High-Resolution ODE Modeling of Nesterov's Accelerated Gradient Method Without Modulus Knowledge

Chris Junchi Li[◇]

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

September 26, 2024

Abstract

This paper presents advancements in the study of Nesterov's Accelerated Gradient Method (AGD) by modeling its continuous-time dynamics through high-resolution second-order ordinary differential equations (ODEs). We offer deeper theoretical insights into the oscillatory behavior, damping effects, and fast inverse quadratic convergence rates of this method. By establishing the continuous-time limit, we simplify the understanding of momentum-driven optimization techniques, bridging discrete algorithms with continuous dynamics. Furthermore, we demonstrate that Nesterov's AGD achieves linear convergence for strongly convex functions without requiring prior knowledge of the convexity modulus, making it highly applicable in real-world settings. We also extend our analysis to composite optimization problems using a unified Lyapunov-based framework, proving the optimal complexity bounds of both AGD and its proximal variant, FISTA. These findings provide a rigorous foundation for accelerated methods, ensuring both practical efficiency and robustness, especially in large-scale or ill-conditioned optimization tasks.

Keywords: Nesterov's Accelerated Gradient (`AGD`); Convex Optimization; High-Resolution ODE Framework; Linear Convergence; Recursive Averaging Gradient Descent (`RAVE-GD`); Strongly Convex Functions

1 Introduction

Optimization has become a cornerstone of modern machine learning and computational mathematics, particularly for solving convex and strongly convex problems. Among the many gradient-based techniques, Nesterov's Accelerated Gradient Method (AGD) stands out for its remarkable efficiency. First introduced in 1983, AGD incorporates a momentum term that accelerates convergence, achieving a convergence rate of $\mathcal{O}(1/k^2)$ for convex functions—far outpacing traditional gradient descent, which converges at $\mathcal{O}(1/k)$. This efficiency has made AGD a fundamental tool in large-scale data analysis, machine learning, and signal processing.

A key development in understanding AGD and other accelerated methods has been the use of continuous-time models, particularly through ordinary differential equations (ODEs). By modeling these algorithms as ODEs, researchers have uncovered deeper insights into their convergence properties, stability, and dynamics. These continuous-time models reveal how momentum affects oscillatory behavior and damping, providing a framework for analyzing optimization methods at a more intuitive and theoretical level. This ODE-based perspective not only simplifies the understanding of accelerated methods but also bridges the gap between discrete-time algorithms and continuous-time dynamics.

Despite its widespread adoption, AGD still faces challenges. For strongly convex functions, implementing AGD often requires knowledge of the convexity modulus, which is not always readily available in practical scenarios. Moreover, extending AGD to composite optimization problems—where the objective function is a combination of smooth and non-smooth components—adds complexity in maintaining acceleration without losing robustness.

In this paper, we address these challenges by investigating the continuous-time dynamics of Nesterov's AGD through a high-resolution ODE framework. Our analysis demonstrates that AGD achieves linear convergence for strongly convex functions without requiring knowledge of the strong convexity modulus, making it applicable in more general real-world scenarios. Additionally, we extend the analysis to composite optimization, showing that AGD and its proximal variant, FISTA, retain their acceleration properties even in non-smooth settings. These contributions provide both theoretical insights and practical improvements, paving the way for more adaptive and robust optimization techniques.

Related work The history of first-order optimization algorithms, particularly those focusing on acceleration, dates back to the Ravine method, a two-step gradient strategy designed to outperform classical gradient descent [GT61]. Nesterov's Accelerated Gradient Method (AGD), introduced in [Nes83, Nes18], revolutionized optimization by achieving an optimal convergence rate of $O(1/k^2)$ for convex functions, significantly improving upon the $O(1/k)$ rate of vanilla gradient descent. This acceleration is achieved by incorporating a momentum term and a tuned averaging coefficient, enabling faster convergence with the same computational complexity.

AGD's dynamics have been further explored using ordinary differential equations (ODEs), offering deeper insights into its continuous-time behavior. The derivation of a low-resolution ODE model for AGD [SBC16] laid the foundation for various research efforts, including analyses of faster convergence rates [AP16], variational methods [WWJ16], and Lyapunov analysis [WRJ21]. High-resolution ODE frameworks have emerged as powerful tools for understanding acceleration in optimization [SDJS22], with gradient correction mechanisms linked to implicit velocity dynamics. These studies bridge the gap between discrete-time algorithms and continuous-time models, explaining key phenomena such as oscillations, damping, and momentum-based acceleration.

The high-resolution ODE framework has also been extended to composite optimization, addressing problems involving both smooth and non-smooth components. These advancements allow for the preservation of acceleration even in non-smooth settings, offering a unified framework for both convex and strongly convex problems.

Moreover, the connection between ODEs and optimization dates back several decades, with numerous studies establishing continuous-time analogs for iterative optimization methods [AMR12, Fio05, DSE12]. By viewing optimization algorithms as discretizations of continuous ODEs, researchers have gained a clearer understanding of the underlying mechanics of gradient-based methods. These insights have influenced a variety of fields, including linear regression [ORX+16], compressed sensing [BBC11], and machine learning applications such as deep neural networks [SMDH13]. The interplay between discrete and continuous optimization remains a rich area of study, providing new theoretical and practical insights into modern optimization methods.

Contributions. The main contributions of this paper are as follows:

- We provide a continuous-time analysis of Nesterov's Accelerated Gradient Method (`AGD`) through a high-resolution ODE framework, offering deeper insights into its oscillatory dynamics and convergence behavior.

- We introduce a novel Lyapunov function with a dynamically adapting coefficient of kinetic energy, proving the linear convergence of both `AGD` and its proximal variant, `FISTA`, in strongly convex optimization without requiring knowledge of the convexity modulus.
- We extend the high-resolution ODE framework to composite optimization problems, ensuring that acceleration is preserved in non-smooth settings and establishing linear convergence for function values and the square of the proximal subgradient norm.
- Our theoretical analysis is supported by extensive numerical experiments, demonstrating the practical effectiveness of these methods in ill-conditioned optimization tasks.

Notation. For two sequences of positive scalars $\{a_n\}$ and $\{b_n\}$, we denote $a_n = \Omega(b_n)$ (resp. $a_n = \mathcal{O}(b_n)$) if $a_n \geq Cb_n$ (resp. $a_n \leq Cb_n$) for all n, and also $a_n = \Theta(b_n)$ if both $\Omega(b_n)$ and $a_n = \mathcal{O}(b_n)$ hold, for some absolute constant $C > 0$, and $\widetilde{\mathcal{O}}$ or $\widetilde{\Omega}$ is adopted in turn when C contains a polylogarithmic factor in problem-dependent parameters. Let $\lambda_{\max}(\mathbf{A})$ and $\lambda_{\min}(\mathbf{A})$ denote the maximal and minimal eigenvalues of a real symmetric matrix $\mathbf{A}$, and $\|\mathbf{A}\|_{\mathrm{op}}$ the operator norm $\sqrt{\lambda_{\max}(\mathbf{A}^\top \mathbf{A})}$. Let vector $z = [\boldsymbol{x}; \boldsymbol{y}] \in \mathbb{R}^{n+m}$ denote the concatenation of $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{y} \in \mathbb{R}^m$. We use $\wedge$ (resp. $\vee$) to denote the bivariate min (resp. max) throughout this paper. For natural number K let $[K]$ denote the set $\{1, \ldots, K\}$. Throughout the paper we also use the standard notation $\|\cdot\|$ to denote the ℓ_2-norm and $\|\cdot\|_{\mathrm{op}}$ to denote the operator norm of a matrix. We will explain other notations at their first appearances.

2 Accelerated Gradient Descent Formulation

We study the unconstrained minimization problem:

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

seeking to minimize $f(x)$ over $x \in \mathbb{R}^d$, where f is a convex function, smooth or non-smooth, and $x \in \mathbb{R}^n$ is the variable. Since Newton, numerous algorithms and methods have been proposed to solve the minimization problem, notably gradient and subgradient descent, Newton's methods, trust region methods, conjugate gradient methods, and interior point methods; see e.g., [Pol87, BV04, NW99, Rus06, BPC+11, Sho85, Bec14], for expositions. And while vanilla gradient descent enjoys an iteration complexity of $\mathcal{O}(\kappa \log(\frac{1}{\epsilon}))$ on L-smooth, μ-strongly convex problems, with $\kappa = \frac{L}{\mu}$ being the condition number, Nesterov's method [Nes83], when equipped with proper restarting, achieves an improved iteration complexity of $\mathcal{O}(\sqrt{\kappa} \log(\frac{1}{\epsilon}))$.

Let y_K denote the output of Nesterov's Accelerated Gradient Method with initialization y_0, x_0, z_0 with $y_0 = x_0 = z_0$. For $k = 0, 1, \ldots,$ AGD can be written in the following three-sequence form [LLF20]:

$$x_k = y_k + \frac{k-1}{k+\lambda} \frac{\alpha_{k-1}}{1-\alpha_{k-1}} (z_k - y_k) \tag{1a}$$

$$y_{k+1} = x_k - s \nabla f(x_k) \tag{1b}$$

$$z_{k+1} = y_k + \frac{1}{\alpha_k} (y_{k+1} - y_k) \tag{1c}$$

where $s \geq 0$ and $0 < \alpha_k < 1$. The scheme can be expressed in the following two-sequence form as well, i.e., it is straightforward to verify (with its proof omitted) that

Proposition 2. *Two methods are equivalent in the sense they generate the same output [Nes83]*

$$\begin{aligned} x_k &= y_k + \frac{k-1}{k+\lambda}\left(y_k - y_{k-1}\right) \\ y_{k+1} &= x_k - s\nabla f\left(x_k\right) \end{aligned} \tag{2}$$

Moreover, The notations employed within this paper mostly aligns with those found in [Nes18], 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 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, $\mathcal{S}_{\mu,L}^2(\mathbb{R}^d)$ refers to a subclass of $\mathcal{S}_{\mu,L}^1(\mathbb{R}^d)$, which encompasses functions possessing a continuous Hessian. Finally, the term z^* is used to denote its unique minimizer of those functions. **In the rest of this paper we assume $\lambda = 2$ unless otherwise stated, and will be assuming, unless otherwise mentioned:**

Assumption 1 (Convexity and smoothness)**.** *We assume that $f(\cdot) : \mathbb{R}^n \to \mathbb{R}$ is μ-strongly convex and L-smooth for some $L \geq \mu \geq 0$. L is dubbed as the Lipschitz constant of f.*

We have under Assumption 1 that for each z

$$f(z) - f(z^*) \geq \frac{\mu}{2}\left\|z - z^*\right\|^2 \tag{3}$$

Remark. If $0 < s \leq \frac{1}{L}$ for all k, then Nesterov's method achieves $f\left(y_k\right) - f^\star = O\left(\prod_{i=0}^{k-1}\left(1-\alpha_i\right)\right)$ [Nes18, Theorem 2.2.1]. A frequent pick of parameter λ is $\lambda = 2$ and constant step size $s = \frac{1}{L}$, where we have momentum parameters $\approx \frac{k}{k+3}$ if $\mu = 0$ [Tse08, Algorithm 2] and $= \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$ if $\mu > 0$. The detailed derivation of this method can be found in the classics, e.g., [Nes18].

3 Low-Resolution Differential Equations: Mathematical Formulations

In this section, we derive a couple of first-order ODEs (or equivalently, a second-order ODE) which is the exact limit of Nesterov's scheme by taking small step sizes in (2); to the best of our knowledge, this work is the first to use ODEs to model Nesterov's scheme or its variants in this

[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.

limit. One surprising fact in connection with this subject is that a *first-order* scheme is modeled by a *second-order* ODE. This ODE takes the following form:

$$\ddot{Y} + \frac{3}{t}\dot{Y} + \nabla f(Y) = 0 \tag{4a}$$

for $t > 0$, with initial conditions $Y(0) = y_0, \dot{Y}(0) = 0$; or equivalently, a cople of first-order ODEs

$$\begin{cases} \dot{Y} = \frac{2}{t}(Z - Y) \\ \dot{Z} = -\frac{t}{2}\nabla f(Y) \end{cases} \tag{4b}$$

with initial conditions $Y(0) = y_0$, $Z(0) = y_0$ such that $Z(t) - Y(t) = o(t)$ as $t \to 0^+$; here, y_0 is the starting point in Nesterov's scheme, $\dot{X} \equiv \frac{\mathrm{d}Y}{\mathrm{d}t}$ denotes the time derivative or velocity and similarly $\ddot{X} \equiv \frac{\mathrm{d}^2Y}{\mathrm{d}t^2}$ denotes the acceleration. The time parameter in this ODE is related to the step size in (2) via $t \approx k\sqrt{s}$. Expectedly, it also enjoys inverse quadratic convergence rate as its discrete analog,

$$f(Y(t)) - f^\star \le O\left(\frac{\|y_0 - z^*\|^2}{t^2}\right)$$

Approximate equivalence between Nesterov's scheme and the ODE is established later in various perspectives, rigorous and intuitive. In the main body of this paper, examples and case studies are provided to demonstrate that the homogeneous and conceptually simpler ODE can serve as a tool for understanding, analyzing and generalizing Nesterov's scheme.

The rest of the section is organized as follows. In Section 3.1, the ODE is rigorously derived from Nesterov's scheme. Section 3.2 exhibits inverse quadratic convergence rate for the ODE solution, and Section 3.3 is devoted to comparing Nesterov's scheme with gradient descent from a numerical perspective.

3.1 Derivation

First, we sketch an informal derivation of the ODE (4a). **Unless otherwise noted, we will be studying its second-order equivalent form ODE.** Assume $f \in \mathcal{F}_L$ for $L > 0$. Combining the two equations of (2) and applying a rescaling gives

$$\frac{y_{k+1} - y_k}{\sqrt{s}} = \frac{k-1}{k+2}\frac{y_k - y_{k-1}}{\sqrt{s}} - \sqrt{s}\nabla f(x_k) \tag{5}$$

Introduce the *Ansatz* $y_k \approx Y(k\sqrt{s})$ for some smooth curve $Y(t)$ defined for $t \ge 0$. Put $k = \frac{t}{\sqrt{s}}$. Then as the step size s goes to zero, $Y(t) \approx y_{\frac{t}{\sqrt{s}}} = y_k$ and $Y(t + \sqrt{s}) \approx y_{\frac{t+\sqrt{s}}{\sqrt{s}}} = y_{k+1}$, and Taylor expansion gives

$$\frac{y_{k+1} - y_k}{\sqrt{s}} = \dot{Y}(t) + \frac{1}{2}\ddot{Y}(t)\sqrt{s} + o(\sqrt{s}) \quad \frac{y_k - y_{k-1}}{\sqrt{s}} = \dot{Y}(t) - \frac{1}{2}\ddot{Y}(t)\sqrt{s} + o(\sqrt{s})$$

and $\sqrt{s}\nabla f(x_k) = \sqrt{s}\nabla f(Y(t)) + o(\sqrt{s})$. Thus (5) can be written as

$$\begin{aligned} &\dot{Y}(t) + \frac{1}{2}\ddot{Y}(t)\sqrt{s} + o(\sqrt{s}) \\ &\qquad = \left(1 - \frac{3\sqrt{s}}{t}\right)\left(\dot{Y}(t) - \frac{1}{2}\ddot{Y}(t)\sqrt{s} + o(\sqrt{s})\right) - \sqrt{s}\nabla f(Y(t)) + o(\sqrt{s}) \end{aligned} \tag{6}$$

By comparing the coefficients of $\sqrt{s}$ in (6), we obtain

$$\ddot{Y} + \frac{3}{t}\dot{Y} + \nabla f(Y) = 0$$

The first initial condition is $Y(0) = y_0$. Taking $k = 1$ in (5) yields

$$\frac{y_2 - y_1}{\sqrt{s}} = -\sqrt{s}\nabla f(x_1) = o(1)$$

Hence, the second initial condition is simply $\dot{Y}(0) = 0$ (vanishing initial velocity).

Classical results in ODE theory do not directly imply the existence or uniqueness of the solution to this ODE because the coefficient $\frac{3}{t}$ is singular at $t = 0$. In addition, ∇f is typically not analytic at y_0, which leads to the inapplicability of the power series method for studying singular ODEs. Nevertheless, the ODE is well posed: the strategy we employ for showing this constructs a series of ODEs approximating (4a), and then chooses a convergent subsequence by some compactness arguments such as the Arzelá-Ascoli theorem. Below, $C^2((0,\infty);\mathbb{R}^n)$ denotes the class of twice continuously differentiable maps from $(0,\infty)$ to $\mathbb{R}^n$; similarly, $C^1([0,\infty);\mathbb{R}^n)$ denotes the class of continuously differentiable maps from $[0,\infty)$ to $\mathbb{R}^n$.

Theorem 1. *For any $f \in \mathcal{F}_\infty := \cup_{L>0}\mathcal{F}_L$ and any $y_0 \in \mathbb{R}^n$, the ODE* (4a) *with initial conditions $Y(0) = y_0, \dot{Y}(0) = 0$ has a unique global solution $X \in C^2((0,\infty);\mathbb{R}^n) \cap C^1([0,\infty);\mathbb{R}^n)$.*

The next theorem, in a rigorous way, guarantees the validity of the derivation of this ODE. The proofs of both theorems are deferred to later subsections.

Theorem 2. *For any $f \in \mathcal{F}_\infty$, as the step size $s \to 0$, Nesterov's scheme* (2) *converges to the ODE* (4a) *in the sense that for all fixed $T > 0$,*

$$\lim_{s\to 0} \max_{0 \le k \le \frac{T}{\sqrt{s}}} \left\| y_k - X\left(k\sqrt{s}\right) \right\| = 0$$

Remark. We collect some elementary properties that are helpful in understanding the ODE.

- **Time Invariance.** If we adopt a linear time transformation, $\widetilde{t} = ct$ for some $c > 0$, by the chain rule it follows that

  $$\frac{\mathrm{d}Y}{\mathrm{d}\widetilde{t}} = \frac{1}{c}\frac{\mathrm{d}Y}{\mathrm{d}t}, \ \frac{\mathrm{d}^2Y}{\mathrm{d}\widetilde{t}^2} = \frac{1}{c^2}\frac{\mathrm{d}^2Y}{\mathrm{d}t^2}$$

  This yields the ODE parameterized by $\widetilde{t}$,

  $$\frac{\mathrm{d}^2Y}{\mathrm{d}\widetilde{t}^2} + \frac{3}{\widetilde{t}}\frac{\mathrm{d}Y}{\mathrm{d}\widetilde{t}} + \frac{1}{c^2}\nabla f(Y) = 0$$

  Also note that minimizing $\frac{f}{c^2}$ is equivalent to minimizing f. Hence, the ODE is invariant under the time change. In fact, it is easy to see that time invariance holds if and only if the coefficient of $\dot{X}$ has the form $\frac{C}{t}$ for some constant C.

- **Rotational Invariance.** Nesterov's scheme and other gradient-based schemes are invariant under rotations. As expected, the ODE is also invariant under orthogonal transformation. To see this, let $Y = QX$ for some orthogonal matrix Q. This leads to $\dot{Y} = Q\dot{X}, \ddot{Y} = Q\ddot{X}$ and $\nabla_Y f = Q\nabla_X f$. Hence, denoting by $Q^\top$ the transpose of Q, the ODE in the new coordinate system reads $Q^\top\ddot{Y} + \frac{3}{t}Q^\top\dot{Y} + Q^\top\nabla_Y f = 0$, which is of the same form as (4a) once multiplying Q on both sides.

- **Initial Asymptotic.** Assume sufficient smoothness of X such that $\lim_{t\to 0}\ddot{Y}(t)$ exists. The mean value theorem guarantees the existence of some $\xi \in (0,t)$ that satisfies $\frac{\dot{Y}(t)}{t} = \frac{\dot{Y}(t)-\dot{Y}(0)}{t} = \ddot{Y}(\xi)$. Hence, from the ODE we deduce $\ddot{Y}(t) + 3\ddot{Y}(\xi) + \nabla f(Y(t)) = 0$. Taking the limit $t \to 0$ gives $\ddot{Y}(0) = -\frac{1}{4}\nabla f(y_0)$. Hence, for small t we have the asymptotic form:

$$Y(t) = -\frac{\nabla f(y_0)t^2}{8} + y_0 + o(t^2)$$

  This asymptotic expansion is consistent with the empirical observation that Nesterov's scheme moves slowly in the beginning.

3.2 Analogous Convergence Rate

The original result from [Nes83] states that, for any $f \in \mathcal{F}_L$, the sequence $\{y_k\}$ given by (2) with step size $s \le \frac{1}{L}$ satisfies

$$f(y_k) - f^\star \le \frac{2\|y_0 - z^*\|^2}{s(k+1)^2} \tag{7}$$

Our next result indicates that the trajectory of (4a) closely resembles the sequence $\{y_k\}$ in terms of the convergence rate to a minimizer z^*. Compared with the discrete case, this proof is shorter and simpler.

Theorem 3. *For any $f \in \mathcal{F}_\infty$, let $Y(t)$ be the unique global solution to* (4a) *with initial conditions $Y(0) = y_0, \dot{Y}(0) = 0$. Then, for any $t > 0$,*

$$f(Y(t)) - f^\star \le \frac{2\left\|y_0 - z^*\right\|^2}{t^2} \tag{8}$$

Proof of Theorem 3. Consider the energy functional[2] defined as $\mathcal{E}(t) = t^2(f(Y(t)) - f^\star) + 2\|Y + \frac{t}{2}\dot{Y} - z^*\|^2$, whose time derivative is

$$\dot{\mathcal{E}} = 2t(f(Y) - f^\star) + t^2\left\langle \nabla f, \dot{Y}\right\rangle + 4\left\langle Y + \frac{t}{2}\dot{Y} - z^*, \frac{3}{2}\dot{Y} + \frac{t}{2}\ddot{Y}\right\rangle$$

Substituting $\frac{3}{2}\dot{Y} + \frac{t}{2}\ddot{Y}$ with $-\frac{t}{2}\nabla f(Y)$, the above equation gives

$$\dot{\mathcal{E}} = 2t(f(Y) - f^\star) + 4\langle Y - z^*, -\frac{t}{2}\nabla f(Y)\rangle = 2t(f(Y) - f^\star) - 2t\langle Y - z^*, \nabla f(Y)\rangle \le 0$$

where the inequality follows from the convexity of f. Hence by monotonicity of $\mathcal{E}$ and non-negativity of $2\left\|Y + \frac{t}{2}\dot{Y} - z^*\right\|^2$, the gap satisfies

$$f(Y(t)) - f^\star \le \frac{\mathcal{E}(t)}{t^2} \le \frac{\mathcal{E}(0)}{t^2} = \frac{2\left\|y_0 - z^*\right\|^2}{t^2}$$

□

[2]We may also view this functional as the negative entropy. Similarly, for the gradient flow $\dot{X} + \nabla f(Y) = 0$, an energy function of form $\mathcal{E}_{\text{gradient}}(t) = t(f(Y(t)) - f^\star) + \frac{1}{2}\|Y(t) - z^*\|^2$ can be used to derive the bound $f(Y(t)) - f^\star \le \frac{\|y_0 - z^*\|^2}{2t}$.

Making use of the approximation $t \approx k\sqrt{s}$, we observe that the convergence rate in (7) is essentially a discrete version of that in (8), providing yet another piece of evidence for the approximate equivalence between the ODE and the scheme.

We finish this subsection by showing that the number 2 appearing in the numerator of the error bound in (8) is optimal. Consider an arbitrary $f \in \mathcal{F}_\infty(\mathbb{R})$ such that $f(x) = x$ for $x \geq 0$. Starting from some $y_0 > 0$, the solution to (4a) is $Y(t) = y_0 - \frac{t^2}{8}$ before hitting the origin. Hence, $t^2(f(Y(t)) - f^\star) = t^2(y_0 - \frac{t^2}{8})$ has a maximum $2y_0^2 = 2|y_0 - 0|^2$ achieved at $t = 2\sqrt{y_0}$. Therefore, we cannot replace 2 by any smaller number, and we can expect that this tightness also applies to the discrete analog (7).

3.3 Nesterov's Scheme Compared with Gradient Descent

The ansatz $t \approx k\sqrt{s}$ in relating the ODE and Nesterov's scheme is formally confirmed in Theorem 2. Consequently, for any constant $t_c > 0$, this implies that y_k does not change much for a range of step sizes s if $k \approx \frac{t_c}{\sqrt{s}}$. To empirically support this claim, we present an example in Figure 1a, where the scheme minimizes $f(x) = \frac{1}{2}\|y - Ax\|^2 + \|x\|_1$ with $y = (4,\ 2,\ 0)$ and $A(:,1) = (0,\ 2,\ 4),\ A(:,2) = (1,\ 1,\ 1)$ starting from $y_0 = (2,\ 0)$ (here $A(:,j)$ is the jth column of A). From this figure, we are delight to observe that y_k with the same t_c are very close to each other.

This interesting square-root scaling has the potential to shed light on the superiority of Nesterov's scheme over gradient descent. Roughly speaking, each iteration in Nesterov's scheme amounts to traveling $\sqrt{s}$ in time along the integral curve of (4a), whereas it is known that the simple gradient descent $y_{k+1} = y_k - s\nabla f(y_k)$ moves s along the integral curve of $\dot{X} + \nabla f(Y) = 0$. We expect that for small s Nesterov's scheme moves more in each iteration since $\sqrt{s}$ is much larger than s. Figure 1b illustrates and supports this claim, where the function minimized is $f = |y_1|^3 + 5|y_2|^3 + 0.001(y_1 + y_2)^2$ with step size $s = 0.05$ (The coordinates are appropriately rotated to allow y_0 and z^* lie on the same horizontal line). The circles are the iterates for $k = 1, 10, 20, 30, 45, 60, 90, 120, 150, 190, 250, 300$. For Nesterov's scheme, the seventh circle has already passed $t = 15$, while for gradient descent the last point has merely arrived at $t = 15$.

A second look at Figure 1b suggests that Nesterov's scheme allows a large deviation from its limit curve, as compared with gradient descent. This raises the question of the stable step size allowed for numerically solving the ODE (4a) in the presence of accumulated errors. The finite difference approximation by the forward Euler method is

$$\frac{Y(t+\Delta t) - 2Y(t) + Y(t-\Delta t)}{\Delta t^2} + \frac{3}{t}\frac{Y(t) - Y(t-\Delta t)}{\Delta t} + \nabla f(Y(t)) = 0 \tag{9}$$

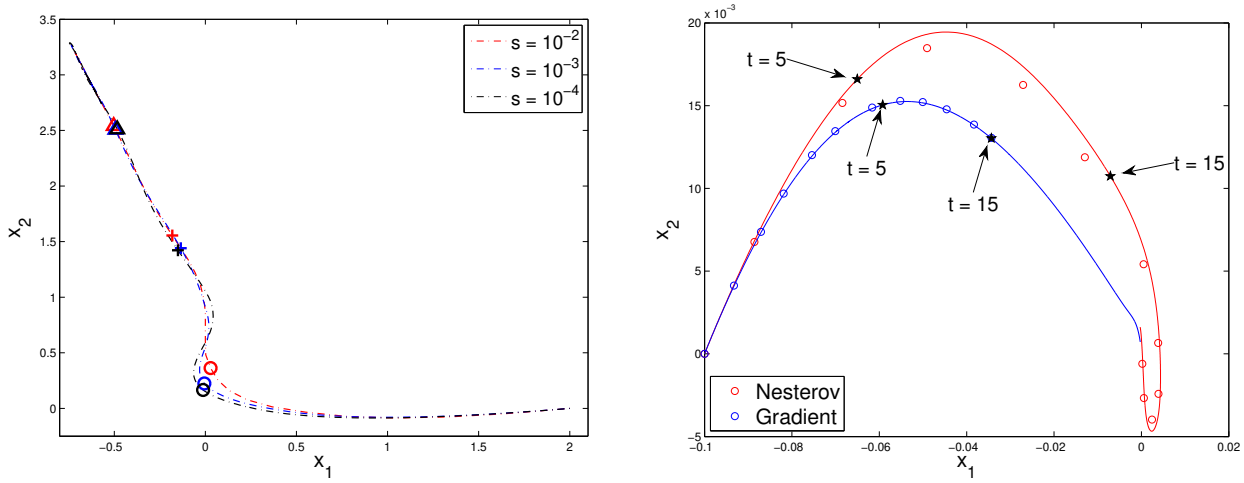
which is equivalent to

$$Y(t+\Delta t) = \Big(2 - \frac{3\Delta t}{t}\Big)Y(t) - \Delta t^2 \nabla f(Y(t)) - \Big(1 - \frac{3\Delta t}{t}\Big)Y(t-\Delta t) \tag{10}$$

Assuming f is sufficiently smooth, we have $\nabla f(x+\delta x) \approx \nabla f(x) + \nabla^2 f(x)\delta x$ for small perturbations δx, where $\nabla^2 f(x)$ is the Hessian of f evaluated at x. Identifying $k = \frac{t}{\Delta t}$, the characteristic equation of this finite difference scheme is approximately

$$\det\left(\overline{\lambda}^2 - \left(2 - \Delta t^2 \nabla^2 f - \frac{3\Delta t}{t}\right)\overline{\lambda} + 1 - \frac{3\Delta t}{t}\right) = 0 \tag{11}$$

The numerical stability of (9) with respect to accumulated errors is equivalent to this: all the roots of (11) lie in the unit circle; see e.g., [Lea22]. When $\nabla^2 f \preceq LI_n$ (i.e. $LI_n - \nabla^2 f$ is positive



(a) Square-root scaling of s.

(b) Race between Nesterov's and gradient.

Figure 1. In (a), the circles, crosses and triangles are y_k evaluated at $k = \left\lceil \frac{1}{\sqrt{s}} \right\rceil, \left\lceil \frac{2}{\sqrt{s}} \right\rceil$ and $\left\lceil \frac{3}{\sqrt{s}} \right\rceil$, respectively. In (b), the circles are iterations given by Nesterov's scheme or gradient descent, depending on the color, and the stars are $Y(t)$ on the integral curves for $t = 5, 15$.

semidefinite), if $\frac{\Delta t}{t}$ small and $\Delta t < \frac{2}{\sqrt{L}}$, we see that all the roots of (11) lie in the unit circle. On the other hand, if $\Delta t > \frac{2}{\sqrt{L}}$, (11) can possibly have a root $\overline{\lambda}$ outside the unit circle, causing numerical instability. Under our identification $s = \Delta t^2$, a step size of $s = \frac{1}{L}$ in Nesterov's scheme (2) is approximately equivalent to a step size of $\Delta t = \frac{1}{\sqrt{L}}$ in the forward Euler method, which is stable for numerically integrating (9).

As a comparison, note that the finite difference scheme of the ODE $\dot{Y}(t) + \nabla f(Y(t)) = 0$, which models gradient descent with updates $y_{k+1} = y_k - s\nabla f(y_k)$, has the characteristic equation $\det(\overline{\lambda} - (1 - \Delta t \nabla^2 f)) = 0$. Thus, to guarantee $-I_n \preceq 1 - \Delta t \nabla^2 f \preceq I_n$ in worst case analysis, one can only choose $\Delta t \leq \frac{2}{L}$ for a fixed step size, which is much smaller than the step size $\frac{2}{\sqrt{L}}$ for (9) when ∇f is very variable, i.e., L is large.

3.4 Proof of Theorem 1

The proof is divided into two parts, namely, existence and uniqueness.

Lemma 1. *For any $f \in \mathcal{F}_\infty$ and any $y_0 \in \mathbb{R}^n$, the ODE* (4a) *has at least one solution X in $C^2(0,\infty) \cap C^1[0,\infty)$.*

Below, some preparatory lemmas are given before turning to the proof of this lemma. To begin with, for any $\delta > 0$ consider the smoothed ODE

$$\ddot{Y} + \frac{3}{t \vee \delta}\dot{Y} + \nabla f(Y) = 0 \tag{12}$$

with $Y(0) = y_0, \dot{Y}(0) = 0$. Denoting by $Z = \dot{Y}$, then (12) is equivalent to

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} Y \\ Z \end{pmatrix} = \begin{pmatrix} Z \\ -\frac{3}{t\vee\delta}Z - \nabla f(Y) \end{pmatrix}$$

with $Y(0) = y_0$, $Z(0) = 0$. As functions of (Y, Z), both Z and $-\frac{3Z}{t\vee\delta} - \nabla f(Y))$ are $\max(1, L) + \frac{3}{\delta}$-Lipschitz continuous. Hence by standard ODE theory, (12) has a unique global solution in $C^2[0, \infty)$, denoted by Y_δ. Note that $\ddot{Y}_\delta$ is also well defined at $t = 0$. Next, introduce $M_\delta(t)$ to be the supremum of $\frac{\|\dot{Y}_\delta(u)\|}{u}$ over $u \in (0, t]$. It is easy to see that $M_\delta(t)$ is finite because $\frac{\|\dot{Y}_\delta(u)\|}{u} = \frac{\|\dot{Y}_\delta(u) - \dot{Y}_\delta(0)\|}{u} = \left\|\ddot{Y}_\delta(0)\right\| + o(1)$ for small u. We give an upper bound for $M_\delta(t)$ in the following lemma.

Lemma 2. *For $\delta < \sqrt{\frac{6}{L}}$, we have*

$$M_\delta(\delta) \le \frac{\|\nabla f(y_0)\|}{1 - \frac{L\delta^2}{6}}$$

The proof of Lemma 2 relies on a simple lemma.

Lemma 3. *For any $u > 0$, the following inequality holds*

$$\|\nabla f(Y_\delta(u)) - \nabla f(y_0)\| \le \frac{1}{2} L M_\delta(u) u^2$$

Proof of Lemma 3. By Lipschitz continuity,

$$\|\nabla f(Y_\delta(u)) - \nabla f(y_0)\| \le L \|Y_\delta(u) - y_0\| = \left\| \int_0^u \dot{Y}_\delta(v) \mathrm{d}v \right\| \le \int_0^u v \frac{\left\|\dot{Y}_\delta(v)\right\|}{v} \mathrm{d}v \le \frac{1}{2} L M_\delta(u) u^2$$

□

Next, we prove Lemma 2.

Proof of Lemma 2. For $0 < t \le \delta$, the smoothed ODE takes the form

$$\ddot{Y}_\delta + \frac{3}{\delta}\dot{Y}_\delta + \nabla f(Y_\delta) = 0$$

which yields

$$\begin{aligned}
\dot{Y}_\delta \exp\left(\frac{3t}{\delta}\right) &= -\int_0^t \nabla f(Y_\delta(u)) \exp\left(\frac{3u}{\delta}\right) \mathrm{d}u \\
&= -\nabla f(y_0) \int_0^t \exp\left(\frac{3u}{\delta}\right) \mathrm{d}u - \int_0^t (\nabla f(Y_\delta(u)) - \nabla f(y_0)) \exp\left(\frac{3u}{\delta}\right) \mathrm{d}u
\end{aligned}$$

Hence, by Lemma 3

$$\begin{aligned}
\frac{\left\|\dot{Y}_\delta(t)\right\|}{t} &\le \frac{1}{t} \exp\left(-\frac{3t}{\delta}\right) \|\nabla f(y_0)\| \int_0^t \exp\left(\frac{3u}{\delta}\right) \mathrm{d}u + \frac{1}{t} \exp\left(-\frac{3t}{\delta}\right) \int_0^t \frac{1}{2} L M_\delta(u) u^2 \exp\left(\frac{3u}{\delta}\right) \mathrm{d}u \\
&\le \|\nabla f(y_0)\| + \frac{L M_\delta(\delta) \delta^2}{6}
\end{aligned}$$

Taking the supremum of $\frac{\|\dot{Y}_\delta(t)\|}{t}$ over $0 < t \le \delta$ and rearranging the inequality give the desired result. □

Next, we give an upper bound for $M_\delta(t)$ when $t > \delta$.

Lemma 4. *For $\delta < \sqrt{\frac{6}{L}}$ and $\delta < t < \sqrt{\frac{12}{L}}$, we have*

$$M_\delta(t) \le \frac{(5 - \frac{L\delta^2}{6}) \left\|\nabla f(y_0)\right\|}{4(1 - \frac{L\delta^2}{6})(1 - \frac{Lt^2}{12})}$$

Proof of Lemma 4. For $t > \delta$, the smoothed ODE takes the form

$$\ddot{Y}_\delta + \frac{3}{t}\dot{Y}_\delta + \nabla f(Y_\delta) = 0$$

which is equivalent to

$$\frac{\mathrm{d}t^3\dot{Y}_\delta(t)}{\mathrm{d}t} = -t^3\nabla f(Y_\delta(t))$$

Hence, by integration, $t^3\dot{Y}_\delta(t)$ is equal to

$$-\int_\delta^t u^3\nabla f(Y_\delta(u))\mathrm{d}u + \delta^3\dot{Y}_\delta(\delta) = -\int_\delta^t u^3\nabla f(y_0)\mathrm{d}u - \int_\delta^t u^3(\nabla f(Y_\delta(u)) - \nabla f(y_0))\mathrm{d}u + \delta^3\dot{Y}_\delta(\delta)$$

Therefore by Lemmas 3 and 2, we get

$$\begin{aligned}\frac{\left\|\dot{Y}_\delta(t)\right\|}{t} &\le \frac{t^4 - \delta^4}{4t^4}\left\|\nabla f(y_0)\right\| + \frac{1}{t^4}\int_\delta^t \frac{1}{2}LM_\delta(u)u^5\mathrm{d}u + \frac{\delta^4}{t^4}\frac{\left\|\dot{Y}_\delta(\delta)\right\|}{\delta} \\ &\le \frac{1}{4}\left\|\nabla f(y_0)\right\| + \frac{1}{12}LM_\delta(t)t^2 + \frac{\left\|\nabla f(Y_0)\right\|}{1 - \frac{L\delta^2}{6}}\end{aligned}$$

where the last expression is an increasing function of t. So for any $\delta < t' < t$, it follows that

$$\frac{\left\|\dot{Y}_\delta(t')\right\|}{t'} \le \frac{1}{4}\left\|\nabla f(y_0)\right\| + \frac{1}{12}LM_\delta(t)t^2 + \frac{\left\|\nabla f(y_0)\right\|}{1 - \frac{L\delta^2}{6}}$$

which also holds for $t' \le \delta$. Taking the supremum over $t' \in (0, t)$ gives

$$M_\delta(t) \le \frac{1}{4}\left\|\nabla f(y_0)\right\| + \frac{1}{12}LM_\delta(t)t^2 + \frac{\left\|\nabla f(Y_0)\right\|}{1 - \frac{L\delta^2}{6}}$$

The desired result follows from rearranging the inequality. □

Lemma 5. *The function class $\mathcal{F} = \{Y_\delta : \left[0, \sqrt{\frac{6}{L}}\right] \to \mathbb{R}^n | \delta = \frac{\sqrt{\frac{3}{L}}}{2^m}, m = 0, 1, \ldots\}$ is uniformly bounded and equicontinuous.*

Proof of Lemma 5. By Lemmas 2 and 4, for any $t \in \left[0, \sqrt{\frac{6}{L}}\right], \delta \in \left(0, \sqrt{\frac{3}{L}}\right)$ the gradient is uniformly bounded as

$$\left\|\dot{Y}_\delta(t)\right\| \le \sqrt{\frac{6}{L}}M_\delta\left(\sqrt{\frac{6}{L}}\right) \le \sqrt{\frac{6}{L}}\max\left\{\frac{\left\|\nabla f(y_0)\right\|}{1 - \frac{1}{2}}, \frac{5\left\|\nabla f(y_0)\right\|}{4(1 - \frac{1}{2})(1 - \frac{1}{2})}\right\} = 5\sqrt{\frac{6}{L}}\left\|\nabla f(y_0)\right\|$$

Thus it immediately implies that $\mathcal{F}$ is equicontinuous. To establish the uniform boundedness, note that

$$\left\|Y_\delta(t)\right\| \le \left\|Y_\delta(0)\right\| + \int_0^t \left\|\dot{Y}_\delta(u)\right\| \mathrm{d}u \le \left\|y_0\right\| + \frac{30\left\|\nabla f(y_0)\right\|}{L}$$

□

We are now ready for the proof of Lemma 1.

Proof of Lemma 1. By the Arzelá-Ascoli theorem and Lemma 5, $\mathcal{F}$ contains a subsequence converging uniformly on $\left[0, \sqrt{\frac{6}{L}}\right]$. Denote by $\{Y_{\delta_{m_i}}\}_{i\in\mathbb{N}}$ the convergent subsequence and $\breve{Y}$ the limit. Above, $\delta_{m_i} = \frac{\sqrt{\frac{3}{L}}}{2^{m_i}}$ decreases as i increases. We will prove that $\breve{Y}$ satisfies (4a) and the initial conditions $\breve{Y}(0) = y_0, \dot{\breve{Y}}(0) = 0$.

Fix an arbitrary $t_0 \in (0, \sqrt{\frac{6}{L}})$. Since $\left\|\dot{Y}_{\delta_{m_i}}(t_0)\right\|$ is bounded, we can pick a subsequence of $\dot{Y}_{\delta_{m_i}}(t_0)$ which converges to a limit, denoted by $X^D_{t_0}$. Without loss of generality, assume the subsequence is the original sequence. Denote by $\widetilde{Y}$ the local solution to (4a) with $Y(t_0) = \breve{Y}(t_0)$ and $\dot{Y}(t_0) = X^D_{t_0}$. Now recall that $Y_{\delta_{m_i}}$ is the solution to (4a) with $Y(t_0) = Y_{\delta_{m_i}}(t_0)$ and $\dot{Y}(t_0) = \dot{Y}_{\delta_{m_i}}(t_0)$ when $\delta_{m_i} < t_0$. Since both $Y_{\delta_{m_i}}(t_0)$ and $\dot{Y}_{\delta_{m_i}}(t_0)$ approach $\breve{Y}(t_0)$ and $X^D_{t_0}$, respectively, there exists $\epsilon_0 > 0$ such that

$$\sup_{t_0-\epsilon_0<t<t_0+\epsilon_0} \left\|Y_{\delta_{m_i}}(t) - \widetilde{Y}(t)\right\| \to 0$$

as $i \to \infty$. However, by definition we have

$$\sup_{t_0-\epsilon_0<t<t_0+\epsilon_0} \left\|Y_{\delta_{m_i}}(t) - \breve{Y}(t)\right\| \to 0$$

Therefore $\breve{Y}$ and $\widetilde{Y}$ have to be identical on $(t_0 - \epsilon_0, t_0 + \epsilon_0)$. So $\breve{Y}$ satisfies (4a) at t_0. Since t_0 is arbitrary, we conclude that $\breve{Y}$ is a solution to (4a) on $\left(0, \sqrt{\frac{6}{L}}\right)$. By extension, $\breve{Y}$ can be a global solution to (4a) on $(0, \infty)$. It only leaves to verify the initial conditions to complete the proof.

The first condition $\breve{Y}(0) = y_0$ is a direct consequence of $Y_{\delta_{m_i}}(0) = y_0$. To check the second, pick a small $t > 0$ and note that

$$\frac{\left\|\breve{Y}(t) - \breve{Y}(0)\right\|}{t} = \lim_{i\to\infty} \frac{\left\|Y_{\delta_{m_i}}(t) - Y_{\delta_{m_i}}(0)\right\|}{t} = \lim_{i\to\infty} \left\|\dot{Y}_{\delta_{m_i}}(\xi_i)\right\|$$
$$\leq \limsup_{i\to\infty} tM_{\delta_{m_i}}(t) \leq 5t\sqrt{\frac{6}{L}}\left\|\nabla f(y_0)\right\|$$

where $\xi_i \in (0, t)$ is given by the mean value theorem. The desired result follows from taking $t \to 0$. □

Next, we aim to prove the uniqueness of the solution to (4a).

Lemma 6. *For any $f \in \mathcal{F}_\infty$, the ODE* (4a) *has at most one local solution in a neighborhood of $t = 0$.*

Suppose on the contrary that there are two solutions, namely, X and Y, both defined on $(0, \alpha)$ for some $\alpha > 0$. Define $\widetilde{M}(t)$ to be the supremum of $\left\|\dot{Y}(u) - \dot{X}(u)\right\|$ over $u \in [0, t)$. To proceed, we need a simple auxiliary lemma.

Lemma 7. *For any $t \in (0, \alpha)$, we have*

$$\|\nabla f(Y(t)) - \nabla f(X(t))\| \leq Lt\widetilde{M}(t)$$

Proof of Lemma 7. By Lipschitz continuity of the gradient, one has

$$\|\nabla f(Y(t)) - \nabla f(X(t))\| \le L\|Y(t) - X(t)\| = L\Big\|\int_0^t \dot{Y}(u) - \dot{X}(u)\mathrm{d}u + Y(0) - X(0)\Big\|$$
$$\le L\int_0^t \Big\|\dot{Y}(u) - \dot{X}(u)\Big\|\,\mathrm{d}u \le Lt\widetilde{M}(t)$$

□

Now we prove Lemma 6.

Proof of Lemma 6. Similar to the proof of Lemma 4, we get

$$t^3(\dot{Y}(t) - \dot{X}(t)) = -\int_0^t u^3(\nabla f(Y(u)) - \nabla f(X(u)))\mathrm{d}u$$

Applying Lemma 7 gives

$$t^3\Big\|\dot{Y}(t) - \dot{X}(t)\Big\| \le \int_0^t Lu^4\widetilde{M}(u)\mathrm{d}u \le \frac{1}{5}Lt^5\widetilde{M}(t)$$

which can be simplified as $\Big\|\dot{Y}(t) - \dot{X}(t)\Big\| \le \frac{Lt^2\widetilde{M}(t)}{5}$. Thus, for any $t' \le t$ it is true that $\Big\|\dot{Y}(t') - \dot{X}(t')\Big\| \le \frac{Lt^2\widetilde{M}(t)}{5}$. Taking the supremum of $\Big\|\dot{Y}(t') - \dot{X}(t')\Big\|$ over $t' \in (0,t)$ gives $\widetilde{M}(t) \le \frac{Lt^2\widetilde{M}(t)}{5}$. Therefore $\widetilde{M}(t) = 0$ for $t < \alpha \wedge \sqrt{\frac{5}{L}}$, which is equivalent to saying $\dot{Y} = \dot{X}$ on $\Big[0, \alpha \wedge \sqrt{\frac{5}{L}}\Big)$. With the same initial value $Y(0) = X(0) = y_0$ and the same gradient, we conclude that X and Y are identical on $\Big(0, \alpha \wedge \sqrt{\frac{5}{L}}\Big)$, a contradiction. □

Given all of the aforementioned lemmas, the proof of Theorem 1 is simply combining 1 and 6.

3.5 Proof of Theorem 2

Identifying $\sqrt{s} = \Delta t$, the comparison between (5) and (10) reveals that Nesterov's scheme is a discrete scheme for numerically integrating the ODE (4a). However, its singularity of the damping coefficient at $t = 0$ leads to the nonexistence of off-the-shelf ODE theory for proving Theorem 2. To address this difficulty, we use the smoothed ODE (12) to approximate the original one; then bound the difference between Nesterov's scheme and the forward Euler scheme of (12), which may take the following form:

$$\begin{aligned} Y_{k+1}^\delta &= Y_k^\delta + \Delta t Z_k^\delta \\ Z_{k+1}^\delta &= \Big(1 - \frac{3\Delta t}{\max\{\delta, k\Delta t\}}\Big)Z_k^\delta - \Delta t\nabla f(Y_k^\delta) \end{aligned} \tag{13}$$

with $Y_0^\delta = y_0$ and $Z_0^\delta = 0$.

Lemma 8. *With step size $\Delta t = \sqrt{s}$, for any $T > 0$ we have*

$$\max_{1 \le k \le \frac{T}{\sqrt{s}}} \|Y_k^\delta - y_k\| \le C\delta^2 + o_s(1)$$

for some constant C.

Proof of Lemma 8. Let $z_k = \frac{y_{k+1}-y_k}{\sqrt{s}}$. Then Nesterov's scheme is equivalent to

$$\begin{aligned} y_{k+1} &= y_k + \sqrt{s} z_k \\ z_{k+1} &= \Big(1 - \frac{3}{k+3}\Big) z_k - \sqrt{s}\nabla f\Big(y_k + \frac{2k+3}{k+3}\sqrt{s} z_k\Big) \end{aligned} \tag{14}$$

Denote by $a_k = \|Y_k^\delta - y_k\|, \quad b_k = \|Z_k^\delta - z_k\|$, whose initial values are $a_0 = 0$ and $b_0 = \|\nabla f(y_0)\|\sqrt{s}$. The idea of this proof is to bound a_k via simultaneously estimating a_k and b_k. By comparing (13) and (14), we get the iterative relationship for a_k: $a_{k+1} \le a_k + \sqrt{s} b_k$. Denoting by $S_k = b_0 + b_1 + \cdots + b_k$, this yields

$$a_k \le \sqrt{s} S_{k-1} \tag{15}$$

Similarly, for sufficiently small s we get

$$\begin{aligned} b_{k+1} &\le \Big|1 - \frac{3}{k \vee \frac{\delta}{\sqrt{s}}}\Big| b_k + L\sqrt{s} a_k + \Big(\Big|\frac{3}{k+3} - \frac{3}{k \vee \frac{\delta}{\sqrt{s}}}\Big| + 2Ls\Big)\|z_k\| \\ &\le b_k + L\sqrt{s} a_k + \Big(\Big|\frac{3}{k+3} - \frac{3}{k \vee \frac{\delta}{\sqrt{s}}}\Big| + 2Ls\Big)\|z_k\| \end{aligned}$$

To upper bound $\|z_k\|$, denoting by C_1 the supremum of $\sqrt{2L(f(x_k) - f^\star)}$ over all k and s, we have

$$\|z_k\| \le \frac{k-1}{k+2}\|z_{k-1}\| + \sqrt{s}\|\nabla f(x_k)\| \le \|z_{k-1}\| + C_1\sqrt{s}$$

which gives $\|z_k\| \le C_1(k+1)\sqrt{s}$. Hence,

$$\Big(\Big|\frac{3}{k+3} - \frac{3}{k \vee \frac{\delta}{\sqrt{s}}}\Big| + 2Ls\Big)\|z_k\| \le \begin{cases} C_2\sqrt{s}, & k \le \frac{\delta}{\sqrt{s}} \\ \frac{C_2\sqrt{s}}{k} < \frac{C_2 s}{\delta}, & k > \frac{\delta}{\sqrt{s}} \end{cases}$$

Making use of (15) gives

$$b_{k+1} \le \begin{cases} b_k + LsS_{k-1} + C_2\sqrt{s}, & k \le \frac{\delta}{\sqrt{s}} \\ b_k + LsS_{k-1} + \frac{C_2 s}{\delta}, & k > \frac{\delta}{\sqrt{s}} \end{cases} \tag{16}$$

By induction on k, for $k \le \frac{\delta}{\sqrt{s}}$ it holds that

$$b_k \le \frac{C_1 Ls + C_2 + (C_1 + C_2)\sqrt{Ls}}{2\sqrt{L}}(1 + \sqrt{Ls})^{k-1} - \frac{C_1 Ls + C_2 - (C_1 + C_2)\sqrt{Ls}}{2\sqrt{L}}(1 - \sqrt{Ls})^{k-1}$$

Hence,

$$S_k \le \frac{C_1 Ls + C_2 + (C_1 + C_2)\sqrt{Ls}}{2L\sqrt{s}}(1 + \sqrt{Ls})^{k} + \frac{C_1 Ls + C_2 - (C_1 + C_2)\sqrt{Ls}}{2L\sqrt{s}}(1 - \sqrt{Ls})^{k} - \frac{C_2}{L\sqrt{s}}$$

Letting $k^\star = \lfloor \frac{\delta}{\sqrt{s}} \rfloor$, we get

$$\limsup_{s \to 0} \sqrt{s} S_{k^\star - 1} \le \frac{C_2 \exp\left(\delta\sqrt{L}\right) + C_2 \exp\left(-\delta\sqrt{L}\right) - 2C_2}{2L} = O(\delta^2)$$

which allows us to conclude that

$$a_k \le \sqrt{s} S_{k-1} = O(\delta^2) + o_s(1) \tag{17}$$

for all $k \le \frac{\delta}{\sqrt{s}}$.

Next, we bound b_k for $k > k^\star = \lfloor \frac{\delta}{\sqrt{s}} \rfloor$. To this end, we consider the worst case of (16), that is,

$$b_{k+1} = b_k + LsS_{k-1} + \frac{C_2 s}{\delta}$$

for $k > k^\star$ and $S_{k^\star} = S_{k^\star+1} = \frac{C_3\delta^2}{\sqrt{s}} + o_s(\frac{1}{\sqrt{s}})$ for some sufficiently large C_3. In this case, $\frac{C_2 s}{\delta} < sS_{k-1}$ for sufficiently small s. Hence, the last display gives

$$b_{k+1} \le b_k + (L+1)sS_{k-1}$$

By induction, we get

$$S_k \le \frac{\frac{C_3\delta^2}{\sqrt{s}} + o_s(\frac{1}{\sqrt{s}})}{2} \left((1 + \sqrt{(L+1)s})^{k-k^\star} + (1 - \sqrt{(L+1)s})^{k-k^\star} \right)$$

Letting $k^\diamond = \lfloor \frac{T}{\sqrt{s}} \rfloor$, we further get

$$\limsup_{s \to 0} \sqrt{s} S_{k^\diamond} \le \frac{C_3\delta^2 (\exp\left((T-\delta)\sqrt{L+1}\right) + \exp\left(-(T-\delta)\sqrt{L+1}\right))}{2} = O(\delta^2)$$

which yields

$$a_k \le \sqrt{s} S_{k-1} = O(\delta^2) + o_s(1)$$

for $k^\star < k \le k^\diamond$. Last, combining (17) and the last display, we get the desired result.

□

Now we turn to the proof of Theorem 2.

Proof of Theorem 2. Note the triangular inequality

$$\|y_k - Y(k\sqrt{s})\| \le \|y_k - Y_k^\delta\| + \|Y_k^\delta - Y_\delta(k\sqrt{s})\| + \|Y_\delta(k\sqrt{s}) - Y(k\sqrt{s})\|$$

where $Y_\delta(\cdot)$ is the solution to the smoothed ODE (12). The proof of Lemma 1 implies that, we can choose a sequence $\delta_m \to 0$ such that

$$\sup_{0 \le t \le T} \|Y_{\delta_m}(t) - Y(t)\| \to 0$$

The second term $\|Y_k^{\delta_m} - Y_{\delta_m}(k\sqrt{s})\|$ will uniformly vanish as $s \to 0$ and so does the first term $\|y_k - X_k^{\delta_m}\|$ if first $s \to 0$ and then $\delta_m \to 0$. This completes the proof. □

4 High-Resolution Differential Equations: Accelerating Forward-Backward Algorithms in Strongly Convex Optimization Without Modulus Knowledge

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. 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. 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`). Additionally, with a small modification, the accelerated convergence rate for both `AGD` 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 `AGD`, proposed in [Nes83, Nes18], 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 its practical utility in real-world contexts. In stark contrast, the original `AGD` 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 `AGD`. Whether these algorithms, `AGD` and `FISTA`, can achieve linear convergence for strongly convex functions is currently recognized as an open question, as outlined in [CP16, Appendix B]. As a starting point for our analysis, we scrutinize the numerical pattern exhibited in Figure 2, which delineates how the square of the proximal subgradient norm varies across successive iterations.

A quadratic analysis: `RAVE-GD` iteration scheme

Consider our algorithm the case $\lambda = 1$—dubbed as Recursive Averaging Gradient Descent (`RAVE-GD`)—for quadratics $\frac{\gamma}{2}(z - z^\star)^2$

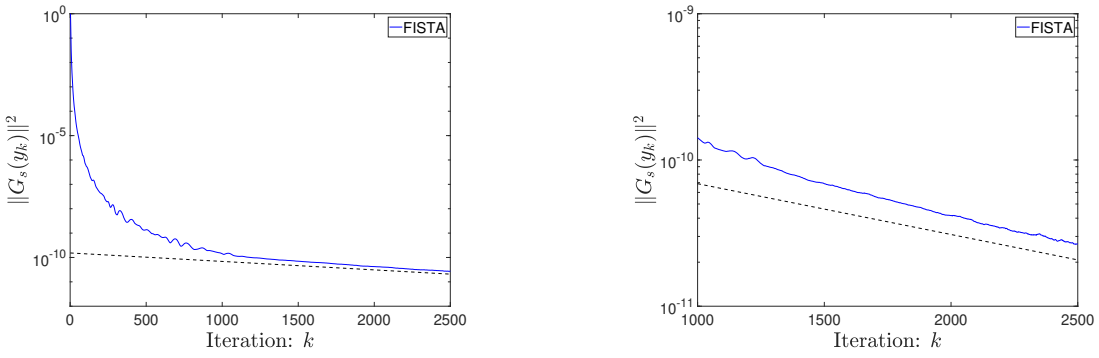
$$\begin{pmatrix} x_k - z^\star \\ y_k - z^\star \end{pmatrix} = \begin{pmatrix} 0 & 1 - \gamma s \\ -\frac{k-1}{k+1} & \frac{2k}{k+1} \cdot (1 - \gamma s) \end{pmatrix} \begin{pmatrix} x_{k-1} - z^\star \\ y_{k-1} - z^\star \end{pmatrix}$$

An eigenvalue analysis yields:

$$\bar{\lambda}_\pm \to 1 - \gamma s \pm i\sqrt{\gamma s(1 - \gamma s)} \quad \text{as} \quad k \to \infty$$

Another formulation for analysis leads to the linear operation:

$$\begin{pmatrix} k(x_k - z^\star) \\ (k+1)(y_k - z^\star) \end{pmatrix} = \begin{pmatrix} 0 & 1 - \gamma s \\ -1 & 2(1 - \gamma s) \end{pmatrix} \begin{pmatrix} (k-1)(x_{k-1} - z^\star) \\ k(y_{k-1} - z^\star) \end{pmatrix}$$



(a) Iteration: Start From $k = 0$ (b) Iteration: Start From $k = 1000$

Figure 2. Iterative progression of the square of the proximal subgradient norm throughout the application of `FISTA` for image deblurring.

The matrix is diagonalizable with complex entries when $1 - \gamma s \in (0, 1)$, leading to eigenvalues:

$$\lambda_{\pm} = 1 - \gamma s \pm i\sqrt{1 - \gamma s - (1 - \gamma s)^2}$$

This style of analysis follows the foundational work found in [Pol87, Chapter 2], demonstrating the robust performance of `RAVE-GD` in quadratic settings:

Proposition 3 ([Pol87])**.** *Let (possibly nonlinear) operator* $\mathbf{A} : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ *be continuously differentiable and let* ω^* *one of its fixed points. Suppose there exists* $\rho^* > 0$ *such that,*

$$\rho\left(\partial \mathbf{A}\left(\mathbf{z}^*\right)\right) \leq \rho^* < 1$$

For $\mathbf{z}_0$ *close to* $\mathbf{z}^*$*,* $\mathbf{z}_t$ *converges linearly to* ω^* *at a rate* $\mathcal{O}\left((\rho^* + \epsilon)^t\right)$*.*

Additionally, refer to the detailed analysis of nonconvex accelerated gradient descent in Appendix C of [JNJ18]. By considering a multidimensional quadratic we have as a corollary

Corollary 1. *We have for any given* $\epsilon > 0$*, for* k *sufficiently large*

$$k^2 \|y_k - z^\star\|^2 \leq \left\| \begin{pmatrix} k(x_k - z^\star) \\ (k+1)(y_k - z^\star) \end{pmatrix} \right\|^2 \leq (1 - \mu s + \epsilon)^k \left\| \begin{pmatrix} 0 \\ y_0 - z^\star \end{pmatrix} \right\|^2 = (1 - \mu s + \epsilon)^k \|y_0 - z^\star\|^2$$

exemplifying a composition of inverse quadratic and linear convergence.

The remainder of this section is organized as follows. In Section 4, we introduce basic notations and preliminaries related to convex and strongly convex functions, as well as the high-resolution ODE framework. Section 4.1 focuses on the derivation and analysis of the Lyapunov function using the gradient-correction high-resolution ODE. In Section 4.2, we prove the linear convergence of `AGD` for strongly convex functions, extending these results to `FISTA` and composite optimization problems, and concludes with potential directions for future research.

Preliminaries on composites

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*

$$\operatorname{prox}_{sg}(x) := \arg\min_{y \in \mathbb{R}^d} \left\{ \frac{1}{2s} \|y - (x - s\nabla f(x))\|^2 + g(y) \right\} \tag{18}$$

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

$$G_s(x) := \frac{x - \operatorname{prox}_{sg}(x)}{s} \tag{19}$$

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

When g simplifies to the ℓ_1-norm, i.e., $g(x) = \overline{\lambda}\|x\|_1$,[3] we can derive the closed-form silution for the s-proximal operator (18) at any $x \in \mathbb{R}^d$ for the particular instance as

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

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

4.1 The gradient-correction high-resolution ODE, $\lambda = 2$

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 `AGD`, the gradient-correction high-resolution ODE is expressed as:

$$\ddot{Y} + \frac{3}{t}\dot{Y} + \sqrt{s}\nabla^2 f(Y)\dot{Y} + \left(1 + \frac{3\sqrt{s}}{2t}\right)\nabla f(Y) = 0 \tag{20}$$

with any initial $(Y(0), \dot{Y}(0)) = (y_0, v_0) \in \mathbb{R}^d \times \mathbb{R}^d$. 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

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

  $$\mathcal{E}_{\mathbf{mix}} = \frac{1}{2}\|t\dot{Y} + 2(Y - z^*) + t\sqrt{s}\nabla f(Y)\|^2 \tag{21}$$

  By using the gradient-correction high-resolution ODE (20), we can calculate its time derivative as

  $$\begin{aligned} \frac{d\mathcal{E}_{\mathbf{mix}}}{dt} &= \left\langle t\dot{Y} + 2(Y - z^*) + t\sqrt{s}\nabla f(Y), -\left(t + \frac{\sqrt{s}}{2}\right)\nabla f(Y) \right\rangle \\ &= \underbrace{-t\left(t + \frac{\sqrt{s}}{2}\right)\left\langle \dot{Y}, \nabla f(Y) \right\rangle}_{\mathbf{I}} - (2t + \sqrt{s})\left\langle \nabla f(Y), Y - z^* \right\rangle \\ &\quad - \sqrt{s}t\left(t + \frac{\sqrt{s}}{2}\right)\|\nabla f(Y)\|^2 \end{aligned} \tag{22}$$

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

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

$$\mathcal{E}_{\mathbf{kin}} = \frac{\tau(t)}{2}\|\dot{Y}\|^2 \tag{23}$$

Taking into account the gradient-correction high-resolution ODE (20), we can calculate the time derivative of the kinetic energy as

$$\begin{aligned}\frac{d\mathcal{E}_{\mathbf{kin}}}{dt} &= \tau(t)\langle \ddot{Y}, \dot{Y}\rangle + \frac{\dot{\tau}(t)}{2}\|\dot{Y}\|^2 \\ &= -\left(\frac{3\tau(t)}{t} - \frac{\dot{\tau}(t)}{2}\right)\|\dot{Y}\|^2 - \sqrt{s}\tau(t)\dot{Y}^\top \nabla^2 f(Y)\dot{Y} \\ &\quad \underbrace{-\tau(t)\left(1+\frac{3\sqrt{s}}{2t}\right)\langle \dot{Y}, \nabla f(Y)\rangle}_{\mathbf{II}}\end{aligned} \tag{24}$$

To simplify the coefficient of $\|\dot{Y}\|^2$, 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 $\alpha = 2$.

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

$$\mathcal{E}_{\mathbf{pot}} = 2t(t+\sqrt{s})\left(f(Y) - f(z^*)\right) \tag{25}$$

Proceeding with the exploration of the gradient-correction high-resolution ODE (20), we calculate its time derivative as follows:

$$\frac{d\mathcal{E}_{\mathbf{pot}}}{dt} = \left(4t + 2\sqrt{s}\right)\left(f(Y) - f(z^*)\right) + \underbrace{2t(t+\sqrt{s})\langle \dot{Y}, \nabla f(Y)\rangle}_{\mathbf{III}} \tag{26}$$

where the resultant term **III** in (26) ensures that the combination of **I**, **II**, and **III** satisfies the equality $\mathbf{I} + \mathbf{II} + \mathbf{III} = 0$.

By integrating the mixed energy (21), the kinetic energy (23) and the potential energy (25), we arrive at the following Lyapunov function as

$$\mathcal{E} = 2t(t+\sqrt{s})\left(f(Y) - f(z^*)\right) + \frac{t^2}{2}\|\dot{Y}\|^2 + \frac{1}{2}\|t\dot{Y} + 2(Y - z^*) + t\sqrt{s}\nabla f(Y)\|^2 \tag{27}$$

Theorem 4. *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 $Y = Y(t)$ to the gradient-correction high-resolution ODE* (20) *satisfies*

$$\begin{cases} f(Y) - f(z^*) \le \dfrac{\mathcal{E}(T)}{2t(t+\sqrt{s})} e^{-\frac{\mu\sqrt{s}}{4}(t-T)} \\ \|\nabla f(Y)\|^2 \le \dfrac{L\mathcal{E}(T)}{t(t+\sqrt{s})} e^{-\frac{\mu\sqrt{s}}{4}(t-T)} \end{cases} \tag{28}$$

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

Proof of Theorem 4. To prove the theorem, we begin with the Lyapunov function (27). Its time derivative can be calculated by summing up (22), (24) and (26) as

$$\begin{aligned}\frac{d\mathcal{E}}{dt} \leq & \left(4t + 2\sqrt{s}\right)\left(f(Y) - f(z^*)\right) - \left(2t + \sqrt{s}\right)\langle \nabla f(Y), Y - z^*\rangle \\ & - 2t\|\dot{Y}\|^2 - \sqrt{s}t^2\dot{Y}^\top \nabla^2 f(Y)\dot{Y} - \sqrt{s}t\left(t + \frac{\sqrt{s}}{2}\right)\|\nabla f(Y)\|^2 \end{aligned} \tag{29}$$

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

$$\langle \nabla f(Y), Y - z^*\rangle \geq f(Y) - f(z^*) + \frac{\mu}{2}\|Y - z^*\|^2 \tag{30}$$

By substituting the μ-strongly convex inequality (30) into the earlier time derivative (29), we obtain

$$\begin{aligned}\frac{d\mathcal{E}}{dt} \leq & \left(2t + \sqrt{s}\right)\left(f(Y) - f(z^*)\right) - \frac{\mu\left(2t + \sqrt{s}\right)}{2}\|Y - z^*\|^2 \\ & - 2t\|\dot{Y}\|^2 - \sqrt{s}t^2\dot{Y}^\top \nabla^2 f(Y)\dot{Y} - \sqrt{s}t\left(t + \frac{\sqrt{s}}{2}\right)\|\nabla f(Y)\|^2 \end{aligned} \tag{31}$$

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

$$\mathcal{E} \leq 2t\left(t + \sqrt{s}\right)\left(f(Y) - f(z^*)\right) + 2t^2\|\dot{Y}\|^2 + 6\|Y - z^*\|^2 + \frac{3}{2}st^2\|\nabla f(Y)\|^2 \tag{32}$$

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

$$\dot{Y}^\top \nabla^2 f(Y)\dot{Y} \geq \mu\|\dot{Y}\|^2 \tag{33a}$$

$$\|\nabla f(Y)\|^2 \geq 2\mu\left(f(Y) - f(z^*)\right) \tag{33b}$$

By substituting (33a) and (33b) into (31), we can obtain the time derivative as

$$\begin{aligned}\frac{d\mathcal{E}}{dt} \leq & -\mu\sqrt{s}\left(t - \frac{1}{\mu\sqrt{s}}\right)\left(t + \sqrt{s}\right)\left(f(Y) - f(z^*)\right) - \mu\sqrt{s}t^2\|\dot{Y}\|^2 \\ & -\mu t\|Y - z^*\|^2 - \frac{\sqrt{s}t}{2}\left(t - \frac{1}{\mu\sqrt{s}}\right)\|\nabla f(Y)\|^2 \\ \leq & -\frac{3\mu t\left(\sqrt{s}t + s\right)}{4}\left(f(Y) - f(z^*)\right) - \mu\sqrt{s}t^2\|\dot{Y}\|^2 \\ & -\frac{4}{\sqrt{s}}\|Y - z^*\|^2 - \frac{3\sqrt{s}t^2}{8}\|\nabla f(Y)\|^2 \end{aligned} \tag{34}$$

where the last inequality is supported by $t \geq T = \frac{4}{\mu\sqrt{s}}$. By matching the corresponding terms in (32) and (34), 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:

$$\|\nabla f(Y)\|^2 \leq 2L\left(f(Y) - f(z^*)\right)$$

Hence, the proof is complete with some elementary calculations.

□

4.2 Linear convergence via a novel discrete Lyapunov function, $\lambda = 1$

In this section, we develop a novel discrete Lyapunov function aimed at deducing the linear convergence properties of the AGD method when implemented on μ-strongly convex functions. Our approach is fundamentally anchored in the theoretical framework of potential and mixed energy. 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 existing techniques.

4.2.1 Smooth optimization with AGD

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

$$\begin{cases} y_{k+1} - y_k = \sqrt{s} v_{k+1} \\ v_{k+1} - v_k = -\dfrac{(1)+1}{k+(1)} \cdot v_k - \sqrt{s}\nabla f(x_k) \end{cases} \tag{35}$$

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

$$x_k = y_k + \frac{k-1}{k+(1)} \cdot \sqrt{s} v_k$$

It is evident that the second iteration of (35) can be neatly reformulated as

$$(k+(1))v_{k+1} - (k-1)v_k = -(k+(1))\sqrt{s}\nabla f(x_k) \tag{36}$$

Building upon this framework, we now proceed to elucidate how the Lyapunov function is systematically constructed by employing the principled approach.

(**I**) We here choose to implement the analogous mixed energy for its computational efficacy. The mixed energy is delineated as:

$$\mathcal{E}_{\mathbf{mix}}(k) = \frac{1}{2}\|\sqrt{s}(k-1)v_k + (1)(y_k - z^*)\|^2 \tag{37}$$

By employing the initial step of the phase-space representation (35) and the reformulated expression provided in (36), we establish the subsequent equality as

$$\begin{aligned} &\sqrt{s}kv_{k+1} + (1)(y_{k+1} - z^*) - \sqrt{s}(k-1)v_k + (1)(y_k - z^*) \\ =&\sqrt{s}kv_{k+1} + (1)(y_{k+1} - y_k) - \sqrt{s}(k-1)v_k \\ =&\sqrt{s}(k+(1))v_{k+1} - \sqrt{s}(k-1)v_k \\ =&-(k+(1))s\nabla f(x_k) \end{aligned} \tag{38}$$

where the second equality is derived from the initial step of the phase-space representation (35) and the last equality follows from equation (36). We are now in a position to determine the iterative difference of mixed energy, which is calculated as follows:

$$\begin{aligned} &\mathcal{E}_{\mathbf{mix}}(k+1) - \mathcal{E}_{\mathbf{mix}}(k) \\ &\qquad = \left\langle \sqrt{s}kv_{k+1} + (1)(y_{k+1} - z^*), -s(k+(1))\nabla f(x_k) \right\rangle \end{aligned}$$

$$
\begin{aligned}
&-\frac{s^2(k+(1))^2}{2}\left\|\nabla f(x_k)\right\|^2 \\
=&-\underbrace{s^{\frac{3}{2}}k(k+(1))\left\langle\nabla f(x_k),v_{k+1}\right\rangle}_{\mathbf{I}}-s(1)(k+(1))\left\langle\nabla f(x_k),x_k-z^*\right\rangle \\
&-\frac{s^2}{2}(k-(1))(k+(1))\left\|\nabla f(x_k)\right\|^2
\end{aligned}
\tag{39}
$$

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

(**II**) We commence our analysis by considering the kinetic energy. We formulate the kinetic energy with an iteration-varying coefficient, which is articulated as follows:

$$
\mathcal{E}_{\mathbf{kin}}(k)=\frac{s\tau(k)}{2}\|v_k\|^2 \tag{40}
$$

By employing the expression from (36), the iterative difference of kinetic energy can be calculated as follows:

$$
\begin{aligned}
&\mathcal{E}_{\mathbf{kin}}(k+1)-\mathcal{E}_{\mathbf{kin}}(k) \\
&\quad=\frac{s\tau(k+1)}{2}\left\|v_{k+1}\right\|^2-\frac{s\tau(k)}{2}\left(\frac{k+(1)}{k-1}\right)^2\left\|v_{k+1}+\sqrt{s}\nabla f(x_k)\right\|^2 \\
&\quad=\frac{s}{2}\left[\tau(k+1)-\tau(k)\left(\frac{k+(1)}{k-1}\right)^2\right]\left\|v_{k+1}\right\|^2 \\
&\qquad-s^{\frac{3}{2}}\tau(k)\left(\frac{k+(1)}{k-1}\right)^2\left\langle\nabla f(x_k),v_{k+1}\right\rangle \\
&\qquad-\frac{s^2\tau(k)}{2}\left(\frac{k+(1)}{k-1}\right)^2\left\|\nabla f(x_k)\right\|^2
\end{aligned}
\tag{41}
$$

To ease computation as indicated by (41), an intuitive choice for $\tau(k)$ would be $\tau(k)=(k-1)^2$. This selection simplifies the iterative difference (41), which can be succinctly expressed as

$$
\begin{aligned}
\mathcal{E}_{\mathbf{kin}}(k+1)-\mathcal{E}_{\mathbf{kin}}(k)=&-\frac{s(1)(2k+(1))}{2}\|v_{k+1}\|^2-\frac{s^2(k+(1))^2}{2}\left\|\nabla f(x_k)\right\|^2 \\
&-\underbrace{s^{\frac{3}{2}}(k+(1))^2\left\langle\nabla f(x_k),v_{k+1}\right\rangle}_{\mathbf{II}}
\end{aligned}
\tag{42}
$$

(**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:

$$
\mathcal{E}_{\mathbf{pot}}(k)=s\gamma(k)\left(f(y_k)-f(z^*)\right) \tag{43}
$$

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 [Nes18]:

$$
f(y-s\nabla f(y))-f(x)
$$

$$\leq \left\langle \nabla f(y), y - x \right\rangle - \frac{\mu}{2}\|y - x\|^2 - \left(s - \frac{Ls^2}{2}\right)\|\nabla f(y)\|^2 \tag{44}$$

Applying (44) with y_{k+1} and y_k, we derive:

$$\begin{aligned}
&f(y_{k+1}) - f(y_k) \\
&\quad \leq \left\langle \nabla f(x_k), x_k - y_k \right\rangle - \frac{\mu}{2}\|x_k - y_k\|^2 - \left(s - \frac{s^2 L}{2}\right)\|\nabla f(x_k)\|^2 \\
&\quad = \left\langle \nabla f(x_k), \sqrt{s} v_{k+1} + s\nabla f(x_k) \right\rangle - \frac{\mu}{2}\left\|\sqrt{s} v_{k+1} + s\nabla f(x_k)\right\|^2 \\
&\qquad - \left(s - \frac{s^2 L}{2}\right)\|\nabla f(x_k)\|^2 \\
&\quad = \sqrt{s}(1 - \mu s)\left\langle \nabla f(x_k), v_{k+1} \right\rangle + \frac{s^2(L - \mu)}{2}\|\nabla f(x_k)\|^2 - \frac{\mu s}{2}\|v_{k+1}\|^2
\end{aligned} \tag{45}$$

where the first equality is justified by the initial iteration in (35) in conjunction with the gradient step of `AGD`. Subsequently, for the potential energy (43), the iterative difference can be estimated as follows:

$$\begin{aligned}
&\mathcal{E}_{\mathbf{pot}}(k+1) - \mathcal{E}_{\mathbf{pot}}(k) \\
&= s\gamma(k+1)\left(f(y_{k+1}) - f(z^*)\right) - s\gamma(k)\left(f(y_k) - f(z^*)\right) \\
&= s\gamma(k)\left(f(y_{k+1}) - f(y_k)\right) + s(\gamma(k+1) - \gamma(k))\left(f(y_{k+1}) - f(z^*)\right) \\
&\leq s\gamma(k)\left(\underbrace{\sqrt{s}(1 - \mu s)\left\langle \nabla f(x_k), v_{k+1} \right\rangle}_{\mathbf{III}} + \frac{s^2(L - \mu)}{2}\|\nabla f(x_k)\|^2 - \frac{\mu s}{2}\|v_{k+1}\|^2\right) \\
&\qquad + s(\gamma(k+1) - \gamma(k))\left(f(y_{k+1}) - f(z^*)\right)
\end{aligned} \tag{46}$$

Coupled with the earlier derivations, (39) and (42), it becomes imperative to fine-tune the coefficient $\gamma(k)$ within the iterative difference (46) to cancel out the term $\left\langle \nabla f(x_k), v_{k+1} \right\rangle$. 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:

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

Integrating the mixed energy (37), the kinetic energy (40) and the potential energy (43) — each bolstered by iteration-varying coefficients — we construct the noval discrete Lyapunov function as depicted below:

$$\begin{aligned}
\mathcal{E}(k) =& \frac{s(k + (1))(2k + (1))}{1 - \mu s}\left(f(y_k) - f(z^*)\right) \\
&+ \frac{s(k-1)^2}{2}\|v_k\|^2 + \frac{1}{2}\left\|\sqrt{s}(k-1)v_k + (1)(y_k - z^*)\right\|^2
\end{aligned} \tag{47}$$

Within this novel Lyapunov function (47) 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 5. *Let $f \in \mathcal{S}^1_{\mu,L}$. Given any step size $0 < s < \frac{1}{L}$, there exists a positive integer $K := K(L, \mu, s, (1))$ such that the iterative sequence $\{(y_k, x_k)\}_{k=0}^{\infty}$ generated by* `AGD` *with any initial $y_0 = x_0 \in \mathbb{R}^d$ satisfies*

$$\begin{cases} f(y_k) - f(z^*) \leq \dfrac{\mathcal{E}(K)}{s(k+(1))(2k+(1))\left[1+(1-Ls)\cdot\frac{\mu s}{4}\right]^{k-K}} \\ \|\nabla f(x_k)\|^2 \leq \dfrac{4\mathcal{E}(K)}{s^2(1-Ls)(k+(1))(2k+(1))\left[1+(1-Ls)\cdot\frac{\mu s}{4}\right]^{k-K}} \end{cases} \tag{48}$$

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 (48) in terms of the dimensionless parameter $\alpha = sL$, which falls within the open interval $(0, 1)$. Accordingly, the established bounds may be recast as

$$\begin{cases} f(y_k) - f(z^*) \leq \dfrac{\mathcal{E}(K)}{(k+(1))(2k+(1))\left(1+\frac{\alpha(1-\alpha)}{4}\cdot\frac{\mu}{L}\right)^{k-K}} \\ \|\nabla f(x_k)\|^2 \leq \dfrac{4L^2\mathcal{E}(K)}{\alpha^2(1-\alpha)(k+(1))(2k+(1))\left[1+\frac{\alpha(1-\alpha)}{4}\cdot\frac{\mu}{L}\right]^{k-K}} \end{cases} \tag{49}$$

for any $k \geq K$. The bounds delineated in (49) 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 [Nes18]. 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 5. Given the discrete Lyapunov function defined in (47), we calculate its iterative difference by summing the three iterative differences laid out in (39), (42) and (46). This amalgamation yields:

$$\begin{aligned} &\mathcal{E}(k+1) - \mathcal{E}(k) \\ &\quad = \underbrace{-s(1)(k+(1))\langle \nabla f(x_k), x_k - z^*\rangle}_{\mathbf{IV}} - \frac{s(1)(2k+(1))}{2}\|v_{k+1}\|^2 \\ &\qquad -s^2k(k+(1))\|\nabla f(x_k)\|^2 \\ &\qquad +\frac{s(k+(1))(2k+(1))}{1-\mu s}\left(\frac{s^2(L-\mu)}{2}\|\nabla f(x_k)\|^2 - \frac{\mu s}{2}\|v_{k+1}\|^2\right) \\ &\qquad +\frac{s(4k+3(1)+2)}{1-\mu s}(f(y_{k+1}) - f(z^*)) \end{aligned} \tag{50}$$

To ensure proportional alignment of terms, we exhibit $\|y_{k+1} - z^*\|^2$ and the other terms in (50), thereby replacing Term **IV**. After inserting y_{k+1} and z^* into (44), we deduce:

$$f(y_{k+1}) - f(z^*) \leq \langle \nabla f(x_k), x_k - z^*\rangle - \frac{s}{2}\|\nabla f(x_k)\|^2 - \frac{\mu}{2}\|x_k - z^*\|^2$$

$$\leq \left\langle \nabla f(x_k), x_k - z^* \right\rangle - \frac{s}{2} \left\| \nabla f(x_k) \right\|^2 - \frac{\mu}{2} \left\| y_{k+1} - z^* \right\|^2 \tag{51}$$

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

$$\begin{aligned} \left\| y_{k+1} - z^* \right\|^2 &= \left\| x_k - z^* - s\nabla f(x_k) \right\|^2 \\ &= \left\| x_k - z^* \right\|^2 - 2s\left\langle \nabla f(x_k), x_k - z^* \right\rangle + s^2 \left\| \nabla f(x_k) \right\|^2 \leq \left\| x_k - z^* \right\|^2 \end{aligned}$$

where the last inequality holds for any $s < \frac{1}{L}$ due to the first inequality of (51). By substituting inequality (51) into the iterative difference (50), we attain:

$$\begin{aligned} \mathcal{E}(k+1) - \mathcal{E}(k) \leq & s\left(\frac{4k + 3(1) + 2}{1 - \mu s} - (1)(k + (1)) \right) \left(f(y_{k+1}) - f(z^*) \right) \\ & - \frac{s}{2} \left(\frac{\mu s}{1 - \mu s} (k + (1))(2k + (1)) + (1)(2k + (1)) \right) \left\| v_{k+1} \right\|^2 \\ & - \frac{\mu s}{2} \cdot (1)(k + (1)) \left\| y_{k+1} - z^* \right\|^2 \\ & - \frac{s^2}{2} \cdot \frac{1 - Ls}{1 - \mu s} \cdot (k + (1))(2k + (1)) \left\| \nabla f(x_k) \right\|^2 \end{aligned} \tag{52}$$

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

$$\begin{aligned} \mathcal{E}(k+1) \leq & \frac{s(k + (1) + 1)(2k + (1) + 2)}{1 - \mu s} (f(y_{k+1}) - f(z^*)) \\ & + \frac{3}{2} s k^2 \left\| v_{k+1} \right\|^2 + (1)^2 \left\| y_{k+1} - z^* \right\|^2 \\ = & \frac{s\left[(k + (1))(2k + (1)) + (4k + 3(1) + 2) \right]}{1 - \mu s} (f(y_{k+1}) - f(z^*)) \\ & + \frac{3}{2} s k^2 \left\| v_{k+1} \right\|^2 + (1)^2 \left\| y_{k+1} - z^* \right\|^2 \end{aligned} \tag{53}$$

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

$$\left\| \nabla f(y) \right\|^2 \geq 2\mu \left(f(y - s\nabla f(y)) - f(z^*) \right) \tag{54}$$

for any $y \in \mathbb{R}^d$. Applying this to x_k of `AGD` within (54), we arrive at:

$$\left\| \nabla f(x_k) \right\|^2 \geq 2\mu \left(f(y_{k+1}) - f(z^*) \right) \tag{55}$$

Incorporating inequality (55) into the iterative difference (52), we deduce:

$$\begin{aligned} & \mathcal{E}(k+1) - \mathcal{E}(k) \\ \leq & - \frac{s}{1 - \mu s} \left(\frac{\mu s(1 - Ls)}{2} (k + (1))(2k + (1)) - (4k + 3(1) + 2) + (1)(1 - \mu s)(k + (1)) \right) \\ & \cdot \left(f(y_{k+1}) - f(z^*) \right) - \frac{\mu s}{2} (1)(k + (1)) \left\| y_{k+1} - z^* \right\|^2 \\ & - \frac{s}{2} \left(\frac{\mu s}{1 - \mu s} (k + (1))(2k + (1)) + (1)(2k + (1)) \right) \left\| v_{k+1} \right\|^2 \end{aligned}$$

$$
\begin{aligned}
&- \frac{s^2}{4}\frac{1-Ls}{1-\mu s}(k+(1))(2k+(1))\left\|\nabla f(x_k)\right\|^2 \\
\leq &- \frac{s}{1-\mu s}\left(\frac{\mu s(1-Ls)}{2}(k+(1))(2k+(1)) - (4k+3(1)+2) + (1)(1-\mu s)(k+(1))\right) \\
&\cdot \left(f(y_{k+1}) - f(z^*)\right) - \frac{\mu s(1)^2}{2}\left\|y_{k+1}-z^*\right\|^2 \\
&- \mu s^2 k^2 \left\|v_{k+1}\right\|^2 - \frac{s^2}{4}(1-Ls)(k+(1))(2k+(1))\left\|\nabla f(x_k)\right\|^2
\end{aligned} \tag{56}
$$

To align the terms between the right-hand side of (56) and (53), we establish the following inequality as

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

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

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

The proof is complete with some elementary operations. □

4.2.2 Composite optimization via FISTA

In this section, we extend the convergence rates of AGD as established in Theorem 5 to include its proximal variant, FISTA. As specified in Definition 1, FISTA utilizes the s-proximal operator (18) and is defined by the following iterative scheme, starting from any initial point $x_0 = y_0 \in \mathbb{R}^d$:

$$
\begin{cases}
y_k = \operatorname{prox}_{sg}\left(x_{k-1} - s\nabla f(x_{k-1})\right) \\
x_k = y_k + \dfrac{k-1}{k+(1)}(y_k - y_{k-1})
\end{cases}
$$

where $s > 0$ denotes the step size. The update on variable x is known as the *forward-backward* (FB) method; see, e.g., [Roc97, FP03]. We can analogously formulate FISTA in a manner akin to AGD using the s-proximal subgradient operator (19), which yields

$$
\begin{cases}
y_k = x_{k-1} - sG_s(x_{k-1}) \\
x_k = y_k + \dfrac{k-1}{k+(1)}(y_k - y_{k-1})
\end{cases} \tag{57}
$$

where the proximal operator $G_s(\cdot)$ replace the gradient operator $\nabla f(\cdot)$ in AGD. Furthermore, by designing the velocity iteration sequence as $v_k = \frac{y_k - y_{k-1}}{\sqrt{s}}$, we reformulate the FISTA updates in a

AGD-esque fashion (57) into a implicit-velocity scheme (phase-space representation), expressed as

$$\begin{cases} y_{k+1} - y_k = \sqrt{s} v_{k+1} \\ v_{k+1} - v_k = -\dfrac{(1)+1}{k+(1)} \cdot v_k - \sqrt{s} G_s(x_k) \end{cases} \tag{58}$$

To derive the convergence rates, we still need to establish a discrete Lyapunov function. Here, we generalize the one previously utilized for smooth functions (47), by substituting the smooth function f with the composite function $\Phi = f + g$, resulting in

$$\begin{aligned} \mathcal{E}(k) =& \frac{s(k+(1))(2k+(1))}{1-\mu s} \left(\Phi(y_k) - \Phi(z^*)\right) \\ &+ \frac{s(k-1)^2}{2} \|v_k\|^2 + \frac{1}{2} \left\| \sqrt{s}(k-1) v_k + (1)(y_k - z^*) \right\|^2 \end{aligned} \tag{59}$$

As outlined in Section 4.2.1, deriving the desired convergence rates for the smooth case predominantly hinges on two key inequalities, (44) and (54). 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 5 into this broader context. For the fundamental inequality (44), its proximal analogue is articulated as

$$\begin{aligned} &\Phi(y - sG_s(y)) - \Phi(x) \\ &\qquad \le \left\langle G_s(y), y - x \right\rangle - \frac{\mu}{2} \|y - x\|^2 - \left(s - \frac{Ls^2}{2} \right) \|G_s(y)\|^2 \end{aligned} \tag{60}$$

The primary challenge lies in extending the μ-strongly convex inequality (54) to the composite function $\Phi = f + g$, for which we establish the subsequent lemma.

Lemma 9. *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* (19)*, satisfies the following inequality:*

$$\|G_s(y)\|^2 \ge 2\mu \left(\Phi(y - sG_s(y)) - \Phi(z^*) \right) \tag{61}$$

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

Proof of Lemma 9. Under the condition that the step size adheres to $0 < s \le \frac{1}{L}$, we invoke the fundamental proximal inequality (60), which simplifies to:

$$\Phi(y - sG_s(y)) - \Phi(x) \le \left\langle G_s(y), y - x \right\rangle - \frac{s}{2} \|G_s(y)\|^2 - \frac{\mu}{2} \|y - x\|^2 \tag{62}$$

for any $x, y \in \mathbb{R}^d$. By reorganizing the terms of (62), we arrive at the expression

$$\Phi(x) \ge \Phi(y - sG_s(y)) + \left\langle G_s(y), x - y \right\rangle + \frac{s}{2} \|G_s(y)\|^2 + \frac{\mu}{2} \|y - x\|^2 \tag{63}$$

for any $x, y \in \mathbb{R}^d$. For succinctness, we denote the right-hand side of (63) as

$$h(x) = \Phi(y - sG_s(y)) + \left\langle G_s(y), x - y \right\rangle + \frac{s}{2} \|G_s(y)\|^2 + \frac{\mu}{2} \|y - x\|^2 \tag{64}$$

for any fixed $y \in \mathbb{R}^d$. Consequently, inequality (63) takes the form $\Phi(x) \ge h(x)$. Considering that z^* is the unique minimizer of the composite function Φ, substituting x with z^* yields:

$$\Phi(z^*) \ge h(z^*) \tag{65}$$

Upon evaluating expression (64), we identify that $h(x)$ embodies a quadratic function whose Hessian is positive definte. This is depicted as:

$$h(x) = \frac{\mu}{2}\left\|x - y + \frac{1}{\mu}G_s(y)\right\|^2 + \Phi(y - sG_s(y)) - \left(\frac{1}{2\mu} - \frac{s}{2}\right)\|G_s(y)\|^2 \tag{66}$$

Incorporating (66) into (65) results in

$$\Phi(z^*) \geq \Phi(y - sG_s(y)) - \left(\frac{1}{2\mu} - \frac{s}{2}\right)\|G_s(y)\|^2 \geq \Phi(y - sG_s(y)) - \frac{1}{2\mu}\|G_s(y)\|^2 \tag{67}$$

By rearranging (67), we complete the proof.

□

With the proximal generalization of the μ-strongly convex inequality, as demonstrated in Lemma 9, we now present the subsequent theorem, which characterizes the convergence rates of `FISTA`.

Theorem 6. *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 $K := K(L, \mu, s, (1))$, such that the iterative sequence $\{(y_k, x_k)\}_{k=0}^{\infty}$ generated by `FISTA`, with any initial $y_0 = x_0 \in \mathbb{R}^d$, satisfies the following inequalities as*

$$\begin{cases} \Phi(y_k) - \Phi(z^*) \leq \dfrac{\mathcal{E}(K)}{s(k+(1))(2k+(1))\left[1 + (1 - Ls)\cdot\frac{\mu s}{4}\right]^{k-K}} \\ \|G_s(x_k)\|^2 \leq \dfrac{4\mathcal{E}(K)}{s^2(1-Ls)(k+(1))(2k+(1))\left[1 + (1 - Ls)\cdot\frac{\mu s}{4}\right]^{k-K}} \end{cases} \tag{68}$$

for any $k \geq K$.

5 Conclusion

In this paper, we explored the continuous-time dynamics of Nesterov's Accelerated Gradient Method (AGD) using a high-resolution ordinary differential equation (ODE) framework. By modeling AGD as a second-order ODE, we provided a deeper understanding of its behavior, explaining key phenomena such as oscillations, damping effects, and fast inverse quadratic convergence. This ODE framework offers a new lens for studying discrete algorithms, drawing approximate equivalence between the discrete scheme and its continuous counterpart.

Our analysis demonstrates that AGD achieves linear convergence in strongly convex settings without requiring prior knowledge of the convexity modulus, making it highly practical for real-world applications. We extended this framework to composite optimization problems, showing that both AGD and its proximal variant, FISTA, retain their acceleration properties even in non-smooth settings. By leveraging a dynamically adapting Lyapunov function and phase-space representation, we provided a robust foundation for understanding and improving the convergence behavior of accelerated methods.

These findings contribute significantly to both the theoretical understanding and practical implementation of gradient-based optimization techniques. The high-resolution ODE framework not only unifies the analysis of smooth and composite optimization but also suggests pathways for the development of new, more efficient algorithms. Future work may focus on refining adaptive step sizes, stopping criteria, and exploring stochastic optimization and large-scale machine learning problems through this continuous-time lens.

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.

[BBC11] Stephen Becker, Jérôme Bobin, and Emmanuel J Candès. NESTA: A fast and accurate first-order method for sparse recovery. *SIAM Journal on Imaging Sciences*, 4(1):1–39, 2011.

[Bec14] Amir Beck. *Introduction to Nonlinear Optimization: Theory, Algorithms, and Applications with MATLAB*. SIAM, 2014.

[BPC+11] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

[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.

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[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.

[DSE12] Hans-Bernd Dürr, Erkin Saka, and Christian Ebenbauer. A smooth vector field for quadratic programming. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pages 2515–2520. IEEE, 2012.

[Fio05] Simone Fiori. Quasi-geodesic neural learning algorithms over the orthogonal group: A tutorial. *Journal of Machine Learning Research*, 6(26):743–781, 2005.

[FP03] Francisco Facchinei and Jong-Shi Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer, 2003.

[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.

[JNJ18] Chi Jin, Praneeth Netrapalli, and Michael I Jordan. Accelerated gradient descent escapes saddle points faster than gradient descent. In *Conference On Learning Theory*, pages 1042–1085. PMLR, 2018.

[Lea22] Jeffery J Leader. *Numerical Analysis and Scientific Computation*. Chapman and Hall/CRC, 2nd edition, 2022.

[LLF20] Zhouchen Lin, Huan Li, and Cong Fang. *Accelerated Optimization for Machine Learning: First-Order Algorithms*. Springer Singapore, 2020.

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

[Nes18] Yurii Nesterov. *Lectures on Convex Optimization*, volume 137 of *Springer Optimization and Its Applications*. Springer Cham, 2018.

[NW99] Jorge Nocedal and Stephen J Wright. *Numerical Optimization*. Springer New York, 1999.

[ORX+16] Stanley Osher, Feng Ruan, Jiechao Xiong, Yuan Yao, and Wotao Yin. Sparse recovery via differential inclusions. *Applied and Computational Harmonic Analysis*, 41(2):436–469, 2016.

[Pol87] Boris T Polyak. *Introduction to Optimization*. Optimization Software, Inc., 1987.

[Roc97] R Tyrrell Rockafellar. *Convex Analysis*. Princeton Landmarks in Mathematics, Reprint of the 1970 original. Princeton University Press, 1997.

[Rus06] Andrzej Ruszczyński. *Nonlinear Optimization*. Princeton University Press, 2006.

[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*, 195:79–148, 2022.

[Sho85] Naum Zuselevich Shor. *Minimization Methods for Non-Differentiable Functions*, volume 3 of *Springer Series in Computational Mathematics*. Springer Berlin Heidelberg, 1985.

[SMDH13] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International Conference on Machine Learning*, pages 1139–1147. PMLR, 2013.

[Tse08] Paul Tseng. On accelerated proximal gradient methods for convex-concave optimization. *SIAM Journal on Optimization, submitted to*, 2008.

[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.