NOTE PAGE

-
-

(1)

Contents

Modeling (Second-Scheme) Nesterov's Accelerated Gradient Method through (Second-Order) Differential Equations: Insights into Theory and Convergence

Chris Junchi Li[◇]

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

September 23, 2024

Abstract

This paper investigates Nesterov's Accelerated Gradient Method through the lens of ordinary differential equations (ODEs). We derive a second-order ODE as the continuous-time limit of Nesterov's scheme by reducing the step size. This new approach provides deeper theoretical insights into the behavior of accelerated first-order methods. Our analysis focuses on the convergence rate, oscillatory behavior, and damping effects that occur during optimization. We demonstrate that the ODE model not only preserves the inverse quadratic convergence rate of Nesterov's discrete scheme but also offers a simplified framework for understanding and generalizing momentum-based methods.

Keywords: Nesterov's Accelerated Gradient; Convex Optimization; First-Order Optimization; Differential Equations; Convergence Rate; Momentum Methods

1 Introduction

Optimization plays a crucial role in many areas of machine learning, particularly in large-scale problems where minimizing convex functions is essential for model estimation. First-order methods, such as gradient descent, have become increasingly popular due to their simplicity and efficiency when handling high-dimensional data. Among these, Nesterov's Accelerated Gradient Method (AGM) method has stood out for its superior convergence properties compared to traditional gradient descent.

Nesterov's method, introduced in 1983, leverages a momentum term to accelerate the convergence rate of gradient-based optimization algorithms. While traditional gradient descent achieves a convergence rate of $O(1/k)$ for convex functions, (second-scheme) AGM achieves a much faster rate of $O(1/k^2)$. Despite its popularity, a thorough understanding of Nesterov's method and its behavior in various settings remains an area of active research. In particular, the oscillatory nature and the momentum-driven overshooting observed in many cases pose interesting theoretical challenges.

Recent research has sought to connect optimization algorithms with continuous-time models, particularly through ordinary differential equations (ODEs). By analyzing optimization schemes in continuous time, deeper insights into the dynamics of the algorithms can be gained, allowing for a more intuitive understanding of their convergence behavior. In this work, we derive a second-order ODE as the continuous limit of Nesterov's Accelerated Gradient method. This provides a novel perspective on the underlying mechanics of the algorithm, revealing connections between discrete-time acceleration and continuous-time damping effects.

Moreover, this ODE formulation allows us to analyze the role of the momentum term in more detail. By interpreting Nesterov's method through the lens of differential equations, we can in-

vestigate the effects of overdamping and underdamping on the convergence trajectory, providing a theoretical explanation for the oscillations frequently observed in practice.

2 Second-Scheme Accelerated Gradient Method in Continuous-Time Dynamics

Mathematically, we consider the minimization problem

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \; f(\boldsymbol{x}) \tag{1}$$

And while vanilla gradient descent enjoys an iteration complexity of $\mathcal{O}(\kappa \log(1/\epsilon))$ on L-smooth, μ-strongly convex problems, with $\kappa = L/\mu$ being the condition number, Nesterov's method [Nesterov(1983)], when equipped with proper restarting, achieves an improved iteration complexity of $\mathcal{O}(\sqrt{\kappa}\log(1/\epsilon))$. We adopt the following version of the Nesterov acceleration, known as the "second scheme" [Tseng(2008)].

Let x_K denote the output of Accelerated Gradient Method with initialization (x_0, y_0, z_0, K). For $k = 0, 1, \ldots, K-1$

$$y_k = (1-\alpha_k)x_k + \alpha_k z_k \tag{2a}$$
$$z_{k+1} = z_k - \eta_k \nabla f(y_k) \tag{2b}$$
$$x_{k+1} = (1-\alpha_k)x_k + \alpha_k z_{k+1} \tag{2c}$$

Here, x_k denotes a α_k-*weighted-averaged* iteration corresponding to z_{k+1}. In other words, compared with vanilla gradient descent, $z_{k+1} = z_k - s\nabla f(z_k)$, Nesterov's acceleration conducts a step at the negated gradient direction evaluated at a *predictive iterate* of the weighted-averaged iterate of the sequence. This enables a larger choice of stepsize, reflecting the enhanced stability. An analogous interpretation has been discussed in work on a heavy-ball-based acceleration method.

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.*

2.1 (Second-Scheme) Nesterov's Accelerated Gradient Method

We set the parameters as

$$\eta_k = \frac{k+\lambda}{\lambda}s \qquad \alpha_k = \frac{\lambda}{k+\lambda}$$

where we assume arbitrarily small $s \leq \frac{1}{L}$. We also assume the initialization is chosen as $z_0 = x_0 = y_0$ throughout. This in all results in

$$z_k = z_{k-1} - \frac{k+\lambda-1}{\lambda}s\nabla f(y_{k-1}) \tag{3a}$$
$$x_k = \frac{k-1}{k+\lambda-1}x_{k-1} + \frac{\lambda}{k+\lambda-1}z_k \tag{3b}$$
$$y_k = \frac{k}{k+\lambda}x_k + \frac{\lambda}{k+\lambda}z_k \tag{3c}$$

Theorem 1. *Two methods are equivalent in the sense they share the same output:*

$$\begin{aligned} x_k &= y_{k-1} - s\nabla f(y_{k-1}) \\ y_k &= x_k + \frac{k-1}{k+\lambda}(x_k - x_{k-1}) \end{aligned} \tag{eq:nesterov-scheme}$$

Proof. Eq. (41c) with index shifted backwards by 1 gives (note this holds automatically for $k = 1$ due to initialization)

$$y_{k-1} = \frac{k-1}{k+\lambda-1}x_{k-1} + \frac{\lambda}{k+\lambda-1}z_{k-1}$$

Subtracting this shifted (41c) from (41b) and combining the resulting equation with (41a), we conclude

$$\begin{aligned} x_k - y_{k-1} &= \frac{\lambda}{k+\lambda-1}(z_k - z_{k-1}) = -s\nabla f(y_{k-1}) \\ \Rightarrow \qquad x_k &= y_{k-1} - s\nabla f(y_{k-1}) \end{aligned} \tag{4}$$

concluding the first line in (3).

Moreover, combining (41c) with (41b) to cancel the z_{k+1} term, we obtain

$$\begin{aligned} \frac{k+\lambda-1}{k+\lambda}x_k - y_k &= \frac{k-1}{k+\lambda}x_{k-1} - \frac{k}{k+\lambda}x_k \\ \Rightarrow \qquad y_k &= x_k + \frac{k-1}{k+\lambda}(x_k - x_{k-1}) \end{aligned} \tag{5}$$

Thus, by a simple notational transformation, (4) plus (5) (and hence the original update rule (3)) is exactly equivalent to the original updates of Nesterov's acceleration scheme [Nesterov(1983)].

The other direction works using an exact reversed argument, by introducing the extrapolation variable $z_k = x_k + \frac{k-1}{\lambda}(x_k - x_{k-1})$. □

In the rest of this paper we assume $\lambda = 2$.

Theorem 2. *Under Assumption 1 we have*

$$\frac{\mu}{2}\|X(T) - z^*\|^2 \le f(X(T)) - f(z^*) \le \frac{2}{T^2}\|z_0 - z^*\|^2 \tag{6}$$

2.2 Understanding the Continuous-Time Dynamics of AGM

We let s be arbitrarily small. Furthermore, we set $\Delta t = \sqrt{s}$ and $t = k\sqrt{s}$. We derive the following proposition:

Theorem 3. *As $\Delta t \to 0^+$ the ODE for the AGM Algorithm 2 under this scaling condition is*

$$\dot{Z} = -\frac{t}{2}\nabla f(X) \qquad \dot{X} = \frac{2}{t}(Z - X) \tag{7}$$

The ODE Eq. (7) for the AGM Algorithm 2 is also equivalent to

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

Proof of Theorem 3. We recall that in Algorithm 2, Line 2b yields:

$$z_{k+1} = z_k - \frac{k+2}{2} s \nabla f(y_k)$$

By shifting indices, we have

$$z_{k+2} = z_{k+1} - \frac{k+3}{2} s \nabla f(y_{k+1}) \tag{8}$$

From (8) we derive the iterative update rule on z_{k+1} that

$$z_{k+2} = z_{k+1} - \frac{k+3}{2} s \nabla f(y_{k+1}) = z_{k+1} - \frac{k+3}{2} s \nabla f(y_{k+1})$$

Moreover, Line 2a and 2c in Algorithm 2 imply that

$$y_{k+1} = x_{k+1} + \frac{k}{k+3} (x_{k+1} - x_k)$$

Thus, we obtain

$$z_{k+2} - z_{k+1} = -\frac{k+3}{2} s \left[\nabla f(x_{k+1})\right] - \frac{k+3}{2} s \left[\nabla f(y_{k+1}) - \nabla f(x_{k+1})\right] \tag{9}$$

Recalling that the stepsize can actually be chosen for any $s \leq \frac{1}{L}$. We let $t = k\Delta t$ where $\Delta t = \sqrt{s}$. Thus, we obtain $\frac{k+2}{2} s = \frac{t\Delta t + 2(\Delta t)^2}{2(\Delta t)^2} s$. Simple algebra yields: $s = (\Delta t)^2$. Combining this with the value of $\frac{k+2}{2} s$ and we obtain that

$$\frac{k+2}{2} s = \frac{t\Delta t + 2(\Delta t)^2}{2} = \frac{(t + 2\Delta t)}{2} \Delta t = \frac{t}{2} \Delta t + (\Delta t)^2 \tag{10}$$

which goes to 0 as $s \to 0^+$ and $\Delta t \to 0^+$. (10) can be shortened as: $\frac{k+2}{2} s = \frac{t}{2} \Delta t + o(\Delta t)$. With this choice of $\frac{k+2}{2} s$, by Taylor expansion we have[1]

$$\nabla f(y_{k+1}) - \nabla f(x_{k+1}) \leq \mathcal{O}\left(y_{k+1} - x_{k+1}\right) \overset{(a)}{\leq} \mathcal{O}\left(x_{k+1} - x_k\right) \leq \mathcal{O}\left(\frac{2}{k+2}(z_{k+1} - x_k)\right) = o(1)$$

We rewrite (9) in continuous dynamics by letting $Z(t) = z_k$ and $X(t) = x_{k+1}$:

$$\begin{aligned} Z(t+1) - Z(t) &= -\frac{k+3}{2} s \nabla f(X(t)) + o(\frac{k+2}{2} s) + o(\frac{k+3}{2} s) \\ &= -\frac{t}{2} \Delta t \nabla f(X(t)) + o(\Delta t) \end{aligned}$$

Dividing both sides by Δt and with $\Delta t \to 0$, we obtain

$$\dot{Z}(t) + \frac{t}{2} \nabla f(X(t)) = 0 \tag{11}$$

In the final step we calculate the relationship between $Z(t)$ and $X(t)$. By Line 2c, we have

$$z_{k+1} = \frac{k}{2} \left(x_{k+1} - x_k\right) + x_{k+1}$$

[1] In fact we have $z_k^{\mathrm{md}} - z_{k+1}^{\mathrm{ag}}$ is $o(\sqrt{s})$.

which is equivalent to

$$Z(t) = \frac{t}{2}\frac{X(t) - X(t-\Delta t)}{\Delta t} + X(t)$$

Letting $\Delta t \to 0$, we have

$$Z(t) = X(t) + \frac{t}{2}\dot{X}(t) \tag{12}$$

Combining (11) with (12) we conclude a single-line higher-order ODE

$$\frac{t}{2}\ddot{X} + \frac{3}{2}\dot{X} + \frac{t}{2}\nabla f(X) = 0 \tag{13}$$

and hence by multiplying both sides by $\frac{2}{t}$, our proof.

□

2.2.1 Proof of Theorem 2

Proof of Theorem 2. We first provide an estimate of the time derivative $\dot{\mathcal{E}}$ of the Lyapunov function corresponding to (7), and the result is shown is Lemma 1:

Lemma 1. *We set the Lyapunov function as defined in the following* (14)*:*

$$\mathcal{E} = t^2\left[f(X) - f(z^*)\right] + 2\|Z - z^*\|^2 \tag{14}$$

Given the dynamics in (7) *starting from* $X(0) = z_0$*, we have*

$$\dot{\mathcal{E}} = \frac{d}{dt}\left[t^2\left[f(X) - f(z^*)\right] + 2\|Z - z^*\|^2\right] \le 0 \tag{15}$$

Proof of Lemma 1. Since $\mathcal{E}(t)$ is set as

$$\mathcal{E} = t^2\left[f(X) - f(z^*)\right] + 2\|Z - z^*\|^2 \tag{14}$$

we have its time derivative

$$\frac{d\mathcal{E}}{dt} = \underbrace{2t\left[f(X) - f(z^*)\right] + t^2\left\langle \nabla f(X) - f(z^*), \dot{X}\right\rangle + 4\left\langle Z - z^*, \dot{Z}\right\rangle}_{\text{“bracketed part”}}$$

We want to show the bracketed part in above is nonpositive, i.e.

$$2t\left[f(X) - f(z^*)\right] + t^2\left\langle \nabla f(X), \dot{X}\right\rangle + 4\left\langle Z - z^*, \dot{Z}\right\rangle \le 0$$

Denote

$$2t\left[f(X) - f(z^*)\right] + t^2\left\langle \nabla f(X), \dot{X}\right\rangle + 4\left\langle Z - z^*, \dot{Z}\right\rangle \equiv \mathrm{I} + \mathrm{II} + \mathrm{III}$$

Then using $Z = X + \frac{t}{2}\dot{X}$ we have

$$\mathrm{III} = 4\left\langle Z - z^*, -\frac{t}{2}\nabla f(X)\right\rangle = -2t\left\langle Z - z^*, \nabla f(X)\right\rangle$$

$$\leq -2t\left\langle Z - z^*, \nabla f(X)\right\rangle = -\left\langle 2t(Z - z^*) + t^2\dot{X}, \nabla f(X)\right\rangle$$

Therefore

$$\mathrm{I} + \mathrm{II} = 2t\left[f(X) - f(z^*)\right] + t^2\left\langle \dot{X}, \nabla f(X)\right\rangle$$

and

$$\begin{aligned}\mathrm{I} + \mathrm{II} + \mathrm{III} &= 2t\left[f(X) - f(z^*)\right] - \left\langle 2t(Z - z^*), \nabla f(X)\right\rangle \\ &= 2t\left[f(X) - f(z^*) - \left\langle Z - z^*, \nabla f(X)\right\rangle\right] \leq 0\end{aligned}$$

where the last step uses the convexity of f. This concludes the desired result of Lemma 1. □

Now bringing (16) in Corollary 1 into (15), we conclude that

$$\dot{\mathcal{E}} = \frac{d}{dt}\left[t^2\left(f(X) - f(z^*)\right) + 2\|Z - z^*\|^2\right] \leq 0$$

Integrating both sides gives

$$T^2\left[f(X) - f(z^*)\right] + 2\|Z - z^*\|^2 - 2\|z_0 - z^*\|^2 \leq 0$$

Rearranging and dividing both sides by T^2, we obtain that

$$f(X) - f(z^*) \leq f(X) - f(z^*) + \frac{2}{T^2}\|Z - z^*\|^2 \leq \frac{2}{T^2}\|z_0 - z^*\|^2$$

First-order necessary condition for optimality gives $\nabla f(z^*) = 0$, and hence μ-strong convexity of F implies

$$f(X) - f(z^*) \geq \left\langle Z - z^*, \nabla f(z^*)\right\rangle + \frac{\mu}{2}\|Z - z^*\|^2 = \frac{\mu}{2}\|Z - z^*\|^2 \geq 0$$

which concludes the entire proof.

□

2.3 A Corollary

Note that both sides of (15) in Lemma 1 presents the quantity $2\|Z - z^*\|^2$ in its original and gradient forms, respectively. By integrating on both sides and applying a Gronwall-type technique, we obtain the following Corollary 1 which shows that the continuous-time dynamics of AGM are non-expansive with respect to the minimax point z^*.

Corollary 1. *We have*

$$\|Z(T) - z^*\| \leq \|z_0 - z^*\| \tag{16}$$

An extension of this result is useful in proving simple accelerated minimax optimization algorithms. We postpone the rigorous proof of Corollary 1 to §2.3.

Proof of Corollary 1. To proceed with proof we adopt a Gronwall-type argument. Note

$$\dot{\mathcal{E}}(t) = \frac{d}{dt}\left[t^2\left(f(X(t)) - f(z^*)\right) + 2\|Z(t) - z^*\|^2\right] \leq 0$$

Taking integrals on both sides $\int_0^T dt$ gives

$$T^2\left(f(X(T)) - f(z^*)\right) + 2\|Z(T) - z^*\|^2 - 2\|z_0 - z^*\|^2 \leq 0$$

Dropping the nonnegative term $T^2\left(f(X(T)) - f(z^*)\right)$ gives

$$\|Z(T) - z^*\|^2 \leq \|z_0 - z^*\|^2$$

completing the proof of Corollary 1. □

3 Mathematical Formulations

In the simplest and most standard form, we are interested in solving

$$\text{minimize} \quad f(x)$$

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., [Polyak(1987), Boyd and Vandenberghe(2004), Nocedal and Wright(2006), Ruszczyński(2006), Boyd et al.(2011), Shor(2012), Beck(2014)], for expositions.

First-order methods have regained popularity as data sets and problems are ever increasing in size and, consequently, there has been much research on the theory and practice of accelerated first-order schemes. Perhaps the earliest first-order method for minimizing a convex function f is the gradient method, which dates back to Euler and Lagrange. Forty years ago, however, in a seminal paper Nesterov proposed an accelerated gradient method [Nesterov(1983)], which may take the following form: starting with x_0, $y_0 = x_0$ and $z_0 = x_0$, inductively define

$$z_k = z_{k-1} - \frac{k+1}{2} s \nabla f(y_{k-1}) \tag{17a}$$

$$x_k = \frac{k-1}{k+1} x_{k-1} + \frac{2}{k+1} z_k \tag{17b}$$

$$y_k = \frac{k}{k+2} x_k + \frac{2}{k+2} z_k \tag{17c}$$

For any fixed step size $s \leq 1/L$, where L is the Lipschitz constant of ∇f, this scheme exhibits the convergence rate

$$f(x_k) - f^\star \leq O\left(\frac{\|x_0 - x^\star\|^2}{sk^2}\right) \tag{18}$$

Above, $x^\star$ is any minimizer of f and $f^\star = f(x^\star)$. It is well-known that this rate is optimal among all methods having only information about the gradient of f at consecutive iterates [Nesterov(2004)]. This is in contrast to vanilla gradient descent methods, which have the same computational complexity but can only achieve a rate of $O(1/k)$. This improvement relies on the introduction of the

momentum term $x_k - x_{k-1}$ as well as the particularly tuned Interpolation coefficient $2/(k+2) \approx 2/k$. Since the introduction of Nesterov's scheme, there has been much work on the development of first-order accelerated methods, see [Nesterov(2004), Nesterov(2005), Nesterov(2013)] for theoretical developments, and [Tseng(2008)] for a unified analysis of these ideas. Notable applications can be found in sparse linear regression [Beck and Teboulle(2009), Qin and Goldfarb(2012)], compressed sensing [Becker et al.(2011)] and, deep and recurrent neural networks [Sutskever et al.(2013)].

In a different direction, there is a long history relating ordinary differential equation (ODEs) to optimization, see [Helmke and Moore(1996)], [Schropp and Singer(2000)], and [Fiori(2005)] for example. The connection between ODEs and numerical optimization is often established via taking step sizes to be very small so that the trajectory or solution path converges to a curve modeled by an ODE. The conciseness and well-established theory of ODEs provide deeper insights into optimization, which has led to many interesting findings. Notable examples include linear regression via solving differential equations induced by linearized Bregman iteration algorithm [Osher et al.(2014)], a continuous-time Nesterov-like algorithm in the context of control design [Dürr and Ebenbauer(2012), Dürr et al.(2012)], and modeling design iterative optimization algorithms as nonlinear dynamical systems [Lessard et al.(2014)].

In this work, we derive a second-order ODE which is the exact limit of Nesterov's scheme by taking small step sizes in (17); to the best of our knowledge, this work is the first to use ODEs to model Nesterov's scheme or its variants in this 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:

$$\begin{cases} \dot{X} = \frac{2}{t}(Z - X) \\ \dot{Z} = -\frac{t}{2}\nabla f(X) \end{cases} \tag{19}$$

for $t > 0$, with initial conditions $X(0) = x_0, \dot{X}(0) = 0$; here, x_0 is the starting point in Nesterov's scheme, $\dot{X} \equiv \mathrm{d}X/\mathrm{d}t$ denotes the time derivative or velocity and similarly $\ddot{X} \equiv \mathrm{d}^2X/\mathrm{d}t^2$ denotes the acceleration. The time parameter in this ODE is related to the step size in (17) via $t \approx k\sqrt{s}$. Expectedly, it also enjoys inverse quadratic convergence rate as its discrete analog,

$$f(X(t)) - f^\star \leq O\left(\frac{\|x_0 - x^\star\|^2}{t^2}\right)$$

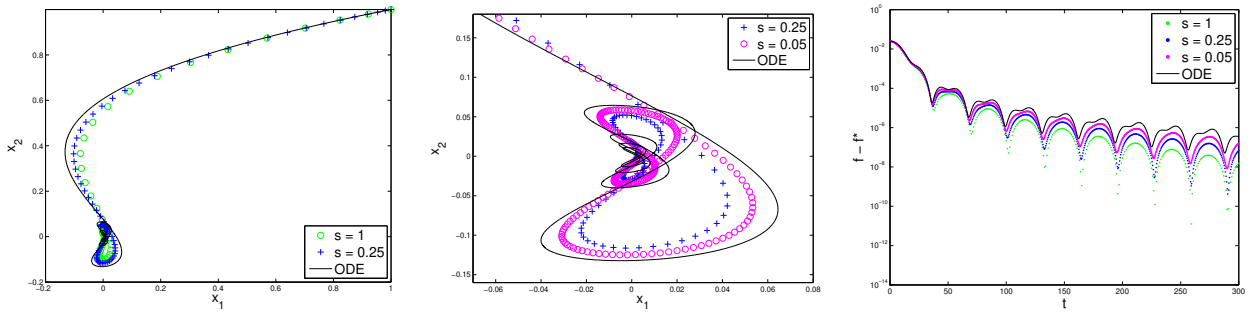
Approximate equivalence between Nesterov's scheme and the ODE is established later in various perspectives, rigorous and intuitive. In the main body of this paper, examples and case studies are provided to demonstrate that the homogeneous and conceptually simpler ODE can serve as a tool for understanding,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

In the following, two insights of Nesterov's scheme are highlighted, the first one on oscillations in the trajectories of this scheme, and the second on the peculiar constant 2 appearing in the ODE.

3.1 From Overdamping to Underdamping

In general, Nesterov's scheme is not monotone in the objective function value due to the introduction of the momentum term. Oscillations or overshoots along the trajectory of iterates approaching the minimizer are often observed when running Nesterov's scheme. Figure 1 presents typical phenomena of this kind, where a two-dimensional convex function is minimized by Nesterov's scheme. Viewing the ODE as a damping system, we obtain interpretations as follows.

- **Small t.** In the beginning, the damping ratio $3/t$ is large. This leads the ODE to be an overdamped system, returning to the equilibrium without oscillating;



(a) Trajectories. (b) Zoomed trajectories. (c) Errors $f - f^\star$.

Figure 1. Minimizing $f = 2 \times 10^{-2} x_1^2 + 5 \times 10^{-3} x_2^2$, starting from $x_0 = (1,\ 1)$. The black and solid curves correspond to the solution to the ODE. In (c), for the x-axis we use the identification between time and iterations, $t = k\sqrt{s}$.

- **Large t.** As t increases, the ODE with a small $3/t$ behaves like an underdamped system, oscillating with the amplitude gradually decreasing to zero.

As depicted in Figure 1a, in the beginning the ODE curve moves smoothly towards the origin, the minimizer $x^\star$. The second interpretation "**Large t**" provides partial explanation for the oscillations observed in Nesterov's scheme at later stage. Although our analysis extends farther, it is similar in spirit to that carried in [O'Donoghue and Candès(2013)]. In particular, the zoomed Figure 1b presents some butterfly-like oscillations for both the scheme and ODE. There, we see that the trajectory constantly moves away from the origin and returns back later. Each overshoot in Figure 1b causes a bump in the function values, as shown in Figure 1c. We observe also from Figure 1c that the periodicity captured by the bumps are very close to that of the ODE solution. In passing, it is worth mentioning that the solution to the ODE in this case can be expressed via Bessel functions, hence enabling quantitative characterizations of these overshoots and bumps, which are given in full detail in Section 5.

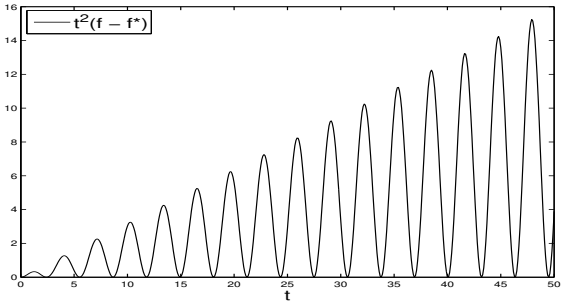
3.2 A Phase Transition

The special constant $\lambda = 2$ is not haphazard. In fact, it is the smallest constant that guarantees $O(1/t^2)$ convergence rate. Specifically, parameterized by a constant r, the generalized ODE

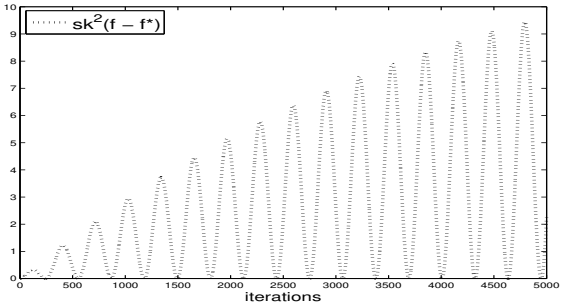
$$\begin{cases} \dot{X} = \frac{\lambda}{t}(Z - X) \\ \dot{Z} = -\frac{t}{\lambda}\nabla f(X) \end{cases}$$

can be translated into a generalized Nesterov's scheme that is the same as the original (17) except for $2/(k+2)$ being replaced by $\lambda/(k+\lambda)$. Surprisingly, for both generalized ODEs and schemes, the inverse quadratic convergence is guaranteed if and only if $\lambda \geq 2$. This phase transition suggests there might be deep causes for acceleration among first-order methods. In particular, for $\lambda \geq 2$, the worst case constant in this inverse quadratic convergence rate is minimized at $r = 3$.

Figure 2 illustrates the growth of $t^2(f(X(t)) - f^\star)$ and $sk^2(f(x_k) - f^\star)$, respectively, for the generalized ODE and scheme with $r = 1$, where the objective function is simply $f(x) = \frac{1}{2}x^2$. Inverse quadratic convergence fails to be observed in both Figures 2a and 2b, where the scaled errors grow with t or iterations, for both the generalized ODE and scheme.



(a) Scaled errors $t^2(f(X(t)) - f^\star)$.



(b) Scaled errors $sk^2(f(x_k) - f^\star)$.

Figure 2. Minimizing $f = \frac{1}{2}x^2$ by the generalized ODE and scheme with $r = 1$, starting from $x_0 = 1$. In (b), the step size $s = 10^{-4}$.

Outline and Notation The rest of the paper is organized as follows. In Section 4, the ODE is rigorously derived from Nesterov's scheme. Connections between the ODE and the scheme, in terms of trajectory behaviors and convergence rates, are summarized in Section 5. In Section 6, we discuss the effect of replacing the constant 2 in (19) by an arbitrary constant on the convergence rate.

Some standard notations used throughout the paper are collected here. We denote by $\mathcal{F}_L$ the class of convex functions f with L–Lipschitz continuous gradients defined on $\mathbb{R}^n$, i.e., f is convex, continuously differentiable, and satisfies

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

for any $x, y \in \mathbb{R}^n$, where $\|\cdot\|$ is the standard Euclidean norm and $L > 0$ is the Lipschitz constant. Next, $\mathcal{S}_\mu$ denotes the class of μ–strongly convex functions f on $\mathbb{R}^n$ with continuous gradients, i.e., f is continuously differentiable and $f(x) - \mu\|x\|^2/2$ is convex. We set $\mathcal{S}_{\mu,L} = \mathcal{F}_L \cap \mathcal{S}_\mu$.

4 Derivation

First, we sketch an informal derivation of the ODE (19). **Unless otherwise noted, we will be studying its one-line equivalent form.** Assume $f \in \mathcal{F}_L$ for $L > 0$. Combining the two equations of (17) and applying a rescaling gives

$$\frac{x_{k+1} - x_k}{\sqrt{s}} = \frac{k-1}{k+2}\frac{x_k - x_{k-1}}{\sqrt{s}} - \sqrt{s}\nabla f(y_k) \tag{20}$$

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

$$(x_{k+1} - x_k)/\sqrt{s} = \dot{X}(t) + \frac{1}{2}\ddot{X}(t)\sqrt{s} + o(\sqrt{s}), \quad (x_k - x_{k-1})/\sqrt{s} = \dot{X}(t) - \frac{1}{2}\ddot{X}(t)\sqrt{s} + o(\sqrt{s})$$

and $\sqrt{s}\nabla f(y_k) = \sqrt{s}\nabla f(X(t)) + o(\sqrt{s})$. Thus (20) can be written as

$$\dot{X}(t) + \frac{1}{2}\ddot{X}(t)\sqrt{s} + o(\sqrt{s})$$

$$= \left(1 - \frac{3\sqrt{s}}{t}\right)\left(\dot{X}(t) - \frac{1}{2}\ddot{X}(t)\sqrt{s} + o(\sqrt{s})\right) - \sqrt{s}\nabla f(X(t)) + o(\sqrt{s}) \quad (21)$$

By comparing the coefficients of $\sqrt{s}$ in (21), we obtain

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

The first initial condition is $X(0) = x_0$. Taking $k = 1$ in (20) yields

$$(x_2 - x_1)/\sqrt{s} = -\sqrt{s}\nabla f(y_1) = o(1)$$

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

One popular alternative momentum coefficient is $\theta_k(\theta_{k-1}^{-1} - 1)$, where θ_k are iteratively defined as $\theta_{k+1} = \left(\sqrt{\theta_k^4 + 4\theta_k^2} - \theta_k^2\right)/2$, starting from $\theta_0 = 1$ [Nesterov(1983), Beck and Teboulle(2009)]. Simple analysis reveals that $\theta_k(\theta_{k-1}^{-1} - 1)$ asymptotically equals $1 - 3/k + O(1/k^2)$, thus leading to the same ODE as (17).

Classical results in ODE theory do not directly imply the existence or uniqueness of the solution to this ODE because the coefficient $3/t$ is singular at $t = 0$. In addition, ∇f is typically not analytic at x_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 (19), 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 4. *For any $f \in \mathcal{F}_\infty := \cup_{L>0}\mathcal{F}_L$ and any $x_0 \in \mathbb{R}^n$, the ODE* (19) *with initial conditions $X(0) = x_0, \dot{X}(0) = 0$ 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 the appendices.

Theorem 5. *For any $f \in \mathcal{F}_\infty$, as the step size $s \to 0$, Nesterov's scheme* (17) *converges to the ODE* (19) *in the sense that for all fixed $T > 0$,*

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

4.1 Simple Properties

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

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

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

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

Also note that minimizing 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 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 (19) once multiplying Q on both sides.

Initial Asymptotic. Assume sufficient smoothness of X such that $\lim_{t\to 0} \ddot{X}(t)$ exists. The mean value theorem guarantees the existence of some $\xi \in (0, t)$ that satisfies $\dot{X}(t)/t = (\dot{X}(t) - \dot{X}(0))/t = \ddot{X}(\xi)$. Hence, from the ODE we deduce $\ddot{X}(t) + 3\ddot{X}(\xi) + \nabla f(X(t)) = 0$. Taking the limit $t \to 0$ gives $\ddot{X}(0) = -\nabla f(x_0)/4$. Hence, for small t we have the asymptotic form:

$$X(t) = -\frac{\nabla f(x_0) t^2}{8} + x_0 + o(t^2)$$

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

4.2 Proof of Theorem 4

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

Lemma 2. *For any $f \in \mathcal{F}_\infty$ and any $x_0 \in \mathbb{R}^n$, the ODE* (19) *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

$$\ddot{X} + \frac{3}{\max(\delta, t)} \dot{X} + \nabla f(X) = 0 \tag{22}$$

with $X(0) = x_0, \dot{X}(0) = 0$. Denoting by $Z = \dot{X}$, then (22) is equivalent to

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

with $X(0) = x_0, Z(0) = 0$. As functions of (X, Z), both Z and $-3Z/\max(\delta, t) - \nabla f(X))$ are $\max(1, L) + 3/\delta$-Lipschitz continuous. Hence by standard ODE theory, (22) has a unique global solution in $C^2[0,\infty)$, denoted by X_δ. Note that $\ddot{X}_\delta$ is also well defined at $t = 0$. Next, introduce $M_\delta(t)$ to be the supremum of $\|\dot{X}_\delta(u)\|/u$ over $u \in (0, t]$. It is easy to see that $M_\delta(t)$ is finite because $\|\dot{X}_\delta(u)\|/u = (\|\dot{X}_\delta(u) - \dot{X}_\delta(0)\|)/u = \|\ddot{X}_\delta(0)\| + o(1)$ for small u. We give an upper bound for $M_\delta(t)$ in the following lemma.

Lemma 3. *For $\delta < \sqrt{6/L}$, we have*

$$M_\delta(\delta) \le \frac{\|\nabla f(x_0)\|}{1 - L\delta^2/6}$$

The proof of Lemma 3 relies on a simple lemma.

Lemma 4. *For any $u > 0$, the following inequality holds*

$$\|\nabla f(X_\delta(u)) - \nabla f(x_0)\| \le \frac{1}{2} L M_\delta(u) u^2$$

Proof of Lemma 4. By Lipschitz continuity,

$$\|\nabla f(X_\delta(u)) - \nabla f(x_0)\| \le L\|X_\delta(u) - x_0\| = \Big\| \int_0^u \dot{X}_\delta(v) \mathrm{d}v \Big\| \le \int_0^u v \frac{\|\dot{X}_\delta(v)\|}{v} \mathrm{d}v \le \frac{1}{2} L M_\delta(u) u^2$$

□

Next, we prove Lemma 3.

Proof of Lemma 3. For $0 < t \le \delta$, the smoothed ODE takes the form

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

which yields

$$\dot{X}_\delta \mathrm{e}^{3t/\delta} = -\int_0^t \nabla f(X_\delta(u)) \mathrm{e}^{3u/\delta} \mathrm{d}u = -\nabla f(x_0) \int_0^t \mathrm{e}^{3u/\delta} \mathrm{d}u - \int_0^t (\nabla f(X_\delta(u)) - \nabla f(x_0)) \mathrm{e}^{3u/\delta} \mathrm{d}u$$

Hence, by Lemma 4

$$\begin{aligned} \frac{\|\dot{X}_\delta(t)\|}{t} &\le \frac{1}{t}\mathrm{e}^{-3t/\delta} \|\nabla f(x_0)\| \int_0^t \mathrm{e}^{3u/\delta} \mathrm{d}u + \frac{1}{t}\mathrm{e}^{-3t/\delta} \int_0^t \frac{1}{2} L M_\delta(u) u^2 \mathrm{e}^{3u/\delta} \mathrm{d}u \\ &\le \|\nabla f(x_0)\| + \frac{L M_\delta(\delta) \delta^2}{6} \end{aligned}$$

Taking the supremum of $\|\dot{X}_\delta(t)\|/t$ 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 5. *For $\delta < \sqrt{6/L}$ and $\delta < t < \sqrt{12/L}$, we have*

$$M_\delta(t) \le \frac{(5 - L\delta^2/6)\|\nabla f(x_0)\|}{4(1 - L\delta^2/6)(1 - Lt^2/12)}$$

Proof of Lemma 5. For $t > \delta$, the smoothed ODE takes the form

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

which is equivalent to

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

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

$$-\int_\delta^t u^3 \nabla f(X_\delta(u)) \mathrm{d}u + \delta^3 \dot{X}_\delta(\delta) = -\int_\delta^t u^3 \nabla f(x_0) \mathrm{d}u - \int_\delta^t u^3 (\nabla f(X_\delta(u)) - \nabla f(x_0)) \mathrm{d}u + \delta^3 \dot{X}_\delta(\delta)$$

Therefore by Lemmas 4 and 3, we get

$$\begin{aligned}\frac{\|\dot{X}_\delta(t)\|}{t} &\le \frac{t^4-\delta^4}{4t^4}\|\nabla f(x_0)\| + \frac{1}{t^4}\int_\delta^t \frac{1}{2}LM_\delta(u)u^5\mathrm{d}u + \frac{\delta^4}{t^4}\frac{\|\dot{X}_\delta(\delta)\|}{\delta} \\ &\le \frac{1}{4}\|\nabla f(x_0)\| + \frac{1}{12}LM_\delta(t)t^2 + \frac{\|\nabla f(X_0)\|}{1-L\delta^2/6}\end{aligned}$$

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

$$\frac{\|\dot{X}_\delta(t')\|}{t'} \le \frac{1}{4}\|\nabla f(x_0)\| + \frac{1}{12}LM_\delta(t)t^2 + \frac{\|\nabla f(x_0)\|}{1-L\delta^2/6}$$

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

$$M_\delta(t) \le \frac{1}{4}\|\nabla f(x_0)\| + \frac{1}{12}LM_\delta(t)t^2 + \frac{\|\nabla f(X_0)\|}{1-L\delta^2/6}$$

The desired result follows from rearranging the inequality. □

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

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

$$\|\dot{X}_\delta(t)\| \le \sqrt{6/L}M_\delta(\sqrt{6/L}) \le \sqrt{6/L}\max\left\{\frac{\|\nabla f(x_0)\|}{1-\frac{1}{2}}, \frac{5\|\nabla f(x_0)\|}{4(1-\frac{1}{2})(1-\frac{1}{2})}\right\} = 5\sqrt{6/L}\|\nabla f(x_0)\|$$

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

$$\|X_\delta(t)\| \le \|X_\delta(0)\| + \int_0^t \|\dot{X}_\delta(u)\|\mathrm{d}u \le \|x_0\| + 30\|\nabla f(x_0)\|/L$$

□

We are now ready for the proof of Lemma 2.

Proof of Lemma 2. By the Arzelá-Ascoli theorem and Lemma 6, $\mathcal{F}$ contains a subsequence converging uniformly on $[0, \sqrt{6/L}]$. Denote by $\{X_{\delta_{m_i}}\}_{i\in\mathbb{N}}$ the convergent subsequence and $\breve{X}$ the limit. Above, $\delta_{m_i} = \sqrt{3/L}/2^{m_i}$ decreases as i increases. We will prove that $\breve{X}$ satisfies (19) and the initial conditions $\breve{X}(0) = x_0, \dot{\breve{X}}(0) = 0$.

Fix an arbitrary $t_0 \in (0, \sqrt{6/L})$. Since $\|\dot{X}_{\delta_{m_i}}(t_0)\|$ is bounded, we can pick a subsequence of $\dot{X}_{\delta_{m_i}}(t_0)$ which converges to a limit, denoted by $X^D_{t_0}$. Without loss of generality, assume the subsequence is the original sequence. Denote by $\widetilde{X}$ the local solution to (19) with $X(t_0) = \breve{X}(t_0)$ and $\dot{X}(t_0) = X^D_{t_0}$. Now recall that $X_{\delta_{m_i}}$ is the solution to (19) with $X(t_0) = X_{\delta_{m_i}}(t_0)$ and $\dot{X}(t_0) = \dot{X}_{\delta_{m_i}}(t_0)$ when $\delta_{m_i} < t_0$. Since both $X_{\delta_{m_i}}(t_0)$ and $\dot{X}_{\delta_{m_i}}(t_0)$ approach $\breve{X}(t_0)$ and $X^D_{t_0}$, respectively, there exists $\epsilon_0 > 0$ such that

$$\sup_{t_0-\epsilon_0<t<t_0+\epsilon_0} \|X_{\delta_{m_i}}(t) - \widetilde{X}(t)\| \to 0$$

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

$$\sup_{t_0-\epsilon_0<t<t_0+\epsilon_0} \|X_{\delta_{m_i}}(t) - \breve{X}(t)\| \to 0$$

Therefore $\breve{X}$ and $\widetilde{X}$ have to be identical on $(t_0 - \epsilon_0, t_0 + \epsilon_0)$. So $\breve{X}$ satisfies (19) at t_0. Since t_0 is arbitrary, we conclude that $\breve{X}$ is a solution to (19) on $(0, \sqrt{6/L})$. By extension, $\breve{X}$ can be a global solution to (19) on $(0, \infty)$. It only leaves to verify the initial conditions to complete the proof.

The first condition $\breve{X}(0) = x_0$ is a direct consequence of $X_{\delta_{m_i}}(0) = x_0$. To check the second, pick a small $t > 0$ and note that

$$\begin{aligned}\frac{\|\breve{X}(t) - \breve{X}(0)\|}{t} = \lim_{i\to\infty} \frac{\|X_{\delta_{m_i}}(t) - X_{\delta_{m_i}}(0)\|}{t} &= \lim_{i\to\infty} \|\dot{X}_{\delta_{m_i}}(\xi_i)\| \\ &\le \limsup_{i\to\infty} tM_{\delta_{m_i}}(t) \le 5t\sqrt{6/L}\|\nabla f(x_0)\|\end{aligned}$$

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 (19).

Lemma 7. *For any $f \in \mathcal{F}_\infty$, the ODE* (19) *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 $\|\dot{X}(u) - \dot{Y}(u)\|$ over $u \in [0, t)$. To proceed, we need a simple auxiliary lemma.

Lemma 8. *For any $t \in (0, \alpha)$, we have*

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

Proof of Lemma 8. By Lipschitz continuity of the gradient, one has

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

□

Now we prove Lemma 7.

Proof of Lemma 7. Similar to the proof of Lemma 5, we get

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

Applying Lemma 8 gives

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

which can be simplified as $\|\dot{X}(t)-\dot{Y}(t)\| \le Lt^2\widetilde{M}(t)/5$. Thus, for any $t' \le t$ it is true that $\|\dot{X}(t')-\dot{Y}(t')\| \le Lt^2\widetilde{M}(t)/5$. Taking the supremum of $\|\dot{X}(t')-\dot{Y}(t')\|$ over $t' \in (0,t)$ gives $\widetilde{M}(t) \le Lt^2\widetilde{M}(t)/5$. Therefore $\widetilde{M}(t)=0$ for $t < \min(\alpha, \sqrt{5/L})$, which is equivalent to saying $\dot{X}=\dot{Y}$ on $[0, \min(\alpha, \sqrt{5/L}))$. With the same initial value $X(0)=Y(0)=x_0$ and the same gradient, we conclude that X and Y are identical on $(0, \min(\alpha, \sqrt{5/L}))$, a contradiction. □

Given all of the aforementioned lemmas, the proof of Theorem 4 is simply combining 2 and 7.

4.3 Proof of Theorem 5

Identifying $\sqrt{s}=\Delta t$, the comparison between (20) and (37) reveals that Nesterov's scheme is a discrete scheme for numerically integrating the ODE (19). However, its singularity of the damping coefficient at $t=0$ leads to the nonexistence of off-the-shelf ODE theory for proving Theorem 5. To address this difficulty, we use the smoothed ODE (22) to approximate the original one; then bound the difference between Nesterov's scheme and the forward Euler scheme of (22), which may take the following form:

$$\begin{aligned} X_{k+1}^{\delta} &= X_k^{\delta} + \Delta t Z_k^{\delta} \\ Z_{k+1}^{\delta} &= \Big(1 - \frac{3\Delta t}{\max\{\delta, k\Delta t\}}\Big) Z_k^{\delta} - \Delta t \nabla f(X_k^{\delta}) \end{aligned} \tag{23}$$

with $X_0^{\delta}=x_0$ and $Z_0^{\delta}=0$.

Lemma 9. *With step size $\Delta t=\sqrt{s}$, for any $T>0$ we have*

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

for some constant C.

Proof of Lemma 9. Let $z_k=(x_{k+1}-x_k)/\sqrt{s}$. Then Nesterov's scheme is equivalent to

$$\begin{aligned} x_{k+1} &= x_k + \sqrt{s} z_k \\ z_{k+1} &= \Big(1-\frac{3}{k+3}\Big) z_k - \sqrt{s}\nabla f\Big(x_k + \frac{2k+3}{k+3}\sqrt{s} z_k\Big) \end{aligned} \tag{24}$$

Denote by $a_k=\|X_k^{\delta}-x_k\|, \quad b_k=\|Z_k^{\delta}-z_k\|$, whose initial values are $a_0=0$ and $b_0=\|\nabla f(x_0)\|\sqrt{s}$. The idea of this proof is to bound a_k via simultaneously estimating a_k and b_k. By comparing (23) and (24), 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

$$a_k \le \sqrt{s} S_{k-1} \tag{25}$$

Similarly, for sufficiently small s we get

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

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

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

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

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

Making use of (25) gives

$$b_{k+1} \le \begin{cases} b_k + LsS_{k-1} + C_2\sqrt{s}, & k \le \delta/\sqrt{s} \\ b_k + LsS_{k-1} + \frac{C_2 s}{\delta}, & k > \delta/\sqrt{s}. \end{cases} \tag{26}$$

By induction on k, for $k \le \delta/\sqrt{s}$ 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 $k^\star = \lfloor \delta/\sqrt{s} \rfloor$, we get

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

which allows us to conclude that

$$a_k \le \sqrt{s} S_{k-1} = O(\delta^2) + o_s(1) \tag{27}$$

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

Next, we bound b_k for $k > k^\star = \lfloor \delta/\sqrt{s} \rfloor$. To this end, we consider the worst case of (26), that is,

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

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

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

By induction, we get

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

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

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

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 (27) and the last display, we get the desired result.

□

Now we turn to the proof of Theorem 5.

Proof of Theorem 5. Note the triangular inequality

$$\|x_k - X(k\sqrt{s})\| \le \|x_k - X_k^\delta\| + \|X_k^\delta - X_\delta(k\sqrt{s})\| + \|X_\delta(k\sqrt{s}) - X(k\sqrt{s})\|$$

where $X_\delta(\cdot)$ is the solution to the smoothed ODE (22). The proof of Lemma 2 implies that, we can choose a sequence $\delta_m \to 0$ such that

$$\sup_{0\le t\le T} \|X_{\delta_m}(t) - X(t)\| \to 0$$

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

5 Connections and Interpretations

In this section, we explore the approximate equivalence between the ODE and Nesterov's scheme, and provide evidence that the ODE can serve as an amenable tool for interpreting and analyzing Nesterov's scheme. The first subsection exhibits inverse quadratic convergence rate for the ODE solution, the next two address the oscillation phenomenon discussed in Section 3.1, and the last subsection is devoted to comparing Nesterov's scheme with gradient descent from a numerical perspective.

5.1 Analogous Convergence Rate

The original result from [Nesterov(1983)] states that, for any $f \in \mathcal{F}_L$, the sequence $\{x_k\}$ given by (17) with step size $s \le 1/L$ satisfies

$$f(x_k) - f^\star \le \frac{2\|x_0 - x^\star\|^2}{s(k+1)^2} \tag{28}$$

Our next result indicates that the trajectory of (19) closely resembles the sequence $\{x_k\}$ in terms of the convergence rate to a minimizer $x^\star$. Compared with the discrete case, this proof is shorter and simpler.

Theorem 6. *For any $f \in \mathcal{F}_\infty$, let $X(t)$ be the unique global solution to* (19) *with initial conditions $X(0) = x_0, \dot{X}(0) = 0$. Then, for any $t > 0$,*

$$f(X(t)) - f^\star \le \frac{2\|x_0 - x^\star\|^2}{t^2} \tag{29}$$

Proof of Theorem 6. Consider the energy functional[2] defined as $\mathcal{E}(t) = t^2(f(X(t)) - f^\star) + 2\|X + t\dot{X}/2 - x^\star\|^2$, whose time derivative is

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

[2]We may also view this functional as the negative entropy. Similarly, for the gradient flow $\dot{X} + \nabla f(X) = 0$, an energy function of form $\mathcal{E}_{\text{gradient}}(t) = t(f(X(t)) - f^\star) + \|X(t) - x^\star\|^2/2$ can be used to derive the bound $f(X(t)) - f^\star \le \frac{\|x_0 - x^\star\|^2}{2t}$.

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

$$\dot{\mathcal{E}} = 2t(f(X) - f^\star) + 4\langle X - x^\star, -t\nabla f(X)/2\rangle = 2t(f(X) - f^\star) - 2t\langle X - x^\star, \nabla f(X)\rangle \le 0$$

where the inequality follows from the convexity of f. Hence by monotonicity of $\mathcal{E}$ and non-negativity of $2\|X + t\dot{X}/2 - x^\star\|^2$, the gap satisfies

$$f(X(t)) - f^\star \le \frac{\mathcal{E}(t)}{t^2} \le \frac{\mathcal{E}(0)}{t^2} = \frac{2\|x_0 - x^\star\|^2}{t^2}$$

□

Making use of the approximation $t \approx k\sqrt{s}$, we observe that the convergence rate in (28) is essentially a discrete version of that in (29), 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 (29) is optimal. Consider an arbitrary $f \in \mathcal{F}_\infty(\mathbb{R})$ such that $f(x) = x$ for $x \ge 0$. Starting from some $x_0 > 0$, the solution to (19) is $X(t) = x_0 - t^2/8$ before hitting the origin. Hence, $t^2(f(X(t)) - f^\star) = t^2(x_0 - t^2/8)$ has a maximum $2x_0^2 = 2|x_0 - 0|^2$ achieved at $t = 2\sqrt{x_0}$. Therefore, we cannot replace 2 by any smaller number, and we can expect that this tightness also applies to the discrete analog (28).

5.2 Quadratic f and Bessel Functions

For quadratic f, the ODE (19) admits a solution in closed form. This closed form solution turns out to be very useful in understanding the issues raised in the introduction.

Let $f(x) = \frac{1}{2}\langle x, Ax\rangle + \langle b, x\rangle$, where $A \in \mathbb{R}^{n\times n}$ is a positive semidefinite matrix and b is in the column space of A because otherwise this function can attain $-\infty$. Then a simple translation in x can absorb the linear term $\langle b, x\rangle$ into the quadratic term. Since both the ODE and the scheme move within the affine space perpendicular to the kernel of A, without loss of generality, we assume that A is positive definite, admitting a spectral decomposition $A = Q^\top \Lambda Q$, where Λ is a diagonal matrix formed by the eigenvalues. Replacing x with Qx, we assume $f = \frac{1}{2}\langle x, \Lambda x\rangle$ from now on. Now, the ODE for this function admits a simple decomposition of form

$$\ddot{X}_i + \frac{3}{t}\dot{X}_i + \lambda_i X_i = 0, \quad i = 1, \dots, n$$

with $X_i(0) = x_{0,i}, \dot{X}_i(0) = 0$. Introduce $Y_i(u) = uX_i(u/\sqrt{\lambda_i})$, which satisfies

$$u^2\ddot{Y}_i + u\dot{Y}_i + (u^2 - 1)Y_i = 0$$

This is Bessel's differential equation of order one. Since Y_i vanishes at $u = 0$, we see that Y_i is a constant multiple of J_1, the Bessel function of the first kind of order one.[3] It has an analytic expansion:

$$J_1(u) = \sum_{m=0}^{\infty} \frac{(-1)^m}{(2m)!!(2m+2)!!} u^{2m+1}$$

[3] Up to a constant multiplier, J_1 is the unique solution to the Bessel's differential equation $u^2\ddot{J}_1 + u\dot{J}_1 + (u^2-1)J_1 = 0$ that is finite at the origin. In the analytic expansion of J_1, $m!!$ denotes the double factorial defined as $m!! = m \times (m-2) \times \cdots \times 2$ for even m, or $m!! = m \times (m-2) \times \cdots \times 1$ for odd m.

which gives the asymptotic expansion

$$J_1(u) = (1 + o(1))\frac{u}{2}$$

when $u \to 0$. Requiring $X_i(0) = x_{0,i}$, hence, we obtain

$$X_i(t) = \frac{2x_{0,i}}{t\sqrt{\lambda_i}} J_1(t\sqrt{\lambda_i}) \tag{30}$$

For large t, the Bessel function has the following asymptotic form; see e.g., [Watson(1995)]:

$$J_1(t) = \sqrt{\frac{2}{\pi t}}\Big(\cos(t - 3\pi/4) + O(1/t)\Big) \tag{31}$$

This asymptotic expansion yields (note that $f^\star = 0$)

$$f(X(t)) - f^\star = f(X(t)) = \sum_{i=1}^{n} \frac{2x_{0,i}^2}{t^2} J_1\left(t\sqrt{\lambda_i}\right)^2 = O\left(\frac{\|x_0 - x^\star\|^2}{t^3\sqrt{\min \lambda_i}}\right) \tag{32}$$

On the other hand, (31) and (32) give a lower bound:

$$\begin{aligned}
\limsup_{t\to\infty} t^3(f(X(t)) - f^\star) &\geq \lim_{t\to\infty} \frac{1}{t}\int_0^t u^3(f(X(u)) - f^\star)\mathrm{d}u \\
&= \lim_{t\to\infty} \frac{1}{t}\int_0^t \sum_{i=1}^{n} 2x_{0,i}^2 u J_1(u\sqrt{\lambda_i})^2 \mathrm{d}u \\
&= \sum_{i=1}^{n} \frac{2x_{0,i}^2}{\pi\sqrt{\lambda_i}} \geq \frac{2\|x_0 - x^\star\|^2}{\pi\sqrt{L}}
\end{aligned} \tag{33}$$

where $L = \|A\|_2$ is the spectral norm of A. The first inequality follows by interpreting $\lim_{t\to\infty} \frac{1}{t}\int_0^t u^3(f(X(u)) - f^\star)\mathrm{d}u$ as the mean of $u^3(f(X(u)) - f^\star)$ on $(0, \infty)$ in certain sense.

In view of (32), Nesterov's scheme might possibly exhibit $O(1/k^3)$ convergence rate for strongly convex functions. This convergence rate is consistent with the second inequality in Theorem 9. In Section 6.3, we prove the $O(1/t^3)$ rate for a generalized version of (19). However, (33) rules out the possibility of a higher order convergence rate.

Recall that the function considered in Figure 1 is $f(x) = 0.02x_1^2 + 0.005x_2^2$, starting from $x_0 = (1,\ 1)$. As the step size s becomes smaller, the trajectory of Nesterov's scheme converges to the solid curve represented via the Bessel function. While approaching the minimizer $x^\star$, each trajectory displays the oscillation pattern, as well-captured by the zoomed Figure 1b. This prevents Nesterov's scheme from achieving better convergence rate. The representation (30) offers excellent explanation as follows. Denote by T_1, T_2, respectively, the approximate periodicities of the first component $|X_1|$ in absolute value and the second $|X_2|$. By (31), we get $T_1 = \pi/\sqrt{\lambda_1} = 5\pi$ and $T_2 = \pi/\sqrt{\lambda_2} = 10\pi$. Hence, as the amplitude gradually decreases to zero, the function $f = 2x_{0,1}^2 J_1(\sqrt{\lambda_1}t)^2/t^2 + 2x_{0,2}^2 J_1(\sqrt{\lambda_2}t)^2/t^2$ has a major cycle of 10π, the least common multiple of T_1 and T_2. A careful look at Figure 1c reveals that within each major bump, roughly, there are $10\pi/T_1 = 2$ minor peaks.

5.3 Fluctuations of Strongly Convex f

The analysis carried out in the previous subsection only applies to convex quadratic functions. In this subsection, we extend the discussion to one-dimensional strongly convex functions. The Sturm-Picone theory; see e.g., [Hinton(2005)] is extensively used all along the analysis.

Let $f \in \mathcal{S}_{\mu,L}(\mathbb{R})$. Without loss of generality, assume f attains minimum at $x^\star = 0$. Then, by definition $\mu \le f'(x)/x \le L$ for any $x \neq 0$. Denoting by X the solution to the ODE (19), we consider the self-adjoint equation,

$$(t^3 Y')' + \frac{t^3 f'(X(t))}{X(t)} Y = 0 \tag{34}$$

which, apparently, admits a solution $Y(t) = X(t)$. To apply the Sturm-Picone comparison theorem, consider

$$(t^3 Y')' + \mu t^3 Y = 0$$

for a comparison. This equation admits a solution $\widetilde{Y}(t) = J_1(\sqrt{\mu}t)/t$. Denote by $\widetilde{t}_1 < \widetilde{t}_2 < \cdots$ all the positive roots of $J_1(t)$, which satisfy; see e.g., [Watson(1995)]

$$3.8317 = \widetilde{t}_1 - \widetilde{t}_0 > \widetilde{t}_2 - \widetilde{t}_3 > \widetilde{t}_3 - \widetilde{t}_4 > \cdots > \pi$$

where $\widetilde{t}_0 = 0$. Then, it follows that the positive roots of $\widetilde{Y}$ are $\widetilde{t}_1/\sqrt{\mu}$, $\widetilde{t}_2/\sqrt{\mu}, \ldots$. Since $t^3 f'(X(t))/X(t) \ge \mu t^3$, the Sturm-Picone comparison theorem asserts that $X(t)$ has a root in each interval $[\widetilde{t}_i/\sqrt{\mu}, \widetilde{t}_{i+1}/\sqrt{\mu}]$.

To obtain a similar result in the opposite direction, consider

$$(t^3 Y')' + L t^3 Y = 0 \tag{35}$$

Applying the Sturm-Picone comparison theorem to (34) and (35), we ensure that between any two consecutive positive roots of X, there is at least one $\widetilde{t}_i/\sqrt{L}$. Now, we summarize our findings in the following. Roughly speaking, this result concludes that the oscillation frequency of the ODE solution is between $O(\sqrt{\mu})$ and $O(\sqrt{L})$.

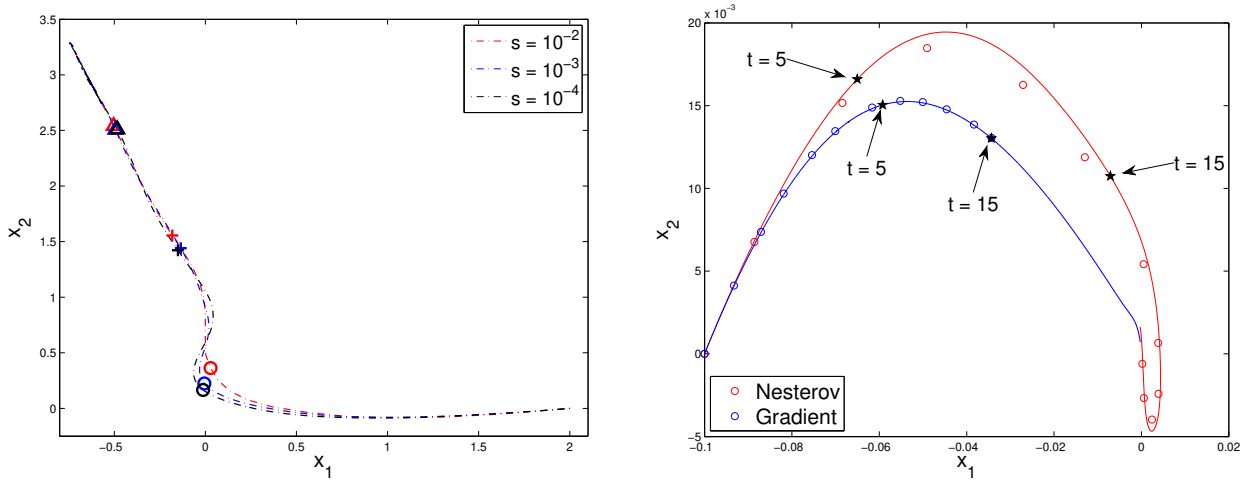
Theorem 7. *Denote by $0 < t_1 < t_2 < \cdots$ all the roots of $X(t) - x^\star$. Then these roots satisfy, for all $i \ge 1$,*

$$t_1 < \frac{7.6635}{\sqrt{\mu}},\ t_{i+1} - t_i < \frac{7.6635}{\sqrt{\mu}},\ t_{i+2} - t_i > \frac{\pi}{\sqrt{L}}$$

5.4 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 5. Consequently, for any constant $t_c > 0$, this implies that x_k does not change much for a range of step sizes s if $k \approx t_c/\sqrt{s}$. To empirically support this claim, we present an example in Figure 3a, where the scheme minimizes $f(x) = \|y - Ax\|^2/2 + \|x\|_1$ with $y = (4,\ 2,\ 0)$ and $A(:,1) = (0,\ 2,\ 4)$, $A(:,2) = (1,\ 1,\ 1)$ starting from $x_0 = (2,\ 0)$ (here $A(:,j)$ is the jth column of A). From this figure, we are delight to observe that x_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 (19), whereas it is known that the simple gradient descent $x_{k+1} = x_k - s\nabla f(x_k)$ moves s along the integral curve of $\dot{X} + \nabla f(X) = 0$. We expect that for small s Nesterov's scheme moves more in each iteration since $\sqrt{s}$ is much



(a) Square-root scaling of s.

(b) Race between Nesterov's and gradient.

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

larger than s. Figure 3b illustrates and supports this claim, where the function minimized is $f = |x_1|^3 + 5|x_2|^3 + 0.001(x_1 + x_2)^2$ with step size $s = 0.05$ (The coordinates are appropriately rotated to allow x_0 and $x^\star$ 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 3b 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 (19) in the presence of accumulated errors. The finite difference approximation by the forward Euler method is

$$\frac{X(t+\Delta t) - 2X(t) + X(t-\Delta t)}{\Delta t^2} + \frac{3}{t}\frac{X(t) - X(t-\Delta t)}{\Delta t} + \nabla f(X(t)) = 0 \tag{36}$$

which is equivalent to

$$X(t+\Delta t) = \left(2 - \frac{3\Delta t}{t}\right) X(t) - \Delta t^2 \nabla f(X(t)) - \left(1 - \frac{3\Delta t}{t}\right) X(t-\Delta t) \tag{37}$$

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 $k = t/\Delta t$, the characteristic equation of this finite difference scheme is approximately

$$\det\left(\lambda^2 - \left(2 - \Delta t^2 \nabla^2 f - \frac{3\Delta t}{t}\right)\lambda + 1 - \frac{3\Delta t}{t}\right) = 0 \tag{38}$$

The numerical stability of (36) with respect to accumulated errors is equivalent to this: all the roots of (38) lie in the unit circle; see e.g., [Leader(2004)]. When $\nabla^2 f \preceq L I_n$ (i.e. $L I_n - \nabla^2 f$ is positive semidefinite), if $\Delta t/t$ small and $\Delta t < 2/\sqrt{L}$, we see that all the roots of (38) lie in the unit circle. On the other hand, if $\Delta t > 2/\sqrt{L}$, (38) can possibly have a root λ outside the unit circle, causing numerical instability. Under our identification $s = \Delta t^2$, a step size of $s = 1/L$ in Nesterov's

scheme (17) is approximately equivalent to a step size of $\Delta t = 1/\sqrt{L}$ in the forward Euler method, which is stable for numerically integrating (36).

As a comparison, note that the finite difference scheme of the ODE $\dot{X}(t) + \nabla f(X(t)) = 0$, which models gradient descent with updates $x_{k+1} = x_k - s\nabla f(x_k)$, has the characteristic equation $\det(\lambda - (1 - \Delta t \nabla^2 f)) = 0$. Thus, to guarantee $-I_n \preceq 1 - \Delta t \nabla^2 f \preceq I_n$ in worst case analysis, one can only choose $\Delta t \leq 2/L$ for a fixed step size, which is much smaller than the step size $2/\sqrt{L}$ for (36) when ∇f is very variable, i.e., L is large.

6 Magic Constant $\lambda = 2$?

Recall that the constant $\lambda = 2$ appearing in the coefficient of $(\dot{X}, \dot{Z})$ in (19). This number leads to the momentum coefficient in (17) taking the form $(k-1)/(k+2) = 1 - 3/k + O(1/k^2)$. In this section, we demonstrate that 3 can be replaced by any larger number, while maintaining the $O(1/k^2)$ convergence rate. To begin with, let us consider the following ODE parameterized by a constant r:

$$\ddot{X} + \frac{1+\lambda}{t}\dot{X} + \nabla f(X) = 0 \tag{39}$$

with initial conditions $X(0) = x_0, \dot{X}(0) = 0$. The proof of Theorem 4, which seamlessly applies here, guarantees the existence and uniqueness of the solution X to this ODE.

Interpreting the damping ratio r/t as a measure of friction[4] in the damping system, our results say that more friction does not end the $O(1/t^2)$ and $O(1/k^2)$ convergence rate. On the other hand, in the lower friction setting, where r is smaller than 3, we can no longer expect inverse quadratic convergence rate, unless some additional structures of f are imposed. We believe that this striking phase transition at 3 deserves more attention as an interesting research challenge.

6.1 High Friction

Here, we study the convergence rate of (39) with $\lambda + 1 > 3$ and $f \in \mathcal{F}_\infty$. Compared with (19), this new ODE as a damping suffers from higher friction. Following the strategy adopted in the proof of Theorem 6, we consider a new energy functional defined as

$$\mathcal{E}(t) = \frac{2t^2}{\lambda}(f(X(t)) - f^\star) + \lambda \left\| X(t) + \frac{t}{\lambda}\dot{X}(t) - x^\star \right\|^2$$

By studying the derivative of this functional, we get the following result.

Theorem 8. *The solution X to (39) satisfies*

$$f(X(t)) - f^\star \leq \frac{\lambda^2 \|x_0 - x^\star\|^2}{2t^2}, \quad \int_0^\infty t(f(X(t)) - f^\star)\mathrm{d}t \leq \frac{\lambda^2 \|x_0 - x^\star\|^2}{2(\lambda - 2)}$$

Proof of Theorem 8. Noting $r\dot{X} + t\ddot{X} = -t\nabla f(X)$, we get $\dot{\mathcal{E}}$ equal to

$$\frac{4t}{\lambda}(f(X) - f^\star) + \frac{2t^2}{\lambda}\langle \nabla f, \dot{X}\rangle + 2\langle X + \frac{t}{\lambda}\dot{X} - x^\star, r\dot{X} + t\ddot{X}\rangle$$

[4] In physics and engineering, damping may be modeled as a force proportional to velocity but opposite in direction, i.e. resisting motion; for instance, this force may be used as an approximation to the friction caused by drag. In our model, this force would be proportional to $-\frac{1+\lambda}{t}\dot{X}$ where $\dot{X}$ is velocity and $\frac{1+\lambda}{t}$ is the damping coefficient.

$$= \frac{4t}{\lambda}(f(X) - f^\star) - 2t\langle X - x^\star, \nabla f(X)\rangle \le -\frac{2(\lambda-2)t}{\lambda}(f(X) - f^\star) \quad (40)$$

where the inequality follows from the convexity of f. Since $f(X) \ge f^\star$, the last display implies that $\mathcal{E}$ is non-increasing. Hence

$$\frac{2t^2}{\lambda}(f(X(t)) - f^\star) \le \mathcal{E}(t) \le \mathcal{E}(0) = \lambda\|x_0 - x^\star\|^2$$

yielding the first inequality of this theorem. To complete the proof, from (40) it follows that

$$\int_0^\infty \frac{2(\lambda-2)t}{\lambda}(f(X) - f^\star)\mathrm{d}t \le -\int_0^\infty \frac{\mathrm{d}\mathcal{E}}{\mathrm{d}t}\mathrm{d}t = \mathcal{E}(0) - \mathcal{E}(\infty) \le \lambda\|x_0 - x^\star\|^2$$

as desired for establishing the second inequality. □

The first inequality is the same as (29) for the ODE (19), except for a larger constant $\lambda^2/2$. The second inequality measures the error $f(X(t)) - f^\star$ in an average sense, and cannot be deduced from the first inequality.

Now, it is tempting to obtain such analogs for the discrete Nesterov's scheme as well. Parametrizing by a constant r, we propose the generalized Nesterov's scheme,

$$z_k = z_{k-1} - \frac{k+\lambda-1}{\lambda}s\nabla f(y_{k-1}) \quad (41a)$$

$$x_k = \frac{k-1}{k+\lambda-1}x_{k-1} + \frac{\lambda}{k+\lambda-1}z_k \quad (41b)$$

$$y_k = \frac{k}{k+\lambda}x_k + \frac{\lambda}{k+\lambda}z_k \quad (41c)$$

starting from $y_0 = x_0$. The discrete analog of Theorem 8 is below.

Theorem 9. *The sequence $\{x_k\}$ given by (41) with $0 < s \le 1/L$ satisfies*

$$f(x_k) - f^\star \le \frac{\lambda^2\|x_0 - x^\star\|^2}{2s(k+\lambda-1)^2}, \quad \sum_{k=1}^{\infty}(k+\lambda)(f(x_k) - f^\star) \le \frac{\lambda^2\|x_0 - x^\star\|^2}{2s(\lambda-2)}$$

The first inequality suggests that the generalized Nesterov's schemes still achieve $O(1/k^2)$ convergence rate. However, if the error bound satisfies $f(x_{k'}) - f^\star \ge c/k'^2$ for some arbitrarily small $c > 0$ and a dense subsequence $\{k'\}$, i.e., $|\{k'\} \cap \{1, \dots, m\}| \ge \alpha m$ for all $m \ge 1$ and some $\alpha > 0$, then the second inequality of the theorem would be violated. To see this, note that if it were the case, we would have $(k'+\lambda)(f(x_{k'}) - f^\star) \gtrsim \frac{1}{k'}$; the sum of the harmonic series $\frac{1}{k'}$ over a dense subset of $\{1, 2, \dots\}$ is infinite. Hence, the second inequality is not trivial because it implies the error bound is, in some sense, $O(1/k^2)$ suboptimal.

Now we turn to the proof of this theorem. It is worth pointing out that, though based on the same idea, the proof below is much more complicated than that of Theorem 8.

Proof of Theorem 9. Consider the discrete energy functional,

$$\mathcal{E}(k) = \frac{2(k+\lambda-1)^2 s}{\lambda}(f(x_k) - f^\star) + \lambda\|z_k - x^\star\|^2$$

where $z_k = (k+\lambda)y_k/\lambda - kx_k/\lambda$. If we have

$$\mathcal{E}(k) + \frac{2s[(\lambda-2)(k+\lambda-1)+1]}{\lambda}(f(x_{k-1}) - f^\star) \le \mathcal{E}(k-1) \tag{42}$$

then it would immediately yield the desired results by summing (42) over k. That is, by recursively applying (42), we see

$$\mathcal{E}(k) + \sum_{i=1}^{k} \frac{2s[(\lambda-2)(i+\lambda-1)+1]}{\lambda}(f(x_{i-1}) - f^\star) \\ \le \mathcal{E}(0) = \frac{2(\lambda-1)^2 s}{\lambda}(f(x_0) - f^\star) + \lambda\|x_0 - x^\star\|^2$$

which is equivalent to

$$\mathcal{E}(k) + \sum_{i=1}^{k-1} \frac{2s[(\lambda-2)(i+\lambda)+1]}{\lambda}(f(x_i) - f^\star) \le \lambda\|x_0 - x^\star\|^2 \tag{43}$$

Noting that the left-hand side of (43) is lower bounded by $2s(k+\lambda-1)^2(f(x_k) - f^\star)/\lambda$, we thus obtain the first inequality of the theorem. Since $\mathcal{E}(k) \ge 0$, the second inequality is verified via taking the limit $k \to \infty$ in (43) and replacing $(\lambda-2)(i+\lambda)+1$ by $(\lambda-2)(i+\lambda)$.

We now establish (42). For $s \le 1/L$, we have the basic inequality,

$$f(y - sG_s(y)) \le f(x) + G_s(y)^\top(y-x) - \frac{s}{2}\|G_s(y)\|^2 \tag{44}$$

for any x and y. Note that $y_{k-1} - sG_s(y_{k-1})$ actually coincides with x_k. Summing of $(k-1)/(k+\lambda-1) \times$ (44) with $x = x_{k-1}, y = y_{k-1}$ and $\lambda/(k+\lambda-1) \times$ (44) with $x = x^\star, y = y_{k-1}$ gives

$$\begin{aligned} f(x_k) &\le \frac{k-1}{k+\lambda-1}f(x_{k-1}) + \frac{\lambda}{k+\lambda-1}f^\star \\ &\quad + \frac{\lambda}{k+\lambda-1}G_s(y_{k-1})^\top\Big(\frac{k+\lambda-1}{\lambda}y_{k-1} - \frac{k-1}{\lambda}x_{k-1} - x^\star\Big) - \frac{s}{2}\|G_s(y_{k-1})\|^2 \\ &= \frac{k-1}{k+\lambda-1}f(x_{k-1}) + \frac{\lambda}{k+\lambda-1}f^\star + \frac{\lambda^2}{2s(k+\lambda-1)^2}\Big(\|z_{k-1} - x^\star\|^2 - \|z_k - x^\star\|^2\Big), \end{aligned}$$

where we use $z_{k-1} - s(k+\lambda-1)G_s(y_{k-1})/\lambda = z_k$. Rearranging the above inequality and multiplying by $2s(k+\lambda-1)^2/\lambda$ gives the desired (42).

□

In closing, we would like to point out this new scheme is equivalent to setting $\theta_k = \lambda/(k+\lambda)$ and letting $\theta_k(\theta_{k-1}^{-1} - 1)$ replace the momentum coefficient $(k-1)/(k+\lambda)$. Then, the equal sign " = " in the update $\theta_{k+1} = (\sqrt{\theta_k^4 + 4\theta_k^2} - \theta_k^2)/2$ has to be replaced by an inequality sign " $\ge$ ". In examining the proof of Theorem 1(b) in [Tseng(2010)], we can get an alternative proof of Theorem 9.

6.2 Low Friction

Now we turn to the case $\lambda < 2$. Then, unfortunately, the energy functional approach for proving Theorem 8 is no longer valid, since the left-hand side of (40) is positive in general. In fact, there

are counterexamples that fail the desired $O(1/t^2)$ or $O(1/k^2)$ convergence rate. We present such examples in continuous time. Equally, these examples would also violate the $O(1/k^2)$ convergence rate in the discrete schemes, and we forego the details.

Let $f(x) = \frac{1}{2}\|x\|^2$ and X be the solution to (39). Then, $Y = t^{\frac{\lambda}{2}}X$ satisfies

$$t^2\ddot{Y} + t\dot{Y} + (t^2 - \lambda^2/4)Y = 0$$

With the initial condition $Y(t) \approx t^{\frac{\lambda}{2}}x_0$ for small t, the solution to the above Bessel equation in a vector form of order $\lambda/2$ is $Y(t) = 2^{\frac{\lambda}{2}}\Gamma((\lambda+2)/2)J_{\lambda/2}(t)x_0$. Thus,

$$X(t) = \frac{2^{\frac{\lambda}{2}}\Gamma((\lambda+2)/2)J_{\lambda/2}(t)}{t^{\frac{\lambda}{2}}}x_0$$

For large t, the Bessel function $J_{\lambda/2}(t) = \sqrt{2/(\pi t)}\big(\cos(t - \lambda\pi/4 - \pi/4) + O(1/t)\big)$. Hence,

$$f(X(t)) - f^\star = O\left(\|x_0 - x^\star\|^2/t^{\lambda+1}\right)$$

where the exponent r is tight. This rules out the possibility of inverse quadratic convergence of the generalized ODE and scheme for all $f \in \mathcal{F}_L$ if $\lambda < 1$. An example with $\lambda = 0$ is plotted in Figure 2.

Next, we consider the case $1 \le \lambda < 2$ and let $f(x) = |x|$ (this also applies to multivariate $f = \|x\|$).[5] Starting from $x_0 > 0$, we get $X(t) = x_0 - \frac{t^2}{2\lambda}$ for $t \le \sqrt{2\lambda x_0}$. Requiring continuity of X and $\dot{X}$ at the change point 0, we get

$$X(t) = \frac{t^2}{2\lambda} + \frac{2(2\lambda x_0)^{\frac{\lambda+2}{2}}}{((\lambda+1)^2 - 1)t^\lambda} - \frac{\lambda+4}{\lambda}x_0$$

for $\sqrt{2\lambda x_0} < t \le \sqrt{2c^\star\lambda x_0}$, where $c^\star$ is the positive root other than 1 of $\lambda c + 4c^{-\frac{\lambda}{2}} = r+3$. Repeating this process solves for X. Note that $t^{-\lambda}$ is in the null space of $\ddot{X} + r\dot{X}/t$ and satisfies $t^2 \times t^{-\lambda} \to \infty$ as $t \to \infty$. For illustration, Figure 4 plots $t^2(f(X(t)) - f^\star)$ and $sk^2(f(x_k) - f^\star)$ with $r = 2, 2.5$, and $r = 4$ for comparison[6]. It is clearly that inverse quadratic convergence does not hold for $r = 2, 2.5$, that is, (18) does not hold for $\lambda < 2$. Interestingly, in Figures 4a and 4d, the scaled errors at peaks grow linearly, whereas for $r = 2.5$, the growth rate, though positive as well, seems sublinear.

However, if f possesses some additional property, inverse quadratic convergence is still guaranteed, as stated below. In that theorem, f is assumed to be a continuously differentiable convex function.

Theorem 10. *Suppose $1 < \lambda < 2$ and let X be a solution to the ODE (39). If $(f - f^\star)^{\frac{\lambda}{2}}$ is also convex, then*

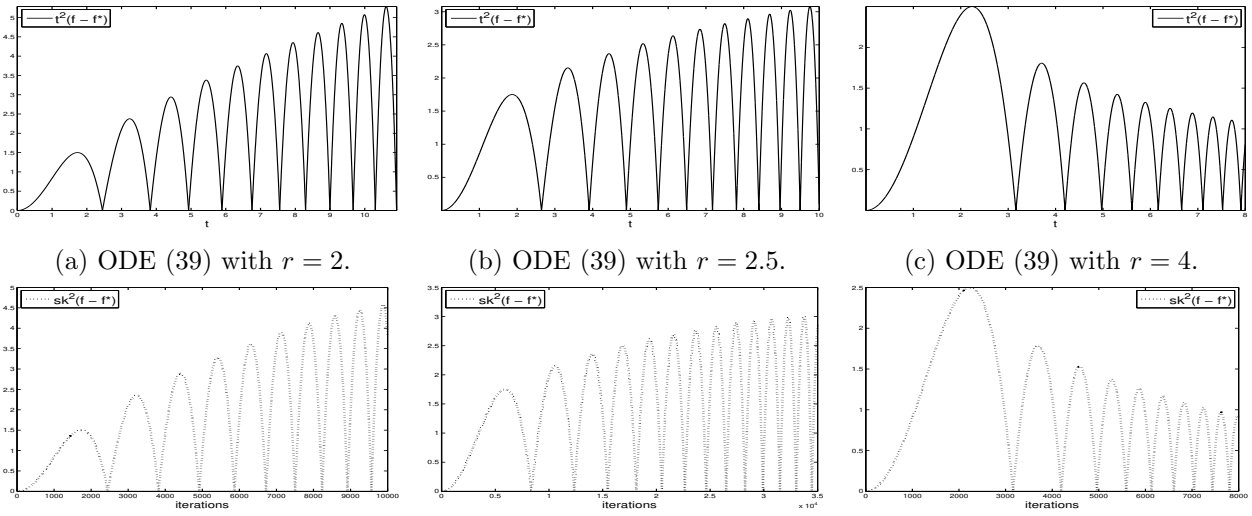
$$f(X(t)) - f^\star \le \frac{\lambda^2\|x_0 - x^\star\|^2}{2t^2}$$

Proof of Theorem 10. Since $(f - f^\star)^{\frac{\lambda}{2}}$ is convex, we obtain

$$(f(X(t)) - f^\star)^{\frac{\lambda}{2}} \le \langle X - x^\star, \nabla(f(X) - f^\star)^{\frac{\lambda}{2}}\rangle = \frac{\lambda}{2}(f(X) - f^\star)^{\frac{\lambda-2}{2}}\langle X - x^\star, \nabla f(X)\rangle$$

[5]This function does not have a Lipschitz continuous gradient. However, a similar pattern as in Figure 2 can be also observed if we smooth $|x|$ at an arbitrarily small vicinity of 0.

[6]For Figures 4d, 4e and 4f, if running generalized Nesterov's schemes with too many iterations (e.g. 10^5), the deviations from the ODE will grow. Taking a sufficiently small s can solve this issue.



(a) ODE (39) with $r = 2$. (b) ODE (39) with $r = 2.5$. (c) ODE (39) with $r = 4$.

(d) Scheme (41) with $r = 2$. (e) Scheme (41) with $r = 2.5$. (f) Scheme (41) with $r = 4$.

Figure 4. Scaled errors $t^2(f(X(t)) - f^\star)$ and $sk^2(f(x_k) - f^\star)$ of generalized ODEs and schemes for minimizing $f = |x|$. In (d), the step size $s = 10^{-6}$, in (e), $s = 10^{-7}$, and in (f), $s = 10^{-6}$.

which can be simplified to $\frac{2}{\lambda}(f(X) - f^\star) \leq \langle X - x^\star, \nabla f(X)\rangle$. This inequality combined with (40) leads to the monotonically decreasing of $\mathcal{E}(t)$ defined for Theorem 8. This completes the proof by noting $f(X) - f^\star \leq \lambda\mathcal{E}(t)/(2t^2) \leq \lambda\mathcal{E}(0)/(2t^2) = \lambda^2\|x_0 - x^\star\|^2/(2t^2)$. □

6.3 Strongly Convex f

Strong convexity is a desirable property for optimization. Making use of this property carefully suggests a generalized Nesterov's scheme that achieves optimal linear convergence [Nesterov(2004)]. In that case, even vanilla gradient descent has a linear convergence rate. Unfortunately, the example given in the previous subsection simply rules out such possibility for (17) and its generalizations (41). However, from a different perspective, this example suggests that $O(t^{-\lambda-1})$ convergence rate can be expected for (39). In the next theorem, we prove a slightly weaker statement of this kind, that is, a provable $O(t^{-\frac{2\lambda+2}{3}})$ convergence rate is established for strongly convex functions. Bridging this gap may require new tools and more careful analysis.

Let $f \in \mathcal{S}_{\mu,L}(\mathbb{R}^n)$ and consider a new energy functional for $\alpha > 2$ defined as

$$\mathcal{E}(t;\alpha) = t^\alpha(f(X(t)) - f^\star) + \frac{(2\lambda + 2 - \alpha)^2 t^{\alpha-2}}{8}\left\|X(t) + \frac{2t}{2\lambda + 2 - \alpha}\dot{X} - x^\star\right\|^2$$

When clear from the context, $\mathcal{E}(t;\alpha)$ is simply denoted as $\mathcal{E}(t)$. For $\lambda + 1 > 3$, taking $\alpha = 2r/3$ in the theorem stated below gives $f(X(t)) - f^\star \lesssim \|x_0 - x^\star\|^2/t^{\frac{2\lambda+2}{3}}$.

Theorem 11. *For any $f \in \mathcal{S}_{\mu,L}(\mathbb{R}^n)$, if $2 \leq \alpha \leq 2r/3$ we get*

$$f(X(t)) - f^\star \leq \frac{C\|x_0 - x^\star\|^2}{\mu^{\frac{\alpha-2}{2}} t^\alpha}$$

for any $t > 0$. Above, the constant C only depends on α and r.

Proof of Theorem 11. Note that $\dot{\mathcal{E}}(t;\alpha)$ equals

$$\alpha t^{\alpha-1}(f(X)-f^\star)-\frac{(2\lambda+2-\alpha)t^{\alpha-1}}{2}\langle X-x^\star,\nabla f(X)\rangle+\frac{(\alpha-2)(2\lambda+2-\alpha)^2t^{\alpha-3}}{8}\|X-x^\star\|^2 \\ +\frac{(\alpha-2)(2\lambda+2-\alpha)t^{\alpha-2}}{4}\langle\dot{X},X-x^\star\rangle \quad (45)$$

By the strong convexity of f, the second term of the right-hand side of (45) is bounded below as

$$\frac{(2\lambda+2-\alpha)t^{\alpha-1}}{2}\langle X-x^\star,\nabla f(X)\rangle\geq\frac{(2\lambda+2-\alpha)t^{\alpha-1}}{2}(f(X)-f^\star)+\frac{\mu(2\lambda+2-\alpha)t^{\alpha-1}}{4}\|X-x^\star\|^2$$

Substituting the last display into (45) with the awareness of $\lambda\geq 3\alpha/2-1$ yields

$$\dot{\mathcal{E}}\leq-\frac{(2\mu(2\lambda+2-\alpha)t^2-(\alpha-2)(2\lambda+2-\alpha)^2)t^{\alpha-3}}{8}\|X-x^\star\|^2+\frac{(\alpha-2)(2\lambda+2-\alpha)t^{\alpha-2}}{8}\frac{\mathrm{d}\|X-x^\star\|^2}{\mathrm{d}t}$$

Hence, if $t\geq t_\alpha:=\sqrt{(\alpha-2)(2\lambda+2-\alpha)/(2\mu)}$, we obtain

$$\dot{\mathcal{E}}(t)\leq\frac{(\alpha-2)(2\lambda+2-\alpha)t^{\alpha-2}}{8}\frac{\mathrm{d}\|X-x^\star\|^2}{\mathrm{d}t}$$

Integrating the last inequality on the interval (t_α,t) gives

$$\mathcal{E}(t)\leq\mathcal{E}(t_\alpha)+\frac{(\alpha-2)(2\lambda+2-\alpha)t^{\alpha-2}}{8}\|X(t)-x^\star\|^2-\frac{(\alpha-2)(2\lambda+2-\alpha)t_\alpha^{\alpha-2}}{8}\|X(t_\alpha)-x^\star\|^2 \\ -\frac{1}{8}\int_{t_\alpha}^t(\alpha-2)^2(2\lambda+2-\alpha)u^{\alpha-3}\|X(u)-x^\star\|^2\mathrm{d}u\leq\mathcal{E}(t_\alpha)+\frac{(\alpha-2)(2\lambda+2-\alpha)t^{\alpha-2}}{8}\|X(t)-x^\star\|^2 \\ \leq\mathcal{E}(t_\alpha)+\frac{(\alpha-2)(2\lambda+2-\alpha)t^{\alpha-2}}{4\mu}(f(X(t))-f^\star) \quad (46)$$

Making use of (46), we apply induction on α to finish the proof. First, consider $2<\alpha\leq 4$. Applying Theorem 8, from (46) we get that $\mathcal{E}(t)$ is upper bounded by

$$\mathcal{E}(t_\alpha)+\frac{(\alpha-2)\lambda^2(2\lambda+2-\alpha)\|x_0-x^\star\|^2}{8\mu t^{4-\alpha}}\leq\mathcal{E}(t_\alpha)+\frac{(\alpha-2)\lambda^2(2\lambda+2-\alpha)\|x_0-x^\star\|^2}{8\mu t_\alpha^{4-\alpha}} \quad (47)$$

Then, we bound $\mathcal{E}(t_\alpha)$ as follows.

$$\mathcal{E}(t_\alpha)\leq t_\alpha^\alpha(f(X(t_\alpha))-f^\star)+\frac{(2\lambda+2-\alpha)^2t_\alpha^{\alpha-2}}{4}\Big\|\frac{2\lambda}{2\lambda+2-\alpha}X(t_\alpha)+\frac{2t_\alpha}{2\lambda+2-\alpha}\dot{X}(t_\alpha)-\frac{2\lambda}{2\lambda+2-\alpha}x^\star\Big\|^2 \\ +\frac{(2\lambda+2-\alpha)^2t_\alpha^{\alpha-2}}{4}\Big\|\frac{\alpha-2}{2\lambda+2-\alpha}X(t_\alpha)-\frac{\alpha-2}{2\lambda+2-\alpha}x^\star\Big\|^2 \\ \leq\lambda^2t_\alpha^{\alpha-2}\|x_0-x^\star\|^2+\frac{(\alpha-2)^2\lambda^2\|x_0-x^\star\|^2}{4\mu t_\alpha^{4-\alpha}} \quad (48)$$

where in the second inequality we use the decreasing property of the energy functional defined for Theorem 8. Combining (47) and (48), we have

$$\mathcal{E}(t)\leq\lambda^2t_\alpha^{\alpha-2}\|x_0-x^\star\|^2+\frac{(\alpha-2)\lambda^2(2\lambda+\alpha-6)\|x_0-x^\star\|^2}{8\mu t_\alpha^{4-\alpha}}=O\Big(\frac{\|x_0-x^\star\|^2}{\mu^{\frac{\alpha-2}{2}}}\Big)$$

For $t \geq t_\alpha$, it suffices to apply $f(X(t)) - f^\star \leq \mathcal{E}(t)/t^3$ to the last display. For $t < t_\alpha$, by Theorem 8, $f(X(t)) - f^\star$ is upper bounded by

$$\begin{aligned}\frac{\lambda^2\|x_0 - x^\star\|^2}{2t^2} &\leq \frac{\lambda^2 \mu^{\frac{\alpha-2}{2}}[(\alpha-2)(2\lambda+2-\alpha)/(2\mu)]^{\frac{\alpha-2}{2}}}{2} \frac{\|x_0 - x^\star\|^2}{\mu^{\frac{\alpha-2}{2}} t^\alpha} \\ &= O\Big(\frac{\|x_0 - x^\star\|^2}{\mu^{\frac{\alpha-2}{2}} t^\alpha}\Big)\end{aligned} \tag{49}$$

Next, suppose that the theorem is valid for some $\widetilde{\alpha} > 2$. We show below that this theorem is still valid for $\alpha := \widetilde{\alpha} + 1$ if still $\lambda \geq 3\alpha/2 - 1$. By the assumption, (46) further induces

$$\mathcal{E}(t) \leq \mathcal{E}(t_\alpha) + \frac{(\alpha-2)(2\lambda+2-\alpha)t^{\alpha-2}}{4\mu} \frac{\widetilde{C}\|x_0 - x^\star\|^2}{\mu^{\frac{\widetilde{\alpha}-2}{2}} t^{\widetilde{\alpha}}} \leq \mathcal{E}(t_\alpha) + \frac{\widetilde{C}(\alpha-2)(2\lambda+2-\alpha)\|x_0 - x^\star\|^2}{4\mu^{\frac{\alpha-1}{2}} t_\alpha}$$

for some constant $\widetilde{C}$ only depending on $\widetilde{\alpha}$ and r. This inequality with (48) implies

$$\begin{aligned}\mathcal{E}(t) &\leq \lambda^2 t_\alpha^{\alpha-2}\|x_0 - x^\star\|^2 + \frac{(\alpha-2)^2\lambda^2\|x_0 - x^\star\|^2}{4\mu t_\alpha^{4-\alpha}} + \frac{\widetilde{C}(\alpha-2)(2\lambda+2-\alpha)\|x_0 - x^\star\|^2}{4\mu^{\frac{\alpha-1}{2}} t_\alpha} \\ &= O\left(\|x_0 - x^\star\|^2/\mu^{\frac{\alpha-2}{2}}\right)\end{aligned}$$

which verify the induction for $t \geq t_\alpha$. As for $t < t_\alpha$, the validity of the induction follows from Theorem 8, similarly to (49). Thus, combining the base and induction steps, the proof is completed. □

It should be pointed out that the constant C in the statement of Theorem 11 grows with the parameter r. Hence, simply increasing r does not guarantee to give a better error bound. While it is desirable to expect a discrete analogy of Theorem 11, i.e., $O(1/k^\alpha)$ convergence rate for (41), a complete proof can be notoriously complicated. That said, we mimic the proof of Theorem 11 for $\alpha = 3$ and succeed in obtaining a $O(1/k^3)$ convergence rate for the generalized Nesterov's schemes, as summarized in the theorem below.

Theorem 12. *Suppose f is written as $f = g + h$, where $g \in \mathcal{S}_{\mu,L}$ and h is convex with possible extended value ∞. Then, the generalized Nesterov's scheme* (41) *with $\lambda \geq 7/2$ and $s = 1/L$ satisfies*

$$f(x_k) - f^\star \leq \frac{CL\|x_0 - x^\star\|^2}{k^2} \frac{\sqrt{L/\mu}}{k}$$

where C only depends on r.

This theorem states that the discrete scheme (41) enjoys the error bound $O(1/k^3)$ without any knowledge of the condition number L/μ. In particular, this bound is much better than that given in Theorem 9 if $k \gg \sqrt{L/\mu}$. The strategy of the proof is fully inspired by that of Theorem 11, though it is much more complicated. The relevant energy functional $\mathcal{E}(k)$ for this Theorem 12 is equal to

$$\begin{aligned}&\frac{s(2k+3\lambda-2)(2k+2\lambda-7)(4k+4\lambda-13)}{16}(f(x_k) - f^\star) \\ &\qquad + \frac{2k+3\lambda-2}{16}\|2(k+\lambda)y_k - (2k+1)x_k - (2\lambda-1)x^\star\|^2\end{aligned} \tag{50}$$

6.4 Proof of Theorem 12

Proof of Theorem 12. Let g be μ–strongly convex and h be convex. For $f = g + h$, we show that (44) can be strengthened to

$$f(y - sG_s(y)) \le f(x) + G_s(y)^\top (y - x) - \frac{s}{2}\|G_s(y)\|^2 - \frac{\mu}{2}\|y - x\|^2 \tag{51}$$

Summing $(4k-3) \times$ (51) with $x = x_{k-1}, y = y_{k-1}$ and $(4r-6) \times$ (51) with $x = x^\star, y = y_{k-1}$ yields

$$\begin{aligned}
(4k+4\lambda-13)f(x_k) &\le (4k-3)f(x_{k-1}) + (4r-6)f^\star \\
&\quad + G_s(y_{k-1})^\top[(4k+4\lambda-13)y_{k-1} - (4k-3)x_{k-1} - (4r-6)x^\star] \\
&\quad - \frac{s(4k+4\lambda-13)}{2}\|G_s(y_{k-1})\|^2 - \frac{\mu(4k-3)}{2}\|y_{k-1} - x_{k-1}\|^2 - \mu(2\lambda-1)\|y_{k-1} - x^\star\|^2 \\
&\le (4k-3)f(x_{k-1}) + (4r-6)f^\star - \mu(2\lambda-1)\|y_{k-1} - x^\star\|^2 \\
&\quad + G_s(y_{k-1})^\top\left[(4k+4\lambda-13)(y_{k-1} - x^\star) - (4k-3)(x_{k-1} - x^\star)\right]
\end{aligned} \tag{52}$$

which gives a lower bound on $G_s(y_{k-1})^\top\left[(4k+4\lambda-13)y_{k-1} - (4k-3)x_{k-1} - (4r-6)x^\star\right]$. Denote by Δ_k the second term of $\widetilde{\mathcal{E}}(k)$ in (50), namely,

$$\Delta_k \triangleq \frac{k+d}{8}\|(2k+2\lambda-4)(y_k - x^\star) - (2k+1)(x_k - x^\star)\|^2$$

where $d := 3r/2 - 5/2$. Then by (52), we get

$$\begin{aligned}
\Delta_k - \Delta_{k-1} &= -\frac{k+d}{8}\Big\langle s(2\lambda+2k-7)G_s(y_{k-1}) + \frac{k-2}{k+\lambda-1}(x_{k-1} - x_{k-2}), (4k+4\lambda-13)(y_{k-1} - x^\star) \\
&\quad - (4k-3)(x_{k-1} - x^\star)\Big\rangle + \frac{1}{8}\|(2k+2\lambda-6)(y_{k-1} - x^\star) - (2k-1)(x_{k-1} - x^\star)\|^2 \\
&\le -\frac{s(k+d)(2k+2\lambda-7)}{8}\big[(4k+4\lambda-13)(f(x_k) - f^\star) \\
&\quad - (4k-3)(f(x_{k-1}) - f^\star) + \mu(2\lambda-1)\|y_{k-1} - x^\star\|^2\big] \\
&\quad - \frac{(k+d)(k-2)}{8(k+\lambda-1)}\langle x_{k-1} - x_{k-2}, (4k+4\lambda-13)(y_{k-1} - x^\star) - (4k-3)(x_{k-1} - x^\star)\rangle \\
&\quad + \frac{1}{8}\|2(k+\lambda-1)(y_{k-1} - x^\star) - (2k-1)(x_{k-1} - x^\star)\|^2.
\end{aligned}$$

Hence,

$$\begin{aligned}
\Delta_k + \frac{s(k+d)(2k+2\lambda-7)(4k+4\lambda-13)}{8}(f(x_k) - f^\star) \\
\le \Delta_{k-1} + \frac{s(k+d)(2k+2\lambda-7)(4k-3)}{8}(f(x_{k-1}) - f^\star) \\
- \frac{s\mu(2\lambda-1)(k+d)(2k+2\lambda-7)}{8}\|y_{k-1} - x^\star\|^2 + \Pi_1 + \Pi_2
\end{aligned} \tag{53}$$

where

$$\Pi_1 \triangleq -\frac{(k+d)(k-2)}{8(k+\lambda-1)}\langle x_{k-1} - x_{k-2}, (4k+4\lambda-13)(y_{k-1} - x^\star) - (4k-3)(x_{k-1} - x^\star)\rangle$$

$$\Pi_2 \triangleq \frac{1}{8}\|2(k+\lambda-1)(y_{k-1}-x^\star)-(2k-1)(x_{k-1}-x^\star)\|^2$$

By the iterations defined in (41), one can show that

$$\begin{aligned}\Pi_1 = &-\frac{(2\lambda-1)(k+d)(k-2)}{8(k+\lambda-1)}(\|x_{k-1}-x^\star\|^2-\|x_{k-2}-x^\star\|^2)\\ &-\frac{(k-2)^2(4k+4\lambda-13)(k+d)+(2\lambda-1)(k-2)(k+\lambda-1)(k+d)}{8(k+\lambda-1)^2}\|x_{k-1}-x_{k-2}\|^2\end{aligned}$$

$$\begin{aligned}\Pi_2 = &\frac{(2\lambda-1)^2}{8}\|y_{k-1}-x^\star\|^2+\frac{(2\lambda-1)(2k-1)(k-2)}{8(k+\lambda-1)}(\|x_{k-1}-x^\star\|^2-\|x_{k-2}-x^\star\|^2)\\ &+\frac{(k-2)^2(2k-1)(2k+4r-7)+(2\lambda-1)(2k-1)(k-2)(k+\lambda-1)}{8(k+\lambda-1)^2}\|x_{k-1}-x_{k-2}\|^2\end{aligned}$$

Although this is a little tedious, it is straightforward to check that $(k-2)^2(4k+4\lambda-13)(k+d)+(2\lambda-1)(k-2)(k+\lambda-1)(k+d) \geq (k-2)^2(2k-1)(2k+4r-7)+(2\lambda-1)(2k-1)(k-2)(k+\lambda-1)$ for any k. Therefore, $\Pi_1+\Pi_2$ is bounded as

$$\Pi_1+\Pi_2 \leq \frac{(2\lambda-1)^2}{8}\|y_{k-1}-x^\star\|^2+\frac{(2\lambda-1)(k-d-1)(k-2)}{8(k+\lambda-1)}(\|x_{k-1}-x^\star\|^2-\|x_{k-2}-x^\star\|^2)$$

which, together with the fact that $s\mu(2\lambda-1)(k+d)(2k+2\lambda-7) \geq (2\lambda-1)^2$ when $k \geq \sqrt{(2\lambda-1)/(2s\mu)}$, reduces (53) to

$$\begin{aligned}&\Delta_k+\frac{s(k+d)(2k+2\lambda-7)(4k+4\lambda-13)}{8}(f(x_k)-f^\star)\\ &\qquad\leq \Delta_{k-1}+\frac{s(k+d)(2k+2\lambda-7)(4k-3)}{8}(f(x_{k-1})-f^\star)\\ &\qquad\quad+\frac{(2\lambda-1)(k-d-1)(k-2)}{8(k+\lambda-1)}(\|x_{k-1}-x^\star\|^2-\|x_{k-2}-x^\star\|^2)\end{aligned}$$

This can be further simplified as

$$\widetilde{\mathcal{E}}(k)+A_k(f(x_{k-1})-f^\star) \leq \widetilde{\mathcal{E}}(k-1)+B_k(\|x_{k-1}-x^\star\|^2-\|x_{k-2}-x^\star\|^2) \tag{54}$$

for $k \geq \sqrt{(2\lambda-1)/(2s\mu)}$, where $A_k = (8\lambda-28)k^2+(20\lambda+1^2-126\lambda+326)k+12(\lambda+1)^3-100\lambda+1^2+288r-281>0$ since $\lambda \geq 7/2$ and $B_k=(2\lambda-1)(k-d-1)(k-2)/(8(k+\lambda-1))$. Denote by $k^\star = \lceil \max\{\sqrt{(2\lambda-1)/(2s\mu)}, 3\lambda/2\}\rceil \asymp 1/\sqrt{s\mu}$. Then B_k is a positive increasing sequence if $k>k^\star$. Summing (54) from k to $k^\star+1$, we obtain

$$\begin{aligned}&\mathcal{E}(k)+\sum_{i=k^\star+1}^{k}A_i(f(x_{i-1})-f^\star) \leq \mathcal{E}(k^\star)+\sum_{i=k^\star+1}^{k}B_i(\|x_{i-1}-x^\star\|^2-\|x_{i-2}-x^\star\|^2)\\ &=\mathcal{E}(k^\star)+B_k\|x_{k-1}-x^\star\|^2-B_{k^\star+1}\|x_{k^\star-1}-x^\star\|^2+\sum_{i=k^\star+1}^{k-1}(B_j-B_{j+1})\|x_{j-1}-x^\star\|^2\\ &\leq \mathcal{E}(k^\star)+B_k\|x_{k-1}-x^\star\|^2\end{aligned}$$

Similarly, as in the proof of Theorem 11, we can bound $\mathcal{E}(k^\star)$ via another energy functional defined from Theorem 8,

$$\begin{aligned}\mathcal{E}(k^\star) \le{}& \frac{s(2k^\star+3\lambda-2)(k^\star+\lambda-1)^2}{2}(f(x_{k^\star})-f^\star)\\ &+\frac{2k^\star+3\lambda-2}{16}\|2(k^\star+\lambda)y_{k^\star}-2k^\star x_{k^\star}-2\lambda x^\star-(x_{k^\star}-x^\star)\|^2\\ &\le \frac{s(2k^\star+3\lambda-2)(k^\star+\lambda-1)^2}{2}(f(x_{k^\star})-f^\star)\\ &+\frac{2k^\star+3\lambda-2}{8}\|2(k^\star+\lambda)y_{k^\star}-2k^\star x_{k^\star}-2\lambda x^\star\|^2\\ &+\frac{2k^\star+3\lambda-2}{8}\|x_{k^\star}-x^\star\|^2 \le \frac{\lambda^2(2k^\star+3\lambda-2)}{2}\|x_0-x^\star\|^2\\ &+\frac{\lambda^2(2k^\star+3\lambda-2)}{8s\mu(k^\star+\lambda-1)^2}\|x_0-x^\star\|^2 \lesssim \frac{\|x_0-x^\star\|^2}{\sqrt{s\mu}} \end{aligned} \tag{55}$$

For the second term, it follows from Theorem 9 that

$$\begin{aligned}B_k\|x_{k-1}-x^\star\|^2 &\le \frac{(2\lambda-1)(2k-3\lambda)(k-2)}{8\mu(k+\lambda-1)}(f(x_{k-1})-x^\star)\\ &\le \frac{(2\lambda-1)(2k-3\lambda)(k-2)}{8\mu(k+\lambda-1)}\frac{\lambda^2\|x_0-x^\star\|^2}{2s(k+\lambda-2)^2}\\ &\le \frac{(2\lambda-1)\lambda^2(2k^\star-3\lambda)(k^\star-2)}{16s\mu(k^\star+\lambda-1)(k^\star+\lambda-2)^2}\|x_0-x^\star\|^2 \lesssim \frac{\|x_0-x^\star\|^2}{\sqrt{s\mu}}\end{aligned} \tag{56}$$

For $k > k^\star$, (55) together with (56) this gives

$$\begin{aligned}f(x_k)-f^\star &\le \frac{16\mathcal{E}(k)}{s(2k+3\lambda-2)(2k+2\lambda-7)(4k+4\lambda-13)}\\ &\le \frac{16(\mathcal{E}(k^\star)+B_k\|x_{k-1}-x^\star\|^2)}{s(2k+3\lambda-2)(2k+2\lambda-7)(4k+4\lambda-13)} \lesssim \frac{\|x_0-x^\star\|^2}{s^{\frac{3}{2}}\mu^{\frac{1}{2}}k^3}\end{aligned}$$

To conclusion, note that by Theorem 9 the gap $f(x_k)-f^\star$ for $k \le k^\star$ is bounded by

$$\frac{\lambda^2\|x_0-x^\star\|^2}{2s(k+\lambda-1)^2} = \frac{\lambda^2\sqrt{s\mu}k^3}{2(k+\lambda-1)^2}\frac{\|x_0-x^\star\|^2}{s^{\frac{3}{2}}\mu^{\frac{1}{2}}k^3} \lesssim \sqrt{s\mu}k^\star\frac{\|x_0-x^\star\|^2}{s^{\frac{3}{2}}\mu^{\frac{1}{2}}k^3} \lesssim \frac{\|x_0-x^\star\|^2}{s^{\frac{3}{2}}\mu^{\frac{1}{2}}k^3}$$

□

7 Conclusion

In this paper, we derived a second-order ordinary differential equation (ODE) that serves as a continuous-time limit for (second-scheme) Nesterov's accelerated gradient method. This ODE provides a new lens for understanding Nesterov's scheme, explaining phenomena such as oscillations in the trajectories and offering a generalized framework for accelerated methods. By drawing an approximate equivalence between the discrete scheme and the ODE, we introduced a family of generalized Nesterov's schemes, all of which guarantee an optimal convergence rate of $O(1/k^2)$.

Our approach highlights the benefits of using ODEs as surrogates for discrete schemes, suggesting that studying simpler ODEs could yield insights into complex discrete algorithms. The close mapping between momentum coefficients in discrete and continuous-time settings reveals a potential for discovering new accelerated schemes. Future work could focus on refining stopping criteria and adaptive step sizes through deeper exploration of the ODE's trajectory properties, such as curvature, thereby further enhancing the practical utility of these methods.

References

[Beck(2014)] A. Beck. *Introduction to Nonlinear Optimization: Theory, Algorithms, and Applications with MATLAB*. SIAM, 2014.

[Beck and Teboulle(2009)] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

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

[Bogdan et al.(2015)] M. Bogdan, E. v. d. Berg, C. Sabatti, W. Su, and E. J. Candès. SLOPE–adaptive variable selection via convex optimization. *The Annals of Applied Statistics*, 9(3):1103–1140, 2015.

[Boyd and Vandenberghe(2004)] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[Boyd et al.(2011)] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

[Dürr and Ebenbauer(2012)] H.-B. Dürr and C. Ebenbauer. On a class of smooth optimization algorithms with applications in control. *Nonlinear Model Predictive Control*, 4(1):291–298, 2012.

[Dürr et al.(2012)] H.-B. Dürr, E. Saka, and C. Ebenbauer. A smooth vector field for quadratic programming. In *51st IEEE Conference on Decision and Control*, pages 2515–2520, 2012.

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

[Helmke and Moore(1996)] U. Helmke and J. Moore. Optimization and dynamical systems. *Proceedings of the IEEE*, 84(6):907, 1996.

[Hinton(2005)] D. Hinton. Sturm's 1836 oscillation results evolution of the theory. In *Sturm-Liouville theory*, pages 1–27. Birkhäuser, Basel, 2005.

[Leader(2004)] J. J. Leader. *Numerical Analysis and Scientific Computation*. Pearson Addison Wesley, 2004.

[Lessard et al.(2014)] L. Lessard, B. Recht, and A. Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *arXiv preprint arXiv:1408.3595*, 2014.

[Monteiro et al.(2012)] R. Monteiro, C. Ortiz, and B. Svaiter. An adaptive accelerated first-order method for convex optimization. Technical report, ISyE, Gatech, 2012.

[Nesterov(1983)] Y. Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983.

[Nesterov(2004)] Y. Nesterov. *Introductory Lectures on Convex Pptimization: A Basic Course*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, Boston, MA, 2004.

[Nesterov(2005)] Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005.

[Nesterov(2013)] Y. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.

[Nocedal and Wright(2006)] J. Nocedal and S. Wright. *Numerical Optimization*. Springer Science & Business Media, 2006.

[O'Donoghue and Candès(2013)] B. O'Donoghue and E. J. Candès. Adaptive restart for accelerated gradient schemes. *Found. Comput. Math.*, 2013.

[Osher et al.(2014)] S. Osher, F. Ruan, J. Xiong, Y. Yao, and W. Yin. Sparse recovery via differential inclusions. *arXiv preprint arXiv:1406.7728*, 2014.

[Polyak(1987)] B. T. Polyak. *Introduction to optimization*. Optimization Software New York, 1987.

[Qin and Goldfarb(2012)] Z. Qin and D. Goldfarb. Structured sparsity via alternating direction methods. *Journal of Machine Learning Research*, 13(1):1435–1468, 2012.

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

[Ruszczyński(2006)] A. P. Ruszczyński. *Nonlinear Optimization*. Princeton University Press, 2006.

[Schropp and Singer(2000)] J. Schropp and I. Singer. A dynamical systems approach to constrained minimization. *Numerical functional analysis and optimization*, 21(3-4):537–551, 2000.

[Shor(2012)] N. Z. Shor. *Minimization Methods for Non-Differentiable Functions*. Springer Science & Business Media, 2012.

[Sutskever et al.(2013)] I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1139–1147, 2013.

[Tseng(2008)] P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. `http://pages.cs.wisc.edu/~brecht/cs726docs/Tseng.APG.pdf`, 2008.

[Tseng(2010)] P. Tseng. Approximation accuracy, gradient methods, and error bound for structured convex optimization. *Mathematical Programming*, 125(2):263–295, 2010.

[Watson(1995)] G. N. Watson. *A Treatise on the Theory of Bessel Functions*. Cambridge Mathematical Library. Cambridge University Press, 1995. Reprint of the second (1944) edition.