Accelerating Forward-Backward Algorithms in Strongly Convex Optimization Without Modulus Knowledge

Chris Junchi Li[◇]

Department of Electrical Engineering and Computer Sciences[◇]
University of California, Berkeley

September 24, 2024

Abstract

In this work, we establish the linear convergence of forward-backward accelerated algorithms applied to strongly convex functions without prior knowledge of the modulus of strong convexity. Using a novel Lyapunov function and leveraging the high-resolution ordinary differential equation (ODE) framework, we rigorously prove the convergence rates for both Nesterov's accelerated gradient (NAG) method and its proximal variant, FISTA, in smooth and composite optimization problems. Our analysis introduces a dynamically adapting coefficient for the kinetic energy, leading to improved theoretical understanding and practical efficiency in optimization tasks, particularly in ill-conditioned scenarios.

Keywords: Nesterov's Accelerated Gradient (NAG); FISTA; Linear Convergence; Strongly Convex Functions; Lyapunov Function; Phase-Space Representation; High-Resolution ODE Framework

1 Introduction

Optimization algorithms have become a cornerstone in modern machine learning and computational mathematics. One of the most prominent challenges in this field is minimizing objective functions, particularly in the context of unconstrained optimization problems. Such problems are mathematically formulated as minimizing a function $f : \mathbb{R}^d \to \mathbb{R}$, where f is generally assumed to be convex or strongly convex. Over the last few decades, gradient-based methods have emerged as the primary tool for solving such problems due to their simplicity and computational efficiency.

Among the many gradient-based techniques, Nesterov's accelerated gradient (NAG) method stands out as one of the most effective approaches. Originally proposed in the 1980s, NAG introduced acceleration to classical gradient descent methods, significantly improving their convergence rates, especially for convex functions. In particular, NAG achieves an impressive $O(\frac{1}{k^2})$ rate for convex functions, far outpacing the $O(\frac{1}{k})$ rate of standard gradient descent.

Despite the well-documented advantages of NAG, several open questions remain, particularly in the context of strongly convex functions. A key area of inquiry is whether NAG can achieve linear convergence without prior knowledge of the modulus of strong convexity. This is of particular importance in real-world applications where such parameters are often unknown. Furthermore, the extension of NAG to composite optimization problems, where the objective function is a sum of smooth and non-smooth components, poses additional challenges.

Recent advancements in high-resolution ODE frameworks have provided new insights into the underlying mechanics of accelerated methods. By utilizing phase-space representation and Lyapunov analysis, researchers have made significant strides in understanding how these methods achieve acceleration. These techniques have been successfully applied to both smooth and composite optimization problems, yielding deeper theoretical results and improved practical algorithms.

In this paper, we build upon this body of work by establishing the linear convergence of NAG and its proximal variant, FISTA, for strongly convex functions without requiring knowledge of the strong convexity modulus. We develop a novel Lyapunov function that adapts dynamically through the iterations, ensuring a robust convergence analysis that applies to a wide range of optimization scenarios.

Formulation The unconstrained optimization problem we are studying is mathematically expressed as

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

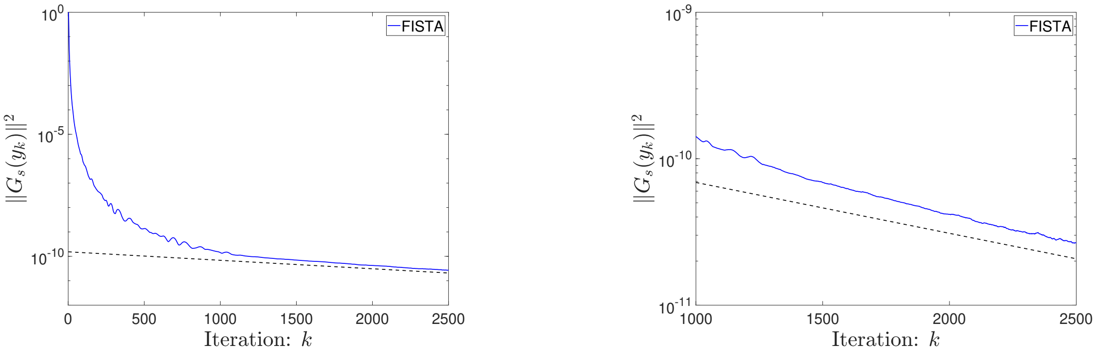
Gradient-based techniques have risen to prominence due to their computational efficiency and minimal storage requirements, , thereby becoming the method of choice for cutting-edge progress in the field.

When we look back at the historical progression of gradient-based methods, a notable landmark that stands out is *Nesterov's accelerated gradient descent* (`NAG`) method. Starting with any initial point $y_0 = x_0 \in \mathbb{R}^d$, `NAG` proceeds through the following iterative scheme:

$$\begin{cases} x_k = y_{k-1} - s\nabla f(y_{k-1}) \\ y_k = x_k + \dfrac{k-1}{k+\lambda^{\ddagger}}(x_k - x_{k-1}) \end{cases}$$

where $s > 0$ denotes the fixed step size. Originally proposed in [Nes83], `NAG` is distinguished by its superior convergence performance, particularly regarding convex functions, where it delivers an accelerated convergence rate of $O(\frac{1}{k^2})$ that represents a substantial improvement over the $O(\frac{1}{k})$ rate observed in classical gradient descent methods. The underlying mechanism behind the acceleration phenomenon was elucidated with the advent of the high-resolution ordinary differential equation (ODE) framework, as proposed in [SDJS22]. This innovative framework employs phase-space representation in conjunction with Lyapunov analysis to decode the enhanced convergence rates applicable to both the function value and the square of the gradient norm. Subsequent refinements, such as those described in [CSY22a], simplify this theoretical framework into a concise step while offering an alternative verification through the concept of implicit-velocity phase-space representation. Furthermore, the discovery of a refined proximal inequality, arising from a key observation, paved the way for generalizing the high-resolution ODE framework to encompass composite optimization problems. These advancements have ultimately led to the development of the *fast iterative shrinkage-thresholding algorithm* (`FISTA`), expounded in [LSY22b]. Additionally, with a small modification, the accelerated convergence rate for both `NAG` and `FISTA` has been applied to the iterates, as further explored in [CD15].

In practical scenarios involving convex objective functions, it is crucial to recognize that their associated Hessian matrices are devoid of zero eigenvalues. More commonly, the spectrum of these matrices is characterized by a small ratio between the smallest and largest eigenvalues, a condition that leads to the functions being labeled as "ill-conditioned" within the lexicon of the optimization community. To be explicit, for optimal outcomes, the objective function should manifest strong convexity rather than mere convexity in general. A variant of `NAG`, proposed in [Nes98], is designed to fast-track convergence for strongly convex functions. It is noteworthy, however, that this iteration is contingent upon the advanced estimation of certain parameters, a requirement that may impede



(a) Iteration: Start From $k = 0$

(b) Iteration: Start From $k = 1000$

Figure 1. Iterative progression of the square of the proximal subgradient norm throughout the application of `FISTA` for image deblurring, as demonstrated in [LSY22b].

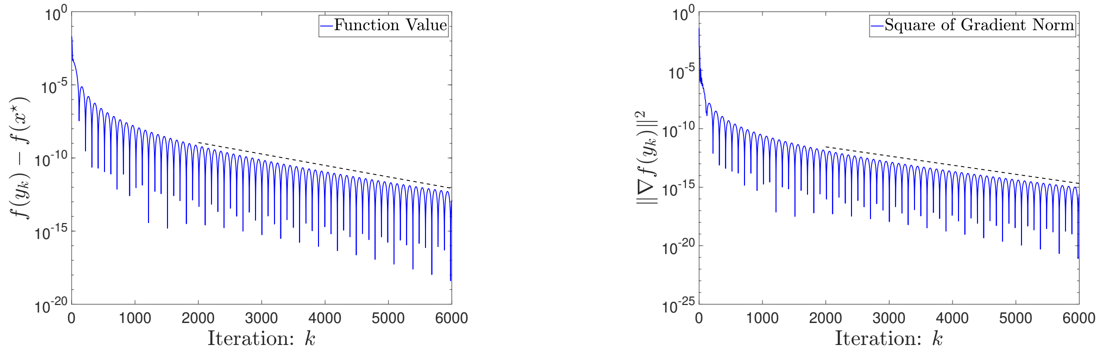
its practical utility in real-world contexts. In stark contrast, the original `NAG` algorithm operates independently of any foreknowledge regarding the modulus of strong convexity, prompting inquiries into its specific convergence rate when applied to strongly convex functions. These inquiries equally pertain to `FISTA`, the proximal generalization of `NAG`. Whether these algorithms, `NAG` and `FISTA`, can achieve linear convergence for strongly convex functions is currently recognized as an open question, as outlined in [CP16, Appendix B]. For a more concrete understanding of the convergence rates, we reference the case study presented in [LSY22b], where `FISTA` is utilized to deblur an image featuring an elephant. As a starting point for our analysis, we scrutinize the numerical pattern exhibited in Figure 1, which delineates how the square of the proximal subgradient norm varies across successive iterations.

1.1 Overview of contribution

In this study, we make use of the high-resolution ODE framework, as previously established in a series of studies [SDJS22, SDSJ19, CSY22a, CSY22b, CSY23, LSY22a, LSY22b], to address the open question posed in [CP16, Appendix B]. Our primary approach revolves around leveraging Lyapunov analysis. The main contributions of our research are outlined below.

- We establish the linear convergence of `NAG` for (smooth) strongly convex functions by formulating an innovative Lyapunov function. What sets our approach apart from previous methods is the incorporation of a dynamically adapting coefficient for the kinetic energy that evolves throughout each iteration, a distinctive feature absent in preceding methods. Importantly, our results indicate that this achieved linear convergence does not depend on the parameter $\lambda^\ddagger$, distinguishing it from the patterns observed in convex scenarios.

- Furthermore, we refine a key inequality associated with strong convexity to encompass the proximal setting. This enhancement effectively reconciles the fields of smooth and composite optimization, mending a theoretical delineation. With the aid of the implicit-velocity phase-space representation, the inventive Lyapunov function that we devise guarantees the linear convergence of function values within `FISTA` and clearly delineates the linear convergence of the square of the proximal subgradient norm.



(a) Function Value

(b) Square of Gradient Norm

Figure 2. Iterative progression of both the function value and the square of the gradient norm throughout the application of `NAG` with a step size of $s = 1$, on the quadratic objective function $f(x_1, x_2) = 2 \times 10^{-2} x_1^2 + 5 \times 10^{-4} x_2^2$.

1.2 Intuitive analysis on a quadratic function

To enhance our intuitive understanding, we initiate our discussion with a visual representation as provided in Figure 2. This demonstrative portrayal reveals how both the function value and the square of the gradient norm converge across successive iterations of `NAG` when employed on an ill-conditioned quadratic objective function. This figure effectively highlights the central attribute of linear convergence as it pertains to both the function value and the square of the gradient norm.

For a quadratic function, its Hessian is invariant with respect to the position x, leading to consistent eigendirections that are mutually orthogonal. For the sake of simplicity, let us consider the objective function to be one-dimensional, taking the general form $f(x) = \frac{1}{2}\mu x^2$, where $\mu > 0$. By incorporating this general form into `NAG`, the iteration scheme is postulated as

$$\begin{cases} x_k = y_{k-1} - \mu s y_{k-1} \\ y_k = x_k + \dfrac{k-1}{k+\lambda^\ddagger}(x_k - x_{k-1}) \end{cases}$$

By performing some elementary operations of linear transformations, we can recast the iteration in vector form as

$$\begin{pmatrix} x_k \\ y_k \end{pmatrix} = \begin{pmatrix} 0 & 1-\mu s \\ -\frac{k-1}{k+\lambda^\ddagger} & \frac{2k+\lambda^\ddagger-1}{k+\lambda^\ddagger}\cdot(1-\mu s) \end{pmatrix} \begin{pmatrix} x_{k-1} \\ y_{k-1} \end{pmatrix}$$

where the eigenvalues of the iterative matrix adhere to the characteristic quadratic polynomial equation:

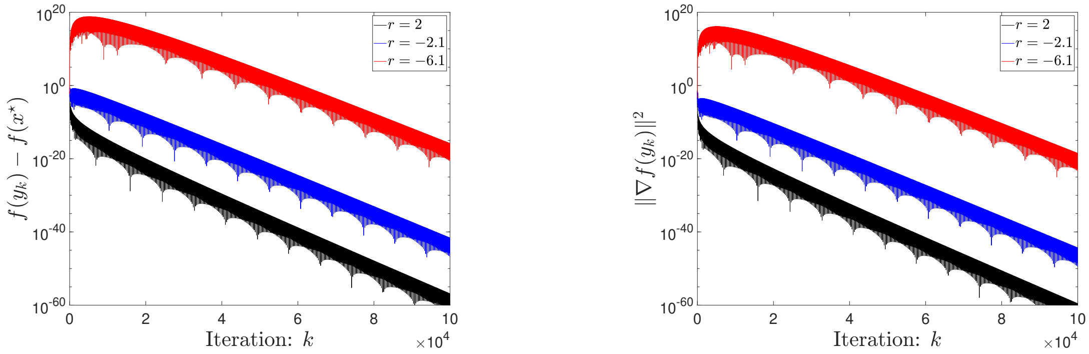
$$\overline{\lambda}^2 - \frac{2k+\lambda^\ddagger-1}{k+\lambda^\ddagger}(1-\mu s)\cdot\overline{\lambda} + \frac{k-1}{k+\lambda^\ddagger}(1-\mu s) = 0$$

As the iterations progress proliferate infinity ($k \to \infty$), the discriminant of this quadratic equation asymptotocally converges to

$$\left(\frac{2k+\lambda^\ddagger-1}{k+\lambda^\ddagger}\right)^2 (1-\mu s)^2 - 4\left(\frac{k-1}{k+\lambda^\ddagger}\right)(1-\mu s) \to -4\mu s(1-\mu s)$$

Given $\mu s \in (0,1)$, the asymptotic expressions for the roots of this quadratic equation become:

$$\overline{\lambda}_{1,2} \to 1 - \mu s \pm \sqrt{\mu s(1-\mu s)}i \qquad \text{as} \quad k \to \infty$$



(a) Function Value

(b) Square of Gradient Norm

Figure 3. Illustration of linear convergence of `NAG`, demonstrated by tracking both the function value and the square of the gradient norm, as it varies with the parameter $\lambda^\ddagger$ in the same scenario as depicted in Figure 2.

This demonstrates that `NAG` consistently achieves linear convergence at an asymptotic rate of $\sqrt{1-\mu s}$, analogous to the convergence rate exhibited in classic gradient descent, as explicated in [SDSJ19, Theorem A.4].

Furthermore, it becomes apparent that the linear convergence rate of `NAG` is rarely influenced by the choice of the parameter $\lambda^\ddagger$. This observation is supported by the results of a numerical experiment in which the parameter $\lambda^\ddagger$ is assigned across three distinct values, with the outcomes depicted in Figure 3. When the parameter $\lambda^\ddagger$ is positive, it is observed that both the function value and the square of the gradient norm always exhibit convergence. In contrast, when the parameter $\lambda^\ddagger$ is negative, there is an initial uptick in both the function value and the square of the gradient norm before they shift toward eventual convergence. As demonstrated in Figure 3, after exceeding $20,000$ iterations, the linear convergence rate retains comparable characteristics across diverse settings of the parameter $\lambda^\ddagger$, regardless of whether assessing the function value or the square of the gradient norm. While distinct patterns of iteration may manifest in the early stages, a consistent rate of linear convergence is eventually achieved as the iterations progress. To preclude complications associated with a zero denominator in the momentum coefficient, which may occur when $k + \lambda^\ddagger$ equals zero, non-integer values are deliberately selected for the negative parameter $\lambda^\ddagger$ in the numerical experiment illustrated in Figure 3. Traditional numerical approaches usually begin with the initial iteration set to zero; nonetheless, to bypass the zero denominator concern, the experiment may alternatively begin at the $(-\lambda^\ddagger+1)$-th iteration, corresponding to $k = -\lambda^\ddagger + 1$. Hence, by adopting this approach, the parameter $\lambda^\ddagger$ can be retained as an integer value for negative instances as well, thereby preventing any operational issues.

1.3 Related works and organization

The history of first-order optimization algorithms, specifically focusing on acceleration, can be traced back to the inception of the Ravine method. This method, a two-step gradient strategy, was designed to outperform the classical single-step gradient descent [GT61]. The renowned `NAG` method and its variant were introduced in [Nes83] and [Nes98], catalyzing a revolution in the phenomena of global acceleration within convex optimization. The discovery of gradient correction, unveiling the underlying mechanism of acceleration, was recognized through the comparison of the `NAG` method with Polyak's heavy-ball method, within the framework of high-resolution ODEs proposed in [SDJS22]. It was further identified in [CSY22a, CSY22b] that this high-resolution ODE framework is specially tailored for NAG and its variant, elucidating how gradient correction is effectively realized through implicit velocity. Advances in the fundamental proximal inequality paved the way for the expansion of this high-resolution ODE framework to address composite optimization, applicable to both convex and strongly convex functions, as reported in [LSY22b, LSY22a]. In addition, the completion of this tailor-made framework to accommodate the underdamped case where $\lambda^{\ddagger} < 2$ was achieved in [CSY23].

Over the past decade, it has attracted considerable attention towards research utilizing ODEs with Hessian-driven damping to explore forward-backward algorithms, with early works published in [AMR12, APR14]. The derivation of a low-resolution ODE as a model for NAG came from employing techniques borrowed from numerical analysis in [SBC16]. This laid the groundwork for a multitude of research pathways, including studies on the faster convergence rate of function values [AP16], the fusion of the low-resolution ODE and Hessian-driven damping to develop continuous dynamics [APR16], the use of variational method [WWJ16], applications of Lyapunov analysis [WRJ21], and investigations into the implicit-velocity perspective [MJ19]. The unveiling of the underlying mechanisms was achieved with the advent of high-resolution numerical approximation techniques, as outlined in [SDJS22]. Utilizing phase-space representation and Lyapunov analysis, this work bridged the gap between continuous dynamics and discrete algorithms, casting light on the interplay between the two areas. Additionally, several studies have inspected the acceleration phenomenon with an eye on numerical stability, as detailed in [LC22, ZOD$^+$21]. Tthe momentum-based scheme has also been extrapolated to the domain of residual neural networks [XSJ$^+$21], further expanding its applicability.

The remainder of this paper is organized as follows. In Section 2, we set forth some basic notations and definitions to serve as preliminaries. In Section 3, we offer a proof predicated on the gradient-correction high-resolution ODE that addresses the continuous dynamics. In Section 4, we are dedicated to proving the linear convergence of NAG for strongly convex objective functions, employing the iteration-varying Lyapunov function. Finally, in Section 5, we discuss the distinctions between the proofs for the continuous ODE and the discrete algorithm and propose potential avenues for future research.

2 Preliminaries

The notations employed within this paper mostly aligns with those found in [Nes98, SDJS22, LSY22a], with slight modifications tailored to our context. Let $\mathcal{F}^0(\mathbb{R}^d)$ denote the class of continuous convex functions on $\mathbb{R}^d$; that is, $g \in \mathcal{F}^0$ if it fulfills the inequality

$$g\left(\alpha x + (1-\alpha)y\right) \leq \alpha g(x) + (1-\alpha)g(y)$$

for any $x, y \in \mathbb{R}^d$ and $\alpha \in [0,1]$. The subclass $\mathcal{F}_L^1(\mathbb{R}^d) \subseteq \mathcal{F}^0(\mathbb{R}^d)$ consists of functions whose gradients are well-defined everywhere and adheres to the global Lipschitz condition. Thus, $f \in \mathcal{F}_L^1$ if $f \in \mathcal{F}^0$ and it satisfies

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

for any $x, y \in \mathbb{R}^d$.[1] We also denote $\mathcal{S}_{\mu,L}^1(\mathbb{R}^d)$ as the subclass of $\mathcal{F}_L^1(\mathbb{R}^d)$ with each member being μ-strongly convex for some $0 < \mu \leq L$. In other words, $f \in \mathcal{S}_{\mu,L}^1(\mathbb{R}^d)$ if $f \in \mathcal{F}_L^1(\mathbb{R}^d)$ and it holds

[1] Throughout this paper, the notation $\|\cdot\|$ specifically refers to the ℓ_2-norm or Euclidean norm, denoted as $\|\cdot\|_2$. It is worth noting that the subscript 2 is often omitted for convenience unless otherwise noted.

that

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

for any $x, y \in \mathbb{R}^d$. Furthermore, $\mathcal{S}^2_{\mu,L}(\mathbb{R}^d)$ refers to a subclass of $\mathcal{S}^1_{\mu,L}(\mathbb{R}^d)$, which encompasses functions possessing a continuous Hessian. Finally, the term $x^\star$ is used to denote its unique minimizer of those functions.

Additionally, we consider a composite function $\Phi = f + g$, where $f \in \mathcal{S}^1_{\mu,L}$ and $g \in \mathcal{F}^0$. This is analogous to [BT09, SBC16], where the s-proximal operator and the s-proximal subgradient operator are defined as follows.

Definition 1. *Let the step size satisfy $s \in (0, \frac{1}{L})$. For any $f \in \mathcal{S}^1_{\mu,L}$ and $g \in \mathcal{F}^0$, the s-proximal operator is defined as*

$$P_s(x) := \arg\min_{y \in \mathbb{R}^d} \left\{ \frac{1}{2s} \left\| y - (x - s\nabla f(x)) \right\|^2 + g(y) \right\} \tag{1}$$

for any $x \in \mathbb{R}^d$. Furthermore, the s-proximal subgradient operator is defined as

$$G_s(x) := \frac{x - P_s(x)}{s} \tag{2}$$

for any $x \in \mathbb{R}^d$.

When g simplifies to the ℓ_1-norm, i.e., $g(x) = \overline{\lambda}\|x\|_1$,[2] we can derive the closed-form silution for the s-proximal operator (1) at any $x \in \mathbb{R}^d$ for the particular instance as

$$P_s(x)_i = \big(|(x - s\nabla f(x))_i| - \overline{\lambda}s \big)_+ \mathrm{sgn}\big((x - s\nabla f(x))_i \big)$$

where $i = 1, \ldots, d$ represents the index.

3 The gradient-correction high-resolution ODE

In this section, we delve into the area of continuous convergence rates. Our primary focus lies on the gradient-correction high-resolution ODE, which is articulated in [SDJS22]. Serving as a continuous analog of `NAG`, the gradient-correction high-resolution ODE is expressed as:

$$\ddot{X} + \frac{3}{t}\dot{X} + \sqrt{s}\nabla^2 f(X)\dot{X} + \left(1 + \frac{3\sqrt{s}}{2t}\right)\nabla f(X) = 0 \tag{3}$$

with any initial $(X(0), \dot{X}(0)) = (x_0, v_0) \in \mathbb{R}^d \times \mathbb{R}^d$. Given that $f \in \mathcal{S}^2_{\mu,L}$, it is certain that the eigenvalue of the Hessian is always greater than or equal to μ, which is denoted as $\overline{\lambda}(\nabla^2 f(x)) \geq \mu$. This observation suggests that the damping term remains consistently substantial, thereby hinting at the potential for linear convergence. To determine the exact convergence rate, we turn to the principled approach of constructing a Lyapunov function, as demonstrated in [CSY22b].

(I) We initiate our analysis with the mixed energy, inspired by the convex case mentioned in [SDJS22, (4.36)], expressed as

$$\mathcal{E}_{\mathbf{mix}} = \frac{1}{2}\|t\dot{X} + 2(X - x^\star) + t\sqrt{s}\nabla f(X)\|^2 \tag{4}$$

[2] The ℓ_1 norm, denoted as $\|\cdot\|_1$, is defined as $\|x\|_1 = \sum_{i=1}^d |x_i|$ for any $x \in \mathbb{R}^d$.

By using the gradient-correction high-resolution ODE (3), we can calculate its time derivative as

$$\begin{aligned}\frac{d\mathcal{E}_{\mathbf{mix}}}{dt} &= \left\langle t\dot{X} + 2(X - x^\star) + t\sqrt{s}\nabla f(X), -\left(t + \frac{\sqrt{s}}{2}\right)\nabla f(X)\right\rangle \\ &= \underbrace{-t\left(t + \frac{\sqrt{s}}{2}\right)\langle \dot{X}, \nabla f(X)\rangle}_{\mathbf{I}} - (2t + \sqrt{s})\big\langle \nabla f(X), X - x^\star\big\rangle \\ &\quad -\sqrt{s}t\left(t + \frac{\sqrt{s}}{2}\right)\|\nabla f(X)\|^2 \end{aligned} \tag{5}$$

(II) Adhering to the principled approach of constructing a Lyapunov function as described in [CSY22b], we consider the kinetic energy. Different from [SDJS22, (2.4)], here the kinetic energy includes a time-varying coefficient as

$$\mathcal{E}_{\mathbf{kin}} = \frac{\tau(t)}{2}\|\dot{X}\|^2 \tag{6}$$

Taking into account the gradient-correction high-resolution ODE (3), we can calculate the time derivative of the kinetic energy as

$$\begin{aligned}\frac{d\mathcal{E}_{\mathbf{kin}}}{dt} &= \tau(t)\big\langle \ddot{X}, \dot{X}\big\rangle + \frac{\dot{\tau}(t)}{2}\|\dot{X}\|^2 \\ &= -\left(\frac{3\tau(t)}{t} - \frac{\dot{\tau}(t)}{2}\right)\|\dot{X}\|^2 - \sqrt{s}\tau(t)\dot{X}^\top\nabla^2 f(X)\dot{X} \\ &\quad \underbrace{-\tau(t)\left(1 + \frac{3\sqrt{s}}{2t}\right)\langle \dot{X}, \nabla f(X)\rangle}_{\mathbf{II}} \end{aligned} \tag{7}$$

To simplify the coefficient of $\|\dot{X}\|^2$, a good choice is to let $\tau(t)$ to be a power of t, such as $\tau(t) = t^\alpha$. In order to combine **I** and **II**, it is appropriate for us to choose $\alpha = 2$.

(III) Following the principled approach, to eliminate the term involving $\big\langle \dot{X}, \nabla f(X)\big\rangle$, the coefficient of the potential energy should be set accordingly. Specifically, amalgamating terms **I** and **II** yields the expression $\mathbf{I} + \mathbf{II} = -2t(t + \sqrt{s})\big\langle \dot{X}, \nabla f(X)\big\rangle$. Consequently, the potential energy is constructed as

$$\mathcal{E}_{\mathbf{pot}} = 2t(t + \sqrt{s})\left(f(X) - f(x^\star)\right) \tag{8}$$

Proceeding with the exploration of the gradient-correction high-resolution ODE (3), we calculate its time derivative as follows:

$$\frac{d\mathcal{E}_{\mathbf{pot}}}{dt} = \left(4t + 2\sqrt{s}\right)\left(f(X) - f(x^\star)\right) + \underbrace{2t(t + \sqrt{s})\langle \dot{X}, \nabla f(X)\rangle}_{\mathbf{III}} \tag{9}$$

where the resultant term **III** in (9) ensures that the combination of **I**, **II**, and **III** satisfies the equality $\mathbf{I} + \mathbf{II} + \mathbf{III} = 0$.

By integrating the mixed energy (4), the kinetic energy (6) and the potential energy (8), we arrive at the following Lyapunov function as

$$\mathcal{E} = 2t(t + \sqrt{s})\left(f(X) - f(x^\star)\right) + \frac{t^2}{2}\|\dot{X}\|^2 + \frac{1}{2}\|t\dot{X} + 2(X - x^\star) + t\sqrt{s}\nabla f(X)\|^2 \tag{10}$$

Equipped with the Lyapunov function (10), we can conclude this section with the following theorem pertaining to the convergence rates of both the function value and the square of the gradient norm, along with its proof.

Theorem 1. *Let $f \in \mathcal{S}^2_{\mu,L}$. For any step size $0 < s < \frac{1}{L}$, there exists some time $T = T(\mu, s) > 0$ such that the solution $X = X(t)$ to the gradient-correction high-resolution ODE* (3) *satisfies*

$$\begin{cases} f(X) - f(x^\star) \le \dfrac{\mathcal{E}(T)}{2t(t+\sqrt{s})} e^{-\frac{\mu\sqrt{s}}{4}(t-T)} \\ \|\nabla f(X)\|^2 \le \dfrac{L\mathcal{E}(T)}{t(t+\sqrt{s})} e^{-\frac{\mu\sqrt{s}}{4}(t-T)} \end{cases} \tag{11}$$

for any $t \ge T = \frac{4}{\mu\sqrt{s}}$.

Proof of Theorem 1. To prove the theorem, we begin with the Lyapunov function (10). Its time derivative can be calculated by summing up (5), (7) and (9) as

$$\begin{aligned} \frac{d\mathcal{E}}{dt} \le & \left(4t + 2\sqrt{s}\right)\left(f(X) - f(x^\star)\right) - (2t + \sqrt{s})\left\langle \nabla f(X), X - x^\star \right\rangle \\ & - 2t\|\dot{X}\|^2 - \sqrt{s}t^2 \dot{X}^\top \nabla^2 f(X) \dot{X} - \sqrt{s}t\left(t + \frac{\sqrt{s}}{2}\right)\|\nabla f(X)\|^2 \end{aligned} \tag{12}$$

For any $f \in \mathcal{S}^2_{\mu,L}$, it satisfies the μ-strongly convex inequality as

$$\left\langle \nabla f(X), X - x^\star \right\rangle \ge f(X) - f(x^\star) + \frac{\mu}{2}\left\|X - x^\star\right\|^2 \tag{13}$$

By substituting the μ-strongly convex inequality (13) into the earlier time derivative (12), we obtain

$$\begin{aligned} \frac{d\mathcal{E}}{dt} \le & \left(2t + \sqrt{s}\right)\left(f(X) - f(x^\star)\right) - \frac{\mu\left(2t + \sqrt{s}\right)}{2}\|X - x^\star\|^2 \\ & - 2t\|\dot{X}\|^2 - \sqrt{s}t^2 \dot{X}^\top \nabla^2 f(X) \dot{X} - \sqrt{s}t\left(t + \frac{\sqrt{s}}{2}\right)\|\nabla f(X)\|^2 \end{aligned} \tag{14}$$

To establish proportionality between the corresponding terms, we can estimate the Lyapunov function using Cauchy-Schwarz inequality as

$$\mathcal{E} \le 2t\left(t + \sqrt{s}\right)\left(f(X) - f(x^\star)\right) + 2t^2\|\dot{X}\|^2 + 6\left\|X - x^\star\right\|^2 + \frac{3}{2}st^2\left\|\nabla f(X)\right\|^2 \tag{15}$$

For any $f \in \mathcal{S}^2_{\mu,L}$, we have two additional μ-strongly convex inequalities as:

$$\dot{X}^\top \nabla^2 f(X) \dot{X} \ge \mu\|\dot{X}\|^2 \tag{16a}$$

$$\|\nabla f(X)\|^2 \ge 2\mu\left(f(X) - f(x^\star)\right) \tag{16b}$$

By substituting (16a) and (16b) into (14), we can obtain the time derivative as

$$\begin{aligned} \frac{d\mathcal{E}}{dt} \le & -\mu\sqrt{s}\left(t - \frac{1}{\mu\sqrt{s}}\right)\left(t + \sqrt{s}\right)\left(f(X) - f(x^\star)\right) - \mu\sqrt{s}t^2\|\dot{X}\|^2 \\ & -\mu t\left\|X - x^\star\right\|^2 - \frac{\sqrt{s}t}{2}\left(t - \frac{1}{\mu\sqrt{s}}\right)\left\|\nabla f(X)\right\|^2 \\ \le & -\frac{3\mu t\left(\sqrt{s}t + s\right)}{4}\left(f(X) - f(x^\star)\right) - \mu\sqrt{s}t^2\|\dot{X}\|^2 \\ & -\frac{4}{\sqrt{s}}\left\|X - x^\star\right\|^2 - \frac{3\sqrt{s}t^2}{8}\left\|\nabla f(X)\right\|^2 \end{aligned} \tag{17}$$

where the last inequality is supported by $t \geq T = \frac{4}{\mu\sqrt{s}}$. By matching the corresponding terms in (15) and (17), we can estimate the time derivative as

$$\frac{d\mathcal{E}}{dt} \leq -\min\left\{\frac{3\mu\sqrt{s}}{8}, \frac{\mu\sqrt{s}}{2}, \frac{2}{3\sqrt{s}}, \frac{1}{4\sqrt{s}}\right\}\mathcal{E} \leq -\frac{\mu\sqrt{s}}{4}\mathcal{E}$$

with the latter inequality following because $\mu \leq L$ and $s \in (0, \frac{1}{L})$. Additionally, owing to the condition $f \in \mathcal{S}^2_{\mu,L}$, it holds that:

$$\|\nabla f(X)\|^2 \leq 2L\left(f(X) - f(x^\star)\right)$$

Hence, the proof is complete with some elementary calculations.

□

4 Linear convergence via a novel discrete Lyapunov function

In this section, we develop a novel discrete Lyapunov function aimed at deducing the linear convergence properties of the NAG method when implemented on μ-strongly convex functions. Our approach is fundamentally anchored in the theoretical framework of potential and mixed energy, as thoroughly detailed in [LSY22b, (5.14)] for convex scenarios. The primary distinction of our method lies in the adoption of an iteration-varying coefficient in the construction of kinetic energy, which ensures that the respective terms remain proportional. This critical alteration marks a significant departure from the techniques described in [LSY22a, (5.3)].

4.1 Smooth optimization with NAG

In the arena of smooth optimization, we leverage the velocity iteration sequence defined by $v_k = \frac{x_k - x_{k-1}}{\sqrt{s}}$, which allows us to recast NAG into the implicit-velocity scheme characterized by a phase-space representation:

$$\begin{cases} x_{k+1} - x_k = \sqrt{s}v_{k+1} \\ v_{k+1} - v_k = -\dfrac{\lambda^\ddagger + 1}{k + \lambda^\ddagger}\cdot v_k - \sqrt{s}\nabla f(y_k) \end{cases} \tag{18}$$

where the sequence $\{y_k\}_{k=0}^\infty$ complies with the relation:

$$y_k = x_k + \frac{k-1}{k+\lambda^\ddagger}\cdot\sqrt{s}v_k$$

It is evident that the second iteration of (18) can be neatly reformulated as

$$(k + \lambda^\ddagger)v_{k+1} - (k-1)v_k = -(k+\lambda^\ddagger)\sqrt{s}\nabla f(y_k) \tag{19}$$

Building upon this framework, we now proceed to elucidate how the Lyapunov function is systematically constructed by employing the principled approach as outlined in [CSY22b].

(**I**) Drawing upon the discrete Lyapunov function established in [LSY22b, (5.14)], we here choose to implement the analogous mixed energy for its computational efficacy as referenced in [CSY22a, Section 4.2]. The mixed energy is delineated as:

$$\mathcal{E}_{\mathbf{mix}}(k) = \frac{1}{2}\|\sqrt{s}(k-1)v_k + \lambda^\ddagger(x_k - x^\star)\|^2 \tag{20}$$

By employing the initial step of the phase-space representation (18) and the reformulated expression provided in (19), we establish the subsequent equality as

$$\begin{aligned}
&\sqrt{s}kv_{k+1} + \lambda^{\ddagger}(x_{k+1} - x^{\star}) - \sqrt{s}(k-1)v_k + \lambda^{\ddagger}(x_k - x^{\star}) \\
=&\sqrt{s}kv_{k+1} + \lambda^{\ddagger}(x_{k+1} - x_k) - \sqrt{s}(k-1)v_k \\
=&\sqrt{s}(k+\lambda^{\ddagger})v_{k+1} - \sqrt{s}(k-1)v_k \\
=& - (k+\lambda^{\ddagger})s\nabla f(y_k)
\end{aligned} \tag{21}$$

where the second equality is derived from the initial step of the phase-space representation (18) and the last equality follows from equation (19). We are now in a position to determine the iterative difference of mixed energy, which is calculated as follows:

$$\begin{aligned}
&\mathcal{E}_{\mathbf{mix}}(k+1) - \mathcal{E}_{\mathbf{mix}}(k) \\
&\quad = \left\langle \sqrt{s}kv_{k+1} + \lambda^{\ddagger}(x_{k+1} - x^{\star}), -s(k+\lambda^{\ddagger})\nabla f(y_k) \right\rangle \\
&\quad\quad - \frac{s^2(k+\lambda^{\ddagger})^2}{2} \left\|\nabla f(y_k)\right\|^2 \\
&\quad = - \underbrace{s^{\frac{3}{2}}k(k+\lambda^{\ddagger})\left\langle \nabla f(y_k), v_{k+1} \right\rangle}_{\mathbf{I}} - s\lambda^{\ddagger}(k+\lambda^{\ddagger})\left\langle \nabla f(y_k), y_k - x^{\star} \right\rangle \\
&\quad\quad - \frac{s^2}{2}(k-\lambda^{\ddagger})(k+\lambda^{\ddagger}) \left\|\nabla f(y_k)\right\|^2
\end{aligned} \tag{22}$$

where the ultimate step incorporates the gradient step inherent in the `NAG` method.

(II) In alignment with the principled approach to constructing the Lyapunov function as outlined in [CSY22b], we commence our analysis by considering the kinetic energy. Departing from [CSY22b, (4.10)], we formulate the kinetic energy with an iteration-varying coefficient, which is articulated as follows:

$$\mathcal{E}_{\mathbf{kin}}(k) = \frac{s\tau(k)}{2}\|v_k\|^2 \tag{23}$$

By employing the expression from (19), the iterative difference of kinetic energy can be calculated as follows:

$$\begin{aligned}
&\mathcal{E}_{\mathbf{kin}}(k+1) - \mathcal{E}_{\mathbf{kin}}(k) \\
&\quad = \frac{s\tau(k+1)}{2}\left\|v_{k+1}\right\|^2 - \frac{s\tau(k)}{2}\left(\frac{k+\lambda^{\ddagger}}{k-1}\right)^2 \left\|v_{k+1} + \sqrt{s}\nabla f(y_k)\right\|^2 \\
&\quad = \frac{s}{2}\left[\tau(k+1) - \tau(k)\left(\frac{k+\lambda^{\ddagger}}{k-1}\right)^2\right]\left\|v_{k+1}\right\|^2 \\
&\quad\quad - s^{\frac{3}{2}}\tau(k)\left(\frac{k+\lambda^{\ddagger}}{k-1}\right)^2 \left\langle \nabla f(y_k), v_{k+1}\right\rangle \\
&\quad\quad - \frac{s^2\tau(k)}{2}\left(\frac{k+\lambda^{\ddagger}}{k-1}\right)^2 \left\|\nabla f(y_k)\right\|^2
\end{aligned} \tag{24}$$

To ease computation as indicated by (24), an intuitive choice for $\tau(k)$ would be $\tau(k) = (k-1)^2$.

This selection simplifies the iterative difference (24), which can be succinctly expressed as

$$
\begin{aligned}
\mathcal{E}_{\mathbf{kin}}(k+1) - \mathcal{E}_{\mathbf{kin}}(k) = & -\frac{s\lambda^{\ddagger}(2k+\lambda^{\ddagger})}{2}\|v_{k+1}\|^2 - \frac{s^2(k+\lambda^{\ddagger})^2}{2}\|\nabla f(y_k)\|^2 \\
& - \underbrace{s^{\frac{3}{2}}(k+\lambda^{\ddagger})^2\big\langle \nabla f(y_k), v_{k+1}\big\rangle}_{\mathbf{II}} \qquad (25)
\end{aligned}
$$

(III) We now turn our attention to the potential energy. It is appropriate for the potential energy to also feature an iteration-varying coefficient, hereafter referred to as $\gamma(k)$. Consequently, we define the potential energy as:

$$
\mathcal{E}_{\mathbf{pot}}(k) = s\gamma(k)\left(f(x_k) - f(x^\star)\right) \qquad (26)
$$

To determine the iterative difference in potential energy, we must refer back to the inequality for $f \in \mathcal{S}^1_{\mu,L}$, as demonstrated in [Nes98]:

$$
\begin{aligned}
& f(y - s\nabla f(y)) - f(x) \\
& \qquad \leq \big\langle \nabla f(y), y - x\big\rangle - \frac{\mu}{2}\|y-x\|^2 - \left(s - \frac{Ls^2}{2}\right)\|\nabla f(y)\|^2 \qquad (27)
\end{aligned}
$$

Applying (27) with x_{k+1} and x_k, we derive:

$$
\begin{aligned}
& f(x_{k+1}) - f(x_k) \\
& \quad \leq \big\langle \nabla f(y_k), y_k - x_k\big\rangle - \frac{\mu}{2}\|y_k - x_k\|^2 - \left(s - \frac{s^2L}{2}\right)\|\nabla f(y_k)\|^2 \\
& \quad = \big\langle \nabla f(y_k), \sqrt{s}v_{k+1} + s\nabla f(y_k)\big\rangle - \frac{\mu}{2}\left\|\sqrt{s}v_{k+1} + s\nabla f(y_k)\right\|^2 \\
& \qquad - \left(s - \frac{s^2L}{2}\right)\|\nabla f(y_k)\|^2 \\
& \quad = \sqrt{s}(1-\mu s)\big\langle \nabla f(y_k), v_{k+1}\big\rangle + \frac{s^2(L-\mu)}{2}\|\nabla f(y_k)\|^2 - \frac{\mu s}{2}\|v_{k+1}\|^2 \qquad (28)
\end{aligned}
$$

where the first equality is justified by the initial iteration in (18) in conjunction with the gradient step of `NAG`. Subsequently, for the potential energy (26), the iterative difference can be estimated as follows:

$$
\begin{aligned}
& \mathcal{E}_{\mathbf{pot}}(k+1) - \mathcal{E}_{\mathbf{pot}}(k) \\
& = s\gamma(k+1)\left(f(x_{k+1}) - f(x^\star)\right) - s\gamma(k)\left(f(x_k) - f(x^\star)\right) \\
& = s\gamma(k)\left(f(x_{k+1}) - f(x_k)\right) + s(\gamma(k+1) - \gamma(k))\left(f(x_{k+1}) - f(x^\star)\right) \\
& \leq s\gamma(k)\left(\underbrace{\sqrt{s}(1-\mu s)\big\langle \nabla f(y_k), v_{k+1}\big\rangle}_{\mathbf{III}} + \frac{s^2(L-\mu)}{2}\|\nabla f(y_k)\|^2 - \frac{\mu s}{2}\|v_{k+1}\|^2\right) \\
& \qquad + s(\gamma(k+1) - \gamma(k))\left(f(x_{k+1}) - f(x^\star)\right) \qquad (29)
\end{aligned}
$$

Coupled with the earlier derivations, (22) and (25), it becomes imperative to fine-tune the coefficient $\gamma(k)$ within the iterative difference (29) to cancel out the term $\big\langle \nabla f(y_k), v_{k+1}\big\rangle$. Explicitly, this necessitates the fulfillment of the relation:

$$
s\gamma(k)\cdot\mathbf{III} - \mathbf{I} - \mathbf{II} = 0
$$

Therefore, the iteration-varying coefficient $\gamma(k)$ should be determined as follows:

$$\gamma(k) = \frac{(k+\lambda^{\ddagger})(2k+\lambda^{\ddagger})}{1-\mu s}$$

Integrating the mixed energy (20), the kinetic energy (23) and the potential energy (26) — each bolstered by iteration-varying coefficients — we construct the noval discrete Lyapunov function as depicted below:

$$\begin{aligned}\mathcal{E}(k) = & \frac{s(k+\lambda^{\ddagger})(2k+\lambda^{\ddagger})}{1-\mu s}\left(f(x_k)-f(x^{\star})\right) \\ & + \frac{s(k-1)^2}{2}\|v_k\|^2 + \frac{1}{2}\left\|\sqrt{s}(k-1)v_k + \lambda^{\ddagger}(x_k - x^{\star})\right\|^2 \end{aligned} \tag{30}$$

Within this novel Lyapunov function (30) at our disposal, we methodically infer the convergence rates for both the function value and the squared gradient norm, as stated in the forthcoming theorem.

Theorem 2. *Let $f \in \mathcal{S}^1_{\mu,L}$. Given any step size $0 < s < \frac{1}{L}$, there exists a positive integer $K := K(L,\mu,s,\lambda^{\ddagger})$ such that the iterative sequence $\{(x_k,y_k)\}_{k=0}^{\infty}$ generated by* `NAG` *with any initial $x_0 = y_0 \in \mathbb{R}^d$ satisfies*

$$\begin{cases} f(x_k) - f(x^{\star}) \leq \dfrac{\mathcal{E}(K)}{s(k+\lambda^{\ddagger})(2k+\lambda^{\ddagger})\left[1+(1-Ls)\cdot\dfrac{\mu s}{4}\right]^{k-K}} \\ \|\nabla f(y_k)\|^2 \leq \dfrac{4\mathcal{E}(K)}{s^2(1-Ls)(k+\lambda^{\ddagger})(2k+\lambda^{\ddagger})\left[1+(1-Ls)\cdot\dfrac{\mu s}{4}\right]^{k-K}} \end{cases} \tag{31}$$

for any $k \geq K$.

Remark. Considering the inverse proportionality between the step size and the Lipshitz constant, denoted as $s \sim \frac{1}{L}$, we articulate the convergence rates specified in (31) in terms of the dimensionless parameter $\alpha = sL$, which falls within the open interval $(0,1)$. Accordingly, the established bounds may be recast as

$$\begin{cases} f(x_k) - f(x^{\star}) \leq \dfrac{\mathcal{E}(K)}{(k+\lambda^{\ddagger})(2k+\lambda^{\ddagger})\left(1+\frac{\alpha(1-\alpha)}{4}\cdot\frac{\mu}{L}\right)^{k-K}} \\ \|\nabla f(y_k)\|^2 \leq \dfrac{4L^2\mathcal{E}(K)}{\alpha^2(1-\alpha)(k+\lambda^{\ddagger})(2k+\lambda^{\ddagger})\left[1+\frac{\alpha(1-\alpha)}{4}\cdot\dfrac{\mu}{L}\right]^{k-K}} \end{cases} \tag{32}$$

for any $k \geq K$. The bounds delineated in (32) indicate that both the function value and the square of the gradient norm adhere to a linear convergence pattern, characterized by a rate akin to $\left(1 + c\cdot\frac{\mu}{L}\right)^{-k}$ where the constant c resides in the range $(0,1)$. Albeit the constant c can be optimized to 1, the ensuing convergence rate of $\left(1+\frac{\mu}{L}\right)^{-k}$ still exhibits a significant gap when compared to the optimal rates of $\left(1-\sqrt{\frac{\mu}{L}}\right)^{k}$ or $\left(1+\sqrt{\frac{\mu}{L}}\right)^{-k}$, as elucidated in [Nes98]. This discrepancy accentuates the sustained pursuit in the realm of computational optimization for devising accelerated methods that circumvent any prerequisites concerning the modulus of strong convexity.

Proof of Theorem 2. Given the discrete Lyapunov function defined in (30), we calculate its iterative difference by summing the three iterative differences laid out in (22), (25) and (29). This amalgamation yields:

$$\begin{aligned}
&\mathcal{E}(k+1) - \mathcal{E}(k) \\
&\quad = - \underbrace{s\lambda^{\ddagger}(k+\lambda^{\ddagger})\big\langle \nabla f(y_k), y_k - x^{\star} \big\rangle}_{\mathbf{IV}} - \frac{s\lambda^{\ddagger}(2k+\lambda^{\ddagger})}{2} \left\| v_{k+1} \right\|^2 \\
&\qquad - s^2 k(k+\lambda^{\ddagger}) \left\| \nabla f(y_k) \right\|^2 \\
&\qquad + \frac{s(k+\lambda^{\ddagger})(2k+\lambda^{\ddagger})}{1-\mu s} \left(\frac{s^2(L-\mu)}{2} \left\| \nabla f(y_k) \right\|^2 - \frac{\mu s}{2} \left\| v_{k+1} \right\|^2 \right) \\
&\qquad + \frac{s(4k+3\lambda^{\ddagger}+2)}{1-\mu s} \left(f(x_{k+1}) - f(x^{\star}) \right)
\end{aligned} \tag{33}$$

To ensure proportional alignment of terms, we exhibit $\|x_{k+1} - x^{\star}\|^2$ and the other terms in (33), thereby replacing Term **IV**. After inserting x_{k+1} and $x^{\star}$ into (27), we deduce:

$$\begin{aligned}
f(x_{k+1}) - f(x^{\star}) &\leq \big\langle \nabla f(y_k), y_k - x^{\star} \big\rangle - \frac{s}{2} \left\| \nabla f(y_k) \right\|^2 - \frac{\mu}{2} \left\| y_k - x^{\star} \right\|^2 \\
&\leq \big\langle \nabla f(y_k), y_k - x^{\star} \big\rangle - \frac{s}{2} \left\| \nabla f(y_k) \right\|^2 - \frac{\mu}{2} \left\| x_{k+1} - x^{\star} \right\|^2
\end{aligned} \tag{34}$$

where the penultimate step is intrinsically connected to the gradient method, such that:

$$\begin{aligned}
\left\| x_{k+1} - x^{\star} \right\|^2 &= \left\| y_k - x^{\star} - s\nabla f(y_k) \right\|^2 \\
&= \left\| y_k - x^{\star} \right\|^2 - 2s\big\langle \nabla f(y_k), y_k - x^{\star} \big\rangle + s^2 \left\| \nabla f(y_k) \right\|^2 \leq \left\| y_k - x^{\star} \right\|^2
\end{aligned}$$

where the last inequality holds for any $s < \frac{1}{L}$ due to the first inequality of (34). By substituting inequality (34) into the iterative difference (33), we attain:

$$\begin{aligned}
\mathcal{E}(k+1) - \mathcal{E}(k) \leq & s \left(\frac{4k+3\lambda^{\ddagger}+2}{1-\mu s} - r(k+\lambda^{\ddagger}) \right) \left(f(x_{k+1}) - f(x^{\star}) \right) \\
& - \frac{s}{2} \left(\frac{\mu s}{1-\mu s}(k+\lambda^{\ddagger})(2k+\lambda^{\ddagger}) + \lambda^{\ddagger}(2k+\lambda^{\ddagger}) \right) \left\| v_{k+1} \right\|^2 \\
& - \frac{\mu s}{2} \cdot r(k+\lambda^{\ddagger}) \left\| x_{k+1} - x^{\star} \right\|^2 \\
& - \frac{s^2}{2} \cdot \frac{1-Ls}{1-\mu s} \cdot (k+\lambda^{\ddagger})(2k+\lambda^{\ddagger}) \left\| \nabla f(y_k) \right\|^2
\end{aligned} \tag{35}$$

To proportionally adjust the corresponding terms within the Lyapunov function $\mathcal{E}(k+1)$, we employ the Cauchy-Schwarz inequality to estimate as follows:

$$\begin{aligned}
\mathcal{E}(k+1) \leq & \frac{s(k+\lambda^{\ddagger}+1)(2k+\lambda^{\ddagger}+2)}{1-\mu s} \left(f(x_{k+1}) - f(x^{\star}) \right) \\
& + \frac{3}{2} s k^2 \left\| v_{k+1} \right\|^2 + \lambda^{\ddagger\,2} \left\| x_{k+1} - x^{\star} \right\|^2 \\
= & \frac{s\big[(k+\lambda^{\ddagger})(2k+\lambda^{\ddagger}) + (4k+3\lambda^{\ddagger}+2)\big]}{1-\mu s} \left(f(x_{k+1}) - f(x^{\star}) \right) \\
& + \frac{3}{2} s k^2 \left\| v_{k+1} \right\|^2 + \lambda^{\ddagger\,2} \left\| x_{k+1} - x^{\star} \right\|^2
\end{aligned} \tag{36}$$

For any $f \in \mathcal{S}^1_{\mu,L}$, the following inequality is satisfied:

$$\|\nabla f(y)\|^2 \geq 2\mu \left(f(y - s\nabla f(y)) - f(x^\star)\right) \tag{37}$$

for any $y \in \mathbb{R}^d$. Applying this to y_k of `NAG` within (37), we arrive at:

$$\|\nabla f(y_k)\|^2 \geq 2\mu \left(f(x_{k+1}) - f(x^\star)\right) \tag{38}$$

Incorporating inequality (38) into the iterative difference (35), we deduce:

$$\begin{aligned}
&\mathcal{E}(k+1) - \mathcal{E}(k) \\
&\leq -\frac{s}{1-\mu s}\left(\frac{\mu s(1-Ls)}{2}(k+\lambda^\ddagger)(2k+\lambda^\ddagger) - (4k+3\lambda^\ddagger+2) + \lambda^\ddagger(1-\mu s)(k+\lambda^\ddagger)\right) \\
&\quad \cdot \left(f(x_{k+1}) - f(x^\star)\right) - \frac{\mu s}{2} r(k+\lambda^\ddagger)\left\|x_{k+1} - x^\star\right\|^2 \\
&\quad - \frac{s}{2}\left(\frac{\mu s}{1-\mu s}(k+\lambda^\ddagger)(2k+\lambda^\ddagger) + \lambda^\ddagger(2k+\lambda^\ddagger)\right)\left\|v_{k+1}\right\|^2 \\
&\quad - \frac{s^2}{4}\frac{1-Ls}{1-\mu s}(k+\lambda^\ddagger)(2k+\lambda^\ddagger)\left\|\nabla f(y_k)\right\|^2 \\
&\leq -\frac{s}{1-\mu s}\left(\frac{\mu s(1-Ls)}{2}(k+\lambda^\ddagger)(2k+\lambda^\ddagger) - (4k+3\lambda^\ddagger+2) + \lambda^\ddagger(1-\mu s)(k+\lambda^\ddagger)\right) \\
&\quad \cdot \left(f(x_{k+1}) - f(x^\star)\right) - \frac{\mu s \lambda^{\ddagger\,2}}{2}\left\|x_{k+1} - x^\star\right\|^2 \\
&\quad - \mu s^2 k^2 \left\|v_{k+1}\right\|^2 - \frac{s^2}{4}(1-Ls)(k+\lambda^\ddagger)(2k+\lambda^\ddagger)\left\|\nabla f(y_k)\right\|^2
\end{aligned} \tag{39}$$

To align the terms between the right-hand side of (39) and (36), we establish the following inequality as

$$\begin{aligned}
(1-Ls)\cdot\frac{\mu s}{4}(k+\lambda^\ddagger)(2k+\lambda^\ddagger) - (4k+3\lambda^\ddagger+2) + \lambda^\ddagger(1-\mu s)(k+\lambda^\ddagger)& \\
\geq \frac{\mu s(1-Ls)}{4}\cdot(4k+3\lambda^\ddagger+2)&
\end{aligned}$$

It is evident that there exists a positive constant $K = K(L, \mu, s, \alpha)$ for which the above inequality holds. Thus, we conclude:

$$\begin{aligned}
&\mathcal{E}(k+1) - \mathcal{E}(k) \\
&\leq -\min\left\{\frac{1-Ls}{4}, \frac{2}{3}, \frac{1}{2}\right\}\mu s \cdot \mathcal{E}(k+1) - \frac{s^2(1-Ls)(k+\lambda^\ddagger)(2k+\lambda^\ddagger)}{4}\left\|\nabla f(y_k)\right\|^2 \\
&= \frac{1-Ls}{4}\cdot\mu s\cdot\mathcal{E}(k+1) - \frac{s^2(1-Ls)(k+\lambda^\ddagger)(2k+\lambda^\ddagger)}{4}\left\|\nabla f(y_k)\right\|^2
\end{aligned}$$

The proof is complete with some elementary operations. □

4.2 Composite optimization via `FISTA`

In this section, we extend the convergence rates of `NAG` as established in Theorem 2 to include its proximal variant, `FISTA`. As specified in Definition 1, `FISTA` utilizes the s-proximal operator (1) and is defined by the following iterative scheme, starting from any initial point $y_0 = x_0 \in \mathbb{R}^d$:

$$\begin{cases} x_k = P_s\left(y_{k-1} - s\nabla f(y_{k-1})\right) \\ y_k = x_k + \dfrac{k-1}{k+\lambda^{\ddagger}}(x_k - x_{k-1}) \end{cases}$$

where $s > 0$ denotes the step size. We can analogously formulate `FISTA` in a manner akin to `NAG` using the s-proximal subgradient operator (2), which yields

$$\begin{cases} x_k = y_{k-1} - sG_s(y_{k-1}) \\ y_k = x_k + \dfrac{k-1}{k+\lambda^{\ddagger}}(x_k - x_{k-1}) \end{cases} \tag{40}$$

where the proximal operator $G_s(\cdot)$ replace the gradient operator $\nabla f(\cdot)$ in `NAG`. Furthermore, by designing the velocity iteration sequence as $v_k = \frac{x_k - x_{k-1}}{\sqrt{s}}$, we reformulate the `FISTA` updates in a `NAG`-esque fashion (40) into a implicit-velocity scheme (phase-space representation), expressed as

$$\begin{cases} x_{k+1} - x_k = \sqrt{s}v_{k+1} \\ v_{k+1} - v_k = -\dfrac{\lambda^{\ddagger}+1}{k+\lambda^{\ddagger}} \cdot v_k - \sqrt{s}G_s(y_k) \end{cases} \tag{41}$$

To derive the convergence rates, we still need to establish a discrete Lyapunov function. Here, we generalize the one previously utilized for smooth functions (30), by substituting the smooth function f with the composite function $\Phi = f + g$, resulting in

$$\begin{aligned} \mathcal{E}(k) = & \frac{s(k+\lambda^{\ddagger})(2k+\lambda^{\ddagger})}{1-\mu s}\left(\Phi(x_k) - \Phi(x^\star)\right) \\ & + \frac{s(k-1)^2}{2}\|v_k\|^2 + \frac{1}{2}\left\|\sqrt{s}(k-1)v_k + \lambda^{\ddagger}(x_k - x^\star)\right\|^2 \end{aligned} \tag{42}$$

As outlined in Section 4.1, deriving the desired convergence rates for the smooth case predominantly hinges on two key inequalities, (27) and (37). If we can successfully adjust these two key inequalities to account for the proximal setting, we shall be able to transpose the results of Theorem 2 into this broader context. For the fundamental inequality (27), its proximal analogue is articulated as

$$\begin{aligned} & \Phi(y - sG_s(y)) - \Phi(x) \\ & \qquad \leq \left\langle G_s(y), y - x \right\rangle - \frac{\mu}{2}\|y - x\|^2 - \left(s - \frac{Ls^2}{2}\right)\|G_s(y)\|^2 \end{aligned} \tag{43}$$

where the proof can be found in [LSY22a, Lemma 3.2]. The primary challenge lies in extending the μ-strongly convex inequality (37) to the composite function $\Phi = f + g$, for which we establish the subsequent lemma.

Lemma 1. *Let $f \in \mathcal{S}^1_{\mu,L}$ and $g \in \mathcal{F}^0$. It then holds that the s-proximal subgradient, as defined in* (2)*, satisfies the following inequality:*

$$\|G_s(y)\|^2 \geq 2\mu\left(\Phi(y - sG_s(y)) - \Phi(x^\star)\right) \tag{44}$$

for any $y \in \mathbb{R}^d$.

Proof of Lemma 1. Under the condition that the step size adheres to $0 < s \leq \frac{1}{L}$, we invoke the fundamental proximal inequality (43), which simplifies to:

$$\Phi(y - sG_s(y)) - \Phi(x) \leq \big\langle G_s(y), y - x \big\rangle - \frac{s}{2}\left\|G_s(y)\right\|^2 - \frac{\mu}{2}\left\|y - x\right\|^2 \tag{45}$$

for any $x, y \in \mathbb{R}^d$. By reorganizing the terms of (45), we arrive at the expression

$$\Phi(x) \geq \Phi(y - sG_s(y)) + \big\langle G_s(y), x - y \big\rangle + \frac{s}{2}\left\|G_s(y)\right\|^2 + \frac{\mu}{2}\left\|y - x\right\|^2 \tag{46}$$

for any $x, y \in \mathbb{R}^d$. For succinctness, we denote the right-hand side of (46) as

$$h(x) = \Phi(y - sG_s(y)) + \big\langle G_s(y), x - y \big\rangle + \frac{s}{2}\left\|G_s(y)\right\|^2 + \frac{\mu}{2}\left\|y - x\right\|^2 \tag{47}$$

for any fixed $y \in \mathbb{R}^d$. Consequently, inequality (46) takes the form $\Phi(x) \geq h(x)$. Considering that $x^\star$ is the unique minimizer of the composite function Φ, substituting x with $x^\star$ yields:

$$\Phi(x^\star) \geq h(x^\star) \tag{48}$$

Upon evaluating expression (47), we identify that $h(x)$ embodies a quadratic function whose Hessian is positive definte. This is depicted as:

$$h(x) = \frac{\mu}{2}\left\|x - y + \frac{1}{\mu}G_s(y)\right\|^2 + \Phi(y - sG_s(y)) - \left(\frac{1}{2\mu} - \frac{s}{2}\right)\left\|G_s(y)\right\|^2 \tag{49}$$

Incorporating (49) into (48) results in

$$\Phi(x^\star) \geq \Phi(y - sG_s(y)) - \left(\frac{1}{2\mu} - \frac{s}{2}\right)\left\|G_s(y)\right\|^2 \geq \Phi(y - sG_s(y)) - \frac{1}{2\mu}\left\|G_s(y)\right\|^2 \tag{50}$$

By rearranging (50), we complete the proof.

□

With the proximal generalization of the μ-strongly convex inequality, as demonstrated in Lemma 1, we now present the subsequent theorem, which characterizes the convergence rates of `FISTA`.

Theorem 3. *Let $f \in \mathcal{S}^1_{\mu,L}$ and $g \in \mathcal{F}^0$. For any step size $0 < s < \frac{1}{L}$, there exists some positive integer $K := K(L, \mu, s, \lambda^\ddagger)$, such that the iterative sequence $\{(x_k, y_k)\}_{k=0}^{\infty}$ generated by `FISTA`, with any initial $x_0 = y_0 \in \mathbb{R}^d$, satisfies the following inequalities as*

$$\begin{cases} \Phi(x_k) - \Phi(x^\star) \leq \dfrac{\mathcal{E}(K)}{s(k + \lambda^\ddagger)(2k + \lambda^\ddagger)\left[1 + (1 - Ls) \cdot \dfrac{\mu s}{4}\right]^{k-K}} \\ \left\|G_s(y_k)\right\|^2 \leq \dfrac{4\mathcal{E}(K)}{s^2(1 - Ls)(k + \lambda^\ddagger)(2k + \lambda^\ddagger)\left[1 + (1 - Ls) \cdot \dfrac{\mu s}{4}\right]^{k-K}} \end{cases} \tag{51}$$

for any $k \geq K$.

5 Conclusion

A significant milestone in modern gradient-based optimization is Nesterov's accelerated gradient descent (`NAG`) method. Its proximal generalization, the fast iterative shrinkage-thresholding algorithm (`FISTA`), has found widespread application in image science and engineering. However, it has remained an open problem whether both `NAG` and `FISTA` exhibit linear convergence on strongly convex functions without requiring knowledge of the modulus of strong convexity, as noted in [CP16, Appendix B].

In this paper, we address this question using the high-resolution ordinary differential equation (ODE) framework. Building on phase-space representation, we introduce a novel Lyapunov function where the coefficient of kinetic energy adapts dynamically with each iteration. Importantly, we demonstrate that the linear convergence of both `NAG` and `FISTA` is independent of the parameter $\lambda^{\ddagger}$, and the square of the proximal subgradient norm also converges linearly. These findings provide new insights into the behavior of accelerated methods in strongly convex settings.

References

[AMR12] Hedy Attouch, Paul-Emile Maingé, and Patrick Redont. A second-order differential system with hessian-driven damping; application to non-elastic shock laws. *Differential Equations and Applications*, 4(1):27–65, 2012.

[AP16] Hedy Attouch and Juan Peypouquet. The rate of convergence of nesterov's accelerated forward-backward method is actually faster than $1/k^2$. *SIAM Journal on Optimization*, 26(3):1824–1834, 2016.

[APR14] Hédy Attouch, Juan Peypouquet, and Patrick Redont. A dynamical approach to an inertial forward-backward algorithm for convex minimization. *SIAM Journal on Optimization*, 24(1):232–256, 2014.

[APR16] Hedy Attouch, Juan Peypouquet, and Patrick Redont. Fast convex optimization via inertial dynamics with hessian driven damping. *Journal of Differential Equations*, 261(10):5734–5783, 2016.

[BT09] 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.

[CD15] Antonin Chambolle and Ch Dossal. On the convergence of the iterates of the "fast iterative shrinkage/thresholding algorithm". *Journal of Optimization theory and Applications*, 166:968–982, 2015.

[CP16] Antonin Chambolle and Thomas Pock. An introduction to continuous optimization for imaging. *Acta Numerica*, 25:161–319, 2016.

[CSY22a] Shuo Chen, Bin Shi, and Ya-Xiang Yuan. Gradient norm minimization of nesterov acceleration: $o(1/k^3)$. *arXiv preprint arXiv:2209.08862*, 2022.

[CSY22b] Shuo Chen, Bin Shi, and Ya-Xiang Yuan. Revisiting the high-resolution phenomenon via high-resolution differential equations. *arXiv preprint arXiv:2212.05700*, 2022.

[CSY23] Shuo Chen, Bin Shi, and Ya-xiang Yuan. On underdamped nesterov's acceleration. *arXiv preprint arXiv:2304.14642*, 2023.

[GT61] I. M. Gelfand and M. L. Tsetlin. Prlnciple of the nonlocal search in the systems of automatic optimization. *Dokl. Akad. Nauk SSSR*, 137(2):295–298, 1961.

[LC22] Hao Luo and Long Chen. From differential equation solvers to accelerated first-order methods for convex optimization. *Mathematical Programming*, 195(1-2):735–781, 2022.

[LSY22a] Bowen Li, Bin Shi, and Ya-Xiang Yuan. Linear convergence of ISTA and FISTA. *arXiv preprint arXiv:2212.06319*, 2022.

[LSY22b] Bowen Li, Bin Shi, and Ya-Xiang Yuan. Proximal subgradient norm minimization of ISTA and FISTA. *arXiv preprint arXiv:2211.01610*, 2022.

[MJ19] Michael Muehlebach and Michael Jordan. A dynamical systems perspective on nesterov acceleration. In *International Conference on Machine Learning*, pages 4656–4662. PMLR, 2019.

[Nes83] Yurii E Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Dokl. Akad. Nauk SSSR*, 269:543–547, 1983.

[Nes98] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer Science & Business Media, 1998.

[SBC16] Weijie Su, Stephen Boyd, and Emmanuel J Candes. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17:1–43, 2016.

[SDJS22] 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.

[SDSJ19] Bin Shi, Simon S Du, Weijie Su, and Michael I Jordan. Acceleration via symplectic discretization of high-resolution differential equations. *Advances in Neural Information Processing Systems*, 32, 2019.

[WRJ21] Ashia C Wilson, Ben Recht, and Michael I Jordan. A lyapunov analysis of accelerated methods in optimization. *J. Mach. Learn. Res.*, 22:113–1, 2021.

[WWJ16] 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.

[XSJ$^+$21] Hedi Xia, Vai Suliafu, Hangjie Ji, Tan Nguyen, Andrea Bertozzi, Stanley Osher, and Bao Wang. Heavy ball neural ordinary differential equations. *Advances in Neural Information Processing Systems*, 34:18646–18659, 2021.

[ZOD$^+$21] Peiyuan Zhang, Antonio Orvieto, Hadi Daneshmand, Thomas Hofmann, and Roy S Smith. Revisiting the role of euler numerical integration on acceleration and stability in convex optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 3979–3987. PMLR, 2021.