{(3)}{\leq} \|x^{k-1} - x^*\|^2 - 2\eta_k (f(x^k) - f^*) + \eta_k^2 \|\nabla f(x^k)\|^2 \end{aligned}$$

$$\begin{aligned}
&= \|x^{k-1} - x^*\|^2 - \eta_k \left(2(f(x^k) - f^*) - \eta_k \|\nabla f(x^k)\|^2 \right) \\
&\leq \|x^{k-1} - x^*\|^2,
\end{aligned}$$

where in the last step, we use

$$\frac{\eta_k \|\nabla f(x^k)\|}{2} \leq \frac{\|\nabla f(x^k)\|^2}{2(L_0 + L_1 \|\nabla f(x^k)\|)} \stackrel{(10)}{\leq} f(x^k) - f^*.$$

Inequality (30), implies

$$\|\nabla f(x^k)\| \geq \frac{f(x^k) - f^*}{R}. \quad (31)$$

Next, we consider two cases: $\|\nabla f(x^k)\| \geq \frac{L_0}{L_1}$ and $\|\nabla f(x^k)\| < \frac{L_0}{L_1}$.

The case of $\|\nabla f(x^k)\| \geq \frac{L_0}{L_1}$. In this case, inequality (29) gives

$$f(x^{k+1}) - f(x^k) \leq -\frac{1}{4L_1} \|\nabla f(x^k)\|. \quad (32)$$

Then, plugging (31) into (32), we obtain:

$$f(x^{k+1}) - f(x^k) \leq -\frac{1}{4L_1 R} (f(x^k) - f^*),$$

which is equivalent to

$$f(x^{k+1}) - f^* \leq \left(1 - \frac{1}{4L_1 R}\right) (f(x^k) - f^*).$$

Since in this case we have the following relation $c \geq \|\nabla f(x^k)\| \geq \frac{L_0}{L_1}$, then for $k = \mathcal{T}_1, \mathcal{T}_1 + 1, \dots, \mathcal{T}_2 - 1$ we have linear convergence:

$$f(x^{\mathcal{T}_2}) - f^* \leq \left(1 - \frac{1}{4L_1 R}\right)^{\mathcal{T}_2 - \mathcal{T}_1} (f(x^{\mathcal{T}_1}) - f^*) \stackrel{(27)}{\leq} \left(1 - \frac{1}{4L_1 R}\right)^{\mathcal{T}_2} (f(x^0) - f^*). \quad (33)$$

The case of $\|\nabla f(x^k)\| < \frac{L_0}{L_1}$. In this case, inequality (29) gives

$$\begin{aligned}
f(x^{k+1}) - f(x^k) &\leq -\frac{1}{2(L_0 + L_1 \|\nabla f(x^k)\|)} \|\nabla f(x^k)\|^2 \\
&< -\frac{1}{4L_0} \|\nabla f(x^k)\|^2.
\end{aligned} \quad (34)$$

Then, plugging (31) into (34) and using the notation $F_k := f(x^k) - f^*$, we obtain:

$$F_{k+1} < F_k - \frac{\|\nabla f(x^k)\|}{4L_0 R} F_k \leq F_k - \frac{1}{4L_0 R^2} F_k^2,$$

which is equivalent to

$$\frac{1}{4L_0 R^2} F_k^2 < F_k - F_{k+1}.$$

Next, we divide both sides by $F_{k+1}F_k$

$$\frac{1}{4L_0R^2} \cdot \frac{F_k}{F_{k+1}} < \frac{1}{F_{k+1}} - \frac{1}{F_k}.$$

and use $F_{k+1} \leq F_k$ due to (34):

$$\frac{1}{4L_0R^2} < \frac{1}{F_{k+1}} - \frac{1}{F_k}. \quad (35)$$

Then, two situations are possible: either $\frac{L_0}{L_1} > c$ or $\frac{L_0}{L_1} \leq c$. We consider each of them separately.

(\mathcal{K}) Considering the scenario $\frac{L_0}{L_1} > c > \|\nabla f(x^k)\|$ and summing up inequality (35) for $k = \mathcal{K}_1, \mathcal{K}_1 + 1, \dots, \mathcal{K}_2 - 1$, we get

$$\frac{\mathcal{K}_2 - \mathcal{K}_1}{4L_0R^2} = \sum_{k=\mathcal{K}_1}^{\mathcal{K}_2-1} \frac{1}{4L_0R^2} < \sum_{k=\mathcal{K}_1}^{\mathcal{K}_2-1} \left(\frac{1}{F_{k+1}} - \frac{1}{F_k} \right) = \frac{1}{F_{\mathcal{K}_2}} - \frac{1}{F_{\mathcal{K}_1}} < \frac{1}{F_{\mathcal{K}_2}},$$

which is equivalent to

$$f(x^{\mathcal{K}_2}) - f^* < \frac{4L_0R^2}{\mathcal{K}_2 - \mathcal{K}_1}. \quad (36)$$

(\mathcal{T}) Considering the scenario $c \geq \frac{L_0}{L_1} > \|\nabla f(x^k)\|$ and summing up inequality (35) for $k = \mathcal{T}_2, \mathcal{T}_2 + 1, \dots, \mathcal{T}_3 - 1$, we get

$$\frac{\mathcal{T}_3 - \mathcal{T}_2}{4L_0R^2} = \sum_{k=\mathcal{T}_2}^{\mathcal{T}_3-1} \frac{1}{4L_0R^2} < \sum_{k=\mathcal{T}_2}^{\mathcal{T}_3-1} \left(\frac{1}{F_{k+1}} - \frac{1}{F_k} \right) = \frac{1}{F_{\mathcal{T}_3}} - \frac{1}{F_{\mathcal{T}_2}} < \frac{1}{F_{\mathcal{T}_3}},$$

which is equivalent to

$$f(x^{\mathcal{T}_3}) - f^* < \frac{4L_0R^2}{\mathcal{T}_3 - \mathcal{T}_2}. \quad (37)$$

Finally, combining (27), (28), (33), (36), (37), and taking into account that $F_{k+1} \leq F_k$, we obtain the convergence rate for Algorithm 3.

- If $c \geq \frac{L_0}{L_1}$, then for $\mathcal{T}_3 = N$ being the total number of iterations of Algorithm 3 the iterates satisfy (see (27), (33) and (37)):

$$f(x^N) - f^* = \mathcal{O} \left(\min \left\{ \frac{L_0R^2}{N - \mathcal{T}_2}, \left(1 - \frac{1}{L_1R} \right)^{\mathcal{T}_2} F_0 \right\} \right),$$

where $\mathcal{T}_2 := \min \left\{ k \in \{0, 1, 2, \dots, N - 1\} \mid \|\nabla f(x^k)\| < \frac{L_0}{L_1} \text{ and } \|\nabla f(x^{k-1})\| \geq \frac{L_0}{L_1} \right\}$.

- If $c < \frac{L_0}{L_1}$, then $\mathcal{K}_2 = N$ being the total number of iterations of Algorithm 3 the iterates satisfy (see (28) and (36)):

$$f(x^N) - f^* = \mathcal{O} \left(\min \left\{ \frac{L_0 R^2}{N - \mathcal{K}_1}, \left(1 - \frac{c}{L_0 R} \right)^{\mathcal{K}_1} F_0 \right\} \right),$$

where $\mathcal{K}_1 := \min \{k \in \{0, 1, 2, \dots, N-1\} \mid \|\nabla f(x^k)\| < c \text{ and } \|\nabla f(x^{k-1})\| \geq c\}$.

It is not difficult to see that these two scenarios can be combined into the following equivalent form:

$$f(x^N) - f^* = \mathcal{O} \left(\min \left\{ \frac{L_0 R^2}{N - T}, \left(1 - \frac{c}{\max\{L_0, L_1 c\} R} \right)^T F_0 \right\} \right),$$

where $T := \min \{k \in \{0, 1, 2, \dots, N-1\} \mid \|\nabla f(x^k)\| < \min \left\{ \frac{L_0}{L_1}, c \right\} \text{ and } \|\nabla f(x^{k-1})\| \geq \min \left\{ \frac{L_0}{L_1}, c \right\}\}$.

D Missing Proofs for Coordinate Descent Type Methods

In this section, we provide missing proofs from Section 4. In particular, see Subsection D.1 for the proof of the convergence results for Algorithm 4, and see Subsection D.2 for Algorithm 5.

D.1 Proof of Theorem 4.1

Using Assumption 1.3, we derive

$$\begin{aligned} f(x^{k+1}) - f(x^k) &= f(x^k - \eta_k \nabla_{i_k} f(x^k) \mathbf{e}_{i_k}) - f(x^k) \\ &\stackrel{(9)}{\leq} -\eta_k \left(\nabla_{i_k} f(x^k) \right)^2 + \eta_k^2 \frac{L_0 + L_1 |\nabla f(x^k)|}{2} \left(\nabla_{i_k} f(x^k) \right)^2 \\ &\stackrel{\textcircled{1}}{\leq} -\eta_k \left(\nabla_{i_k} f(x^k) \right)^2 + \frac{\eta_k}{2} \left(\nabla_{i_k} f(x^k) \right)^2 \\ &= -\frac{\eta_k}{2} \left(\nabla_{i_k} f(x^k) \right)^2, \end{aligned} \tag{38}$$

where in $\textcircled{1}$ we used $\eta_k \leq \frac{1}{L_0 + L_1 |\nabla_{i_k} f(x^k)|}$. Next, we take the expectation w.r.t. i_k and use $\eta_k = \frac{1}{L_0 + L_1 |\nabla_{i_k} f(x^k)|}$:

$$\begin{aligned} \mathbb{E}_{i_k}[f(x^{k+1})] - f(x^k) &\leq -\frac{1}{2d} \sum_{i=1}^d \frac{|\nabla_i f(x^k)|^2}{L_0 + L_1 |\nabla_i f(x^k)|} \\ &\leq -\frac{1}{4d} \sum_{i=1}^d \min \left\{ \frac{|\nabla_i f(x^k)|^2}{L_0}, \frac{|\nabla_i f(x^k)|}{L_1} \right\} \end{aligned} \tag{39}$$

$$= -\frac{1}{4d} \left(\sum_{i \in I_k} \frac{|\nabla_i f(x^k)|}{L_1} + \sum_{i \in [d] \setminus I_k} \frac{|\nabla_i f(x^k)|^2}{L_0} \right), \tag{40}$$

where $I_k := \left\{ i \in [d] \mid |\nabla_i f(x^k)| \geq \frac{L_0}{L_1} \right\}$. Next, we introduce the set of indices \mathcal{K} as

$$\mathcal{K} := \left\{ k \in [N-1] \mid \sum_{i \in I_k} |\nabla_i f(x^k)|^2 > \sum_{i \in [d] \setminus I_k} |\nabla_i f(x^k)|^2 \right\}$$

and consider two possible situations.

The case of $k \in \mathcal{K}$. In this case, we continue our derivation as follows:

$$\mathbb{E}_{i_k}[f(x^{k+1})] - f(x^k) \leq -\frac{1}{4dL_1} \sum_{i \in I_k} |\nabla_i f(x^k)|. \quad (41)$$

Using the convexity assumption and notation $F_k := f(x^k) - f^*$, we derive

$$\begin{aligned} F_k &\leq \left\langle \nabla f(x^k), x^k - x^* \right\rangle \stackrel{(7)}{\leq} \left\| \nabla f(x^k) \right\| \underbrace{\left\| x^k - x^* \right\|}_R \\ &= R \sqrt{\sum_{i \in I_k} |\nabla_i f(x^k)|^2 + \sum_{i \in [d] \setminus I_k} |\nabla_i f(x^k)|^2} \leq R \sqrt{2 \sum_{i \in I_k} |\nabla_i f(x^k)|^2} \leq \sqrt{2} R \sum_{i \in I_k} |\nabla_i f(x^k)| \end{aligned}$$

that implies

$$\sum_{i \in I_k} |\nabla_i f(x^k)| \geq \frac{F_k}{\sqrt{2}R}. \quad (42)$$

Plugging (42) in (41), we obtain

$$\mathbb{E}_{i_k}[F_{k+1}] \leq \left(1 - \frac{1}{4\sqrt{2}dL_1R} \right) F_k. \quad (43)$$

The case of $k \notin \mathcal{K}$. In this case, we continue (40) as follows:

$$\mathbb{E}_{i_k}[f(x^{k+1})] - f(x^k) \leq -\frac{1}{4dL_0} \sum_{i \in [d] \setminus I_k} |\nabla_i f(x^k)|^2. \quad (44)$$

Using the convexity assumption and notation $F_k := f(x^k) - f^*$, we derive

$$\begin{aligned} F_k &\leq \left\langle \nabla f(x^k), x^k - x^* \right\rangle \stackrel{(7)}{\leq} \left\| \nabla f(x^k) \right\| \underbrace{\left\| x^k - x^* \right\|}_R \\ &= R \sqrt{\sum_{i \in I_k} |\nabla_i f(x^k)|^2 + \sum_{i \in [d] \setminus I_k} |\nabla_i f(x^k)|^2} \leq R \sqrt{2 \sum_{i \in [d] \setminus I_k} |\nabla_i f(x^k)|^2} \end{aligned}$$

that implies

$$\sum_{i \in I_k} |\nabla_i f(x^k)|^2 \geq \frac{F_k^2}{2R^2}. \quad (45)$$

Plugging (45) in (44), we obtain

$$\mathbb{E}_{i_k}[F_{k+1}] \leq F_k - \frac{1}{8dL_0R^2}F_k^2. \quad (46)$$

To get the final bound, let us specify the indices belonging to \mathcal{K} : let $\mathcal{K} := \{k_1, k_2, \dots, k_r\}$ and $\mathcal{T} := [N-1] \setminus \mathcal{K} := \{t_1, t_2, \dots, t_{N-r}\}$, where $0 \leq k_1 \leq k_2 \leq \dots \leq k_r \leq N-1$ and $0 \leq t_1 \leq t_2 \leq \dots \leq t_{N-r} \leq N-1$. Note that $\mathcal{K} \cap \mathcal{T} = \emptyset$, $\mathcal{K} \cup \mathcal{T} = [N-1]$, and $|\mathcal{K}| = r$ is random variable. There exist two possible situations: either $r > N/2$ or $r \leq N/2$. If $r > N/2$, then we use (43) together with $F_{k+1} \leq F_k$ following from (38):

$$\begin{aligned} \mathbb{E}_{i \in \mathcal{K}}[F_N] &\stackrel{(38)}{\leq} \mathbb{E}_{i \in \mathcal{K}}[F_{k_r+1}] \stackrel{(43)}{\leq} \left(1 - \frac{1}{4\sqrt{2}dL_1R}\right) \mathbb{E}_{i \in \mathcal{K} \setminus \{k_r\}}[F_{k_r}] \\ &\stackrel{(38)}{\leq} \left(1 - \frac{1}{4\sqrt{2}dL_1R}\right) \mathbb{E}_{i \in \mathcal{K} \setminus \{k_r\}}[F_{k_{r-1}+1}] \stackrel{(43)}{\leq} \left(1 - \frac{1}{4\sqrt{2}dL_1R}\right)^2 \mathbb{E}_{i \in \mathcal{K} \setminus \{k_r, k_{r-1}\}}[F_{k_{r-1}}] \\ &\leq \dots \leq \left(1 - \frac{1}{4\sqrt{2}dL_1R}\right)^r F_0 \stackrel{r > N/2}{\leq} \left(1 - \frac{1}{4\sqrt{2}dL_1R}\right)^{N/2} F_0 \mathbb{1}_{\{r > N/2\}}, \end{aligned} \quad (47)$$

where $\mathbb{E}_{i \in \mathcal{K}}$ denotes the expectation w.r.t. all indices in set \mathcal{K} and $\mathbb{1}_{\{r > N/2\}}$ is an indicator of the event $\{r > N/2\}$.

Next, we consider the situation when $r \leq N/2$. In this case, we first notice that (46) gives

$$\begin{aligned} \mathbb{E}_{i \in \mathcal{T}}[F_{t_{N-r}+1}] &\stackrel{(46)}{\leq} \mathbb{E}_{i \in \mathcal{T} \setminus \{t_{N-r}\}}[F_{t_{N-r}}] - \frac{1}{8dL_0R^2} \mathbb{E}_{i \in \mathcal{T} \setminus \{t_{N-r}\}}[F_{t_{N-r}}^2] \\ &\stackrel{\textcircled{1}}{\leq} \mathbb{E}_{i \in \mathcal{T} \setminus \{t_{N-r}\}}[F_{t_{N-r}}] - \frac{1}{8dL_0R^2} \mathbb{E}_{i \in \mathcal{T} \setminus \{t_{N-r}\}}[F_{t_{N-r}}]^2 \end{aligned}$$

where in $\textcircled{1}$ we used $\mathbb{E}_{i \in \mathcal{T} \setminus \{t_{N-r}\}}[F_{t_{N-r}}]^2 \leq \mathbb{E}_{i \in \mathcal{T} \setminus \{t_{N-r}\}}[F_{t_{N-r}}^2]$. Dividing both sides by $\mathbb{E}_{i \in \mathcal{T}}[F_{t_{N-r}+1}] \mathbb{E}_{i \in \mathcal{T} \setminus \{t_{N-r}\}}[F_{t_{N-r}}]$ and rearranging the terms, we get

$$\frac{1}{8dL_0R^2} \frac{\mathbb{E}_{i \in \mathcal{T} \setminus \{t_{N-r}\}}[F_{t_{N-r}}]}{\mathbb{E}_{i \in \mathcal{T}}[F_{t_{N-r}+1}]} \leq \frac{1}{\mathbb{E}_{i \in \mathcal{T}}[F_{t_{N-r}+1}]} - \frac{1}{\mathbb{E}_{i \in \mathcal{T} \setminus \{t_{N-r}\}}[F_{t_{N-r}}]}. \quad (48)$$

In view of (38), we have $\mathbb{E}_{i \in \mathcal{T}}[F_{t_{N-r}+1}] \leq \mathbb{E}_{i \in \mathcal{T}}[F_{t_{N-r}}] = \mathbb{E}_{i \in \mathcal{T} \setminus \{t_{N-r}\}}[F_{t_{N-r}}]$ and $-\frac{1}{\mathbb{E}_{i \in \mathcal{T} \setminus \{t_{N-r}\}}[F_{t_{N-r}}]} \leq -\frac{1}{\mathbb{E}_{i \in \mathcal{T} \setminus \{t_{N-r}\}}[F_{t_{N-r-1}+1}]} = -\frac{1}{\mathbb{E}_{i \in \mathcal{T}}[F_{t_{N-r-1}+1}]}$. Using these inequalities in (48), we obtain

$$\frac{1}{8dL_0R^2} \leq \frac{1}{\mathbb{E}_{i \in \mathcal{T}}[F_{t_{N-r}+1}]} - \frac{1}{\mathbb{E}_{i \in \mathcal{T}}[F_{t_{N-r-1}+1}]}.$$

Following the same arguments, we can also show

$$\begin{aligned} \frac{1}{8dL_0R^2} &\leq \frac{1}{\mathbb{E}_{i \in \mathcal{T}}[F_{t_{N-r-1}+1}]} - \frac{1}{\mathbb{E}_{i \in \mathcal{T}}[F_{t_{N-r-2}+1}]} \\ &\dots \\ \frac{1}{8dL_0R^2} &\leq \frac{1}{\mathbb{E}_{i \in \mathcal{T}}[F_{t_1+1}]} - \frac{1}{\mathbb{E}_{i \in \mathcal{T}}[F_{t_1}]} \leq \frac{1}{\mathbb{E}_{i \in \mathcal{T}}[F_{t_1+1}]} - \frac{1}{F_0}. \end{aligned}$$

Summing up all of them, we arrive at

$$\frac{N-r}{8dL_0R^2} \leq \frac{1}{\mathbb{E}_{i \in \mathcal{T}}[F_{t_{N-r}+1}]} - \frac{1}{F_0} \leq \frac{1}{\mathbb{E}_{i \in \mathcal{T}}[F_{t_{N-r}+1}]},$$

implying

$$\mathbb{E}_{i \in \mathcal{T}}[F_N] \leq \mathbb{E}_{i \in \mathcal{T}}[F_{t_{N-r}+1}] \leq \frac{8dL_0R^2}{N-r} \stackrel{r \leq N/2}{\leq} \frac{16dL_0R^2}{N} \mathbb{1}_{\{r \leq N/2\}}. \quad (49)$$

Combining (47) and (49) and taking the full expectation, we get

$$\begin{aligned} \mathbb{E}[F_N] &\leq \left(1 - \frac{1}{4\sqrt{2}dL_1R}\right)^{N/2} F_0 \mathbb{E}[\mathbb{1}_{\{r > N/2\}}] + \frac{16dL_0R^2}{N} \mathbb{E}[\mathbb{1}_{\{r \leq N/2\}}] \\ &\leq \max \left\{ \left(1 - \frac{1}{4\sqrt{2}dL_1R}\right)^{N/2} F_0, \frac{16dL_0R^2}{N} \right\}, \end{aligned}$$

which concludes the proof.

D.2 Proof of Theorem 4.3

Algorithm 5, presented in Section 4, uses the Golden Ratio Method (GRM) once per iteration. This method utilizes the oracle concept (5) (see Algorithm 6).

Algorithm 6 Golden Ratio Method (GRM)

```

1: Input: interval  $[a, b]$ , accuracy  $\hat{\epsilon}$ 
2: Initialization: define constant  $\rho = \frac{1}{\Phi} = \frac{\sqrt{5}-1}{2}$ 
3:  $y \leftarrow a + (1 - \rho)(b - a)$ 
4:  $z \leftarrow a + \rho(b - a)$ 
5: while  $b - a > \hat{\epsilon}$  do
6:   if  $\phi(y, z) = -1$  then
7:      $b \leftarrow z$ 
8:      $z \leftarrow y$ 
9:      $y \leftarrow a + (1 - \rho)(b - a)$ 
10:  else
11:     $a \leftarrow y$ 
12:     $y \leftarrow z$ 
13:     $z \leftarrow a + \rho(b - a)$ 
14:  end if
15: end while
16: Return:  $\frac{a+b}{2}$ 

```

We utilize the Golden Ratio Method to find a solution to the following one-dimensional problem (see line 2 in Algorithm 5):

$$\zeta_k = \arg \min_{\zeta \in \mathbb{R}} f(x^k + \zeta \mathbf{e}_{i_k}).$$

Using the well-known fact about the golden ratio method that GRM is required to do $N = \mathcal{O}(\log \frac{1}{\epsilon})$ (where ϵ is the accuracy of the solution to the linear search problem in terms of the function value; due to (9), it is sufficient to take $\hat{\epsilon} = 2\epsilon/L_0$), we derive the following corollaries from the solution of this problem: for simplicity, we consider the scenario when the golden ratio method solves the inner problem exactly ($\epsilon \simeq 0$). Then, we have the following:

$$f(x_k + \zeta_k \mathbf{e}_{i_k}) \leq f(x_k + \zeta \mathbf{e}_{i_k}), \quad \forall \zeta \in \mathbb{R}. \quad (50)$$

Using the above inequality with $\zeta = \eta_k \nabla_{i_k} f(x^k)$, $\eta_k := \frac{1}{L_0 + L_1 |\nabla_{i_k} f(x^k)|}$, and applying Assumption 1.3, we get

$$\begin{aligned} f(x^{k+1}) - f(x^k) &= f(x^k + \zeta_k \mathbf{e}_{i_k}) - f(x^k) \\ &\stackrel{(50)}{\leq} f(x^k - \eta_k \nabla_{i_k} f(x^k) \mathbf{e}_{i_k}) - f(x^k) \\ &\stackrel{(38)}{\leq} -\frac{\eta_k}{2} \left(\nabla_{i_k} f(x^k) \right)^2. \end{aligned} \quad (51)$$

The rest of the proof is identical to the proof given in Appendix D.1 and leads to the same bound:

$$\mathbb{E}[f(x^N)] - f^* \leq \max \left\{ \left(1 - \frac{1}{4\sqrt{2}dL_1R} \right)^{N/2} F_0, \frac{16dL_0R^2}{N} \right\}.$$

The above upper bound implies that to achieve $\mathbb{E}[f(x^N)] - f^* \leq \varepsilon$, OrderRCD needs to perform

$$N = \mathcal{O} \left(\max \left\{ \frac{dL_0R^2}{\varepsilon}, dL_1R \log \frac{F_0}{\varepsilon} \right\} \right) \quad \text{iterations and}$$

$$T = \mathcal{O} \left(\max \left\{ \frac{dL_0R^2}{\varepsilon}, dL_1R \log \frac{F_0}{\varepsilon} \right\} \cdot \log \frac{1}{\epsilon} \right) \quad \text{Order Oracle calls,}$$

where ϵ is the accuracy of the solution of the auxiliary optimization problem (see line 2 in Algorithm 5), and it has to be sufficiently small.

E Missing Proof for GD in the Strongly Convex Setup

From the analysis of the convex case, we have

$$f(x^{k+1}) - f(x^k) \stackrel{(14)}{\leq} -\frac{1}{2(L_0 + L_1 \|\nabla f(x^k)\|)} \|\nabla f(x^k)\|^2. \quad (52)$$

Next, let us consider two cases: $\|\nabla f(x^k)\| \geq \frac{L_0}{L_1}$ and $\|\nabla f(x^k)\| < \frac{L_0}{L_1}$.

The case of $\|\nabla f(x^k)\| \geq \frac{L_0}{L_1}$. In this case, we have $L_0 + L_1 \|\nabla f(x^k)\| \leq 2L_1 \|\nabla f(x^k)\|$. Using this inequality in (52), we obtain

$$f(x^{k+1}) - f(x^k) \leq -\frac{\|\nabla f(x^k)\|}{4L_1}. \quad (53)$$

To continue the derivation, we also consider two possible situations depending on $F_k := f(x^k) - f^*$.

i) If $F_k \geq 1$, then we proceed as in the convex case and get

$$F_{k+1} \stackrel{(16)}{\leq} \left(1 - \frac{1}{4L_1R}\right) F_k. \quad (54)$$

In view of (52) and Lemma B.2, we have $F_{k+1} \leq F_k$ and $\|\nabla f(x^{k+1})\| \leq \|\nabla f(x^k)\|$. Therefore, if $F_k \geq 1$ and $\|\nabla f(x^k)\| \geq \frac{L_0}{L_1}$, then $F_t \geq 1$ and $\|\nabla f(x^t)\| \geq \frac{L_0}{L_1}$ for all $t = 0, 1, \dots, k$. Let \mathcal{T}_1 be the largest $k \in [N-1]$ such that $F_k \geq 1$ and $\|\nabla f(x^k)\| \geq \frac{L_0}{L_1}$ (if there is no such k for given initialization, then $\mathcal{T}_1 := -1$). Then, we have

$$F_{\mathcal{T}_1+1} \stackrel{(54)}{\leq} \left(1 - \frac{1}{4L_1R}\right)^{\mathcal{T}_1+1} F_0. \quad (55)$$

Using the above inequality, we can upper bound \mathcal{T}_1 as

$$\mathcal{T}_1 \leq 4L_1R \log(F_0).$$

ii) If $F_k < 1$, we use Polyak-Łojasiewicz (Polyak, 1963; Łojasiewicz, 1963) inequality

$$\|\nabla f(x^k)\|^2 \geq 2\mu F_k \quad (56)$$

$$\stackrel{F_k < 1}{>} 2\mu(F_k)^2, \quad (57)$$

which follows from strong convexity Nesterov (2018). Then, we can continue the derivation as follows:

$$\begin{aligned} F_{k+1} &\stackrel{(53)}{\leq} F_k - \frac{1}{4L_1} \|\nabla f(x^k)\| \\ &\stackrel{(57)}{\leq} \left(1 - \frac{\sqrt{\mu}}{2\sqrt{2}L_1}\right) F_k. \end{aligned} \quad (58)$$

Moreover, since (54) holds whenever $\|\nabla f(x^k)\| \geq \frac{L_0}{L_1}$, we can tighten the above inequality as

$$F_{k+1} \leq \left(1 - \max\left\{\frac{\sqrt{\mu}}{2\sqrt{2}L_1}, \frac{1}{4L_1R}\right\}\right) F_k. \quad (59)$$

In view of Lemma B.2, we have $\|\nabla f(x^{k+1})\| \leq \|\nabla f(x^k)\|$. Therefore, if $\|\nabla f(x^k)\| \geq \frac{L_0}{L_1}$, then $\|\nabla f(x^t)\| \geq \frac{L_0}{L_1}$ for all $t = 0, 1, \dots, k$. Let \mathcal{T}_2 be the largest $k \in [N-1]$ such that $\|\nabla f(x^k)\| \geq \frac{L_0}{L_1}$ (if there is no such k for given initialization, then $\mathcal{T}_2 := -1$). Then, we have

$$\begin{aligned} F_{\mathcal{T}_2+1} &\leq \left(1 - \max\left\{\frac{\sqrt{\mu}}{2\sqrt{2}L_1}, \frac{1}{4L_1R}\right\}\right)^{\mathcal{T}_2-\mathcal{T}_1} F_{\mathcal{T}_1+1} \\ &\stackrel{(55)}{\leq} \left(1 - \max\left\{\frac{\sqrt{\mu}}{2\sqrt{2}L_1}, \frac{1}{4L_1R}\right\}\right)^{\mathcal{T}_2-\mathcal{T}_1} \left(1 - \frac{1}{4L_1R}\right)^{\mathcal{T}_1+1} F_0. \end{aligned} \quad (60)$$

The case of $\|\nabla f(x^k)\| < \frac{L_0}{L_1}$. In this case, we have $L_0 + L_1\|\nabla f(x^k)\| \leq 2L_0$. Using this inequality in (52), we obtain

$$\begin{aligned} F_{k+1} &\stackrel{(52)}{\leq} F_k - \frac{\|\nabla f(x^k)\|^2}{4L_0} \\ &\stackrel{(56)}{\leq} \left(1 - \frac{\mu}{2L_0}\right) F_k. \end{aligned} \tag{61}$$

Since Algorithm 1 converges monotonically in terms of the gradient norm (see Appendix B), the above inequality holds for $k = \mathcal{T}_2 + 1, \mathcal{T}_2 + 2, \dots, N - 1$ iterations and gives

$$\begin{aligned} F_N &\leq \left(1 - \frac{\mu}{2L_0}\right)^{N-\mathcal{T}_2} F_{\mathcal{T}_2+1} \\ &\stackrel{(60)}{\leq} \left(1 - \frac{\mu}{2L_0}\right)^{N-\mathcal{T}_2} \left(1 - \max\left\{\frac{\sqrt{\mu}}{2\sqrt{2}L_1}, \frac{1}{4L_1R}\right\}\right)^{\mathcal{T}_2-\mathcal{T}_1} \left(1 - \frac{1}{4L_1R}\right)^{\mathcal{T}_1+1} F_0. \end{aligned}$$

This concludes the proof.

F Motivation Strong Growth Conditions on the Example of Logistic Regression

We know from Section 6 that the strong growth condition for smoothness (see Assumption 1.2 when $L_0 = 0$) is satisfied by the logistic regression problem, which is often used in the machine learning community. However, this problem does not reach a minimum (hence $R = \arg \inf f(x) = +\infty$). Therefore, in this section we show that, for example, gradient descent (Algorithm 1) will achieve the desired accuracy ε in a finite number of iterations with linear rate.

We introduce the hyperparameter of the algorithm $s : f(s) - f^* \leq \varepsilon$. Then we show linear convergence to the desired accuracy by the example of gradient descent.

Using the strong growth condition for smoothness (see Assumption 1.2 with $L_0 = 0$) we have:

$$\begin{aligned} f(x^{k+1}) - f(x^k) &= f(x^k - \eta_k \nabla f(x^k)) - f(x^k) \\ &\stackrel{(8)}{\leq} -\eta_k \langle \nabla f(x^k), \nabla f(x^k) \rangle + \eta_k^2 \frac{L_1 \|\nabla f(x^k)\|}{2} \|\nabla f(x^k)\|^2 \\ &\stackrel{\textcircled{1}}{\leq} -\eta_k \|\nabla f(x^k)\|^2 + \frac{\eta_k}{2} \|\nabla f(x^k)\|^2 \\ &= -\frac{\eta_k}{2} \|\nabla f(x^k)\|^2 \\ &= -\frac{1}{2L_1 \|\nabla f(x^k)\|} \|\nabla f(x^k)\|^2 \\ &= -\frac{1}{2L_1} \|\nabla f(x^k)\|, \end{aligned} \tag{62}$$

where in $\textcircled{1}$ we used $\eta_k \leq \frac{1}{L_1 \|\nabla f(x^k)\|}$.

Then using the convexity assumption of the function (see Assumption 1.4, $\mu = 0$), we have the following:

$$\begin{aligned} f(x^k) - f(s) &\leq \langle \nabla f(x^k), x^k - s \rangle \\ &\stackrel{(7)}{\leq} \|\nabla f(x^k)\| \|x^k - s\| \\ &\leq \|\nabla f(x^k)\| \underbrace{\|x^0 - s\|}_{R_s}. \end{aligned}$$

Hence we have:

$$\|\nabla f(x^k)\| \geq \frac{f(x^k) - f(s)}{R_s}. \quad (63)$$

Then substituting (63) into (62) we obtain:

$$f(x^{k+1}) - f(x^k) \leq -\frac{1}{2L_1} \|\nabla f(x^k)\| \leq -\frac{1}{2L_1 R_s} (f(x^k) - f(s)).$$

This inequality is equivalent to the trailing inequality:

$$f(x^{k+1}) - f^* \leq \left(1 - \frac{1}{2L_1 R_s}\right) (f(x^k) - f^*) + \frac{1}{2L_1 R_s} (f(s) - f^*). \quad (64)$$

Applying recursion to (64) we obtain:

$$f(x^N) - f^* \leq \left(1 - \frac{1}{2L_1 R_s}\right)^N (f(x^0) - f^*) + \underbrace{\frac{1}{2L_1 R_s} (f(s) - f^*)}_{\varepsilon}.$$

Therefore, we have shown that Algorithm 1 will achieve the desired accuracy ε in a finite number of iterations: $N = \mathcal{O}\left(L_1 R_s \log \frac{1}{\varepsilon}\right)$. R_s is a finite number and increases as the desired accuracy improves. The same can be shown for other algorithms.