

An Adaptive and Parameter-Free Nesterov’s Accelerated Gradient Method for Convex Optimization

Jaewook J. Suh*

Shiqian Ma*

Abstract

We propose AdaNAG, an adaptive accelerated gradient method based on Nesterov’s accelerated gradient method. AdaNAG is line-search-free, parameter-free, and achieves the accelerated convergence rates $f(x_k) - f_\star = \mathcal{O}(1/k^2)$ and $\min_{i \in \{1, \dots, k\}} \|\nabla f(x_i)\|^2 = \mathcal{O}(1/k^3)$ for L -smooth convex function f . We provide a Lyapunov analysis for the convergence proof of AdaNAG, which additionally enables us to propose a novel adaptive gradient descent (GD) method, AdaGD. AdaGD achieves the non-ergodic convergence rate $f(x_k) - f_\star = \mathcal{O}(1/k)$, like the original GD. The analysis of AdaGD also motivated us to propose a generalized AdaNAG that includes practically useful variants of AdaNAG. Numerical results demonstrate that our methods outperform some other recent adaptive methods for representative applications.

1 Introduction

With the growing prevalence of large-scale optimization problems, there has been increasing interest in adaptive optimization methods that are line-search-free and parameter-free. Initiated by [26] and further developed in [23, 22, 27, 40], the study of adaptive methods for convex optimization has been actively explored in recent years. However, research on adaptive accelerated gradient methods remains limited, and the first such algorithm was recently proposed in [25].

In this paper, we propose a novel adaptive Nesterov’s accelerated gradient method (AdaNAG) for solving the minimization problem

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

where $f: \mathbb{R}^d \rightarrow \mathbb{R}$ is a convex and (locally) smooth function. We use x_\star to denote a minimizer of (1) and f_\star to denote the optimal value. Our method achieves the accelerated convergence rates $f(x_k) - f_\star = \mathcal{O}(1/k^2)$ and $\min_{i \in \{1, \dots, k\}} \|\nabla f(x_i)\|^2 = \mathcal{O}(1/k^3)$, similar to the original Nesterov’s accelerated gradient method (NAG) [28, 17, 33]. We establish the convergence of the method using a novel Lyapunov analysis, which also provides intuition for developing new adaptive methods. Based on this intuition, we further propose a novel adaptive gradient descent method (AdaGD) that achieves the non-ergodic convergence rate $f(x_k) - f_\star = \mathcal{O}(1/k)$. In addition, we provide a family of algorithms that generalize AdaNAG, and present instances that are both practically useful and theoretically sound. We demonstrate the efficiency of our algorithms through several practical scenarios in which they outperform existing adaptive methods.

1.1 Prior works

Nesterov acceleration. Nesterov’s accelerated gradient (NAG) method was introduced in his seminal work [28]. NAG achieves an accelerated convergence rate $f(x_k) - f_\star = \mathcal{O}(1/k^2)$ when

*Department of Computational Applied Mathematics and Operations Research, Rice University, Houston, TX, USA.

f is a globally smooth convex function, while using only first-order information. This rate has been proven to be optimal for this problem setup up to order, as shown in [29, Theorem 2.1.7]. As the demand for first-order methods grew due to the need for large-scale optimization problems, NAG has been studied in various aspects [34, 35, 39, 24, 10, 33], and numerous variants have been proposed in recent decades [5, 7, 30, 14, 2, 18, 20], to name a few. Optimized gradient method (OGM) [14, 15] improved the constant of the bound by 2, and is proven to be the optimal method for L -smooth convex function minimization [11]. FISTA [5] extended NAG to the proximal gradient (prox-grad) algorithm with the same convergence rate. The works in [34, 35, 7, 2] extended the original parameter choice $\theta_k = \frac{k+2}{2}$ in NAG to $\theta_k = \frac{k+p}{p}$ with $p \geq 2$, and still achieved the accelerated rate $f(x_k) - f_* = \mathcal{O}(1/k^2)$. In addition, [7, 2] further established iterate convergence when $p > 2$. An improved convergence rate for the gradient norm square $\min_{i \in \{1, \dots, k\}} \|\nabla f(x_i)\|^2 = \mathcal{O}(1/k^3)$ was provided in [17, 33].

Recently, OptISTA [13] improved the convergence rate of FISTA by a constant factor of 2, and was proven to be optimal for the prox-grad optimization setup. Unlike NAG and FISTA, which achieve the same rate, the optimal methods OGM and OptISTA have different rates, with the rate of OGM having a slightly better constant than that of OptISTA. This difference suggests that we may achieve improved convergence rates by focusing on single-function minimization when designing an algorithm. This motivated us to develop a new accelerated adaptive algorithm refined on the single function minimization problem, taking OGM as the reference.

Adaptive methods for convex function minimization. The demand for adaptive and parameter-free methods has grown as data size increases and as large-scale optimization problems become more prevalent. The seminal work [26] initiated the study of adaptive methods for convex minimization problems, without using additional expensive computations such as line search [12, 21, 30]. They introduced the adaptive gradient descent (AdGD) method, which achieves an ergodic convergence rate of $\mathcal{O}(1/k)$ for the function value error. Without requiring knowledge of the smoothness parameter or performing line-search computations, AdGD selects the step size based on an approximation of the local smoothness parameter using the last two iterates.

Introducing AdaPGM, an extension to the prox-grad setup was considered in [31]. A generalized version of AdaPGM, AdaPG^{q,r} was introduced in [22]. The analysis of the original AdGD was refined in [27], where an advanced method with a larger step size (AdGD2) was proposed, along with an extension to the prox-grad setup based on the refined proof. An adaptive method based on the Barzilai-Borwein method [4], AdaBB, was introduced in [40]. However, none of these adaptive gradient descent-based methods achieve a non-ergodic convergence rate, which was identified as a challenging problem in [27].

Recently, an adaptive method that achieves the accelerated convergence rate $f(x_k) - f_* = \mathcal{O}(1/k^2)$, called the auto-conditioned fast gradient method (AC-FGM), was introduced by [25]. Focusing on the prox-grad setup, they introduced an additional auxiliary sequence called “prox-centers,” which does not appear in the original Nesterov’s acceleration. Furthermore, they extended their analysis to convex problems with Hölder continuous gradients. We focus on the single-function convex minimization problem and propose a novel adaptive accelerated method that does not require such an auxiliary sequence.

Lyapunov analysis. A refined proof technique plays an important role in analyzing adaptive methods, particularly in deriving new step sizes and designing new algorithms as observed in [27, 22]. Among such techniques, Lyapunov analysis is a powerful framework that leverages a nonincreasing potential function, and it has been widely used to analyze a variety of opti-

mization methods, including accelerated methods [5, 3, 38, 10, 24, 32, 37] and adaptive methods [26, 22, 31, 27, 40]. Lyapunov analysis is especially crucial for adaptive methods, since such analyses typically use only recent iterates x_k and x_{k+1} , which provides strong motivation to approximate the local smoothness parameter using recent iterates. In this work, we present a novel Lyapunov analysis for an adaptive accelerated method. Our proof is inspired by the Lyapunov analysis of OGM [32, 10]. As a consequence of our analysis, we derive a novel convergence result $\min_{i \in \{1, \dots, k\}} \|\nabla f(x_i)\|^2 = \mathcal{O}(1/k^3)$, which is, to our knowledge, the first such rate established for adaptive methods. Furthermore, we design a new gradient-descent-type algorithm that achieves a non-ergodic convergence rate $f(x_k) - f_\star = \mathcal{O}(1/k)$, as an application of our Lyapunov analysis.

	$f(x) - f_\star$	non-ergodic	$\min_{i \in \{1, \dots, k\}} \ \nabla f(x_i)\ ^2$	without momentum
AdGD2 [27]	$\mathcal{O}(1/k)$	\times	$\mathcal{O}(1/k)$	\checkmark
AdaPGM [31, 22]	$\mathcal{O}(1/k)$	\times	$\mathcal{O}(1/k)$	\checkmark
AdaBB [40]	$\mathcal{O}(1/k)$	\times	$\mathcal{O}(1/k)$	\checkmark
AdaGD (ours)	$\mathcal{O}(1/k)$	\checkmark	$\mathcal{O}(1/k^2)$	\checkmark
AC-FGM [25]	$\mathcal{O}(1/k^2)$	\checkmark	$\mathcal{O}(1/k^2)$	\times
AdaNAG (ours)	$\mathcal{O}(1/k^2)$	\checkmark	$\mathcal{O}(1/k^3)$	\times

Table 1: Comparison between our methods and previous adaptive methods. In cases where the convergence rate of $\min_{i \in \{1, \dots, k\}} \|\nabla f(x_i)\|^2$ is not explicitly mentioned, we apply $\|\nabla f(x)\|^2 \leq 2L(f(x) - f_\star)$, assuming the smoothness parameter L .

1.2 Contributions

The main contributions of this work are as follows.

- We provide a novel accelerated adaptive algorithm, **AdaNAG**, that is line-search-free, parameter-free, and achieves the accelerated convergence rate $f(x_k) - f_\star = \mathcal{O}(1/k^2)$. Compared to AC-FGM, we have a better coefficient of $\|x_0 - x_\star\|^2$ in the upper bound of the convergence rate of **AdaNAG**. Moreover, **AdaNAG** achieves a convergence rate for the minimum of the squared gradient norm, $\min_{i \in \{1, \dots, k\}} \|\nabla f(x_i)\|^2 = \mathcal{O}(1/k^3)$. To the best of our knowledge, the latter is the first result of its kind for adaptive algorithms.
- We provide a Lyapunov analysis that allows for a concise understanding of the proof for **AdaNAG**. Such understanding motivates us to consider a strategy to develop a novel adaptive method, which leads to a novel adaptive gradient descent algorithm, **AdaGD**, with a non-ergodic $\mathcal{O}(1/k)$ convergence rate. This is the first non-ergodic convergence result for adaptive methods without momentum. **AdaGD** also achieves the convergence rate $\min_{i \in \{1, \dots, k\}} \|\nabla f(x_i)\|^2 = \mathcal{O}(1/k^2)$.
- Finally, we provide a family of algorithms that generalize AdaNAG and enable practically useful parameter selection. We introduce our ultimate algorithms, **AdaNAG-G₁₂** and **AdaNAG-G^{1/2}**, which are both theoretically sound and practically useful. We demonstrate the efficiency of our algorithms by showing that they outperform prior algorithms, including AC-FGM, when applied to some representative applications.

Organization. The rest of the paper is organized as follows. In Section 2, we introduce the preliminary concepts for our analysis. In Section 3, we introduce our adaptive accelerated algorithm **AdaNAG** and present the main convergence results, whose proofs, based on a Lyapunov

analysis, are provided in Section 4. In Section 5, we introduce a novel family of GD-type adaptive algorithms, [AdaGD](#). In Section 6, we introduce a family of adaptive algorithms that generalizes [AdaNAG](#) and allows for practically useful parameter selection. As concrete instances derived from this family, we present [AdaNAG-G₁₂](#) and [AdaNAG-G^{1/2}](#), both of which are theoretically sound and practically useful. Numerical results are presented in Section 7. Finally, we draw some concluding remarks in Section 8.

2 Preliminaries

f is called L -smooth if it is differentiable and ∇f is L -Lipschitz continuous, i.e.,

$$\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|, \quad \forall x, y \in \mathbb{R}^d.$$

When f is an L -smooth convex function, it is known [29, Theorem 2.1.5] that the following inequality holds for all $x, y \in \mathbb{R}^d$

$$f(y) - f(x) + \langle \nabla f(y), x - y \rangle + \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|^2 \leq 0. \quad (2)$$

We say f is *locally smooth* if ∇f is locally Lipschitz continuous, i.e., if for every compact set $K \subset \mathbb{R}^d$ there exists $L_K > 0$ such that

$$\|\nabla f(x) - \nabla f(y)\| \leq L_K \|x - y\|, \quad \forall x, y \in K.$$

We refer to L_K as the (*local*) *smoothness parameter* of f on K . It is clear that L -smoothness implies local smoothness. It is natural to ask whether a locally smooth convex function also satisfies an inequality similar to (2). The answer is positive. However, to the best of our knowledge, this is not immediate from known facts for globally smooth functions. Therefore, we state it as a lemma, along with a useful corollary that will be used when we approximate the local smoothness parameter in our later arguments. We provide the proofs in Appendix A.

Lemma 1. *Suppose f is a locally smooth convex function and $K \subset \mathbb{R}^d$ is a compact set. Let $K \subset \bar{B}_R(c)$, where $\bar{B}_R(c) = \{x \mid \|x - c\| \leq R\}$. Let $\bar{L}_K > 0$ be a smoothness parameter of f on $\bar{B}_{3R}(c)$. Then the following inequality is true*

$$f(y) - f(x) + \langle \nabla f(y), x - y \rangle + \frac{1}{2\bar{L}_K} \|\nabla f(x) - \nabla f(y)\|^2 \leq 0.$$

Note that when f is a (global) L -smooth convex function, the above statements hold with $\bar{L}_K = L$.

Corollary 2. *Let f be a locally smooth convex function. Then following statements are true:*

- (i) *For all $x, y \in \mathbb{R}^d$, if $f(y) - f(x) + \langle \nabla f(y), x - y \rangle = 0$, then $\|\nabla f(x) - \nabla f(y)\|^2 = 0$.*
- (ii) *Let $K \subset \mathbb{R}^d$ be a compact set. Then there exists $\bar{L}_K > 0$ such that*

$$0 \leq -\frac{\frac{1}{2} \|\nabla f(x) - \nabla f(y)\|^2}{f(y) - f(x) + \langle \nabla f(y), x - y \rangle} \leq \bar{L}_K, \quad \forall x, y \in K \text{ with } f(y) - f(x) + \langle \nabla f(y), x - y \rangle \neq 0.$$

In our analysis, we need an initial guess (say, L_0) to the smoothness parameter L such that $0 < L_0 \leq L$. This can be done by choosing two arbitrary points x_0, \tilde{x}_0 with $x_0 \neq \tilde{x}_0$ and defining

$$L_0 = \frac{\|\nabla f(x_0) - \nabla f(\tilde{x}_0)\|}{\|x_0 - \tilde{x}_0\|}. \quad (3)$$

Moreover, we adopt the convention $\frac{0}{0} = 0$ and $\frac{a}{0} = +\infty$ for all $a > 0$ when the denominator is zero.

3 Our AdaNAG Algorithm

One typical iteration of OGM and NAG can be written in the following form:

$$y_{k+1} = x_k - s_k \nabla f(x_k), \quad z_{k+1} = z_k - s_k \alpha_k \theta_k \nabla f(x_k), \quad x_{k+1} = \left(1 - \frac{1}{\theta_{k+1}}\right) y_{k+1} + \frac{1}{\theta_{k+1}} z_{k+1}.$$

Both methods consider constant choice $s_k = \frac{1}{L}$. As studied in [14, 32], this reduces to OGM when $\alpha_k = 2$, and reduces to NAG when $\alpha_k = 1$. We want to find proper s_k, α_k without knowing L such that similar proof structure of OGM works, so to achieve similar accelerated convergence rate.

3.1 AdaNAG

Our *Adaptive Nesterov's Accelerated Gradient (AdaNAG)* method is described in Algorithm 1. The convergence results of [AdaNAG](#) are summarized in Theorem 1, and its proof is given in Section 4.

Algorithm 1 Adaptive Nesterov Accelerated Gradient (AdaNAG)

1: **Input:** $x^0 = z^0 \in \mathbb{R}^d, s_0 > 0$.

2: **Define**

$$\theta_k = \begin{cases} 1 & \text{if } k = 0 \\ \frac{1}{2} (1 + \sqrt{1 + 4\theta_{k-1}^2}) & \text{if } k \geq 1, \end{cases} \quad \alpha_k = \begin{cases} \frac{2\theta_2}{\theta_2 - 1} \left(\frac{1}{\alpha_3} + \frac{1}{\alpha_2^2} - \frac{1}{\alpha_1} \right)^{-1} & \text{if } k = 0 \\ \frac{1}{2} \left(1 - \frac{1}{\theta_{k+2}} \right) & \text{if } k \geq 1 \end{cases} \quad (4)$$

3: **for** $k = 0, 1, \dots$ **do**

4:

$$\begin{aligned} y_{k+1} &= x_k - s_k \nabla f(x_k) \\ z_{k+1} &= z_k - s_k \alpha_k \theta_{k+2} \nabla f(x_k) \\ x_{k+1} &= \left(1 - \frac{1}{\theta_{k+3}}\right) y_{k+1} + \frac{1}{\theta_{k+3}} z_{k+1} \end{aligned} \quad (5)$$

$$L_{k+1} = - \frac{\frac{1}{2} \|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2}{f(x_{k+1}) - f(x_k) + \langle \nabla f(x_{k+1}), x_k - x_{k+1} \rangle} \quad (6)$$

$$s_{k+1} = \begin{cases} \min \left\{ \frac{\alpha_0}{\alpha_1} \frac{\theta_2}{\theta_3 - 1} s_0, \frac{\alpha_2^2 \alpha_3}{\alpha_3 + \alpha_2^2} \frac{1}{\alpha_1} \frac{1}{L_1} \right\} & \text{if } k = 0 \\ \min \left\{ \frac{\alpha_k}{\alpha_{k+1}} s_k, \frac{\alpha_k^2}{\alpha_{k+1} + \alpha_k^2} \frac{1}{L_{k+1}} \right\} & \text{if } k \geq 1 \end{cases} \quad (7)$$

5: **end for**

Theorem 1. Let f be an L -smooth convex function. Suppose $\{x_k\}_{k \geq 0}$ is a sequence generated by [AdaNAG](#) (Algorithm 1) with $s_0 = 0.4255 \frac{1}{L_0}$ where L_0 defined as in (3). Denote $R = \|x_0 - x_\star\|^2 + 0.14 \frac{1}{L_0} \left(\frac{1}{L_0} - \frac{2}{L} \right) \|\nabla f(x_0)\|^2$. Then the following convergence rate results hold.

$$\begin{aligned} f(x_k) - f_\star &\leq \frac{5.5L}{\theta_{k+2}^2} R \leq \frac{22L}{(k+4)^2} R = \mathcal{O} \left(\frac{L}{k^2} \right), \\ \min_{i \in \{1, \dots, k\}} \|\nabla f(x_i)\|^2 &\leq \frac{120L^2}{\sum_{i=1}^k \theta_{i+2}^2} R \leq \frac{1440L^2}{k(k^2 + 12k + 47)} R = \mathcal{O} \left(\frac{L^2}{k^3} \right). \end{aligned} \quad (8)$$

Note that when the denominator of L_{k+1} in (6) is zero, the numerator is also zero by Corollary 2, so we have $L_{k+1} = 0$ from our arithmetic convention. There are several useful properties of θ_k , α_k and s_k and we summarize them in Lemma 3, whose proof is given in Appendix B.1.

Lemma 3. Suppose $\{\theta_k\}_{k \geq 0}$, $\{\alpha_k\}_{k \geq 0}$ and $\{s_k\}_{k \geq 0}$ are defined as in (4) and (7). Then θ_k is an increasing sequence that satisfies $\theta_k \geq \frac{k+2}{2}$ for $k \geq 0$. Furthermore, the following inequality holds

$$\theta_{k+1}(\theta_{k+1} - 1) - \theta_k^2 \leq 0, \quad \forall k \geq 0. \quad (9)$$

Thus, $\alpha_{k+1} \geq \alpha_k$ for $k \geq 1$ and $\lim_{k \rightarrow \infty} \alpha_k = \frac{1}{2}$. Moreover, $s_{k+1} \leq s_k$ for $k \geq 0$.

Note that the definition of θ_k in (4) is the same as the one used for the original NAG [28], which satisfies (9) as an equality. The inequality (9) is often considered a core property of θ_k in studies of variants of NAG [5, 7, 32], and will play a crucial role in our convergence analysis. As an additional comment, we note that

$$\theta_k = \frac{k+2}{2} \quad (10)$$

is also often considered an alternative to (4), as it behaves similarly but is simpler and more intuitive. We show that a simplified variant of AdaNAG obtained by replacing (4) with (10) can also achieve a similar convergence rate in Appendix B.2.

3.2 Comparison with AC-FGM

AC-FGM is another adaptive method that achieves accelerated convergence rate, studied in [25]. When $h = 0$, $\mathcal{X} = \mathbb{R}^d$, AC-FGM reduces to:

$$\begin{aligned} z_{k+1} &= y_k - \eta_{k+1} \nabla f(x_k) \\ y_{k+1} &= (1 - \beta_{k+1})y_k + \beta_{k+1}z_{k+1} \\ x_{k+1} &= \frac{\tau_{k+1}}{1 + \tau_{k+1}}x_k + \frac{1}{1 + \tau_{k+1}}z_{k+1}. \end{aligned} \quad (\text{AC-FGM})$$

Note that AC-FGM has a different form from our AdaNAG, and therefore, the mechanism of the coefficients differs. The concrete choice of parameters in AC-FGM are ([25, Corollary 1]):

$$\tau_k = \begin{cases} 0 & \text{if } k = 1 \\ \frac{k}{2} & \text{if } k \geq 2, \end{cases} \quad \beta_k = \begin{cases} 0 & \text{if } k = 1 \\ \beta & \text{if } k \geq 2, \end{cases} \quad \eta_k = \begin{cases} \min \left\{ (1 - \beta)\eta_1, \frac{1}{4L_1} \right\} & \text{if } k = 2 \\ \min \left\{ \eta_2, \frac{1}{4L_2} \right\} & \text{if } k = 3 \\ \min \left\{ \frac{k}{k-1}\eta_{k-1}, \frac{k-1}{8L_{k-1}} \right\} & \text{if } k \geq 4, \end{cases}$$

with L_k defined in the same way as (6). Then AC-FGM has the convergence rate

$$f(x_k) - f_* \leq \frac{(12/\beta)\hat{L}_k}{(k+1)(k+2)} \left(\|x_0 - x_*\|^2 + \beta \left(\frac{5\eta_2 L_1}{2} - \frac{\eta_2}{\eta_1} \right) \eta_1^2 \|\nabla f(x_0)\|^2 \right), \quad (11)$$

where $\eta_1 > 0$ is arbitrary and $\hat{L}_k = \max \left\{ \frac{1}{4(1-\beta)\eta_1}, L_1, \dots, L_k \right\}$, $\beta \in (0, 1 - \sqrt{6}/3]$. The coefficient of $\|\nabla f(x_0)\|^2$ differs depending on the relationship between η_1 , L_1 , and η_2 . Thus, this coefficient fundamentally contains some uncertainty, depending on how well we estimated the initial step size η_1 and the point x_1 , which we do not have full control over. The authors provided a line search technique to adjust this coefficient, but in [25, § 4.4], they presented an ablation study showing that such an adjustment does not make a significant difference in various practical scenarios. In

this context, we believe there is no absolute way to compare this coefficient, or that this coefficient is not crucial to the algorithm's behavior. Therefore, we focus more on the coefficient of $\|x_0 - x_\star\|^2$. In [25], they choose $\beta = 1 - \frac{\sqrt{6}}{3}$ for the experiments, which also results in the tightest coefficient of $\|x_0 - x_\star\|^2$, which is $12/(1 - \frac{\sqrt{6}}{3}) \approx 65.39$. This is nearly three times larger compared to the coefficient of $\|x_0 - x_\star\|^2$ in (8), which is 22.

4 Convergence analysis of AdaNAG

We note that the proofs in this section only assume f to be locally smooth rather than globally smooth, except for Corollary 9 and Proposition 11. On the other hand, Corollary 9 and Proposition 11 can even be achieved under the locally smooth condition if we consider a small perturbation to (7) as $s_{k+1} = \min \left\{ \frac{\alpha_k}{\alpha_{k+1}} s_k, \frac{\alpha_k^2}{\alpha_{k+1} + \alpha_k^2(1+10^{-6})} \frac{1}{L_{k+1}} \right\}$. Since the value 10^{-6} is negligibly small compared to α_k^2 and α_{k+1} , the overall behavior of the perturbed variant remains the same as that of AdaNAG, including the convergence guarantee. However, since the analysis requires more technical arguments, we provide the formal discussion of this variant and its proof in Appendix B.3.

4.1 Motivation of the Lyapunov function and L_{k+1}

Since we want to design an adaptive method that mimics OGM [14, 15], it is worth briefly reviewing the proof structure of OGM. OGM considers Lyapunov function of the form

$$V_k^{\text{OGM}} = \frac{1}{L} 2\theta_{k+1}(\theta_{k+1} - 1) (f(x_k) - f_\star) - \frac{1}{L^2} \theta_k^2 \|\nabla f(x_k)\|^2 + \frac{1}{2} \|z_{k+1} - x_\star\|^2. \quad (12)$$

The core part of the convergence proof for OGM is to show:

$$V_{k+1}^{\text{OGM}} - V_k^{\text{OGM}} = \frac{2}{L} \theta_{k+1}(\theta_{k+1} - 1) \mathcal{I}_1 + \frac{2}{L} \theta_{k+1} \mathcal{I}_2 \leq 0, \quad (13)$$

where $\mathcal{I}_1 = f(x_{k+1}) - f(x_k) + \langle \nabla f(x_{k+1}), x_k - x_{k+1} \rangle + \frac{1}{2L} \|\nabla f(x_k) - \nabla f(x_{k+1})\|^2$ and $\mathcal{I}_2 = f(x_{k+1}) - f_\star - \langle \nabla f(x_{k+1}), x_{k+1} - x_\star \rangle + \frac{1}{2L} \|\nabla f(x_{k+1})\|^2$. For details, see [10, § 4.3] or [32]. Note \mathcal{I}_1 and \mathcal{I}_2 are nonpositive since f is L -smooth convex by (2). We want to set the parameters to construct a similar proof for (13) when L is not known and is replaced by some estimation L_{k+1} . A good way of doing it is to simply *define* it as nonpositive. That is, define L_{k+1} as in (6) and make \mathcal{I}_1 zero. The idea of defining an adaptive step size motivated by the proof of an optimal non-adaptive method was also considered in [36, § 6]. We note that this specific definition of L_{k+1} was also previously considered in [25].

However, we cannot treat \mathcal{I}_2 using the same approach because it requires information about the optimal point and value, f_\star and x_\star . Thus instead, we forgo the term $\frac{2}{L} \theta_{k+1} \times \frac{1}{2L} \|\nabla f(x_{k+1})\|^2$ and use the relaxed inequality $f(x_{k+1}) - f_\star - \langle \nabla f(x_{k+1}), x_{k+1} - x_\star \rangle \leq 0$. Then, the $\|\nabla f(x_{k+1})\|^2$ term should be accounted for somewhere. This requires us to adjust the definition of the Lyapunov function by flipping the sign of the $\|\nabla f(x_k)\|^2$ term. This ultimately leads to our Lyapunov function for AdaNAG. The details will be provided in the next section.

4.2 Convergence analysis of AdaNAG

The following theorem is the core part of our convergence analysis and one of the main contributions of our paper.

Theorem 2. For *AdaNAG* (Algorithm 1), define Lyapunov function as

$$V_k = s_{k+1}A_k (f(x_k) - f_\star) + \frac{1}{2}s_k^2B_k \|\nabla f(x_k)\|^2 + \frac{1}{2}\|z_{k+1} - x_\star\|^2, \quad \forall k \geq 0, \quad (14)$$

where A_k, B_k are defined as

$$A_k := \alpha_{k+1}\theta_{k+3}(\theta_{k+3} - 1), \quad B_k := \begin{cases} \tilde{\alpha}_0\alpha_0\theta_2^2 & \text{if } k = 0 \\ \alpha_k^2\theta_{k+2}^2 & \text{if } k \geq 1, \end{cases} \quad \text{with } \tilde{\alpha}_0 := \frac{1}{2}\left(1 - \frac{1}{\theta_2}\right). \quad (15)$$

Then for a locally smooth function f , we have

$$V_{k+1} - V_k \leq -\frac{1}{2} \min\left\{s_k^2B_k, \frac{s_{k+1}}{L_{k+1}}A_k\right\} \|\nabla f(x_k)\|^2 \leq 0, \quad k \geq 0.$$

Proof. We first consider the case when $\nabla f(x_k) \neq \nabla f(x_{k+1})$, which implies that the denominator of L_{k+1} is nonzero by Corollary 2. The proof can be broadly summarized into four parts:

Part 1. (The proof is given in Lemma 4). Show the following inequalities hold for $k \geq 0$. Note that these are the core properties of the step size and parameters.

$$s_{k+2} \leq \frac{A_k + \alpha_{k+1}\theta_{k+3}}{A_{k+1}} s_{k+1} \quad (16a)$$

$$s_{k+1} \leq \rho_k^{-1} s_k \quad (16b)$$

$$s_{k+1} \leq \left(\frac{A_k}{B_k} + \rho_k\right)^{-1} \frac{1}{L_{k+1}}, \quad (16c)$$

where

$$\rho_k = \frac{B_{k+1} + \alpha_{k+1}^2\theta_{k+3}^2}{A_k}. \quad (17)$$

Part 2. (The proof is given in Lemma 5). Show the following equality holds for $k \geq 0$:

$$\begin{aligned} & V_{k+1} - V_k \\ &= (s_{k+1}(A_k + \alpha_{k+1}\theta_{k+3}) - s_{k+2}A_{k+1})(f_\star - f(x_{k+1})) \\ & \quad + s_{k+1}\alpha_{k+1}\theta_{k+3}(f(x_{k+1}) - f_\star - \langle \nabla f(x_{k+1}), x_{k+1} - x_\star \rangle) \\ & \quad + s_{k+1}A_k \left(f(x_{k+1}) - f(x_k) + \langle \nabla f(x_{k+1}), x_k - x_{k+1} \rangle + \frac{1}{2L_{k+1}} \|\nabla f(x_k) - \nabla f(x_{k+1})\|^2 \right) \\ & \quad + Q_k, \end{aligned} \quad (18)$$

where Q_k is defined as

$$\begin{aligned} Q_k = & -\frac{s_{k+1}A_k}{2L_{k+1}} \left[(1 - \rho_k s_{k+1}L_{k+1}) \|\nabla f(x_{k+1})\|^2 \right. \\ & \left. - 2(1 - s_k L_{k+1}) \langle \nabla f(x_{k+1}), \nabla f(x_k) \rangle + \left(1 + \frac{B_k}{A_k} \frac{s_k^2}{s_{k+1}} L_{k+1}\right) \|\nabla f(x_k)\|^2 \right]. \end{aligned} \quad (19)$$

Part 3. (The proof is given in Lemma 6). Show the following inequality holds

$$Q_k \leq \bar{Q}_k, \quad \forall k \geq 0, \quad (20)$$

where

$$\bar{Q}_k := \begin{cases} -\frac{1}{2}s_k^2 B_k \|\nabla f(x_k)\|^2 & \text{if } s_k L_{k+1} \leq 1 \\ -\frac{1}{2} \frac{s_{k+1}}{L_{k+1}} A_k \|\nabla f(x_k)\|^2 & \text{if } s_k L_{k+1} \geq 1. \end{cases}$$

Part 4. Finally, we show that the first three lines on the right hand side of (18) are nonpositive:

- $s_{k+1}(A_k + \alpha_{k+1}\theta_{k+3}) - s_{k+2}A_{k+1} \geq 0$ follows from (16a).
- $f(x_{k+1}) - f_\star - \langle \nabla f(x_{k+1}), x_{k+1} - x_\star \rangle \leq 0$ holds by the convexity of f .
- $f(x_{k+1}) - f(x_k) + \langle \nabla f(x_{k+1}), x_k - x_{k+1} \rangle + \frac{1}{2L_{k+1}} \|\nabla f(x_k) - \nabla f(x_{k+1})\|^2 = 0$ holds by (6).

Hence, we obtain that

$$V_{k+1} - V_k \leq Q_k, \quad \forall k \geq 0, \quad (21)$$

and the desired conclusion for the case $\nabla f(x_k) \neq \nabla f(x_{k+1})$ follows from **Part 3**.

When $\nabla f(x_k) = \nabla f(x_{k+1})$. We have $L_{k+1} = 0$ in this case, and hence $s_k L_{k+1} = 0 \leq 1$. Now, by conducting the same calculation as in **Part 2** but without the term $\frac{s_{k+1}A_k}{2L_{k+1}} \|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2$, we obtain the same equation as (18), except that this term is omitted:

$$\begin{aligned} V_{k+1} - V_k &= (s_{k+1}(A_k + \alpha_{k+1}\theta_{k+3}) - s_{k+2}A_{k+1})(f_\star - f(x_{k+1})) \\ &\quad + s_{k+1}\alpha_{k+1}\theta_{k+3}(f(x_{k+1}) - f_\star - \langle \nabla f(x_{k+1}), x_{k+1} - x_\star \rangle) \\ &\quad + s_{k+1}A_k(f(x_{k+1}) - f(x_k) + \langle \nabla f(x_{k+1}), x_k - x_{k+1} \rangle) \\ &\quad + \frac{1}{2}s_{k+1}^2(B_{k+1} + \alpha_{k+1}^2\theta_{k+3}^2)\|\nabla f(x_{k+1})\|^2 - \frac{1}{2}s_k^2B_k\|\nabla f(x_k)\|^2 - s_k s_{k+1}A_k \langle \nabla f(x_{k+1}), \nabla f(x_k) \rangle \\ &\leq \frac{1}{2}s_{k+1}^2(B_{k+1} + \alpha_{k+1}^2\theta_{k+3}^2)\|\nabla f(x_{k+1})\|^2 - \frac{1}{2}s_k^2B_k\|\nabla f(x_k)\|^2 - s_k s_{k+1}A_k \langle \nabla f(x_{k+1}), \nabla f(x_k) \rangle \\ &= \frac{1}{2}s_{k+1}A_k \left(\frac{B_{k+1} + \alpha_{k+1}^2\theta_{k+3}^2}{A_k} s_{k+1} - s_k \right) \|\nabla f(x_k)\|^2 - \frac{1}{2}s_k^2B_k\|\nabla f(x_k)\|^2 - \frac{1}{2}s_k s_{k+1}A_k \|\nabla f(x_k)\|^2 \\ &\leq -\frac{1}{2}s_k^2B_k\|\nabla f(x_k)\|^2, \end{aligned}$$

where the second inequality is due to (16b). □

Lemma 4. Inequalities (16) hold for $k \geq 0$.

Proof. Recall that A_k and B_k are defined as in (15). From (7) and (9), we have

$$s_{k+2} \leq \frac{\alpha_{k+1}}{\alpha_{k+2}} s_{k+1} \leq \frac{\alpha_{k+1}}{\alpha_{k+2}} \frac{\theta_{k+3}^2}{\theta_{k+4}(\theta_{k+4} - 1)} s_{k+1} = \frac{A_k + \alpha_{k+1}\theta_{k+3}}{A_{k+1}} s_{k+1}, \quad \forall k \geq 0,$$

which proves (16a). Next, from (4), we have

$$\rho_k = \frac{A_k}{B_{k+1} + \alpha_{k+1}^2\theta_{k+3}^2} = \frac{\alpha_{k+1}\theta_{k+3}(\theta_{k+3} - 1)}{2\alpha_{k+1}^2\theta_{k+3}^2} = \frac{\frac{1}{2}\left(1 - \frac{1}{\theta_{k+3}}\right)}{\alpha_{k+1}} = 1. \quad (22)$$

Thus (16b) is equivalent to $s_{k+1} \leq s_k$ for $k \geq 0$, which holds true by Lemma 3. This proves (16b). Lastly, from (9) we have

$$\frac{A_k}{B_k} = \frac{\alpha_{k+1}}{\alpha_k^2} \frac{\theta_{k+3}(\theta_{k+3} - 1)}{\theta_{k+2}^2} \leq \frac{\alpha_{k+1}}{\alpha_k^2}, \quad \forall k \geq 1,$$

which, together with (22), yields

$$\left(\frac{A_k}{B_k} + \rho_k \right)^{-1} = \left(\frac{A_k}{B_k} + 1 \right)^{-1} \geq \left(\frac{\alpha_{k+1}}{\alpha_k^2} + 1 \right)^{-1} = \frac{\alpha_k^2}{\alpha_{k+1} + \alpha_k^2}, \quad \forall k \geq 1.$$

For $k = 0$, from (4) and (15) we know $\alpha_0 = \frac{1}{\alpha_0} \left(\frac{1}{\alpha_3} + \frac{1}{\alpha_2^2} - \frac{1}{\alpha_1} \right)^{-1}$, which, together with (9), yields

$$\frac{\alpha_2^2 \alpha_3}{\alpha_3 + \alpha_2^2} \frac{1}{\alpha_1} = \left(\frac{\alpha_1}{\tilde{\alpha}_0 \alpha_0} + 1 \right)^{-1} \leq \left(\frac{\alpha_1}{\tilde{\alpha}_0 \alpha_0} \frac{\theta_3(\theta_3 - 1)}{\theta_2^2} + 1 \right)^{-1} = \left(\frac{A_0}{B_0} + \rho_0 \right)^{-1}.$$

Combining the above two inequalities with (7) proves (16c). \square

Lemma 5. Equality (18) holds for $k \geq 0$.

Proof. First, from (5) we know

$$z_{k+2} - z_{k+1} = -s_{k+1} \alpha_{k+1} \theta_{k+3} \nabla f(x_{k+1}),$$

and thus we have

$$\begin{aligned} \frac{1}{2} \|z_{k+2} - x_\star\|^2 - \frac{1}{2} \|z_{k+1} - x_\star\|^2 &= \frac{1}{2} \langle z_{k+2} - z_{k+1}, z_{k+1} + z_{k+2} - 2x_\star \rangle \\ &= \frac{1}{2} \langle z_{k+2} - z_{k+1}, 2(z_{k+1} - x_\star) + (z_{k+2} - z_{k+1}) \rangle \\ &= -s_{k+1} \alpha_{k+1} \theta_{k+3} \langle \nabla f(x_{k+1}), z_{k+1} - x_\star \rangle + \frac{1}{2} s_{k+1}^2 \alpha_{k+1}^2 \theta_{k+3}^2 \|\nabla f(x_{k+1})\|^2. \end{aligned} \quad (23)$$

Combining (23) with the following identity,

$$\begin{aligned} s_{k+2} A_{k+1} (f(x_{k+1}) - f_\star) - s_{k+1} A_k (f(x_k) - f_\star) &= (s_{k+1} (A_k + \alpha_{k+1} \theta_{k+3}) - s_{k+2} A_{k+1}) (f_\star - f(x_{k+1})) \\ &\quad + s_{k+1} \alpha_{k+1} \theta_{k+3} (f(x_{k+1}) - f_\star - \langle \nabla f(x_{k+1}), x_{k+1} - x_\star \rangle) \\ &\quad + s_{k+1} A_k (f(x_{k+1}) - f_\star) - s_{k+1} A_k (f(x_k) - f_\star) \\ &\quad + s_{k+1} \alpha_{k+1} \theta_{k+3} \langle \nabla f(x_{k+1}), x_{k+1} - x_\star \rangle, \end{aligned}$$

we have

$$\begin{aligned} s_{k+2} A_{k+1} (f(x_{k+1}) - f_\star) + \frac{1}{2} \|z_{k+2} - x_\star\|^2 &- \left(s_{k+1} A_k (f(x_k) - f_\star) + \frac{1}{2} \|z_{k+1} - x_\star\|^2 \right) \\ &= (s_{k+1} (A_k + \alpha_{k+1} \theta_{k+3}) - s_{k+2} A_{k+1}) (f_\star - f(x_{k+1})) \\ &\quad + s_{k+1} \alpha_{k+1} \theta_{k+3} (f(x_{k+1}) - f_\star - \langle \nabla f(x_{k+1}), x_{k+1} - x_\star \rangle) \\ &\quad + s_{k+1} A_k (f(x_{k+1}) - f(x_k)) + s_{k+1} \alpha_{k+1} \theta_{k+3} \langle \nabla f(x_{k+1}), x_{k+1} - z_{k+1} \rangle + \frac{1}{2} s_{k+1}^2 \alpha_{k+1}^2 \theta_{k+3}^2 \|\nabla f(x_{k+1})\|^2. \end{aligned} \quad (24)$$

The first two lines of (24) are the desired terms in (18). We focus on reformulating the terms in the last line. From (5), we have

$$x_{k+1} - z_{k+1} = (\theta_{k+3} - 1) (y_{k+1} - x_{k+1}) = -(\theta_{k+3} - 1) (x_{k+1} - x_k + s_k \nabla f(x_k)).$$

Plugging this into the last line of (24), we have

$$\begin{aligned}
& s_{k+1}A_k (f(x_{k+1}) - f(x_k)) + s_{k+1}\alpha_{k+1}\theta_{k+3} \langle \nabla f(x_{k+1}), x_{k+1} - z_{k+1} \rangle + \frac{1}{2}s_{k+1}^2\alpha_{k+1}^2\theta_{k+3}^2 \|\nabla f(x_{k+1})\|^2 \\
&= s_{k+1}A_k \left(f(x_{k+1}) - f(x_k) + \frac{1}{2L_{k+1}} \|\nabla f(x_k) - \nabla f(x_{k+1})\|^2 \right) \\
&\quad - s_{k+1} \underbrace{\alpha_{k+1}\theta_{k+3}(\theta_{k+3} - 1)}_{=A_k} \langle \nabla f(x_{k+1}), x_{k+1} - x_k + s_k \nabla f(x_k) \rangle \\
&\quad - \frac{s_{k+1}A_k}{2L_{k+1}} \|\nabla f(x_k) - \nabla f(x_{k+1})\|^2 + \frac{1}{2}s_{k+1}^2\alpha_{k+1}^2\theta_{k+3}^2 \|\nabla f(x_{k+1})\|^2 \tag{25}
\end{aligned}$$

$$\begin{aligned}
&= s_{k+1}A_k \left(f(x_{k+1}) - f(x_k) + \langle \nabla f(x_{k+1}), x_k - x_{k+1} \rangle + \frac{1}{2L_{k+1}} \|\nabla f(x_k) - \nabla f(x_{k+1})\|^2 \right) \\
&\quad + s_{k+1}A_k \left(\frac{1}{L_{k+1}} - s_k \right) \langle \nabla f(x_{k+1}), \nabla f(x_k) \rangle \\
&\quad - \frac{s_{k+1}A_k}{2L_{k+1}} \left(1 - \frac{\alpha_{k+1}^2\theta_{k+3}^2}{A_k} s_{k+1}L_{k+1} \right) \|\nabla f(x_{k+1})\|^2 - \frac{s_{k+1}A_k}{2L_{k+1}} \|\nabla f(x_k)\|^2. \tag{26}
\end{aligned}$$

From the definition of V_k we know

$$\begin{aligned}
& s_{k+2}A_{k+1} (f(x_{k+1}) - f_\star) + \frac{1}{2} \|z_{k+2} - x_\star\|^2 - \left(s_{k+1}A_k (f(x_k) - f_\star) + \frac{1}{2} \|z_{k+1} - x_\star\|^2 \right) \\
&= V_{k+1} - \frac{1}{2}s_{k+1}^2B_{k+1} \|\nabla f(x_{k+1})\|^2 - \left(V_k - \frac{1}{2}s_k^2B_k \|\nabla f(x_k)\|^2 \right), \quad \forall k \geq 0,
\end{aligned}$$

which, together with (24) and (26), yields the desired equation (18). \square

Lemma 6. *Inequality (20) holds for $k \geq 0$.*

Proof. Note that (20) is equivalent to

$$\begin{aligned}
\frac{2L_{k+1}}{s_{k+1}A_k}Q_k + \frac{B_k}{A_k} \frac{s_k^2}{s_{k+1}}L_{k+1} \|\nabla f(x_k)\|^2 &\leq 0 & \text{if } s_kL_{k+1} \leq 1 \\
\frac{2L_{k+1}}{s_{k+1}A_k}Q_k + \|\nabla f(x_k)\|^2 &\leq 0 & \text{if } s_kL_{k+1} \geq 1.
\end{aligned}$$

Moreover, by recalling the definition of Q_k in (19), we know:

$$\begin{aligned}
\frac{2L_{k+1}}{s_{k+1}A_k}Q_k &= -(1 - \rho_k s_{k+1}L_{k+1}) \|\nabla f(x_{k+1})\|^2 \\
&\quad + 2(1 - s_kL_{k+1}) \langle \nabla f(x_{k+1}), \nabla f(x_k) \rangle - \left(1 + \frac{B_k}{A_k} \frac{s_k^2}{s_{k+1}}L_{k+1} \right) \|\nabla f(x_k)\|^2.
\end{aligned}$$

First, we focus on the coefficient of the term $\|\nabla f(x_{k+1})\|^2$ in the above expression. From (16c):

$$1 - \rho_k s_{k+1}L_{k+1} \geq 1 - \rho_k \left(\frac{A_k}{B_k} + \rho_k \right)^{-1} \geq 0, \quad \forall k \geq 0.$$

Therefore, our goal reduces to showing that the discriminant of the quadratic form is nonpositive:

$$\begin{aligned}
(1 - s_kL_{k+1})^2 - (1 - \rho_k s_{k+1}L_{k+1}) &\leq 0 & \text{if } s_kL_{k+1} \leq 1, \\
(1 - s_kL_{k+1})^2 - (1 - \rho_k s_{k+1}L_{k+1}) \frac{B_k}{A_k} \frac{s_k^2}{s_{k+1}}L_{k+1} &\leq 0 & \text{if } s_kL_{k+1} \geq 1.
\end{aligned}$$

The proof for each case is as follows.

- $s_k L_{k+1} \leq 1$. From (16b) we obtain

$$\rho_k s_{k+1} L_{k+1} \leq s_k L_{k+1}, \quad \forall k \geq 0.$$

For arbitrary $\delta_1, \delta_2 \in [0, 1]$, if $\delta_2 \leq \delta_1$ we know

$$(1 - \delta_1)^2 - (1 - \delta_2) \leq (1 - \delta_1) - (1 - \delta_2) = \delta_2 - \delta_1 \leq 0. \quad (27)$$

We obtain the desired result by substituting $\delta_1 = s_k L_{k+1}$ and $\delta_2 = \rho_k s_{k+1} L_{k+1}$.

- $s_k L_{k+1} \geq 1$. From (16c) we know

$$\frac{1}{s_{k+1} L_{k+1}} \geq \frac{A_k}{B_k} + \rho_k, \quad \forall k \geq 0,$$

which, together with the fact that $1 - 2s_k L_{k+1} \leq 0$, yields

$$(1 - \rho_k s_{k+1} L_{k+1}) \frac{B_k}{A_k} \frac{s_k^2}{s_{k+1}} L_{k+1} = \frac{B_k}{A_k} \left(\frac{1}{s_{k+1} L_{k+1}} - \rho_k \right) s_k^2 L_{k+1}^2 \geq s_k^2 L_{k+1}^2 \geq (1 - s_k L_{k+1})^2.$$

This completes the proof. \square

Since V_0 involves x_1 and z_1 , which are not the initial points, we need to define V_{-1} and prove $V_0 \leq V_{-1}$. This is done in Lemma 7.

Lemma 7. Define

$$V_{-1} = \frac{1}{2} \|x_0 - x_\star\|^2 + \frac{1}{2} s_0^2 (\tilde{\alpha}_0 + \alpha_0) \alpha_0 \theta_2^2 \|\nabla f(x_0)\|^2.$$

Let f be a locally smooth convex function, and L be a smoothness parameter of f on $\bar{B}_{3\|x_0 - x_\star\|}(x_\star)$. Then, the following inequality holds

$$V_0 \leq V_{-1} - \frac{s_0 \alpha_0 \theta_2}{2L} \|\nabla f(x_0)\|^2.$$

Proof. Using (23) with $k = -1$, we have (note $x_0 = z_0$):

$$\frac{1}{2} \|z_1 - x_\star\|^2 - \frac{1}{2} \|x_0 - x_\star\|^2 = -s_0 \alpha_0 \theta_2 \langle \nabla f(x_0), x_0 - x_\star \rangle + \frac{1}{2} s_0^2 \alpha_0^2 \theta_2^2 \|\nabla f(x_0)\|^2,$$

which implies

$$\begin{aligned} V_0 - V_{-1} &= s_1 \alpha_1 \theta_3 (\theta_3 - 1) (f(x_0) - f_\star) + \frac{1}{2} \|z_1 - x_\star\|^2 + \frac{1}{2} s_0^2 \tilde{\alpha}_0 \alpha_0 \theta_2^2 \|\nabla f(x_0)\|^2 \\ &\quad - \left(\frac{1}{2} \|x_0 - x_\star\|^2 + \frac{1}{2} s_0^2 \tilde{\alpha}_0 \alpha_0 \theta_2^2 \|\nabla f(x_0)\|^2 + \frac{1}{2} s_0^2 \alpha_0^2 \theta_2^2 \|\nabla f(x_0)\|^2 \right) \\ &= (s_0 \alpha_0 \theta_2 - s_1 \alpha_1 \theta_3 (\theta_3 - 1)) (f_\star - f(x_0)) \\ &\quad + s_0 \alpha_0 \theta_2 (f(x_0) - f_\star - \langle \nabla f(x_0), x_0 - x_\star \rangle) \\ &\leq s_0 \alpha_0 \theta_2 (f(x_0) - f_\star - \langle \nabla f(x_0), x_0 - x_\star \rangle) \\ &\leq -\frac{s_0 \alpha_0 \theta_2}{2L} \|\nabla f(x_0)\|^2, \end{aligned}$$

where the first inequality is due to (7), and the second inequality is due to Lemma 1. \square

Next, we establish a lower bound for $s_k \alpha_k$.

Lemma 8. *Let f be a locally smooth convex function. Suppose $\{s_k\}_{k \geq 0}$ is generated by AdaNAG. Denote $r_0 = \frac{\theta_3(\theta_3-1)}{\theta_2} \frac{1}{\alpha_0} \frac{\alpha_2^2 \alpha_3}{\alpha_3 + \alpha_2^2}$ and set $s_0 = r_0 \frac{1}{L_0}$ for some arbitrary $L_0 > 0$. Define*

$$S_k = \min \left\{ \frac{1}{L_0}, \frac{1}{L_1}, \dots, \frac{1}{L_k} \right\}. \quad (28)$$

Then

$$s_k \geq \frac{\alpha_2^2 \alpha_3}{\alpha_3 + \alpha_2^2} \frac{1}{\alpha_k} S_k, \quad \forall k \geq 1.$$

Proof. Proof by induction.

- $k = 1$. Applying (7) and the assumption $s_0 = r_0 \frac{1}{L_0}$, we obtain:

$$s_1 = \min \left\{ \frac{\alpha_0}{\alpha_1} \frac{\theta_2}{\theta_3(\theta_3-1)} s_0, \frac{\alpha_2^2 \alpha_3}{\alpha_3 + \alpha_2^2} \frac{1}{\alpha_1} \frac{1}{L_1} \right\} = \frac{\alpha_2^2 \alpha_3}{\alpha_3 + \alpha_2^2} \frac{1}{\alpha_1} \min \left\{ \frac{1}{L_0}, \frac{1}{L_1} \right\} = \frac{\alpha_2^2 \alpha_3}{\alpha_3 + \alpha_2^2} \frac{1}{\alpha_1} S_1.$$

- $k \geq 2$. We first show $\frac{\alpha_{k+1}}{\frac{\alpha_{k+1}^2}{\alpha_k^2} + 1}$ is increasing for $k \geq 2$. By Lemma 3, the numerator α_{k+1} is increasing, it is suffices to show the denominator $\frac{\alpha_{k+1}^2}{\alpha_k^2} + 1$ is decreasing. Since $\frac{1}{\alpha_k}$ is decreasing, it is enough to show $\frac{\alpha_{k+1}}{\alpha_k}$ is decreasing. Observe that for $k \geq 2$, we have

$$\frac{\alpha_{k+1}}{\alpha_k} = \frac{1 - \frac{1}{\theta_{k+3}}}{1 - \frac{1}{\theta_{k+2}}} = 1 + \frac{\frac{1}{\theta_{k+2}} - \frac{1}{\theta_{k+3}}}{1 - \frac{1}{\theta_{k+2}}} = 1 + \frac{1 - \frac{\theta_{k+2}}{\theta_{k+3}}}{\theta_{k+2} - 1}.$$

Since $\theta_{k+2} - 1$ is increasing, it is suffices to show $1 - \frac{\theta_{k+2}}{\theta_{k+3}}$ is decreasing. Observing

$$\frac{\theta_{k+2}}{\theta_{k+3}} = \frac{\theta_{k+2}}{\frac{1}{2} + \sqrt{\frac{1}{4} + \theta_{k+2}^2}} = \left(\frac{1}{2\theta_{k+2}} + \sqrt{\frac{1}{4\theta_{k+2}^2} + 1} \right)^{-1},$$

we see $\frac{\theta_{k+2}}{\theta_{k+3}}$ is increasing since θ_k is increasing by Lemma 3. Therefore, $1 - \frac{\theta_{k+2}}{\theta_{k+3}}$ is decreasing, we conclude $\frac{\alpha_{k+1}}{\frac{\alpha_{k+1}^2}{\alpha_k^2} + 1}$ is increasing. As a result, we obtain

$$\frac{\alpha_2^2 \alpha_3}{\alpha_3 + \alpha_2^2} = \frac{\alpha_3}{\frac{\alpha_3^2}{\alpha_2^2} + 1} \leq \frac{\alpha_{k+1}}{\frac{\alpha_{k+1}^2}{\alpha_k^2} + 1} = \frac{\alpha_k^2}{\alpha_{k+1} + \alpha_k^2} \alpha_{k+1}, \quad \forall k \geq 2. \quad (29)$$

Applying the induction hypothesis $s_k \geq \frac{\alpha_2^2 \alpha_3}{\alpha_3 + \alpha_2^2} \frac{1}{\alpha_k} S_k$ and (29), we conclude:

$$\begin{aligned} s_{k+1} &= \min \left\{ \frac{\alpha_k}{\alpha_{k+1}} s_k, \frac{\alpha_k^2}{\alpha_{k+1} + \alpha_k^2} \frac{1}{L_{k+1}} \right\} \\ &\geq \min \left\{ \frac{\alpha_2^2 \alpha_3}{\alpha_3 + \alpha_2^2} \frac{1}{\alpha_{k+1}} S_k, \frac{\alpha_k^2}{\alpha_{k+1} + \alpha_k^2} \frac{1}{L_{k+1}} \right\} = \frac{\alpha_2^2 \alpha_3}{\alpha_3 + \alpha_2^2} \frac{1}{\alpha_{k+1}} S_{k+1}. \end{aligned}$$

This completes the proof. \square

Corollary 9. Suppose f is an L -smooth convex function. Set s_0 as in Lemma 8 with L_0 as in (3). Then

$$s_k \geq \frac{\alpha_2^2 \alpha_3}{\alpha_3 + \alpha_2^2} \frac{1}{\alpha_k} \frac{1}{L}, \quad \forall k \geq 1.$$

Proof. From Lemma 1, we have $\frac{1}{L_k} \geq \frac{1}{L}$ for $k \geq 1$ and also $\frac{1}{L_0} \geq \frac{1}{L}$ by its definition (3). Recalling the definition (28), we obtain $S_k \geq \frac{1}{L}$ for $k \geq 0$. We obtain the desired conclusion by Lemma 8. \square

Finally, we are ready to prove the statements of our main theorem. We state and prove generalized results that directly imply it.

Proposition 10. Suppose f is a locally smooth convex function. Define L as in Lemma 7 and set s_0 as in Lemma 8. Denote $R = 2(V_{-1} - \frac{s_0 \alpha_0 \theta_2}{2L} \|\nabla f(x_0)\|^2)$, which is equivalent to

$$R = \|x_0 - x_\star\|^2 + r_0^2 \alpha_0 (\tilde{\alpha}_0 + \alpha_0) \theta_2^2 \frac{1}{L_0} \left(\frac{1}{L_0} - \frac{1}{r_0 (\tilde{\alpha}_0 + \alpha_0) \theta_2} \frac{1}{L} \right) \|\nabla f(x_0)\|^2,$$

when $s_0 = r_0 \frac{1}{L_0}$. Suppose $\{x_k\}_{k \geq 0}$ is generated by AdaNAG. Then, the following inequality is true:

$$f(x_k) - f_\star \leq \frac{\alpha_3 + \alpha_2^2}{2\alpha_2^2 \alpha_3} \frac{1/S_{k+1}}{\theta_{k+3} (\theta_{k+3} - 1)} R.$$

Proof. From Theorem 2, Lemma 7, and $s_0 = r_0 \frac{1}{L_0}$, we know that

$$s_{k+1} \alpha_{k+1} \theta_{k+3} (\theta_{k+3} - 1) (f(x_k) - f_\star) \leq V_k \leq \dots \leq V_0 \leq V_{-1} - \frac{s_0 \alpha_0 \theta_2}{2L} \|\nabla f(x_0)\|^2 = \frac{1}{2} R.$$

Thus we obtain

$$f(x_k) - f_\star \leq \frac{1}{2s_{k+1} \alpha_{k+1} \theta_{k+3} (\theta_{k+3} - 1)} R,$$

which, combining with Lemma 8, yields the desired result. \square

Proposition 11. Let f be an L -smooth convex function. Set s_0 as in Corollary 9 and suppose $\{x_k\}_{k \geq 0}$ is generated by AdaNAG. Denote R as in Proposition 10. Then, the following inequality is true:

$$\sum_{i=1}^k \theta_{i+3} (\theta_{i+3} - 1) \|\nabla f(x_i)\|^2 \leq L^2 \left(\frac{\alpha_3 + \alpha_2^2}{\alpha_2^2 \alpha_3} \right)^2 R. \quad (30)$$

As a result, we have

$$\min_{i \in \{1, \dots, k\}} \|\nabla f(x_i)\|^2 \leq \frac{L^2}{\sum_{i=1}^k \theta_{i+3} (\theta_{i+3} - 1)} \left(\frac{\alpha_3 + \alpha_2^2}{\alpha_2^2 \alpha_3} \right)^2 R = \mathcal{O} \left(\frac{L^2}{k^3} \right).$$

Proof. Plugging in (15) and applying (9) to the result of Theorem 2, we have

$$V_{k+1} - V_k \leq \begin{cases} -\frac{1}{2} s_k^2 \alpha_k^2 \theta_{k+3} (\theta_{k+3} - 1) \|\nabla f(x_k)\|^2 & \text{if } s_k L_{k+1} \leq 1 \\ -\frac{1}{2} \frac{s_{k+1} \alpha_{k+1}}{L_{k+1}} \theta_{k+3} (\theta_{k+3} - 1) \|\nabla f(x_k)\|^2 & \text{if } s_k L_{k+1} \geq 1 \end{cases}$$

for $k \geq 0$. Here, we have used $\tilde{\alpha}_0 \approx 0.2706 < 0.4707 \approx \alpha_0$. Now, since f is an L -smooth convex function, applying Corollary 9, we have:

$$s_k^2 \alpha_k^2 \geq \left(\frac{\alpha_2^2 \alpha_3}{\alpha_3 + \alpha_2^2} \right)^2 \frac{1}{L^2}, \quad \frac{s_{k+1} \alpha_{k+1}}{L_{k+1}} \geq \frac{\alpha_2^2 \alpha_3}{\alpha_3 + \alpha_2^2} \frac{1}{L^2}.$$

From Table 5, we know that $\frac{\alpha_2^2 \alpha_3}{\alpha_3 + \alpha_2^2} < 1$. Collecting these observations, we have

$$V_{k+1} - V_k \leq -\frac{1}{2} \left(\frac{\alpha_2^2 \alpha_3}{\alpha_3 + \alpha_2^2} \right)^2 \frac{1}{L^2} \theta_{k+3} (\theta_{k+3} - 1) \|\nabla f(x_k)\|^2.$$

Summing up from 0 to k , and applying Lemma 7, we obtain (30). Note that since $\theta_k = \Omega(k)$, we have $\sum_{i=1}^k \theta_{i+3} (\theta_{i+3} - 1) = \Omega(k^3)$. \square

Proof of Theorem 1. The proofs immediately follow from Proposition 10 and Proposition 11. Note, we have $\theta_{k+3}(\theta_{k+3} - 1) = \theta_{k+2}^2$ by (4) and $\theta_{k+2}^2 \geq \frac{1}{4}(k+4)^2$ by Lemma 3. From Lemma 1 and $L_0 \leq L$, we know $1/S_{k+1} \leq L$. Finally, by plugging in the numerical values from Table 5 with appropriate rounding, we obtain the convergence results in Theorem 1. \square

4.3 Discussion: Core points of the proof

Designing a new algorithm with theoretical guarantees means finding a new proof. Organizing the core elements of a proof often enables us to consider a more general strategy for finding a new proof. As mentioned in [27, § 2.1], bounding the term $\|\nabla f(x_{k+1})\|^2$ that appears in $V_{k+1} - V_k$ is an important point and arguably a tricky part of the proof for adaptive algorithms. In [27], the authors were able to propose AdGD2 with a refined step size by treating this point in a novel way in Lemma 1. We consider the following to be the core elements of our proof related to this issue. Recalling (21), we note that the term Q_k defined in (19) is the core quantity we need to bound.

(i) *Definition of L_{k+1} .*

Observing (25), we see that a negative term $-\frac{s_{k+1} A_k}{2L_{k+1}} \|\nabla f(x_k) - \nabla f(x_{k+1})\|^2$ appears, which later subsumed into Q_k . Such a term can appear because $f(x_{k+1}) - f(x_k) + \langle \nabla f(x_{k+1}), x_k - x_{k+1} \rangle$ cancels out the term $\frac{1}{2L_{k+1}} \|\nabla f(x_k) - \nabla f(x_{k+1})\|^2$. This is a consequence of the definition of L_{k+1} in (6), and also the motivation for the definition as discussed in Section 4.1. Note that the well-definedness of L_{k+1} is based on Corollary 2, which crucially exploits both the locally smooth and convex assumptions of f .

(ii) *The positive part of the coefficient of $\|\nabla f(x_{k+1})\|^2$ contains s_{k+1}^2 .*

Taking a closer look at the coefficient of the $\|\nabla f(x_{k+1})\|^2$ term in (19), and recalling the definition of ρ_k in (17), we find that its positive part is $\frac{1}{2} (B_{k+1} + \alpha_{k+1}^2 \theta_{k+3}^2) s_{k+1}^2$. The important point is that it contains s_{k+1}^2 , which we can control to be small since it is defined after s_k and L_{k+1} are determined. Tracing its origin, B_{k+1} originates directly from the definition of V_{k+1} in (14), and $\alpha_{k+1}^2 \theta_{k+3}^2$ originates from the difference in the $\|z_{k+1} - x_\star\|^2$ term as observed in (23). A crucial point is that the index of z is $k+1$ rather than k . Note that this shifted index of z in V_k was inherited from the Lyapunov function of OGM (12).

In the next section, we introduce a novel adaptive gradient descent method obtained by applying the above observations.

5 A novel gradient-descent-type adaptive method

In this section, we propose a gradient-descent-type (GD-type) adaptive algorithm whose update rule is of the following form:

$$x_{k+1} = x_k - s_k \nabla f(x_k),$$

for some step size schedule s_k . That is, there is no momentum term.

5.1 The AdaGD Algorithm

Our GD-type adaptive method [AdaGD](#) is described in [Algorithm 2](#).

Algorithm 2 Adaptive Gradient Descent (AdaGD)

- 1: **Input:** $x_0 \in \mathbb{R}^d$, $s_0 > 0$, $\{A_k\}_{k \geq 0}$, $\{B_k\}_{k \geq 0}$, $A_{-1} = 0$
- 2: **for** $k = 0, 1, \dots$ **do**
- 3:

$$\begin{aligned}
 x_{k+1} &= x_k - s_k \nabla f(x_k) \\
 L_{k+1} &= -\frac{\frac{1}{2} \|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2}{f(x_{k+1}) - f(x_k) + \langle \nabla f(x_{k+1}), x_k - x_{k+1} \rangle} \\
 s_{k+1} &= \min \left\{ \frac{A_{k-1} + 1}{A_k} s_k, \left(\frac{A_k}{B_k} + \frac{B_{k+1} + 1}{A_k} \right)^{-1} \frac{1}{L_{k+1}} \right\}
 \end{aligned} \tag{31}$$

- 4: **end for**
-

Our main result in this section is given in [Theorem 3](#). We provide its proof in [Section 5.2](#).

Theorem 3. *Let f be a locally smooth convex function. Suppose $\{x_k\}_{k \geq 0}$ is generated by [AdaGD](#) ([Algorithm 2](#)) with $s_0 = r A_0 \frac{1}{L_0}$ where L_0 defined as in [\(3\)](#), and positive $\{A_k\}_{k \geq 0}$, $\{B_k\}_{k \geq 0}$ satisfying*

$$\frac{A_k + 1}{A_{k+1}} \geq 1, \quad \left(\frac{A_k}{B_k} + \frac{B_{k+1} + 1}{A_k} \right)^{-1} \geq r, \quad \forall k \geq 0, \tag{32}$$

with some $r \in (0, 1]$ and

$$B_{k+2} \leq \frac{A_{k+1}^2}{A_k + 1} - 1, \quad \forall k \geq 0. \tag{33}$$

Let L be a smoothness parameter of f on $\bar{B}_{3\bar{R}}(x_*) \cup \bar{B}_{3\|\tilde{x}_0 - x_*\|}(x_*)$, where \tilde{x}_0 is the vector used in [\(3\)](#) and $\bar{R} = \|x_0 - x_*\|^2 + (B_0 + 1)s_0^2 \|\nabla f(x_0)\|^2$. Define $R = \bar{R} - \frac{s_0}{L} \|\nabla f(x_0)\|^2$.

- Then the following holds, which indicates a non-ergodic convergence rate when $\lim_{k \rightarrow \infty} A_k = \infty$.

$$f(x_k) - f_* \leq \frac{L}{2r A_k} R = \mathcal{O} \left(\frac{L}{A_k} \right). \tag{34}$$

- Also, x_k achieves minimum selection convergence rate

$$\min_{i \in \{1, \dots, k\}} (f(x_i) - f_*) \leq \frac{1}{2 \sum_{i=1}^k s_i} R = \mathcal{O} \left(\frac{L}{k} \right) \tag{35}$$

and

$$\min_{i \in \{1, \dots, k\}} \|\nabla f(x_i)\|^2 \leq \frac{L^2}{r \sum_{i=1}^k (1 + r \min \{A_i, B_i\})} R = \mathcal{O} \left(\frac{L^2}{k + \sum_{i=1}^k A_i} \right).$$

Note that from the first condition of [\(32\)](#), we see that:

$$\frac{A_k + 1}{A_{k+1}} \geq 1 \iff A_{k+1} - A_k \leq 1 \implies A_k \leq A_0 + k.$$

Therefore, the fastest rate we can obtain from [\(34\)](#) is $\mathcal{O}(1/k)$, which matches the order of non-adaptive gradient descent method. In this context, the choice of $A_k = \Theta(k)$ is worth emphasizing.

Corollary 4. Suppose $\{x_k\}_{k \geq 0}$ is generated by [AdaGD](#) with $A_k = \gamma(k+1) + 2$ and $B_k = \gamma(k+1)$, where $\gamma \in (0, 1]$. Then

$$f(x_k) - f_\star \leq \frac{L}{2r} \frac{1}{\gamma(k+1) + 2} R = \mathcal{O}\left(\frac{L}{k}\right), \quad (36)$$

where $r = \left(\frac{A_0}{B_0} + \frac{B_1+1}{A_0}\right)^{-1} = \frac{\gamma(\gamma+2)}{3\gamma^2+5\gamma+4} \leq 1$, and L, R are defined as in [Theorem 3](#). Also,

$$\min_{i \in \{1, \dots, k\}} \|\nabla f(x_i)\|^2 \leq \frac{2L^2}{r} \frac{1}{k(r\gamma(k+3) + 2)} R = \mathcal{O}\left(\frac{L^2}{k^2}\right).$$

Proof. It is sufficient to check the two conditions (32) and (33). Since $A_{k+1} - A_k = \gamma \leq 1$, the first inequality of (32) holds. Observe that

$$\frac{A_k}{B_k} + \frac{B_{k+1} + 1}{A_k} = \frac{B_k + 2}{B_k} + \frac{B_k + \gamma + 1}{B_k + 2} = 2 + \frac{(\gamma + 1)B_k + 4}{B_k(B_k + 2)} = 2 + \frac{\gamma + 1 + \frac{4}{B_k}}{B_k + 2}$$

is decreasing for $k \geq 0$, since B_k is increasing. Hence, the second inequality of (32) holds with $r = \left(\frac{A_0}{B_0} + \frac{B_1+1}{A_0}\right)^{-1}$. Next, (33) follows from

$$B_{k+2} - \left(\frac{A_{k+1}^2}{A_k + 1} - 1\right) = A_k + 2(\gamma - 1) - \frac{(A_k + \gamma)^2}{A_k + 1} + 1 = -\frac{(1 - \gamma)^2}{A_k + 1} \leq 0.$$

Finally,

$$\sum_{i=1}^k (1 + r \min\{A_i, B_i\}) = \sum_{i=1}^k (1 + r\gamma(i+1)) = \frac{1}{2}k(r\gamma(k+3) + 2).$$

So the desired conclusion follows by applying [Theorem 3](#). \square

Non-ergodic convergence. Note that both (34) and (36) are non-ergodic convergence results. To the best of our knowledge, these are the first non-ergodic results for GD-type (i.e., no momentum) adaptive algorithms. In contrast, only ergodic convergence results are available for existing GD-type adaptive algorithms [26, 23, 27, 40].

5.2 Convergence analysis of AdaGD (Proof of [Theorem 3](#))

We use the Lyapunov function motivated by the discussion in [Section 4.3](#). The overall proof structure resembles that of [AdaNAG](#). We first prove $V_{k+1} \leq V_k$ for $k \geq -1$.

Proposition 12. Let f be a locally smooth convex function. Suppose $\{x_k\}_{k \geq 0}$ is generated by [AdaGD](#) with $\{A_k\}_{k \geq 0}$ and $\{B_k\}_{k \geq 0}$ satisfying the assumptions in [Theorem 3](#). Define $B_{-1} = B_0 + 1$, $x_{-1} = x_0$, $s_{-1} = s_0$, and the Lyapunov function

$$V_k = s_{k+1}A_k(f(x_k) - f_\star) + \frac{1}{2}s_k^2B_k\|\nabla f(x_k)\|^2 + \frac{1}{2}\|x_{k+1} - x_\star\|^2, \quad \forall k \geq -1.$$

Denote $\tilde{s}_k = \min\{s_k, s_{k+1}, \frac{1}{L_{k+1}}\}$ for $k \geq -1$. Then the following inequality holds:

$$\begin{aligned} V_{k+1} - V_k &\leq (s_{k+1}A_k + s_{k+1} - s_{k+2}A_{k+1})(f_\star - f(x_{k+1})) \\ &\quad - \frac{s_{k+1}}{2L}\|\nabla f(x_{k+1})\|^2 - \frac{\tilde{s}_k^2}{2}\min\{A_k, B_k\}\|\nabla f(x_k)\|^2, \quad \forall k \geq -1. \end{aligned} \quad (37)$$

Also, $x_k \in \bar{B}_{\bar{R}}(x_\star)$ holds. Here, both L and \bar{R} are defined as in [Theorem 3](#).

Proof. Suppose we have shown

$$\begin{aligned} V_{k+1} - V_k &\leq (s_{k+1}A_k + s_{k+1} - s_{k+2}A_{k+1})(f_\star - f(x_{k+1})) \\ &\quad + s_{k+1}(f(x_{k+1}) - f_\star - \langle \nabla f(x_{k+1}), x_{k+1} - x_\star \rangle) - \frac{\tilde{s}_k^2}{2} \min\{A_k, B_k\} \|\nabla f(x_k)\|^2 \end{aligned} \quad (38)$$

for $k \geq -1$. With (38), we have

$$\frac{1}{2} \|x_k - x_\star\|^2 \leq V_{k-1} \leq \dots \leq V_{-1} = \frac{1}{2} \bar{R}, \quad \forall k \geq 0.$$

This implies $x_k \in \bar{B}_{\bar{R}}(x_\star)$. Then we have $f(x_{k+1}) - f_\star - \langle \nabla f(x_{k+1}), x_{k+1} - x_\star \rangle \leq -\frac{1}{2L} \|\nabla f(x_{k+1})\|^2$ from Lemma 1, therefore, (38) implies (37). The proof of (38) resembles the proof of Theorem 2, and we place it in Appendix C.1. \square

Leveraging the boundedness of the iterate $x_k \in \bar{B}_{\bar{R}}(x_\star)$, we can achieve a lower bound for s_k . We provide it in Lemma 13, and the proof in Appendix C.2.

Lemma 13. *Let f be a locally smooth convex function. Suppose $\{s_k\}_{k \geq 0}$ is generated by AdaGD with $\{A_k\}_{k \geq 0}$ and $\{B_k\}_{k \geq 0}$ satisfying the assumptions in Theorem 3. For s_0 and L defined in Theorem 3, we have:*

$$s_k \geq \frac{r}{L}, \quad \forall k \geq 1. \quad (39)$$

We are now ready to prove Theorem 3.

Proof of Theorem 3. For notation simplicity, denote $m_k = \frac{1}{2} \tilde{s}_k^2 \min\{A_k, B_k\}$. First, summing up (37) from 0 to $k-1$ we have

$$V_k + \sum_{i=0}^{k-1} \left((s_{i+1}A_i - s_{i+2}A_{i+1} + s_{i+1})(f(x_{i+1}) - f_\star) + \frac{s_{i+1}}{2L} \|\nabla f(x_{i+1})\|^2 + m_i \|\nabla f(x_i)\|^2 \right) \leq V_0.$$

Plugging in $k = -1$ to (37), we have

$$\begin{aligned} V_0 - V_{-1} &\leq (s_0A_{-1} + s_0 - s_1A_0)(f_\star - f(x_0)) \\ &\quad - \frac{s_0}{2L} \|\nabla f(x_0)\|^2 - \frac{\tilde{s}_{-1}^2}{2} \min\{A_{-1}, B_{-1}\} \|\nabla f(x_0)\|^2 \leq -\frac{s_0}{2L} \|\nabla f(x_0)\|^2. \end{aligned}$$

The second inequality follows from the fact $s_1 \leq \frac{1}{A_0} s_0$ which follows from (31), and $\min\{A_{-1}, B_{-1}\} = 0$ since $A_{-1} = 0$. Therefore, we obtain $V_0 \leq V_{-1} - \frac{s_0}{2L} \|\nabla f(x_0)\|^2 = \frac{1}{2}R$. Combining with $s_{k+1}A_k(f(x_k) - f_\star) \leq V_k$, we have

$$\begin{aligned} s_{k+1}A_k(f(x_k) - f_\star) &+ \sum_{i=1}^k (s_iA_{i-1} - s_{i+1}A_i + s_i)(f(x_i) - f_\star) \\ &+ \sum_{i=1}^k \frac{s_i}{2L} \|\nabla f(x_i)\|^2 + \sum_{i=0}^{k-1} m_i \|\nabla f(x_i)\|^2 \leq \frac{1}{2}R. \end{aligned} \quad (40)$$

Since the summations in (40) are nonnegative, from (40) we have $s_{k+1}A_k(f(x_k) - f_\star) \leq \frac{1}{2}R$, dividing both sides by $s_{k+1}A_k$ and applying Lemma 13, we obtain

$$f(x_k) - f_\star \leq \frac{L}{2rA_k} R = \mathcal{O}\left(\frac{L}{A_k}\right).$$

Next, considering only the function value terms on the left-hand side of (40), we obtain

$$s_{k+1}A_k(f(x_k) - f_\star) + \sum_{i=1}^k (s_i A_{i-1} - s_{i+1} A_i + s_i)(f(x_i) - f_\star) \leq \frac{1}{2}R. \quad (41)$$

Observing the sum of the coefficients of the $f(x_i) - f_\star$ terms, we see that

$$s_{k+1}A_k + \sum_{i=1}^k (s_i A_{i-1} - s_{i+1} A_i + s_i) = s_1 A_0 + \sum_{i=1}^k s_i \geq \sum_{i=1}^k s_i \geq \frac{r}{L}k,$$

where the second inequality is from Lemma 13. From (31), we know $s_{i+1}A_i - s_{i+2}A_{i+1} + s_{i+1} \geq 0$. Gathering these observations, from (41) we conclude

$$\min_{i \in \{1, \dots, k\}} (f(x_i) - f_\star) \leq \frac{L}{2rk} R = \mathcal{O}\left(\frac{L}{k}\right).$$

Lastly, from Lemma 13 we know $s_k, \tilde{s}_k \geq r/L$. Focusing on the squared gradient norm terms in (40), and recalling the definition $m_k = \frac{1}{2}\tilde{s}_k^2 \min\{A_k, B_k\}$, we obtain

$$\frac{r}{2L^2} \sum_{i=1}^k (1 + r \min\{A_i, B_i\}) \|\nabla f(x_i)\|^2 \leq \sum_{i=1}^{k+1} \frac{s_i}{2L} \|\nabla f(x_i)\|^2 + \sum_{i=0}^k m_i \|\nabla f(x_i)\|^2 \leq \frac{1}{2}R. \quad (42)$$

Therefore, we conclude

$$\min_{i \in \{1, \dots, k\}} \|\nabla f(x_i)\|^2 \leq \frac{L^2}{r \sum_{i=1}^k (1 + r \min\{A_i, B_i\})} R.$$

Lastly, from the second inequality of (32), it follows that $\limsup_{k \rightarrow \infty} \frac{A_k}{B_k} < \infty$. Therefore, we obtain $A_k = \mathcal{O}(\min\{A_k, B_k\})$, and thus $\frac{1}{\sum_{i=1}^k (1 + r \min\{A_i, B_i\})} \leq \frac{1}{k + r \sum_{i=1}^k \min\{A_i, B_i\}} = \mathcal{O}\left(\frac{1}{k + \sum_{i=1}^k A_i}\right)$. \square

Previous Lyapunov analysis also implies the pointwise convergence.

Theorem 5. *Let f be a locally smooth convex function. Suppose $\{s_k\}_{k \geq 0}$ is generated by AdaGD with $\{A_k\}_{k \geq 0}$ and $\{B_k\}_{k \geq 0}$ satisfying the assumptions in Theorem 3. Then, for some optimal point \bar{x}_\star ,*

$$\lim_{k \rightarrow \infty} x_k = \bar{x}_\star.$$

Proof. Recall that $x_k \in \bar{B}_{\bar{R}}(x_\star)$ holds by Proposition 12, we know $\{x_k\}_{k \geq 0}$ is a bounded sequence. From (42), we have $\sum_{k=1}^\infty \|\nabla f(x_k)\|^2 \leq \frac{L^2}{r} R < \infty$, which implies that $\lim_{k \rightarrow \infty} \|\nabla f(x_k)\|^2 = 0$. Therefore, all limit points converge to an optimal point. Now, we leverage [26, Lemma 2] by setting $a_k = 2s_k A_{k-1}(f(x_{k-1}) - f_\star) + s_{k-1}^2 B_{k-1} \|\nabla f(x_{k-1})\|^2$, and conclude the desired result. For convenience, we restate the lemma below. \square

Lemma. ([26, Lemma 2]) *Let $\{x_k\}_{k \geq 0}$ and $\{a_k\}_{k \geq 0}$ be two sequences in \mathbb{R}^d and \mathbb{R}_+ respectively. Suppose that $\{x_k\}_{k \geq 0}$ is bounded, its cluster points belong to $\mathcal{X} \subset \mathbb{R}^d$ and it also holds that*

$$\|x^{k+1} - \bar{x}\|^2 + a_{k+1} \leq \|x^k - \bar{x}\|^2 + a_k, \quad \forall \bar{x} \in \mathcal{X}.$$

Then $\{x_k\}_{k \geq 0}$ converges to some element in \mathcal{X} .

5.3 Discussion: Trade-off between the step size and the theoretical guarantee

Theorem 3 covers a family of algorithms. We provide some representative examples of AdaGD and compare their behaviors. For notational simplicity, we rewrite (31) as

$$s_{k+1} = \min \left\{ r_k^s s_k, r_k^L \frac{1}{L_{k+1}} \right\}, \quad \text{where} \quad r_k^s = \frac{A_{k-1} + 1}{A_k}, \quad r_k^L = \left(\frac{A_k}{B_k} + \frac{B_{k+1} + 1}{A_k} \right)^{-1}.$$

Note that when $s_k \ll \frac{1}{L_{k+1}}$, we have $s_{k+1} = r_k^s s_k$. Thus, r_k^s represents the rate at which the adaptive step size catches up to the local smoothness parameter $\frac{1}{L_{k+1}}$. On the other hand, when $\frac{1}{L_{k+1}} \ll s_k$, we obtain $s_{k+1} = r_k^L \frac{1}{L_{k+1}}$. In this case, we can take a larger step size when r_k^L is large. The following table gives examples of AdaGD in which A_k is $\Theta(k)$, $\Theta(\sqrt{k})$ and $\Theta(1)$, respectively.

	A_k	B_k	r_k^s	r_k^L	$f(x_k) - f_\star$
AdaGD ¹	$\frac{1}{2}(k+5)$	$\frac{1}{2}(k+1)$	$\frac{k+6}{k+5}$	$\frac{1}{2} + \Theta(1/k)$	$\mathcal{O}(1/k)$
AdaGD ^{1/2}	$2\sqrt{k+4}$	$2\sqrt{k+2} - 2$	$\frac{2\sqrt{k+3+1}}{2\sqrt{k+4}}$	$\frac{1}{2} + \Theta(1/\sqrt{k})$	$\mathcal{O}(1/\sqrt{k})$
AdaGD ⁰	3	$\frac{5}{4}$	$\frac{4}{3}$	$\frac{20}{63}$	$o(1)$

(i) AdaGD¹: $A_k, B_k = \Theta(k)$.

For this case, we have $\lim_{k \rightarrow \infty} r_k^s = 1$. Thus, the growth rate of the step size slows down as the iteration proceeds. This does not mean there is an upper bound on the growth induced by r_k^s , since $\prod_{k=N}^{\infty} \frac{k+1}{k} = \infty$, but its growth becomes very slow compared to other cases. On the other hand, $\lim_{k \rightarrow \infty} r_k^L = \frac{1}{2}$, which is larger than the value in AdaGD⁰, where $r_k^L = \frac{20}{63}$.

(ii) AdaGD^{1/2}: $A_k, B_k = \Theta(k^q)$ with $0 < q < 1$.

It still holds that $\lim_{k \rightarrow \infty} r_k^s = 1$ and $\lim_{k \rightarrow \infty} r_k^L = \frac{1}{2}$. However, comparing r_k^s for AdaGD^{1/2} and AdaGD¹, we see that $1 + \Theta(1/\sqrt{k}) = \frac{2\sqrt{k+3+1}}{2\sqrt{k+4}} > \frac{k+6}{k+5} = 1 + \Theta(1/k)$. In this context, AdaGD^{1/2} has a larger r_k^s compared to AdaGD¹, while the guaranteed convergence rate is slower.

(iii) AdaGD⁰: $A_k, B_k = \Theta(1)$.

For this case, $\lim_{k \rightarrow \infty} r_k^s > 1$ since it remains constant. Thus, its growth is larger compared to other cases. However, we do not have a non-ergodic convergence rate for this case. Note that, though AdaGD⁰ still has an ergodic $\mathcal{O}(1/k)$ rate by (35), like other GD-type adaptive algorithms in the prior works [26, 23, 27, 40].

In many practical scenarios, $\frac{1}{L_{k+1}}$ becomes large near the optimum, so a larger r_k^s often allows a larger step size. However, from the above context, a larger step size is not free. It requires a trade-off between the theoretical guarantee on the convergence rate.* Note that the choice $\gamma = 1$ in Corollary 4 may guarantee the fastest convergence rate, but it is practically not useful since $r_k^s = \frac{(k+2)+1}{k+3} = 1$, implying that the step size s_k is nonincreasing.

Recall from (7), for AdaNAG we see $r_k^s = \frac{\alpha_k}{\alpha_{k+1}} \leq 1$ by Lemma 3. Above discussion motivates us to consider different choice of θ_k of AdaNAG that makes $r_k^s > 1$. Therefore, we consider such generalization of AdaNAG in the next section.

*To clarify, at least for our algorithm family AdaGD and under our proof argument.

6 Generalized AdaNAG

The parameter selection of [AdaNAG](#) was motivated by the goal of tightening the rate of the theoretical guarantee. However, the discussion in Section 5.3 suggests that we should consider a broader range of parameter choices to enlarge the step size. Previous studies [35, 7, 2] considered generalized versions of NAG with $\theta_k = \frac{k+p}{p}$ for $p \geq 2$, which achieve a slower rate compared to $p = 2$ but maintain an $\mathcal{O}(1/k^2)$ convergence rate. Note that [AdaNAG](#) can be thought of as the $p = 2$ case, since the simplified variant that uses (10) exhibits similar behavior, as shown in Appendix B.2. As we can benefit from a larger step size when considering adaptive methods, we do not need to restrict ourselves to the choices of θ_k that achieve the tightest convergence guarantee.

Note, the core properties of the parameters that make the proof of Theorem 2 work were (16) and the definition $A_k = \alpha_{k+1}\theta_{k+3}(\theta_{k+3} - 1)$. Motivated by this observation, we now introduce the generalized version of AdaNAG, [AdaNAG-G](#), in Algorithm 3. The convergence results are summarized in Theorem 6. The proof resembles the proof of Theorem 1, and is provided in Appendix D.1.

Algorithm 3 Generalized AdaNAG (AdaNAG-G)

1: **Input:** $x^0 = z^0 \in \mathbb{R}^d$, $\{\tau_k\}_{k \geq 0}$, $\{\alpha_k\}_{k \geq 0}$, $s_0 > 0$, $B_0 > 0$, $A_{-1} = 0$.

2: **Define**

$$A_k = \alpha_{k+1}\tau_{k+1}(\tau_{k+1} - 1), \quad B_k = \begin{cases} B_0 & \text{if } k = 0 \\ \alpha_k^2 \tau_k^2 \left(\frac{(\tau_k - 1)^2}{\alpha_{k-1} \tau_{k-1}^2} - 1 \right) & \text{if } k \geq 1 \end{cases}$$

3: **for** $k = 0, 1, \dots$ **do**

4:

$$\begin{aligned} y_{k+1} &= x_k - s_k \nabla f(x_k) \\ z_{k+1} &= z_k - s_k \alpha_k \tau_k \nabla f(x_k) \\ x_{k+1} &= \left(1 - \frac{1}{\tau_{k+1}}\right) y_{k+1} + \frac{1}{\tau_{k+1}} z_{k+1} \\ L_{k+1} &= -\frac{\frac{1}{2} \|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2}{f(x_{k+1}) - f(x_k) + \langle \nabla f(x_{k+1}), x_k - x_{k+1} \rangle} \\ s_{k+1} &= \min \left\{ \frac{A_{k-1} + \alpha_k \tau_k}{A_k} s_k, \left(\frac{A_k}{B_k} + \frac{B_{k+1} + \alpha_{k+1}^2 \tau_{k+1}^2}{A_k} \right)^{-1} \frac{1}{L_{k+1}} \right\} \end{aligned} \quad (43)$$

5: **end for**

Theorem 6. Suppose $\{\tau_k\}_{k \geq 0}$, $\{\alpha_k\}_{k \geq 0}$ satisfy

$$0 < \alpha_k \leq 1, \quad \frac{\tau_k^2}{\tau_{k+1}(\tau_{k+1} - 1)} \geq 1, \quad \left(\frac{A_k}{B_k} + \frac{B_{k+1} + \alpha_{k+1}^2 \tau_{k+1}^2}{A_k} \right)^{-1} \geq \frac{r}{\alpha_{k+1}}, \quad \forall k \geq 0 \quad (44)$$

with some $r > 0$. Set $s_0 = \frac{A_0}{\alpha_0 \tau_0} \frac{r}{\alpha_1} \frac{1}{L_0}$ with L_0 defined as in (3). Suppose f is an L -smooth convex function. Denote $R = \|x_0 - x_\star\|^2 + \left(s_0^2 (B_0 + \alpha_0^2 \theta_0^2) - \frac{s_0 \alpha_0 \theta_2}{2L} \right) \|\nabla f(x_0)\|^2$. Then the sequence $\{x_k\}_{k \geq 0}$ generated by [AdaNAG-G](#) satisfies $s_{k+1} \alpha_{k+1} \geq \frac{r}{L}$ for $k \geq 0$ and

$$f(x_k) - f_\star \leq \frac{1}{2s_{k+1} \alpha_{k+1}} R = \mathcal{O} \left(\frac{L}{\alpha_k \tau_k^2} \right), \quad \min_{i \in \{1, \dots, k\}} \|\nabla f(x_i)\|^2 \leq \mathcal{O} \left(\frac{L^2}{\sum_{i=1}^k \alpha_i \tau_i^2} \right).$$

The important point of [AdaNAG-G](#) compared to [AdaNAG](#) is, it covers the choice of τ_k that makes s_k possible to increase. The worst possible scenario for [AdaNAG](#) is that the initial step size s_0 is chosen too small, and thus the first option of (7) is activated at every iteration. Then the information about the local smoothness parameter L_{k+1} is never used. However, this cannot occur for [AdaNAG-G](#) with a proper choice of τ_k . Observing the first option of (43), we see

$$\prod_{k=1}^N \frac{A_{k-1} + \alpha_k \tau_k}{A_k} = \prod_{k=1}^N \frac{\alpha_k}{\alpha_{k+1}} \frac{\tau_k^2}{\tau_{k+1}(\tau_{k+1} - 1)} \geq \alpha_1 \prod_{k=1}^N \frac{\tau_k^2}{\tau_{k+1}(\tau_{k+1} - 1)},$$

where the inequality comes from the fact $\alpha_N \leq 1$. When $\tau_k = \frac{k+p}{p}$ and $p > 2$, we can verify

$$\prod_{k=1}^{\infty} \frac{\tau_k^2}{\tau_{k+1}(\tau_{k+1} - 1)} = \prod_{k=1}^{\infty} \frac{(k+p+1)^2}{(k+2)(k+p+2)} \geq \prod_{k=1}^{\infty} \left(1 + \frac{p-2}{k+2}\right) \geq 1 + (p-2) \sum_{k=3}^{\infty} \frac{1}{k} = \infty. \quad (45)$$

Therefore, if only the first option of (43) is always activated, the step size would diverge to ∞ . Thus the second option would be activated before that happens. On the other hand, the lower bound r appearing in Theorem 6, does not capture this growth and misrepresents the asymptotic behavior of the step size s_k . This motivates us to consider the following lemma. The proof can be done by formalizing the discussion above, and we provide it in Appendix D.2.

Lemma 14. *Suppose the limit $\bar{r} = \lim_{k \rightarrow \infty} \left(\frac{A_k}{B_k} + \frac{B_{k+1} + \alpha_{k+1}^2 \tau_{k+1}^2}{A_k} \right)^{-1}$ exists. Assume that there exists $N > 0$ such that $\frac{A_{k-1} + \alpha_k \tau_k}{A_k} \geq 1$ holds for all $k \geq N$. Also, assume that $\prod_{k=1}^{\infty} \frac{\tau_k^2}{\tau_{k+1}(\tau_{k+1} - 1)} = \infty$. Suppose $\{s_k\}_{k \geq 0}$ is generated by [AdaNAG-G](#) under the same assumptions as in Theorem 6. Then,*

$$\liminf_{k \rightarrow \infty} s_{k+1} \geq \frac{\bar{r}}{L}.$$

Intuitively speaking, Lemma 14 states that if the growth rate $\frac{\tau_k^2}{\tau_{k+1}(\tau_{k+1} - 1)}$ is large enough, the step size eventually catches up to a certain ratio of the smoothness parameter. However, similar to [AdaGD](#), the growth rate has a trade-off between the convergence guarantee. We have empirically identified two parameter choices that balance both practical performance and theoretical guarantees, which constitute the final main results of this paper. The proofs can be completed by verifying that the assumptions of Theorem 6 and Lemma 14 hold, and we defer them to Appendix D.3 and Appendix D.4. In the proof, we consider the generalized parameter choice $\tau_k = \frac{(k+2)+p}{p}$ with $p > 2$ for Corollary 7, following the spirit of non-adaptive NAG variants [35, 7, 2].

Corollary 7. *Let $\tau_k = \frac{(k+2)+12}{12}$, $\alpha_k = \frac{1}{2} \frac{(\tau_{k+1}-1)^2}{\tau_k^2}$, $B_0 = \alpha_0^2 \tau_0^2 \left(\frac{(\tau_0-1)^2}{\alpha_{-1} \tau_{-1}^2} - 1 \right)$, set s_0, L as Theorem 6. Suppose f is an L -smooth convex function. Then for $\{x_k\}_{k \geq 0}$ generated by [AdaNAG-G](#), we have:*

$$f(x_k) - f_{\star} \leq \frac{1}{s_{k+1}} \frac{144(k+15)}{(k+3)(k+4)^2} R = \mathcal{O}\left(\frac{L}{k^2}\right), \quad \min_{i \in \{1, \dots, k\}} \|\nabla f(x_i)\|^2 = \mathcal{O}\left(\frac{L^2}{k^3}\right).$$

Here, $s_k \geq \frac{1}{250} \frac{1}{L}$ and $\liminf_{k \rightarrow \infty} s_k \geq \frac{1}{3} \frac{1}{L}$. We name the algorithm as **AdaNAG-G₁₂**.

Corollary 8. *Let $\tau_k = 2\sqrt{k+3}$, $\alpha_k = \frac{1}{2}$, $B_0 = \alpha_0^2 \tau_0^2 \left(\frac{(\tau_0-1)^2}{\alpha_{-1} \tau_{-1}^2} - 1 \right)$ and set s_0, L as in Theorem 6. Suppose f is an L -smooth convex function. Then for $\{x_k\}_{k \geq 0}$ generated by [AdaNAG-G](#), we have:*

$$f(x_k) - f_{\star} \leq \frac{1}{s_{k+1}} \frac{1}{4 \left(k + 4 - \frac{\sqrt{k+4}}{2} \right)} R = \mathcal{O}\left(\frac{L}{k}\right), \quad \min_{i \in \{1, \dots, k\}} \|\nabla f(x_i)\|^2 = \mathcal{O}\left(\frac{L^2}{k^2}\right).$$

Here, $s_k \geq \frac{1}{5} \frac{1}{L}$ and $\liminf_{k \rightarrow \infty} s_k \geq \frac{1}{3} \frac{1}{L}$. We name the algorithm as **AdaNAG-G^{1/2}**.

[AdaNAG-G₁₂](#) has a larger constant in the guaranteed convergence upper bound compared to [AdaNAG](#), but it still achieves an accelerated $\mathcal{O}(1/k^2)$ rate. Although [AdaNAG-G^{1/2}](#) does not achieve the accelerated convergence rate, it still attains a non-ergodic $\mathcal{O}(1/k)$ rate. The reason for introducing both methods is that there were specific experimental cases where each method performed well. From our empirical observations, these algorithms performed well compared to prior methods in various practical scenarios, which we present in the next section.

7 Numerical Experiments

In this section, we apply our methods to two problems: *logistic regression* and *least squares problem*. The objective function f is L -smooth and convex for both problems. We primarily focus on two algorithms, [AdaNAG-G₁₂](#) and [AdaNAG-G^{1/2}](#), which were considered in Corollary 7 and Corollary 8. Among the parameter choices that satisfy Theorem 6 and that we have tried, these two algorithms performed the best overall.

We compare our algorithms with recently developed adaptive algorithms, mostly focusing on the accelerated adaptive algorithm AC-FGM [25]. Specifically, we compare with AdaBB [40], AdGD [27], AdaPG^{q,r} [31], and the original Nesterov accelerated gradient method [28], which is a non-adaptive method. For AC-FGM, we use the parameters from [25, Corollary 2], with the α chosen in their experiments. We note that AC-FGM guarantees an $\mathcal{O}(1/k^2)$ rate for $\alpha > 0$, but only an $\mathcal{O}(1/k)$ ergodic rate when $\alpha = 0$. Overall, our methods performed the best in the scenarios we considered.

Algorithm	AdaNAG	AdaNAG-G		AdaGD		
		AdaNAG-G ₁₂	AdaNAG-G ^{1/2}	AdaGD ¹	AdaGD ^{1/2}	AdaGD ⁰
$f(x_k) - f_*$	$\mathcal{O}(1/k^2)$	$\mathcal{O}(1/k^2)$	$\mathcal{O}(1/k)$	$\mathcal{O}(1/k)$	$\mathcal{O}(1/\sqrt{k})$	$\mathcal{O}(1)$
Key Properties	Theoretically tighter	Practically useful		Without momentum		

Table 2: Summary of our algorithms. Note that the convergence rates of $f(x_k) - f_*$ are non-ergodic, and both [AdaGD^{1/2}](#) and [AdaGD⁰](#) still achieve an $\mathcal{O}(1/k)$ ergodic convergence rate.

Our codes were written in Python 3.13.1 and leveraged the implementations provided by [26] and [25]. All numerical experiments were conducted on a personal computer with an Intel Core i5-8265U processor, Intel UHD Graphics 620, and 16GB of memory. Details of the dataset and the parameters used for each algorithm in the experiments are provided in the respective subsections. For convenience, we summarize our algorithms in Table 2.

7.1 Logistic regression

In this subsection, we present several experimental cases of the logistic regression problem

$$\min_{x \in \mathbb{R}^n} f(x) = -\frac{1}{m} \sum_{i=1}^m (y_i \log(\sigma(a_i^\top x)) + (1 - y_i) \log(1 - \sigma(a_i^\top x))) + \frac{\gamma}{2} \|x\|^2,$$

where our algorithm outperforms others. Here, m denotes the number of observations, $a_i \in \mathbb{R}^n$, $y_i \in \{0, 1\}$, $\sigma(z) = \frac{1}{1+e^{-z}}$ and $\gamma > 0$ is a regularization parameter. The gradient of f is given by $\nabla f(x) = \frac{1}{m} \sum_{i=1}^m a_i (\sigma(a_i^\top x) - y_i) + \gamma x$. As discussed in [26], the function f is L -smooth with $L = \frac{1}{4} \lambda_{\max}(A^\top A) + \gamma$, where $A = [a_1^\top \ \dots \ a_m^\top]^\top \in \mathbb{R}^{m \times n}$, and λ_{\max} denotes the largest eigenvalue.

We follow the experimental setup described in [40, § 6.1]. The value f_* is defined as the lowest objective function value achieved among all tested algorithms. We used the *mushrooms*, *w8a*, and *covtype* datasets from LIBSVM [8]. The parameter details for each dataset are given in Table 3.

dataset	m	n	L	γ	max iteration
mushrooms	8,124	112	2.59	L/m	600
w8a	49,749	300	0.66	L/m	1,000
covtype	581,012	54	5.04×10^6	$L/(10m)$	9,000

Table 3: Parameters used in the logistic regression problems

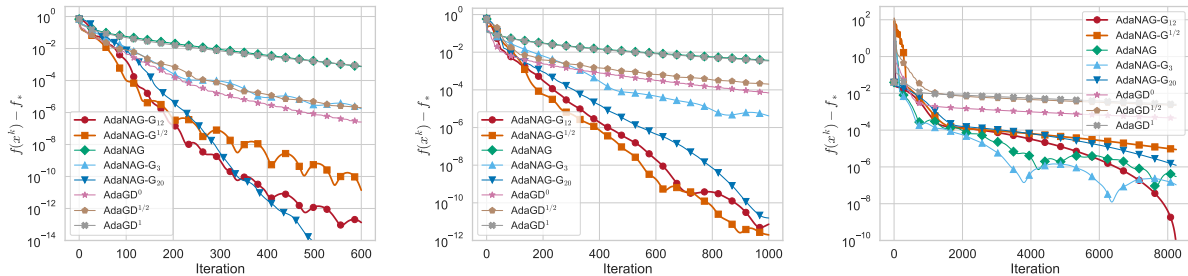
7.1.1 Comparison between our algorithms

Before comparing our algorithms with prior methods, we first analyze the comparison among our own algorithms. This analysis justifies selecting [AdaNAG-G₁₂](#) and [AdaNAG-G^{1/2}](#) as the representative algorithms. The algorithms under comparison include [AdaNAG](#) from Section 3, [AdaGD¹](#), [AdaGD^{1/2}](#), and [AdaGD⁰](#) from Section 5, as well as [AdaNAG-G₁₂](#) and [AdaNAG-G^{1/2}](#) from Section 6. We also include the algorithms defined by $\tau_k = \frac{(k+2)+p}{p}$ with $p = 3, 20$ for [AdaNAG-G](#), to justify the choice $p = 12$ for [AdaNAG-G₁₂](#). We label them as [AdaNAG-G₃](#) and [AdaNAG-G₂₀](#), respectively. As shown in the proof of Corollary 7, provided in Appendix D.3, [AdaNAG-G₃](#) and [AdaNAG-G₂₀](#) also achieve a convergence rate of $\mathcal{O}(1/k^2)$ in (63).

We use L_0 as defined in (3) to define the initial step size for all our algorithms, as stated in the theorems. Specifically, we set

$$\tilde{x}_0 = x_0 + u, \quad (46)$$

where the entries of u are chosen randomly from the uniform distribution over $[0, 1]$. Also, we set x_0 as the zero vector. The corresponding plots are presented in Figure 1.



(a) mushrooms, iteration 6×10^2

(b) w8a, iteration 10^3

(c) covtype, iteration 9×10^3

Figure 1: Logistic regression problem, comparison between our algorithms

First, we focus on the ones with the best performance. Comparing the three accelerated algorithms with $\mathcal{O}(1/k^2)$ convergence rate, [AdaNAG-G₃](#), [AdaNAG-G₁₂](#) and [AdaNAG-G₂₀](#), the best-performing one varies depending on the case. However, [AdaNAG-G₁₂](#) never placed the third. [AdaNAG-G^{1/2}](#), which has an $\mathcal{O}(1/k)$ convergence rate, outperformed other [AdaNAG](#) instances in the *w8a* dataset experiment. [AdaNAG](#) and [AdaGD¹](#), which are of the best theoretical guarantees for accelerated and GD-type algorithms, respectively, performed poorly in practice.

Comparing the GD-type algorithms, [AdaGD⁰](#) always performed the best, while [AdaGD¹](#) always performed the worst. Note that this is the opposite order of the speed of their theoretical guar-

antees. Overall, GD-type algorithms never performed the best compared to other algorithms for logistic regression problems.

Based on these observations, we will only compare [AdaNAG-G₁₂](#) and [AdaNAG-G^{1/2}](#) with the previous algorithms from now on.

7.1.2 Scenarios that our algorithms perform the best

The comparisons of [AdaNAG-G₁₂](#) and [AdaNAG-G^{1/2}](#) with prior algorithms are given in Figures 2 and 3. We observe that our algorithms outperform previous methods for the three tested datasets. In particular, [AdaNAG-G₁₂](#) consistently outperformed all prior algorithms across all three cases. Meanwhile, [AdaNAG-G^{1/2}](#) achieved the best performance for the *w8a* dataset, while in the other cases, it performed similarly to the AC-FGM algorithms.

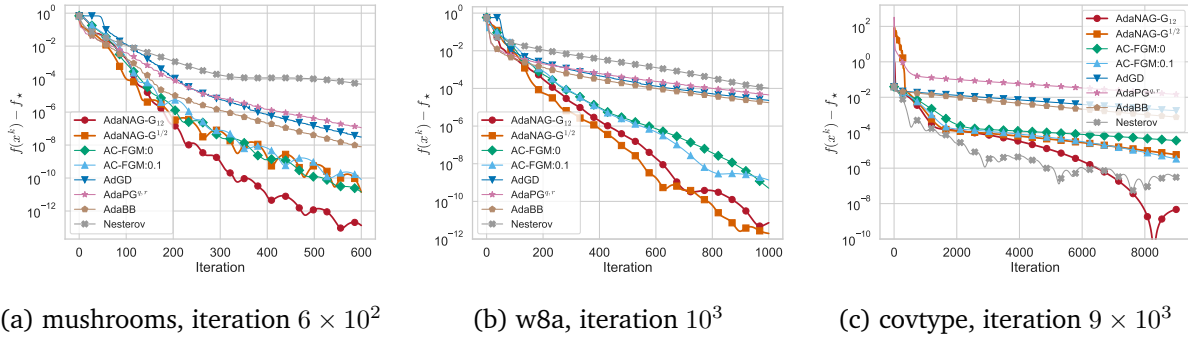


Figure 2: Logistic regression problem: iteration number

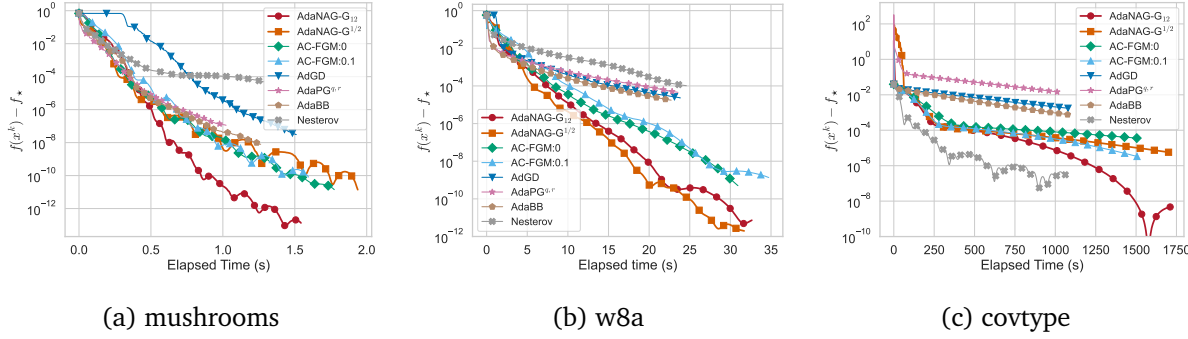


Figure 3: Logistic regression problem: CPU time

The details of the algorithm parameters are as follows. For our algorithms [AdaNAG-G₁₂](#) and [AdaNAG-G^{1/2}](#), we use L_0 as defined in (3), with \tilde{x}_0 as considered in (46). Overall, for the parameter settings of algorithms from prior works, we selected the best-performing variant introduced in each respective paper. For AC-FGM, we set $\beta = 1 - \frac{\sqrt{6}}{3}$ and $\eta_1 = \frac{2}{5L_0}$, as in [25, § 4], and consider $\alpha = 0$ and 0.1 . In the figures, AC-FGM:0 and AC-FGM:0.1 mean AC-FGM with $\alpha = 0$ and $\alpha = 0.1$, respectively. For AdGD, we chose one of Algorithms 1 and 2 from [27], and we performed one line search to ensure $\alpha_0 L_1 \in [1/\sqrt{2}, 2]$ for Algorithm 2. For AdaPG^{q,r}, we set $q = 5/3$ and $r = 5/6$, based on comparisons of the suggested values in Table 1 of [31]. We use $\gamma_0 = \|\tilde{x}_0 - x_0\|^2 / \langle \tilde{x}_0 - x_0, \nabla f(\tilde{x}_0) - \nabla f(x_0) \rangle = \gamma_{-1}$, where \tilde{x}_0 is defined as in (46). Finally, for AdaBB, we adopt AdaBB3 as introduced in Table 2 of [40, § 6], and set $\alpha_0 = 10^{-10}$, $\theta_0 = \max \{ \lambda_1^2 / (2\alpha_0^2) - 1, 0 \}$, and $\theta_1 = 1$.

Overall, the accelerated adaptive algorithms and their variants (AdaNAG, AC-FGM) demonstrated superior performance compared to GD-type adaptive algorithms (AdaBB, AdGD, AdaPG^{q,r}). This performance gap is especially noticeable in the *covtype* dataset, which has a significantly larger m compared to the others, requiring more iterations and a longer running time. Interestingly, the *covtype* dataset is also the only case where the non-adaptive Nesterov algorithm outperformed most of the other algorithms.

7.2 Least squares problem

In this subsection, we present several experimental cases of the classical least squares problem

$$\min_{x \in \mathbb{R}^n} f(x) = \frac{1}{m} \|Ax - b\|^2,$$

where our algorithm outperforms others or achieves the best performance. Here, $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$. We provide two cases with randomly generated synthetic data and two cases with real-world datasets, *bodyfat* and *cadata*, from LIBSVM [8].

To ensure a fair comparison, we follow most of the experimental details in [25, § 4.1], using the same settings as our benchmark algorithm AC-FGM. We consider both synthetic and real-world datasets, comparing two cases with relatively larger and smaller values of m , which corresponds to the number of data points. In the synthetic data setting, A is generated with entries uniformly distributed in $[0, 1]$. An optimal point x_* is randomly chosen from the unit ball $B_1(0)$, and we set $b = Ax_*$, which implies $f_* = 0$. The parameter details for the algorithms are the same as in Section 7.1.2 and parameters for each dataset are summarized in Table 4.

dataset	m	n	L	max iteration
random-small	1,000	4,000	2,000	6×10^3
random-large	4,000	8,000	4,000	10^4
bodyfat	252	14	156,269	2×10^4
cadata	20,640	8	20	10^5

Table 4: Parameters used in the least squares problem

As captured in Figure 4 and Figure 5, our algorithms AdaNAG-G₁₂ and AdaNAG-G^{1/2} outperform others or achieve the best performance. For synthetic datasets in Figure 4, we observe that AdaNAG-G^{1/2} outperforms other algorithms for smaller datasets, while AdaNAG-G₁₂ outperforms others for larger datasets. For the real-world datasets in Figure 5, AdaNAG-G₁₂ achieves the best performance on both datasets, tying with AC-FGM:0.1 on *bodyfat* but performing noticeably better on *cadata*. Note that the performance gap between AdaNAG-G₁₂ and AdaNAG-G^{1/2} is larger on *cadata*, which is the larger dataset.

To summarize, AdaNAG-G₁₂, which achieves the accelerated rate of $\mathcal{O}(1/k^2)$, tends to perform the best overall and performs even better on larger datasets that require more iterations. Another noteworthy observation is that, for the least squares problem, variants of accelerated adaptive algorithms (AdaNAG, AC-FGM) perform significantly better than GD-type adaptive algorithms (AdaGD, AdaPG^{q,r}, AdaBB).

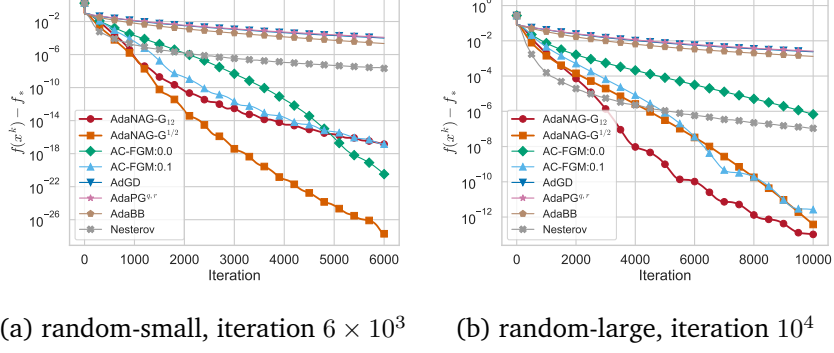


Figure 4: Least squares problem, synthetic datasets

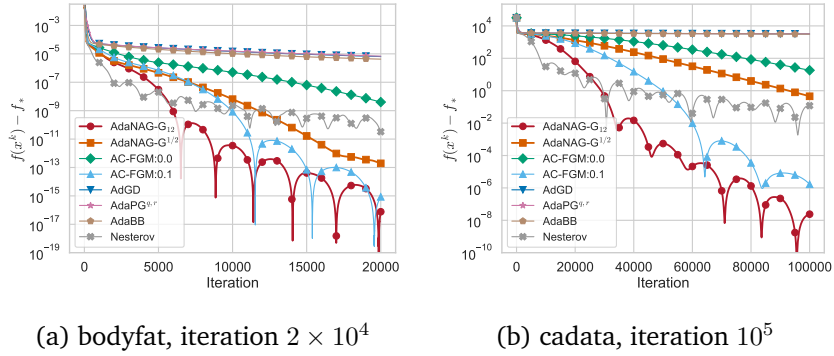


Figure 5: Least squares problem, real-world datasets

8 Conclusion

This paper introduces a novel adaptive accelerated algorithm based on Nesterov’s accelerated gradient method, named **AdaNAG**. The algorithm is line-search-free, parameter-free, and achieves the accelerated convergence rate $f(x_k) - f_* = \mathcal{O}(1/k^2)$. We establish this result via a Lyapunov analysis, and also achieve $\min_{i \in \{1, \dots, k\}} \|\nabla f(x_i)\|^2 = \mathcal{O}(1/k^3)$, as in the original Nesterov’s algorithm, which is the first of its kind. Applying a similar proof idea, we further introduce a family of novel GD-type adaptive algorithms, denoted as **AdaGD**, and provide an instance that achieves a non-ergodic rate $f(x_k) - f_* = \mathcal{O}(1/k)$, analogous to classical gradient descent. We observe that, within a family of adaptive algorithms, choosing a parameter associated with a slower convergence rate can allow for a larger step size. Inspired by the previous observation, we propose another family of adaptive algorithms that generalizes **AdaNAG**, which includes instances that both theoretically sound and practically useful, **AdaNAG-G₁₂** and **AdaNAG-G^{1/2}**.

Our adaptive method **AdaNAG** resembles the original Nesterov’s method in the form of the algorithm and the Lyapunov-based proof structure, and achieves the same order of convergence rates. A natural next step would be to investigate whether other properties and extensions developed for the original Nesterov’s method can also be applied to **AdaNAG**. Studying whether **AdaNAG** or its variants can achieve iterate convergence [7, 2], extend to the prox-grad setup [5, 16, 13], or extend to other advanced setups [1, 19, 9, 6], would be interesting directions for future work.

References

- [1] Kwangjun Ahn and Suvrit Sra. From Nesterov’s estimate sequence to Riemannian acceleration. *International Conference on Machine Learning*, 2020.
- [2] Hedy Attouch, Zaki Chbani, Juan Peypouquet, and Patrick Redont. Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity. *Mathematical Programming*, 168(1):123–175, 2018.
- [3] Nikhil Bansal and Anupam Gupta. Potential-function proofs for gradient methods. *Theory of Computing*, 15(4):1–32, 2019.
- [4] Jonathan Barzilai and Jonathan M. Borwein. Two-point step size gradient methods. *IMA Journal of Numerical Analysis*, 8(1):141–148, 1988.
- [5] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.
- [6] Jincheng Cao, Ruichen Jiang, Erfan Yazdandoost Hamedani, and Aryan Mokhtari. An accelerated gradient method for convex smooth simple bilevel optimization. *Neural Information Processing Systems*, 2024.
- [7] A. Chambolle and Ch. Dossal. On the convergence of the iterates of the “fast iterative shrinkage/thresholding algorithm”. *Journal of Optimization Theory and Applications*, 166(3):968–982, 2015.
- [8] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3), 2011.
- [9] Pengyu Chen, Xu Shi, Rujun Jiang, and Jiulin Wang. Penalty-based methods for simple bilevel optimization under Hölderian error bounds. *Neural Information Processing Systems*, 2024.
- [10] Alexandre d’Aspremont, Damien Scieur, and Adrien Taylor. Acceleration methods. *Foundations and Trends® in Optimization*, 5(1-2):1–245, 2021.
- [11] Yoel Drori. The exact information-based complexity of smooth convex minimization. *Journal of Complexity*, 39:1–16, 2017.
- [12] A. A. Goldstein. Cauchy’s method of minimization. *Numerische Mathematik*, 4(1):146–150, 1962.
- [13] Uijeong Jang, Shuvomoy Das Gupta, and Ernest K. Ryu. Computer-assisted design of accelerated composite optimization methods: OptISTA. *arXiv:2305.15704*, 2024.
- [14] Donghwan Kim and Jeffrey A. Fessler. Optimized first-order methods for smooth convex minimization. *Mathematical Programming*, 159(1-2):81–107, 2016.
- [15] Donghwan Kim and Jeffrey A. Fessler. On the convergence analysis of the optimized gradient method. *Journal of Optimization Theory and Applications*, 172(1):187–205, 2017.
- [16] Donghwan Kim and Jeffrey A. Fessler. Another look at the fast iterative shrinkage/thresholding algorithm (FISTA). *SIAM Journal on Optimization*, 28(1):223–250, 2018.

- [17] Donghwan Kim and Jeffrey A. Fessler. Generalizing the optimized gradient method for smooth convex minimization. *SIAM Journal on Optimization*, 28(2):1920–1950, 2018.
- [18] Donghwan Kim and Jeffrey A. Fessler. Optimizing the efficiency of first-order methods for decreasing the gradient of smooth convex functions. *Journal of Optimization Theory and Applications*, 188(1):192–219, 2021.
- [19] Jungbin Kim and Insoon Yang. Accelerated gradient methods for geodesically convex optimization: Tractable algorithms and convergence analysis. *International Conference on Machine Learning*, 2022.
- [20] Jungbin Kim and Insoon Yang. Unifying Nesterov’s accelerated gradient methods for convex and strongly convex objective functions. *International Conference on Machine Learning*, 2023.
- [21] Larry Armijo. Minimization of functions having Lipschitz continuous first partial derivatives. *Pacific Journal of Mathematics*, 16(1):1–3, 1966.
- [22] Puya Latafat, Andreas Themelis, and Panagiotis Patrinos. On the convergence of adaptive first order methods: Proximal gradient and alternating minimization algorithms. *Learning for Dynamics & Control Conference*, 2024.
- [23] Puya Latafat, Andreas Themelis, Lorenzo Stella, and Panagiotis Patrinos. Adaptive proximal algorithms for convex optimization under local Lipschitz continuity of the gradient. *Mathematical Programming*, 2024.
- [24] Jongmin Lee, Chanwoo Park, and Ernest K. Ryu. A geometric structure of acceleration and its role in making gradients small fast. *Neural Information Processing Systems*, 2021.
- [25] Tianjiao Li and Guanghui Lan. A simple uniformly optimal method without line search for convex optimization. *arXiv:2310.10082*, 2024.
- [26] Yura Malitsky and Konstantin Mishchenko. Adaptive gradient descent without descent. *International Conference on Machine Learning*, 2020.
- [27] Yura Malitsky and Konstantin Mishchenko. Adaptive proximal gradient method for convex optimization. *Neural Information Processing Systems*, 2024.
- [28] Yurii Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Doklady Akademii Nauk SSSR*, 269(3):543–547, 1983.
- [29] Yurri Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer, 2004.
- [30] Yurri Nesterov. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 152(1):381–404, 2015.
- [31] Konstantinos Oikonomidis, Emanuel Laude, Puya Latafat, Andreas Themelis, and Panagiotis Patrinos. Adaptive proximal gradient methods are universal without approximation. *International Conference on Machine Learning*, 2024.
- [32] Chanwoo Park, Jisun Park, and Ernest K. Ryu. Factor- $\sqrt{2}$ acceleration of accelerated gradient methods. *Applied Mathematics & Optimization*, 88(3):77, 2023.

- [33] Bin Shi, Simon S. Du, Michael I. Jordan, and Weijie J. Su. Understanding the acceleration phenomenon via high-resolution differential equations. *Mathematical Programming*, 195(1):79–148, 2022.
- [34] Weijie Su, Stephen Boyd, and Emmanuel J. Candès. A differential equation for modeling Nesterov’s accelerated gradient method: Theory and insights. *Neural Information Processing Systems*, 2014.
- [35] Weijie Su, Stephen Boyd, and Emmanuel J. Candès. A differential equation for modeling Nesterov’s accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17(153):1–43, 2016.
- [36] Jaewook J. Suh, Jisun Park, and Ernest K. Ryu. Continuous-time analysis of anchor acceleration. *Neural Information Processing Systems*, 2023.
- [37] Jaewook J. Suh, Gyumin Roh, and Ernest K. Ryu. Continuous-time analysis of AGM via conservation laws in dilated coordinate systems. *International Conference on Machine Learning*, 2022.
- [38] Adrien Taylor and Francis Bach. Stochastic first-order methods: Non-asymptotic and computer-aided analyses via potential functions. *Conference on Learning Theory*, 2019.
- [39] Andre Wibisono, Ashia C. Wilson, and Michael I. Jordan. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, 113(47):E7351–E7358, 2016.
- [40] Danqing Zhou, Shiqian Ma, and Junfeng Yang. AdaBB: Adaptive Barzilai-Borwein method for convex optimization. *Mathematics of Operations Research*, 2025.

A Omitted proofs in Section 2

A.1 Proof of Lemma 1

We use arguments similar to those for L -smooth functions that are provided in [29, Theorem 2.1.5]. Let $K \subset \mathbb{R}^d$ be a compact set. Let $R > 0$ and $c \in \mathbb{R}^d$ satisfy $K \subset \bar{B}_R(c) = \{x \in \mathbb{R}^d \mid \|x - c\| \leq R\}$. For notation simplicity, denote $\bar{K} = \bar{B}_{3R}(c)$. Since f is locally smooth, there is $\bar{L}_K > 0$ satisfying

$$\|\nabla f(\bar{x}) - \nabla f(\bar{y})\| \leq \bar{L}_K \|\bar{x} - \bar{y}\|, \quad \forall \bar{x}, \bar{y} \in \bar{K}.$$

Now, take $x, y \in K$ such that $x \neq y$. Define $\tilde{f}_y: \mathbb{R}^d \rightarrow \mathbb{R}$ as

$$\tilde{f}_y(z) = f(z) - \langle \nabla f(y), z \rangle. \quad (47)$$

Note

$$\nabla \tilde{f}_y(z) = \nabla f(z) - \nabla f(y). \quad (48)$$

Form (48) we know that \tilde{f}_y has the same smoothness parameter as f , that is

$$\|\nabla \tilde{f}_y(\bar{x}) - \nabla \tilde{f}_y(\bar{y})\| = \|\nabla f(\bar{x}) - \nabla f(\bar{y})\| \leq \bar{L}_K \|\bar{x} - \bar{y}\|, \quad \forall \bar{x}, \bar{y} \in \bar{K}.$$

Also, we know $\tilde{f}_y(y) = \min_{u \in \mathbb{R}^d} \tilde{f}_y(u)$ since $\nabla \tilde{f}_y(y) = 0$ and \tilde{f}_y is a convex function. Now, define

$$x^+ = x - \frac{1}{\bar{L}_K} \nabla \tilde{f}_y(x). \quad (49)$$

Then from (48) and the fact $x, y \in \bar{K}$, we have

$$\|x^+ - x\| = \frac{1}{\bar{L}_K} \|\nabla \tilde{f}_y(x)\| = \frac{1}{\bar{L}_K} \|\nabla f(x) - \nabla f(y)\| \leq \|x - y\|.$$

By triangular inequality and the fact $x, y \in \bar{B}_R(c)$, we have

$$\|x^+ - c\| \leq \|x - c\| + \|x^+ - x\| \leq \|x - c\| + \|x - y\| \leq 3R,$$

and we can conclude $x^+ \in \bar{B}_{3R}(c) = \bar{K}$. Thus for $t \in [0, 1]$ we have

$$\|\nabla \tilde{f}_y(x + t(x^+ - x)) - \nabla \tilde{f}_y(x)\| = \|\nabla f(x + t(x^+ - x)) - \nabla f(x)\| \leq t\bar{L}_K \|x^+ - x\|, \quad (50)$$

where the equality comes from (48) and the inequality comes from the fact $x, x^+ \in \bar{K}$.

Now we can proceed with the standard argument of [29, Lemma 1.2.3]. From the Cauchy-Schwarz inequality, (50), and (49), we obtain:

$$\begin{aligned} \tilde{f}_y(x^+) &= \tilde{f}_y(x) + \langle \nabla \tilde{f}_y(x), x^+ - x \rangle + \int_0^1 \langle \nabla \tilde{f}_y(x + t(x^+ - x)) - \nabla \tilde{f}_y(x), x^+ - x \rangle dt \\ &\leq \tilde{f}_y(x) + \langle \nabla \tilde{f}_y(x), x^+ - x \rangle + \int_0^1 \|\nabla \tilde{f}_y(x + t(x^+ - x)) - \nabla \tilde{f}_y(x)\| \|x^+ - x\| dt \\ &\leq \tilde{f}_y(x) + \langle \nabla \tilde{f}_y(x), x^+ - x \rangle + \int_0^1 t\bar{L}_K \|x^+ - x\|^2 dt \\ &= \tilde{f}_y(x) - \frac{1}{2\bar{L}_K} \|\nabla \tilde{f}_y(x)\|^2. \end{aligned}$$

Now recalling the fact $\tilde{f}_y(y) = \min_{u \in \mathbb{R}^d} \tilde{f}_y(u)$ and applying the above inequality, we have

$$\tilde{f}_y(y) \leq \tilde{f}_y(x^+) \leq \tilde{f}_y(x) - \frac{1}{2\bar{L}_K} \|\nabla \tilde{f}_y(x)\|^2.$$

Plugging the definition of \tilde{f}_y in (47) and (48) we have

$$f(y) - \langle \nabla f(y), y \rangle \leq f(x) - \langle \nabla f(y), x \rangle - \frac{1}{2\bar{L}_K} \|\nabla f(x) - \nabla f(y)\|^2,$$

which implies

$$f(y) - f(x) + \langle \nabla f(y), x - y \rangle + \frac{1}{2\bar{L}_K} \|\nabla f(x) - \nabla f(y)\|^2 \leq 0. \quad (51)$$

Since $x, y \in K$ were arbitrary, this completes the proof. \square

A.2 Proof of Corollary 2

- (i) Suppose $f(y) - f(x) + \langle \nabla f(y), x - y \rangle = 0$. For any compact set K that contains x and y , the desired conclusion follows immediately from Lemma 1.
- (ii) Suppose $f(y) - f(x) + \langle \nabla f(y), x - y \rangle \neq 0$. Then we have $f(y) - f(x) + \langle \nabla f(y), x - y \rangle < 0$ by the convexity of f . Multiplying both sides of (51) by $-\frac{1}{f(y) - f(x) + \langle \nabla f(y), x - y \rangle} \bar{L}_K$ yields the desired inequalities. \square

B Further Discussions on AdaNAG

B.1 Proof of Lemma 3

From (4), we have $\theta_{k+1} = \frac{1}{2} \left(1 + \sqrt{1 + 4\theta_k^2} \right) \geq \frac{1}{2} + \theta_k$. Thus θ_k is increasing, and the inequality $\theta_k \geq \frac{k+2}{2}$ for $k \geq 0$ follows by induction and the fact $\theta_0 = 1$. Next, from $\theta_{k+1} = \frac{1}{2} \left(1 + \sqrt{1 + 4\theta_k^2} \right)$ we have $(2\theta_{k+1} - 1)^2 = 1 + 4\theta_k^2$ and so $\theta_{k+1}^2 - \theta_{k+1} = \theta_k^2$. Therefore θ_k satisfies (9) as an equality.

Moreover, from (4) we have $\alpha_{k+1} = \frac{1}{2} \left(1 - \frac{1}{\theta_{k+3}} \right) \geq \frac{1}{2} \left(1 - \frac{1}{\theta_{k+2}} \right) = \alpha_k$ for $k \geq 1$ since θ_k is increasing, and $\lim_{k \rightarrow \infty} \alpha_k = \lim_{k \rightarrow \infty} \frac{1}{2} \left(1 - \frac{1}{\theta_{k+2}} \right) = \frac{1}{2}$ since $\lim_{k \rightarrow \infty} \theta_k = \infty$. Finally, we have $s_{k+1} \leq \frac{\alpha_k}{\alpha_{k+1}} s_k \leq s_k$ since $\alpha_{k+1} \geq \alpha_k$ for $k \geq 1$. When $k = 0$, leveraging the numerical values in the first column of Table 5, we can obtain $s_1 \leq \frac{\alpha_0}{\alpha_1} \frac{\theta_2}{\theta_3(\theta_3-1)} s_0 < s_0$. \square

B.2 Simple AdaNAG

For NAG, θ_k defined in (4) is of a rather complex form. A simpler alternative to it is $\theta_k = \frac{k+2}{2}$ as defined in (10). It is easy to verify that $\theta_k = \frac{k+2}{2}$ also satisfies (9). Other properties mentioned in Lemma 3 can also be verified using the same argument as in Appendix B.1. In this section, we provide a simplified version of AdaNAG with this θ_k . In this case, α_k defined in (4) becomes:

$$\alpha_k = \frac{1}{2} \left(1 - \frac{2}{k+4} \right), \text{ for } k \geq 1, \quad \text{and} \quad \alpha_0 = \frac{60}{127}.$$

Plugging in the above, we can check that the coefficients in (7) become $\frac{\alpha_k}{\alpha_{k+1}} = \frac{k(k+3)}{(k+1)(k+2)}$ and $\frac{\alpha_k^2}{\alpha_{k+1} + \alpha_k^2} = \frac{k^2(k+3)}{3k^3 + 13k^2 + 16k + 8}$. Calculating other coefficients in AdaNAG with similar manner, we get a simplified version of AdaNAG described in Algorithm 4.

Algorithm 4 Simple AdaNAG

1: **Input:** $x^0 = z^0 \in \mathbb{R}^d$, $s_0 > 0$,

2:

$$\begin{aligned} y_1 &= x_0 - s_0 \nabla f(x_k) \\ z_1 &= z_0 - s_0 \frac{60}{127} \nabla f(x_0) \\ x_1 &= \frac{3}{5} y_1 + \frac{2}{5} z_1 \\ L_1 &= -\frac{\frac{1}{2} \|\nabla f(x_1) - \nabla f(x_0)\|^2}{f(x_1) - f(x_0) + \langle \nabla f(x_1), x_0 - x_1 \rangle} \\ s_1 &= \min \left\{ \frac{320}{381} s_0, \frac{50}{177} \frac{1}{L_1} \right\} \end{aligned}$$

3: **for** $k = 1, 2, \dots$ **do**

4:

$$\begin{aligned} y_{k+1} &= x_k - s_k \nabla f(x_k) \\ z_{k+1} &= z_k - \frac{k+2}{4} s_k \nabla f(x_k) \\ x_{k+1} &= \frac{k+3}{k+5} y_{k+1} + \frac{2}{k+5} z_{k+1} \\ L_{k+1} &= -\frac{\frac{1}{2} \|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2}{f(x_{k+1}) - f(x_k) + \langle \nabla f(x_{k+1}), x_k - x_{k+1} \rangle} \\ s_{k+1} &= \min \left\{ \frac{k(k+3)}{(k+1)(k+2)} s_k, \frac{k^2(k+3)}{3k^3 + 13k^2 + 16k + 8} \frac{1}{L_{k+1}} \right\} \end{aligned}$$

5: **end for**

Using the same proof argument, we can also obtain the convergence result similar to Theorem 1 for Algorithm 4.

Proposition 15. *Let f be an L -smooth convex function. In Algorithm 4, set $s_0 = \frac{635}{1888} \frac{1}{L_0}$ with L_0 defined as in (3). Define $R = \|x_0 - x_\star\|^2 + 0.15 \frac{1}{L_0} \left(\frac{1}{L_0} - \frac{2}{L} \right) \|\nabla f(x_0)\|^2$. Let $\{x_k\}_{k \geq 0}$ be a sequence generated by Algorithm 4. Then the following convergence rate results hold.*

$$\begin{aligned} f(x_k) - f_\star &\leq \frac{24L}{(k+3)(k+5)} R = \mathcal{O} \left(\frac{L}{k^2} \right), \\ \min_{i \in \{1, \dots, k\}} \|\nabla f(x_i)\|^2 &\leq \frac{1671L^2}{k(k^2 + 10k + 33)} R = \mathcal{O} \left(\frac{L^2}{k^3} \right). \end{aligned}$$

Proof. The proof follows the same argument as that of Theorem 1. The only difference is that since $\theta_{k+3}(\theta_{k+3} - 1) = \theta_{k+2}^2$ does not hold, we instead use $\theta_{k+3}(\theta_{k+3} - 1) = \frac{1}{4}(k+3)(k+5)$. \square

B.3 Extension of AdaNAG to locally smooth f

Various previous adaptive methods [26, 23, 27, 40, 25] achieve convergence for locally smooth convex functions f , which are highly dependent on the boundedness of the iterates. In this section, we show that we can achieve boundedness with a locally smooth function f and the same convergence rate, if we consider a small perturbation to the step size selection (7) of AdaNAG.

Theorem 9. *Let f be a locally smooth convex function. Set s_0, R as in Theorem 1. Suppose $\{x_k\}_{k \geq 0}$ is generated by AdaNAG but with (7) replaced by the following rule when $k \geq 3$:*

$$s_{k+1} = \min \left\{ \frac{\alpha_k}{\alpha_{k+1}} s_k, \frac{\alpha_k^2}{\alpha_{k+1} + \alpha_k^2(1 + 10^{-6})} \frac{1}{L_{k+1}} \right\}, \quad \text{if } k \geq 3. \quad (52)$$

Then the sequence $\{x_k\}_{k \geq 0}$ is bounded. Let $\bar{R} > 0$ such that $x_k \in \bar{B}_{\bar{R}}(x_)$. Define L as a smoothness parameter of f on $\bar{B}_{3\bar{R}}(x_*) \cup \bar{B}_{3\|\tilde{x}_0 - x_*\|}(x_*)$, where \tilde{x}_0 is the vector in (3). Then the convergence rates stated in Theorem 1 continue to hold with the (local) smoothness parameter L , i.e.,*

$$f(x_k) - f_* \leq \frac{5.5L}{\theta_{k+2}^2} R = \mathcal{O} \left(\frac{L}{k^2} \right), \quad \min_{i \in \{1, \dots, k\}} \|\nabla f(x_i)\|^2 \leq \frac{120L^2}{\sum_{i=1}^k \theta_{i+2}^2} R = \mathcal{O} \left(\frac{L^2}{k^3} \right). \quad (53)$$

We note that the value 10^{-6} in (52) is negligibly small compared to α_k^2 and α_{k+1} . To check this point, recall that α_k is increasing for $k \geq 1$ and $\lim_{k \rightarrow \infty} \alpha_k = 0.5$, as stated in Lemma 3. Also, from numerical calculation, we find that $\alpha_1 \approx 0.3182$. Therefore, we can view this method as a perturbed version of AdaNAG, with a negligibly small relaxation in the coefficient of $\frac{1}{L_{k+1}}$.

The selection of the value 10^{-6} is simply an arbitrary example of a small number, chosen to help clear understanding. This choice can be generalized, as stated in Proposition 16.

Proposition 16. *Choose $N \geq 3$. Define α_k as in (4), and define*

$$\bar{\epsilon}_N = \left(\frac{1}{\alpha_2^2} + \frac{1}{\alpha_3} - \frac{1}{\alpha_N^2} \right) \alpha_{N+1} - 1. \quad (54)$$

Take an arbitrary $\epsilon \in (0, \bar{\epsilon}_N)$. Suppose $\{x_k\}_{k \geq 0}$ is generated by AdaNAG but with (7) replaced by

$$s_{k+1} = \min \left\{ \frac{\alpha_k}{\alpha_{k+1}} s_k, \frac{\alpha_k^2}{\alpha_{k+1} + \alpha_k^2(1 + \epsilon)} \frac{1}{L_{k+1}} \right\}, \quad \text{if } k \geq N. \quad (55)$$

Then the sequence $\{x_k\}_{k \geq 0}$ is bounded. Define L as in Theorem 9. Then the inequalities (53) hold.

We postpone the proof of Proposition 16 to Appendix B.3.1. Here, we first assume Proposition 16 and prove Theorem 9.

Proof of Theorem 9. Let $N = 3$. Using numerical calculation, we check that $\bar{\epsilon}_3 \approx 0.399 > 10^{-6}$. Therefore, we obtain the desired conclusion by Proposition 16. \square

Note that when $\epsilon = 0$ or $N = \infty$, (55) reduces to the original step size update rule (7) of AdaNAG. Therefore, the variant in Proposition 16 resembles AdaNAG more closely as ϵ becomes smaller and N becomes larger. Nevertheless, regardless of how small ϵ is or how large N becomes, Proposition 16 tells us that we still achieve the same convergence rates for locally smooth convex f , possibly with a larger local smoothness parameter L .

B.3.1 Proof of Proposition 16

As mentioned at the beginning of Section 4, the statements in the convergence proof of [AdaNAG](#) that require global smoothness are only [Corollary 9](#) and [Proposition 11](#). Since the reason that [Proposition 11](#) requires global smoothness is to leverage [Corollary 9](#), our goal is to extend [Corollary 9](#) to the locally smooth case.

Recall that [Corollary 9](#) is about the upper bound of L_k . Unlike the globally smooth case, when f is only locally smooth, the boundedness of the iterates becomes crucial for obtaining an upper bound on L_k . To establish the boundedness of the iterates, we first prove a statement similar to [Theorem 2](#), but with a sharper bound involving the coefficient $s_k^2 B_k$ on the right-hand side.

Lemma 17. *Suppose $\{x_k\}_{k \geq 0}$ is generated by [AdaNAG](#) but with (7) replaced by (55) for some $N \geq 3$. Let f be a locally smooth convex function. Define V_k , A_k , and B_k as in [Theorem 2](#). Then there exists $\delta > 0$ such that*

$$V_{k+1} - V_k \leq -\frac{\delta}{2} s_k^2 B_k \|\nabla f(x_k)\|^2, \quad \forall k \geq N.$$

Proof. Using the same argument as in [Lemma 4](#), from (22), we have:

$$\left(\frac{A_k}{B_k} + \rho_k + \epsilon\right)^{-1} = \left(\frac{A_k}{B_k} + 1 + \epsilon\right)^{-1} \geq \left(\frac{\alpha_{k+1}}{\alpha_k^2} + 1 + \epsilon\right)^{-1} = \frac{\alpha_k^2}{\alpha_{k+1} + \alpha_k^2(1 + \epsilon)}.$$

Therefore, for $k \geq N$, using (55), we obtain a tighter inequality compared to (16c):

$$s_{k+1} \leq \left(\frac{A_k}{B_k} + \rho_k + \epsilon\right)^{-1} \frac{1}{L_{k+1}}. \quad (16c')$$

Next, for notational simplicity, we define:

$$\delta_k = \epsilon s_{k+1} L_{k+1} (1 - \rho_k s_{k+1} L_{k+1})^{-1}. \quad (56)$$

With Q_k defined as in [Lemma 6](#), we now aim to prove a tighter inequality than (20), stated below:

$$Q_k \leq \begin{cases} -\frac{1}{2} s_k^2 B_k \|\nabla f(x_k)\|^2 & \text{if } s_k L_{k+1} \leq 1 \\ -\left(\frac{1}{2} \frac{s_{k+1}}{L_{k+1}} A_k + \frac{\delta_k}{2} s_k^2 B_k\right) \|\nabla f(x_k)\|^2 & \text{if } s_k L_{k+1} \geq 1. \end{cases} \quad (20')$$

Note that the difference from (20) is the additional term $\frac{\delta_k}{2} s_k^2 B_k \|\nabla f(x_k)\|^2$ when $s_k L_{k+1} \geq 1$.

We first show that the above inequality implies our desired conclusion. Suppose (20') is true. Recall that $\lim_{k \rightarrow \infty} \alpha_k = \frac{1}{2}$ from [Lemma 3](#). Thus $\frac{\alpha_k}{\alpha_{k+1}}$ and $\frac{\alpha_k^2}{\alpha_{k+1} + \alpha_k^2(1 + \epsilon)}$ have positive limits, namely 1 and $\frac{1}{3 + \epsilon}$ respectively, thus there exists $\delta \in (0, 1)$ such that

$$\epsilon \min \left\{ \frac{\alpha_k}{\alpha_{k+1}}, \frac{\alpha_k^2}{\alpha_{k+1} + \alpha_k^2(1 + \epsilon)} \right\} > \delta, \quad \forall k \geq N.$$

For such δ , when $s_k L_{k+1} \geq 1$, recalling the definition of δ_k in (56), we have:

$$\delta_k \geq \epsilon s_{k+1} L_{k+1} = \epsilon \min \left\{ \frac{\alpha_k}{\alpha_{k+1}} s_k L_{k+1}, \frac{\alpha_k^2}{\alpha_{k+1} + \alpha_k^2(1 + \epsilon)} \right\} > \delta, \quad \forall k \geq N.$$

From the same argument as in [Theorem 2](#), we have $V_{k+1} - V_k \leq Q_k$. Therefore, we obtain:

$$V_{k+1} - V_k \leq Q_k \leq -\frac{1}{2} \min \{1, \delta_k\} s_k^2 B_k \|\nabla f(x_k)\|^2 \leq -\frac{\delta}{2} s_k^2 B_k \|\nabla f(x_k)\|^2, \quad \forall k \geq N.$$

It now remains to show (20'). The overall argument follows the same structure as the proof of Lemma 6, except that we use a slightly tighter argument in the case $s_k L_{k+1} \geq 1$. With the same argument of Lemma 6, showing (20') reduces to showing that the discriminant of the quadratic form is nonpositive:

$$(1 - s_k L_{k+1})^2 - (1 - \rho_k s_{k+1} L_{k+1}) \frac{B_k}{A_k} \frac{s_k^2}{s_{k+1}} L_{k+1} (1 - \delta_k) \leq 0, \quad \text{if } s_k L_{k+1} \geq 1.$$

Note that from (16c'), we have $\frac{1}{s_{k+1} L_{k+1}} \geq \frac{A_k}{B_k} + \rho_k + \epsilon$. Together with the definition of δ_k in (56), we can prove the above inequality as follows:

$$\begin{aligned} (1 - \rho_k s_{k+1} L_{k+1}) \frac{B_k}{A_k} \frac{s_k^2}{s_{k+1}} L_{k+1} (1 - \delta_k) &= \left(\frac{1}{s_{k+1} L_{k+1}} - \rho_k \right) \frac{B_k}{A_k} s_k^2 L_{k+1}^2 - \epsilon \frac{B_k}{A_k} s_k^2 L_{k+1}^2 \\ &\geq \frac{B_k}{A_k} \left(\frac{1}{s_{k+1} L_{k+1}} - \rho_k \right) s_k^2 L_{k+1}^2 - \epsilon \frac{B_k}{A_k} s_k^2 L_{k+1}^2 + 2(1 - s_k L_{k+1}) \\ &\geq \frac{B_k}{A_k} \left(\frac{A_k}{B_k} + \epsilon \right) s_k^2 L_{k+1}^2 - \epsilon \frac{B_k}{A_k} s_k^2 L_{k+1}^2 + 2(1 - s_k L_{k+1}) = 2 + s_k^2 L_{k+1}^2 - 2s_k L_{k+1} \geq (1 - s_k L_{k+1})^2. \end{aligned}$$

This concludes the proof. \square

We now establish the boundedness of the iterates by leveraging Lemma 17.

Proposition 18. *Let f be a locally smooth convex function. Suppose $\{x_k\}_{k \geq 0}$ is generated by AdaNAG but with (7) replaced by (55) for some $N \geq 3$. Define $R_k = \sqrt{2V_{-1}} + \sum_{i=0}^{k-1} s_k \|\nabla f(x_i)\|$, where V_{-1} defined as in Proposition 7. Then, $\bar{R} = \lim_{k \rightarrow \infty} R_k$ exists and*

$$x_k \in \bar{B}_{\bar{R}}(x_*), \quad \forall k \geq 0. \quad (57)$$

Proof. By summing the inequality obtained in Lemma 17, we have $\frac{\delta}{2} \sum_{k=N}^{\infty} s_k^2 B_k \|\nabla f(x_k)\|^2 \leq V_N$. Recall that from (15) we know $B_k = \alpha_k^2 \theta_{k+2}^2$ for $k \geq N$. Since $\theta_{k+2} \geq \frac{1}{2}(k+4) \geq \frac{1}{2}k$ and $\alpha_k \geq \alpha_N$ for $k \geq N$ by Lemma 3, we have

$$\sum_{k=N}^{\infty} s_k^2 k^2 \|\nabla f(x_k)\|^2 \leq \frac{4}{\alpha_N^2} \sum_{k=N}^{\infty} s_k^2 B_k \|\nabla f(x_k)\|^2 \leq \frac{4}{\alpha_N^2} \frac{2}{\delta} V_N < \infty.$$

Therefore, $\sum_{k=1}^{\infty} s_k^2 k^2 \|\nabla f(x_k)\|^2 < \infty$. Now, from the Cauchy-Schwarz inequality in the ℓ^2 space,

$$\sum_{k=1}^{\infty} s_k \|\nabla f(x_k)\| = \sum_{k=1}^{\infty} \left(\frac{1}{k} \times s_k k \|\nabla f(x_k)\| \right) \leq \left(\sum_{k=1}^{\infty} \frac{1}{k^2} \right)^{\frac{1}{2}} \left(\sum_{k=1}^{\infty} s_k^2 k^2 \|\nabla f(x_k)\|^2 \right)^{\frac{1}{2}} < \infty.$$

This implies that $R_k = \|x_0 - x_*\| + \sum_{i=0}^{k-1} s_i \|\nabla f(x_i)\| < \infty$. Since R_k is a nondecreasing sequence, we conclude that R_k converges.

To prove (57), since $\bar{B}_{R_k}(x_*) \subset \bar{B}_{\bar{R}}(x_*)$, it suffices to show $x_k \in \bar{B}_{R_k}(x_*)$ for all $k \geq 0$. We prove this by induction. By definition, $x_0 \in \bar{B}_{R_0}(x_*)$ is clear. Now, assume $x_k \in \bar{B}_{R_k}(x_*)$ is true, and we want to prove $x_{k+1} \in \bar{B}_{R_{k+1}}(x_*)$. From (5), we have $y_{k+1} = x_k - s_k \nabla f(x_k)$. Hence,

$$\|y_{k+1} - x_*\| \leq \|x_k - x_*\| + s_k \|\nabla f(x_k)\| \leq R_k + s_k \|\nabla f(x_k)\| = R_{k+1}.$$

From Theorem 2, we have $\|z_{k+1} - x_*\| \leq \sqrt{2V_k} \leq \sqrt{2V_{-1}} \leq R_{k+1}$. Using (5) again, we have:

$$\|x_{k+1} - x_*\| \leq \left(1 - \frac{1}{\theta_{k+3}} \right) \|y_{k+1} - x_*\| + \frac{1}{\theta_{k+3}} \|z_{k+1} - x_*\| \leq R_{k+1}.$$

Therefore $x_{k+1} \in \bar{B}_{R_{k+1}}(x_*)$, and this completes the proof. \square

The following lemma is an extension of Lemma 8 to the relaxed step size (55). The definition of $\bar{\epsilon}_N$ in (54) plays a crucial role in keeping the coefficient in front of S_k the same.

Lemma 19. Suppose $\{x_k\}_{k \geq 0}$ is generated by AdaNAG but with (7) replaced by (55) for some $N \geq 3$. Let f be a locally smooth convex function. Set s_0 as in Lemma 8 and define S_k as (28). Then

$$s_k \geq \frac{\alpha_2^2 \alpha_3}{\alpha_3 + \alpha_2^2} \frac{1}{\alpha_k} S_k, \quad \forall k \geq 1.$$

Proof. Note that $\bar{\epsilon}_N$ in (54) satisfies the equality $\frac{\alpha_N^2 \alpha_{N+1}}{\alpha_{N+1} + \alpha_N^2(1 + \bar{\epsilon}_N)} = \frac{\alpha_2^2 \alpha_3}{\alpha_3 + \alpha_2^2}$. Since $\epsilon \mapsto \frac{\alpha_N^2 \alpha_{N+1}}{\alpha_{N+1} + \alpha_N^2(1 + \epsilon)}$ is decreasing in ϵ when $\epsilon > 0$, we have:

$$\frac{\alpha_N^2 \alpha_{N+1}}{\alpha_{N+1} + \alpha_N^2(1 + \epsilon)} > \frac{\alpha_2^2 \alpha_3}{\alpha_3 + \alpha_2^2}, \quad \forall \epsilon \in (0, \bar{\epsilon}_N).$$

Note that when $1 \leq k \leq N - 1$, the inequality $s_k \geq \frac{\alpha_2^2 \alpha_3}{\alpha_3 + \alpha_2^2} \frac{1}{\alpha_k} S_k$ is immediate from Lemma 8. In particular, the inequality is true for $k = N - 1$. Now, we proceed by induction for the case $k \geq N - 1$. Repeating the same argument as in Lemma 8, we can show that for every fixed $\epsilon > 0$, the function $k \mapsto \frac{\alpha_k^2 \alpha_{k+1}}{\alpha_{k+1} + \alpha_k^2(1 + \epsilon)} = \frac{\alpha_{k+1}}{\frac{\alpha_k^2}{\alpha_k} + 1 + \epsilon}$ is increasing. Therefore, we have

$$\frac{\alpha_k^2 \alpha_{k+1}}{\alpha_{k+1} + \alpha_k^2(1 + \epsilon)} \geq \frac{\alpha_k^2 \alpha_{k+1}}{\alpha_{k+1} + \alpha_k^2(1 + \bar{\epsilon}_N)} \geq \frac{\alpha_N^2 \alpha_{N+1}}{\alpha_{N+1} + \alpha_N^2(1 + \bar{\epsilon}_N)} \geq \frac{\alpha_2^2 \alpha_3}{\alpha_3 + \alpha_2^2}, \quad \forall k \geq N.$$

Applying the above inequality and the induction hypothesis $s_k \geq \frac{\alpha_2^2 \alpha_3}{\alpha_3 + \alpha_2^2} \frac{1}{\alpha_k} S_k$, we have

$$\begin{aligned} s_{k+1} &= \min \left\{ \frac{\alpha_k}{\alpha_{k+1}} s_k, \frac{\alpha_k^2}{\alpha_{k+1} + \alpha_k^2(1 + \epsilon)} \frac{1}{L_{k+1}} \right\} \\ &\geq \min \left\{ \frac{\alpha_2^2 \alpha_3}{\alpha_3 + \alpha_2^2} \frac{1}{\alpha_{k+1}} S_k, \frac{\alpha_2^2 \alpha_3}{\alpha_3 + \alpha_2^2} \frac{1}{\alpha_{k+1}} \frac{1}{L_{k+1}} \right\} = \frac{\alpha_2^2 \alpha_3}{\alpha_3 + \alpha_2^2} \frac{1}{\alpha_{k+1}} S_{k+1}. \end{aligned}$$

We conclude the desired result by induction. \square

The following corollary is our promised goal, an extension of Corollary 9 to a locally smooth function f . The purpose of the relaxation (55) and Proposition 18 was to prove this corollary.

Corollary 20. Suppose f is a locally smooth convex function and define $\{s_k\}_{k \geq 0}$ as in Lemma 19. Let \bar{R} be as in Proposition 18, and define L as in Theorem 9. Then

$$s_k \geq \frac{\alpha_2^2 \alpha_3}{\alpha_3 + \alpha_2^2} \frac{1}{\alpha_k} \frac{1}{L}, \quad \forall k \geq 1.$$

Proof. Note that we have $x_k \in \bar{B}_{\bar{R}}(x_*)$ for all $k \geq 0$ from Proposition 18. Since $\tilde{x}_0 \in \bar{B}_{\|\tilde{x}_0 - x_*\|}(x_*)$, we obtain the conclusion from Lemma 1. \square

Since we have established a lower bound for s_k , the proof of Proposition Lemma 16 follows immediately by repeating the previous arguments.

Proof of Proposition 16. We obtain the boundedness of $\{x_k\}_{k \geq 0}$ from Proposition 18. By applying Lemma 19 and Corollary 20 in place of Lemma 8 and Corollary 9, and repeating the same arguments used in the proof of Theorem 1, we obtain the convergence rate (53). \square

B.3.2 Boundedness of AdaNAG for globally smooth f

As an independent and interesting observation, we show that AdaNAG also achieves the boundedness of the iterates for globally smooth f , using an argument similar to that of Proposition 18. This implies that L_k is bounded by some local smoothness parameter smaller than the global smoothness constant L , and thus we actually have a tighter guarantee than Theorem 1.

Lemma 21. *Let f be an L -smooth convex function. Suppose $\{x_k\}_{k \geq 0}$ is generated by AdaNAG. Define R_k as in Proposition 18. Then, $\bar{R} = \lim_{k \rightarrow \infty} R_k$ exists and $x_k \in \bar{B}_{\bar{R}}(x_*)$ for all $k \geq 0$.*

Proof. Recalling the proof of Proposition 18, it suffices to show that $\sum_{i=1}^{\infty} s_i^2 i^2 \|\nabla f(x_i)\|^2 < \infty$. Considering $k \rightarrow \infty$ in (30) and using the fact $\theta_{i+3} \geq \frac{1}{2}(i+5)$ for $i \geq 1$ from Lemma 3, we obtain $\frac{1}{4} \sum_{i=1}^{\infty} i^2 \|\nabla f(x_i)\|^2 < \infty$. From Lemma 3, we know that $s_{i+1} \leq s_i$ holds for $i \geq 1$. Therefore,

$$\sum_{i=1}^{\infty} s_i^2 i^2 \|\nabla f(x_i)\|^2 \leq s_1^2 \sum_{i=1}^{\infty} i^2 \|\nabla f(x_i)\|^2 < \infty.$$

The rest of the proof can be obtained with the same argument of Proposition 18. \square

B.4 Numerical values used in the proof of Theorem 1 and Proposition 15

The numerical values used in the proof of Theorem 1 and Proposition 15 are provided in Table 5.

Recall that we denote $r_0 = \frac{\theta_3(\theta_3-1)}{\theta_2} \frac{1}{\alpha_0} \frac{\alpha_2^2 \alpha_3}{\alpha_3 + \alpha_2^2}$.

θ_k	$\frac{\alpha_0}{\alpha_1} \frac{\theta_2}{\theta_3(\theta_3-1)}$	r_0	$\frac{2(\alpha_3 + \alpha_2^2)}{\alpha_2^2 \alpha_3}$	$r_0^2 \alpha_0 (\tilde{\alpha}_0 + \alpha_0) \theta_2^2$	$\frac{1}{r_0(\tilde{\alpha}_0 + \alpha_0) \theta_2}$
AdaNAG	0.6745	0.4255	21.9032	0.3045	1.4424
Algorithm 4	0.8399	0.3363	23.6	0.1544	2.0578

Table 5: Numerical values (rounded to the fourth decimal place) used in the proof of Theorems 1 and 15.

C Omitted details in Section 5

C.1 Remaining Proof of Proposition 12

By similar calculation as (23), for $k \geq -1$ we have

$$\frac{1}{2} \|x_{k+2} - x_*\|^2 - \frac{1}{2} \|x_{k+1} - x_*\|^2 = -s_{k+1} \langle \nabla f(x_{k+1}), x_{k+1} - x_* \rangle + \frac{s_{k+1}^2}{2} \|\nabla f(x_{k+1})\|^2.$$

When $k = -1$, we can verify (38) as follows:

$$V_0 - V_{-1}$$

$$\begin{aligned} &= s_1 A_0 (f(x_0) - f_*) + \frac{1}{2} \|x_1 - x_*\|^2 + \frac{1}{2} s_0^2 B_0 \|\nabla f(x_0)\|^2 - \left(\frac{1}{2} \|x_0 - x_*\|^2 + \frac{1}{2} s_0^2 (B_0 + 1) \|\nabla f(x_0)\|^2 \right) \\ &= (s_0 - s_1 A_0) (f_* - f(x_0)) + s_0 (f(x_0) - f_* - \langle \nabla f(x_0), x_0 - x_* \rangle). \end{aligned}$$

Now suppose $k \geq 0$. We first consider the case $L_{k+1} \neq 0$. It is easy to check

$$\begin{aligned}
& s_{k+2}A_{k+1}(f(x_{k+1}) - f_\star) - s_{k+1}A_k(f(x_k) - f_\star) \\
&= (s_{k+1}A_k + s_{k+1} - s_{k+2}A_{k+1})(f_\star - f(x_{k+1})) + s_{k+1}(f(x_{k+1}) - f_\star) \\
&+ s_{k+1}A_k \underbrace{\left(f(x_{k+1}) - f(x_k) - \langle \nabla f(x_{k+1}), x_{k+1} - x_k \rangle + \frac{1}{2L_{k+1}} \|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2 \right)}_{=0} \\
&+ s_{k+1}A_k \langle \nabla f(x_{k+1}), x_{k+1} - x_k \rangle - \frac{s_{k+1}A_k}{2L_{k+1}} \|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2.
\end{aligned}$$

Applying $x_{k+1} - x_k = -s_k \nabla f(x_{k+1})$ to the last line, we obtain

$$\begin{aligned}
V_{k+1} - V_k &= (s_{k+1}A_k + s_{k+1} - s_{k+2}A_{k+1})(f_\star - f(x_{k+1})) \\
&+ s_{k+1}(f(x_{k+1}) - f_\star - \langle \nabla f(x_{k+1}), x_{k+1} - x_\star \rangle) + Q_k,
\end{aligned} \tag{58}$$

where

$$\begin{aligned}
Q_k &= -\frac{s_{k+1}A_k}{2L_{k+1}} \left[\left(1 - \frac{B_{k+1}+1}{A_k} s_{k+1}L_{k+1} \right) \|\nabla f(x_{k+1})\|^2 \right. \\
&\quad \left. - 2(1 - s_k L_{k+1}) \langle \nabla f(x_{k+1}), \nabla f(x_k) \rangle + \left(1 + \frac{B_k}{A_k} \frac{s_k^2}{s_{k+1}} L_{k+1} \right) \|\nabla f(x_k)\|^2 \right].
\end{aligned}$$

For simplicity, denote $\rho_k = \frac{B_{k+1}+1}{A_k}$. We now repeat the argument in the proof of Lemma 6. Since

$$\tilde{s}_k^2 \min \{A_k, B_k\} \leq \min \left\{ \frac{s_{k+1}}{L_{k+1}} A_k, s_k^2 B_k \right\},$$

where $\tilde{s}_k = \min \{s_k, s_{k+1}, \frac{1}{L_{k+1}}\}$, it suffices to show:

$$\begin{aligned}
Q_k + \frac{B_k}{2} s_k^2 \|\nabla f(x_k)\|^2 &\leq 0 & \text{if } s_k L_{k+1} \leq 1 \\
Q_k + \frac{A_k}{2} \frac{s_{k+1}}{L_{k+1}} \|\nabla f(x_k)\|^2 &\leq 0 & \text{if } s_k L_{k+1} \geq 1.
\end{aligned}$$

First, we focus on the coefficient of the $\|\nabla f(x_{k+1})\|^2$ term. From (31), we have

$$1 - \rho_k s_{k+1} L_{k+1} \geq 1 - \left(\frac{A_k}{B_k} + \rho_k \right) s_{k+1} L_{k+1} \geq 0.$$

Considering the discriminant of the quadratic form, our goal reduces to showing:

$$\begin{aligned}
(1 - s_k L_{k+1})^2 - (1 - \rho_k s_{k+1} L_{k+1}) &\leq 0 & \text{if } s_k L_{k+1} \leq 1, \\
(1 - s_k L_{k+1})^2 - (1 - \rho_k s_{k+1} L_{k+1}) \frac{B_k}{A_k} \frac{s_k^2}{s_{k+1}} L_{k+1} &\leq 0 & \text{if } s_k L_{k+1} \geq 1.
\end{aligned}$$

Now, we divide into two cases, as we have done in the proof of Lemma 6.

- $s_k L_{k+1} \leq 1$. From (33) and (31) we have $\rho_k s_{k+1} L_{k+1} \leq \frac{A_k}{A_{k-1}+1} s_{k+1} L_{k+1} \leq s_{k+1} L_{k+1}$. Again leveraging (27), we conclude the desired inequality.

- $s_k L_{k+1} \geq 1$. From (31) we know $\frac{1}{s_{k+1} L_{k+1}} \geq \frac{A_k}{B_k} + \rho_k$ holds for $k \geq 0$. Therefore, recalling that $1 - 2s_k L_{k+1} \leq 0$, we obtain:

$$(1 - \rho_k s_{k+1} L_{k+1}) \frac{B_k}{A_k} \frac{s_k^2}{s_{k+1}} L_{k+1} = \frac{B_k}{A_k} \left(\frac{1}{s_{k+1} L_{k+1}} - \rho_k \right) s_k^2 L_{k+1}^2 \geq s_k^2 L_{k+1}^2 \geq (1 - s_k L_{k+1})^2.$$

Now we move on to the case $L_{k+1} = 0$. Note that this implies $s_k L_{k+1} = 0 \leq 1$. From Corollary 2 we have $\nabla f(x_k) = \nabla f(x_{k+1})$. Then from the same calculation done to (58) but without $\frac{1}{L_{k+1}} \|\nabla f(x_k) - \nabla f(x_{k+1})\|^2$, we have

$$\begin{aligned} V_{k+1} - V_k &= (s_{k+1} A_k + s_{k+1} - s_{k+2} A_{k+1})(f_\star - f(x_{k+1})) \\ &\quad + s_{k+1}(f(x_{k+1}) - f_\star - \langle \nabla f(x_{k+1}), x_{k+1} - x_\star \rangle) \\ &\quad + s_{k+1} A_k (f(x_{k+1}) - f(x_k) - \langle \nabla f(x_{k+1}), x_{k+1} - x_k \rangle) \\ &\quad - s_k s_{k+1} A_k \langle \nabla f(x_{k+1}), \nabla f(x_k) \rangle + \frac{1}{2} s_{k+1}^2 (B_{k+1} + 1) \|\nabla f(x_{k+1})\|^2 - \frac{1}{2} s_k^2 B_k \|\nabla f(x_k)\|^2 \\ &\leq -s_k s_{k+1} A_k \langle \nabla f(x_{k+1}), \nabla f(x_k) \rangle + \frac{1}{2} s_{k+1}^2 (B_{k+1} + 1) \|\nabla f(x_{k+1})\|^2 - \frac{1}{2} s_k^2 B_k \|\nabla f(x_k)\|^2 \\ &= \frac{1}{2} s_{k+1} (s_{k+1} (B_{k+1} + 1) - s_k A_k) \|\nabla f(x_k)\|^2 - \left(\frac{1}{2} s_k s_{k+1} A_k + \frac{1}{2} s_k^2 B_k \right) \|\nabla f(x_k)\|^2 \\ &\leq -\frac{1}{2} s_k^2 B_k \|\nabla f(x_k)\|^2. \end{aligned}$$

The second inequality holds since the following inequalities follow from (33) and (31):

$$s_{k+1} (B_{k+1} + 1) - s_k A_k \leq s_{k+1} \frac{A_k^2}{1 + A_{k-1}} - s_k A_k \leq 0.$$

This concludes the proof. \square

C.2 Proof of Lemma 13

Define S_k as in (28), i.e., $S_k = \min \left\{ \frac{1}{L_0}, \frac{1}{L_1}, \dots, \frac{1}{L_k} \right\}$. We first prove the following by induction:

$$s_k \geq r S_k, \quad \forall k \geq 1. \quad (59)$$

From (31), we have $s_1 = \min \left\{ \frac{1}{A_0} s_0, r \frac{1}{L_1} \right\} = \min \left\{ r \frac{1}{L_0}, r \frac{1}{L_1} \right\} = r S_1$. Thus (59) holds for $k = 1$. Assume (59) holds for k , now we show that it also holds for $k+1$. Applying the induction hypothesis $s_k \geq r S_k$ and (32), we obtain the following inequality, which completes the proof of (59):

$$s_{k+1} \geq \min \left\{ \frac{A_{k-1} + 1}{A_k} r S_k, \left(\frac{A_k}{B_k} + \frac{B_{k+1} + 1}{A_k} \right)^{-1} \frac{1}{L_{k+1}} \right\} \geq \min \left\{ r S_k, r \frac{1}{L_{k+1}} \right\} = r S_{k+1}.$$

To prove (39), recall that we have $x_k \in \bar{B}_R(x_\star)$ from Proposition 12 and $\tilde{x}_0 \in \bar{B}_{\|\tilde{x}_0 - x_\star\|}(x_\star)$. Therefore, from Lemma 1, we know that $L_k \leq L$ for $k \geq 0$, where L is a smoothness parameter of f on $\bar{B}_{3\bar{R}}(x_\star) \cup \bar{B}_{3\|\tilde{x}_0 - x_\star\|}(x_\star)$. Therefore, we have $S_k \geq \frac{1}{L}$. From (59), we conclude (39). \square

D Omitted proofs in Section 6

D.1 Proof of Theorem 6

Similar to the proof of the convergence of [AdaNAG](#), we state and prove lemmas similar to Theorem 2, Lemma 7 and Lemma 8.

Lemma 22. Suppose $\{x_k\}_{k \geq 0}$ is generated by [AdaNAG-G](#) with $\{\tau_k\}_{k \geq 0}$, $\{\alpha_k\}_{k \geq 0}$ satisfying the assumptions of Theorem 6. Define Lyapunov function as

$$V_k = s_{k+1}A_k(f(x_k) - f_\star) + \frac{1}{2}s_k^2B_k\|\nabla f(x_k)\|^2 + \frac{1}{2}\|z_{k+1} - x_\star\|^2, \quad \forall k \geq -1$$

where $x_{-1} = x_0$, $s_{-1} = s_0$ and $B_{-1} = B_0 + \alpha_0^2\tau_0^2$. Then

$$V_{k+1} - V_k \leq -\frac{1}{2} \min \left\{ s_k^2B_k, \frac{s_{k+1}}{L_{k+1}}A_k \right\} \|\nabla f(x_k)\|^2, \quad \forall k \geq 0$$

holds, and

$$V_0 \leq V_{-1} + s_0\alpha_0\tau_0(f(x_0) - f_\star - \langle \nabla f(x_0), x_0 - x_\star \rangle)$$

holds, when f is a locally smooth convex function.

Proof. For $k \geq 0$, the proof follows by repeating the proof argument of Theorem 2, with θ_{k+2} replaced by τ_k . The only thing we need to verify is that (16) holds when θ_{k+2} is replaced with τ_k . We now verify this point. Starting with (16a) and (16c), we can see they follow immediately from the step size rule of (43). Next, observe that

$$B_k = \alpha_k^2\tau_k^2 \left(\frac{(\tau_k - 1)^2}{\alpha_{k-1}\tau_{k-1}^2} - 1 \right) = \frac{A_{k-1}^2}{A_{k-2} + \alpha_{k-1}\tau_{k-1}} - \alpha_k^2\tau_k^2 \iff \frac{A_{k-2} + \alpha_{k-1}\tau_{k-1}}{A_{k-1}} = \frac{A_{k-1}}{B_k + \alpha_k^2\tau_k^2},$$

so the definition of B_k implies $\frac{A_{k-2} + \alpha_{k-1}\tau_{k-1}}{A_{k-1}} = \frac{A_{k-1}}{B_k + \alpha_k^2\tau_k^2}$. Combining with (16a), we obtain (16b).

Now we move on to $k = -1$. Recalling $x_0 = z_0$ and (23), we have

$$\|z_1 - x_\star\|^2 - \|x_0 - x_\star\|^2 = -2s_0\alpha_0\tau_0 \langle \nabla f(x_0), x_0 - x_\star \rangle + s_0^2\alpha_0^2\tau_0^2 \|\nabla f(x_0)\|^2.$$

Therefore

$$\begin{aligned} V_0 - V_{-1} &= s_1A_0(f(x_0) - f_\star) + \frac{1}{2}\|z_1 - x_\star\|^2 + \frac{1}{2}s_0^2B_0\|\nabla f(x_0)\|^2 - \left(\frac{1}{2}\|x_0 - x_\star\|^2 + \frac{1}{2}s_0^2B_{-1}\|\nabla f(x_0)\|^2 \right) \\ &= (s_0\alpha_0\tau_0 - s_1A_0)(f_\star - f(x_0)) + s_0\alpha_0\tau_0(f(x_0) - f_\star - \langle \nabla f(x_0), x_0 - x_\star \rangle) \\ &\leq s_0\alpha_0\tau_0(f(x_0) - f_\star - \langle \nabla f(x_0), x_0 - x_\star \rangle). \end{aligned}$$

The inequality follows from $s_0\alpha_0\tau_0 - s_1A_0 = A_0 \left(\frac{\alpha_0\tau_0}{A_0}s_0 - s_1 \right) \geq 0$, which holds true by (43). This concludes the proof. \square

Lemma 23. Suppose $\{s_k\}_{k \geq 0}$ is generated by [AdaNAG-G](#) with $\{\tau_k\}_{k \geq 0}$, $\{\alpha_k\}_{k \geq 0}$ satisfying the assumptions in Theorem 6. Set s_0 as in Theorem 6 and let f be an L -smooth convex function. Then

$$s_k \geq \frac{r}{\alpha_k} \frac{1}{L}, \quad \forall k \geq 1.$$

Proof. Define $S_k = \min \left\{ \frac{1}{L_0}, \frac{1}{L_1}, \dots, \frac{1}{L_k} \right\}$, as in (28). We first prove the following by induction.

$$s_k \geq \frac{r}{\alpha_k} S_k, \quad \forall k \geq 1. \quad (60)$$

- $k = 1$. From (43), the assumption $s_0 = \frac{A_0}{\alpha_0 \tau_0} \frac{r}{\alpha_1} \frac{1}{L_0}$, we have $s_1 \geq \min \left\{ \frac{\alpha_0 \tau_0}{A_0} s_0, \frac{r}{\alpha_1} \frac{1}{L_1} \right\} = \frac{r}{\alpha_1} S_1$.
- $k \geq 2$. From the second inequality of (44), we have $\frac{A_{k-1} + \alpha_k \tau_k}{A_k} = \frac{\alpha_k}{\alpha_{k+1}} \frac{\tau_k}{\tau_{k+1}(\tau_{k+1} - 1)} \geq \frac{\alpha_k}{\alpha_{k+1}}$. Applying the third inequality of (44), the induction hypothesis $s_k \geq \frac{r}{\alpha_k} S_k$, and above inequality, we have

$$s_{k+1} \geq \min \left\{ \frac{\alpha_k}{\alpha_{k+1}} \frac{r}{\alpha_k} S_k, \frac{r}{\alpha_{k+1}} \frac{1}{L_{k+1}} \right\} = \frac{r}{\alpha_{k+1}} \min \left\{ S_k, \frac{1}{L_{k+1}} \right\} = \frac{r}{\alpha_{k+1}} S_{k+1}.$$

We conclude the result of (60) by induction.

Next, from Lemma 1, we have $\frac{1}{L_k} \geq \frac{1}{L}$ for $k \geq 1$ and $\frac{1}{L_0} \geq \frac{1}{L}$ by its definition (3). Recalling the definition of S_k , we obtain $S_k \geq \frac{1}{L}$ for $k \geq 0$. We obtain the desired conclusion by (60). \square

Proof of Theorem 6. The proof follows from Lemma 22 and Lemma 23. From Lemma 22 and (2) we have

$$s_{k+1} A_k (f(x_k) - f_\star) \leq V_k \leq \dots \leq V_0 \leq V_{-1} - \frac{s_0 \alpha_0 \tau_0}{2L} \|\nabla f(x_0)\|^2 = \frac{1}{2} R,$$

and so $f(x_k) - f_\star \leq \frac{1}{s_{k+1} A_k} R$. From Lemma 23, we have $s_{k+1} \alpha_{k+1} \geq \frac{r}{L}$ for $k \geq 0$.

The last statement can be proved with a similar argument as in Proposition 11. Since $\alpha_k \in (0, 1]$, by Lemma 23 we have $s_k \geq \frac{r}{L}$ for $k \geq 1$. Thus, we have $\min \{s_k^2, \frac{s_{k+1}}{L_{k+1}}\} \geq \frac{r^2}{L^2}$. Therefore, from Lemma 22 we obtain

$$V_{k+1} - V_k \leq -\frac{r^2}{2L^2} \min \{A_k, B_k\} \|\nabla f(x_k)\|^2, \quad \forall k \geq 0.$$

Summing up from 0 to $k-1$, applying $V_0 \leq \frac{1}{2} R$, we obtain

$$\left(\frac{r^2}{2L^2} \sum_{i=1}^k \min \{A_i, B_i\} \right) \min_{i \in \{1, \dots, k\}} \|\nabla f(x_i)\|^2 \leq \frac{r^2}{2L^2} \sum_{i=1}^k \min \{A_i, B_i\} \|\nabla f(x_i)\|^2 \leq \frac{1}{2} R,$$

which implies our desired result.

Lastly, from the last inequality of (44), we have $\limsup_{k \rightarrow \infty} \frac{A_k}{B_k} < \infty$. Therefore, we have $A_k = \mathcal{O}(\min \{A_k, B_k\})$, we conclude $\frac{1}{\sum_{i=1}^k \min \{A_i, B_i\}} = \mathcal{O}\left(\frac{1}{\sum_{i=1}^k A_i}\right) = \mathcal{O}\left(\frac{1}{\sum_{i=1}^k \alpha_i \tau_i^2}\right)$. \square

D.2 Proof of Lemma 14

Take $\epsilon > 0$. For notation simplicity, name

$$r_k^s = \frac{A_{k-1} + \alpha_k \tau_k}{A_k}, \quad r_k^L = \left(\frac{A_k}{B_k} + \frac{B_{k+1} + \alpha_{k+1}^2 \tau_{k+1}^2}{A_k} \right)^{-1}.$$

Since $\lim_{k \rightarrow \infty} r_k^L = \bar{r}$, and by the assumption on $\frac{A_{k-1} + \alpha_k \tau_k}{A_k}$, there exists $N > 0$ such that

$$k \geq N \implies r_k^L \geq \bar{r} - \epsilon, \quad r_k^s \geq 1. \quad (61)$$

Suppose there exists $K \geq N$ such that $s_K \geq (\bar{r} - \epsilon) \frac{1}{L}$. Since $r_K^s \geq 1$ holds by (61) and $\frac{1}{L_{K+1}} \geq \frac{1}{L}$ holds by Lemma 1, this implies

$$s_{K+1} = \min \left\{ r_K^s s_K, r_K^L \frac{1}{L_{K+1}} \right\} \geq \min \left\{ s_K, (\bar{r} - \epsilon) \frac{1}{L_{K+1}} \right\} \geq (\bar{r} - \epsilon) \frac{1}{L}.$$

By induction, we can conclude $s_k \geq (\bar{r} - \epsilon) \frac{1}{L}$ for all $k \geq K$. Thus, it suffices to show such K exists. Proof by contradiction. Assume that such a K does not exist, i.e., assume that

$$K \geq N \quad \implies \quad s_K < (\bar{r} - \epsilon) \frac{1}{L}. \quad (62)$$

Suppose $k \geq N$. Then from (62), (61) and the fact $\frac{1}{L} \leq \frac{1}{L_{k+1}}$, we obtain $s_{k+1} < (\bar{r} - \epsilon) \frac{1}{L} \leq r_k^L \frac{1}{L_{k+1}}$. This implies that $s_{k+1} \neq r_k^L \frac{1}{L_{k+1}}$, which means that the first option of (43) has been activated, and thus $s_{k+1} = r_k^s s_k$ holds. Therefore, we conclude that $k \geq N$ implies $s_{k+1} = r_k^s s_k$. By applying this fact recursively, we obtain:

$$s_K = \prod_{k=N}^K r_k^s s_k = \frac{\alpha_N}{\alpha_K} \prod_{k=N}^K \frac{\tau_k^2}{\tau_{k+1}(\tau_{k+1} - 1)}, \quad \forall K > N.$$

Applying the assumption $\prod_{k=1}^{\infty} \frac{\tau_k^2}{\tau_{k+1}(\tau_{k+1} - 1)} = \infty$, we get $\lim_{K \rightarrow \infty} s_K = \infty$. This implies that there exists K such that $s_K \geq (\bar{r} - \epsilon) \frac{1}{L}$, which contradicts (62). This completes the proof. \square

D.3 Proof of Corollary 7

Here, we state and prove the result for the generalized parameter choice $\tau_k = \frac{(k+2)+p}{p}$, where $p > 2$.

Proposition 24. Let $\tau_k = \frac{(k+2)+p}{p}$, where $p > 2$. Set other parameters and assumptions as Corollary 7. Then for $\{x_k\}_{k \geq 0}$ generated by AdaNAG-G, we have:

$$f(x_k) - f_* \leq \frac{1}{s_{k+1}} \frac{p^2(k+p+3)}{(k+3)(k+4)^2} R = \mathcal{O} \left(\frac{L}{k^2} \right), \quad \min_{i \in \{1, \dots, k\}} \|\nabla f(x_i)\|^2 \leq \mathcal{O} \left(\frac{L^2}{k^3} \right). \quad (63)$$

Here, $s_k \geq \frac{27}{(p+3)(2p^2+8p+17)} \frac{1}{L}$ and

$$\liminf_{k \rightarrow \infty} s_k \geq \frac{1}{3} \frac{1}{L}.$$

Proof. We first show that (44) holds. Name $r_k^L = \left(\frac{A_k}{B_k} + \frac{B_{k+1} + \alpha_{k+1}^2 \tau_{k+1}^2}{A_k} \right)^{-1}$. Using elementary calculations, we can verify:

$$\alpha_k = \frac{(k+3)^2}{2(k+p+2)^2} \in (0, 1] \quad (64a)$$

$$\frac{\tau_k^2}{\tau_{k+1}(\tau_{k+1} - 1)} = \frac{(k+p+2)^2}{(k+3)(k+p+3)} = 1 + \frac{(p-2)(k+p+3)+1}{(k+3)(k+p+3)} \geq 1 \quad (64b)$$

$$\alpha_{k+1} r_k^L = \frac{k+3}{2(k+p+3)} \left(2 \left(\frac{k+p+2}{k+3} \right)^2 + 1 \right)^{-1} \geq \alpha_1 r_0^L =: r. \quad (64c)$$

(64b) holds since $p > 2$. (64c) can be shown by proving that $\alpha_{k+1} r_k^L$ is an increasing sequence, which follows from the facts that $\frac{k+3}{k+p+3} = 1 - \frac{p}{k+p+3}$ is increasing and $\frac{k+p+2}{k+3} = 1 + \frac{p-1}{k+3}$ is decreasing.

Next, note that $\alpha_k \tau_k^2 = \Theta(k^2)$. Therefore, we obtain the convergence rates (63) by Theorem 6. Also, from Theorem 6 and (64a), we have $s_k \alpha_k \geq \frac{r}{L}$ and $\alpha_k \leq \frac{1}{2}$, which together imply $s_k \geq \frac{2r}{L}$. We obtain $r = \frac{27}{2(p+3)(2p^2+8p+17)}$ by substituting $k = 0$ to (64c).

Next, we check the assumptions of Lemma 14. Since $\bar{\alpha} := \lim_{k \rightarrow \infty} \alpha_k = \frac{1}{2}$ and $\lim_{k \rightarrow \infty} \frac{\tau_{k+1}}{\tau_k} = 1$, we obtain:

$$\bar{r} = \lim_{k \rightarrow \infty} \left(\frac{A_k}{B_k} + \frac{B_{k+1} + \alpha_{k+1}^2 \tau_{k+1}^2}{A_k} \right)^{-1} = \left(\frac{1}{\bar{\alpha}} + \bar{\alpha} + \bar{\alpha} \right)^{-1} = \left(2 + \frac{1}{2} + \frac{1}{2} \right)^{-1} = \frac{1}{3}.$$

The condition $\prod_{k=1}^{\infty} \frac{\tau_k^2}{\tau_{k+1}(\tau_{k+1}-1)} = \infty$ follows from (45). Finally, since $p > 2$, there exists $N > 0$ such that $(N+3)(p-2) > 1$. For such N , we can verify that

$$\frac{A_{k-1} + \alpha_k \tau_k}{A_k} = \frac{(k+3)(k+p+3)}{(k+4)^2} = 1 + \frac{(k+3)(p-2)-1}{(k+4)^2} \geq 1, \quad \forall k \geq N.$$

We conclude $\liminf_{k \rightarrow \infty} s_k \geq \frac{1}{3} \frac{1}{L}$ by Lemma 14. □

Proof of Corollary 7. The proof can be obtained by substituting $p = 12$ into Proposition 24. □

D.4 Proof of Corollary 8

We first show that (44) holds. Name $r_k^L = \left(\frac{A_k}{B_k} + \frac{B_{k+1} + \alpha_{k+1}^2 \tau_{k+1}^2}{A_k} \right)^{-1}$. Using elementary calculations, we can verify that:

$$\begin{aligned} \alpha_k &= \frac{1}{2} \in (0, 1] \\ \frac{\tau_k^2}{\tau_{k+1}(\tau_{k+1}-1)} &= \frac{2(k+3)}{2(k+4)-\sqrt{k+4}} = 1 + \frac{\sqrt{k+4}-2}{2(k+4)-\sqrt{k+4}} \geq 1 \\ \alpha_{k+1} r_k^L &\geq \alpha_1 r_0^L \approx 0.102 > \frac{1}{10} =: r. \end{aligned}$$

Therefore, we obtain $s_k \alpha_k \geq \frac{r}{L} \geq \frac{1}{10} \frac{1}{L}$ from Theorem 6, and we conclude that $s_k \geq \frac{1}{5} \frac{1}{L}$. Also, from Theorem 6 and the fact $\alpha_k \tau_k^2 = \Theta(k)$, we obtain the convergence results of Corollary 8.

Next, we verify the assumptions of Lemma 14. Since $\alpha_k = \frac{1}{2}$ and $\lim_{k \rightarrow \infty} \frac{\tau_{k+1}}{\tau_k} = 1$, we have:

$$\bar{r} = \lim_{k \rightarrow \infty} \left(\frac{A_k}{B_k} + \frac{B_{k+1} + \alpha_{k+1}^2 \tau_{k+1}^2}{A_k} \right)^{-1} = \left(2 + \frac{1}{2} + \frac{1}{2} \right)^{-1} = \frac{1}{3}.$$

Also, since $\alpha_k = \alpha_{k+1}$, from the second inequality of (44) we obtain:

$$\frac{A_{k-1} + \alpha_k \tau_k}{A_k} = \frac{\alpha_k}{\alpha_{k+1}} \frac{\tau_k^2}{\tau_{k+1}(\tau_{k+1}-1)} = \frac{\tau_k^2}{\tau_{k+1}(\tau_{k+1}-1)} \geq 1, \quad \forall k \geq 0.$$

Finally, from $\frac{\sqrt{k+4}-2}{2(k+4)-\sqrt{k+4}} \geq 0$ we have:

$$\prod_{k=1}^{\infty} \frac{\tau_k^2}{\tau_{k+1}(\tau_{k+1}-1)} = \prod_{k=1}^{\infty} \left(1 + \frac{\sqrt{k+4}-2}{2(k+4)-\sqrt{k+4}} \right) \geq 1 + \sum_{k=1}^{\infty} \frac{\sqrt{k+4}-2}{2(k+4)-\sqrt{k+4}} = \infty.$$

This completes the proof. □