A Differential Equation for Modeling Nesterov's Accelerated Gradient Method: Theory and Insights

Weijie Su[1]   **Stephen Boyd**[2]   **Emmanuel J. Candès**[1,3]

[1]Department of Statistics, Stanford University, Stanford, CA 94305
[2]Department of Electrical Engineering, Stanford University, Stanford, CA 94305
[3]Department of Mathematics, Stanford University, Stanford, CA 94305
{wjsu, boyd, candes}@stanford.edu

Abstract

We derive a second-order ordinary differential equation (ODE), which is the limit of Nesterov's accelerated gradient method. This ODE exhibits approximate equivalence to Nesterov's scheme and thus can serve as a tool for analysis. We show that the continuous time ODE allows for a better understanding of Nesterov's scheme. As a byproduct, we obtain a family of schemes with similar convergence rates. The ODE interpretation also suggests restarting Nesterov's scheme leading to an algorithm, which can be rigorously proven to converge at a linear rate whenever the objective is strongly convex.

1   Introduction

As data sets and problems are ever increasing in size, accelerating first-order methods is both of practical and theoretical interest. Perhaps the earliest first-order method for minimizing a convex function f is the gradient method, which dates back to Euler and Lagrange. Thirty years ago, in a seminar paper [11] Nesterov proposed an accelerated gradient method, which may take the following form: starting with x_0 and $y_0 = x_0$, inductively define

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

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

$$
f(x_k) - f^\star \leq O\Big(\frac{L\|x_0 - x^\star\|^2}{k^2}\Big).
$$

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 [12]. This is in contrast to vanilla gradient descent methods, which can only achieve a rate of $O(1/k)$ [18]. This improvement relies on the introduction of the momentum term $x_k - x_{k-1}$ as well as the particularly tuned coefficient $(k-1)/(k+2) \approx 1 - 3/k$. Since the introduction of Nesterov's scheme, there has been much work on the development of first-order accelerated methods, see [12, 13, 14, 1, 2] for example, and [20] for a unified analysis of these ideas.

In a different direction, there is a long history relating ordinary differential equations (ODE) to optimization, see [6, 4, 8, 19] for references. 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 [5, 7, 16].

In this work, we derive a second-order ordinary differential equation, which is the exact limit of Nesterov's scheme by taking small step sizes in (1). This ODE reads

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

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, $\dot{X}$ denotes the time derivative or velocity $\mathrm{d}X/\mathrm{d}t$ and similarly $\ddot{X} = \mathrm{d}^2 X/\mathrm{d}t^2$ denotes the acceleration. The time parameter in this ODE is related to the step size in (1) via $t \approx k\sqrt{s}$. Case studies are provided to demonstrate that the homogeneous and conceptually simpler ODE can serve as a tool for analyzing and generalizing Nesterov's scheme. To the best of our knowledge, this work is the first to model Nesterov's scheme or its variants by ODEs.

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 obeys

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

for any $x, y \in \mathbb{R}^n$, where $\|\cdot\|$ is the standard Euclidean norm and $L > 0$ is the Lipschitz constant throughout this paper. 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. Last, we set $\mathcal{S}_{\mu,L} = \mathcal{F}_L \cap \mathcal{S}_\mu$.

2 Derivation of the ODE

Assume $f \in \mathcal{F}_L$ for $L > 0$. Combining the two equations of (1) and applying a rescaling give

$$\frac{x_{k+1} - x_k}{\sqrt{s}} = \frac{k-1}{k+2}\frac{x_k - x_{k-1}}{\sqrt{s}} - \sqrt{s}\nabla f(y_k). \tag{3}$$

Introduce the *ansatz* $x_k \approx X(k\sqrt{s})$ for some smooth curve $X(t)$ defined for $t \geq 0$. For fixed t, as the step size s goes to zero, $X(t) \approx x_{t/\sqrt{s}} = x_k$ and $X(t + \sqrt{s}) \approx x_{(t+\sqrt{s})/\sqrt{s}} = x_{k+1}$ with $k = t/\sqrt{s}$. With these approximations, we get Taylor expansions:

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

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

$$\sqrt{s}\nabla f(y_k) = \sqrt{s}\nabla f(X(t)) + o(\sqrt{s}),$$

where in the last equality we use $y_k - X(t) = o(1)$. Thus (3) can be written as

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

$$= \left(1 - \frac{3\sqrt{s}}{t}\right)\left(\dot{X}(t) - \frac{1}{2}\ddot{X}(t)\sqrt{s} + o(\sqrt{s})\right) - \sqrt{s}\nabla f(X(t)) + o(\sqrt{s}). \tag{4}$$

By comparing the coefficients of $\sqrt{s}$ in (4), we obtain

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

for $t > 0$. The first initial condition is $X(0) = x_0$. Taking $k = 1$ in (3) 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). In the formulation of [1] (see also [21]), the momentum coefficient $(k-1)/(k+2)$ is replaced by $\theta_k(\theta_{k-1}^{-1} - 1)$, where θ_k are iteratively defined as

$$\theta_{k+1} = \frac{\sqrt{\theta_k^4 + 4\theta_k^2} - \theta_k^2}{2} \tag{5}$$

starting from $\theta_0 = 1$. A bit of 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 (1).

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 (2) and then chooses a convergent subsequence by some compactness arguments such as the Arzelá-Ascoli theorem. A proof of this theorem can be found in the supplementary material for this paper.

Theorem 2.1. *For any $f \in \mathcal{F}_\infty \triangleq \cup_{L>0}\mathcal{F}_L$ and any $x_0 \in \mathbb{R}^n$, the ODE (2) with initial conditions $X(0) = x_0, \dot{X}(0) = 0$ has a unique global solution $X \in C^2((0, \infty); \mathbb{R}^n) \cap C^1([0, \infty); \mathbb{R}^n)$.*

3 Equivalence between the ODE and Nesterov's scheme

We study the stable step size allowed for numerically solving the ODE in the presence of accumulated errors. The finite difference approximation of (2) by the forward Euler method is

$$\frac{X(t + \Delta t) - 2X(t) + X(t - \Delta t)}{\Delta t^2} + \frac{3}{t}\frac{X(t) - X(t - \Delta t)}{\Delta t} + \nabla f(X(t)) = 0, \qquad (6)$$

which is equivalent to

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

Assuming that f is sufficiently smooth, for small perturbations δx, $\nabla f(x + \delta x) \approx \nabla f(x) + \nabla^2 f(x)\delta x$, 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

$$\det\left(\lambda^2 - \left(2 - \Delta t^2 \nabla^2 f - \frac{3\Delta t}{t}\right)\lambda + 1 - \frac{3\Delta t}{t}\right) = 0. \qquad (7)$$

The numerical stability of (6) with respect to accumulated errors is equivalent to this: all the roots of (7) lie in the unit circle [9]. When $\nabla^2 f \preceq LI_n$ (i.e., $LI_n - \nabla^2 f$ is positive semidefinite), if $\Delta t/t$ small and $\Delta t < 2/\sqrt{L}$, we see that all the roots of (7) lie in the unit circle. On the other hand, if $\Delta t > 2/\sqrt{L}$, (7) 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 (1) is approximately equivalent to a step size of $\Delta t = 1/\sqrt{L}$ in the forward Euler method, which is stable for numerically integrating (6).

As a comparison, note that the corresponding ODE for gradient descent with updates $x_{k+1} = x_k - s\nabla f(x_k)$, is

$$\dot{X}(t) + \nabla f(X(t)) = 0,$$

whose finite difference scheme 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 (6) when ∇f is very variable, i.e., L is large.

Next, we exhibit approximate equivalence between the ODE and Nesterov's scheme in terms of convergence rates. We first recall the original result from [11].

Theorem 3.1 (Nesterov). *For any $f \in \mathcal{F}_L$, the sequence $\{x_k\}$ in (1) with step size $s \leq 1/L$ obeys*

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

Our first result indicates that the trajectory of ODE (2) closely resembles the sequence $\{x_k\}$ in terms of the convergence rate to a minimizer $x^\star$.

Theorem 3.2. *For any $f \in \mathcal{F}_\infty$, let $X(t)$ be the unique global solution to (2) with initial conditions $X(0) = x_0, \dot{X}(0) = 0$. For any $t > 0$,*

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

Proof of Theorem 3.2. Consider the energy functional defined as

$$\mathcal{E}(t) \triangleq t^2(f(X(t)) - f^\star) + 2\|X + \frac{t}{2}\dot{X} - x^\star\|^2,$$

whose time derivative is

$$\dot{\mathcal{E}} = 2t(f(X) - f^\star) + t^2\langle \nabla f, \dot{X}\rangle + 4\langle X + \frac{t}{2}\dot{X} - x^\star, \frac{3}{2}\dot{X} + \frac{t}{2}\ddot{X}\rangle. \tag{8}$$

Substituting $3\dot{X}/2 + t\ddot{X}/2$ with $-t\nabla f(X)/2$, (8) gives

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

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 obeys $f(X(t)) - f^\star \leq \mathcal{E}(t)/t^2 \leq \mathcal{E}(0)/t^2 = 2\|x_0 - x^\star\|^2/t^2$.

$\square$

4 A family of generalized Nesterov's schemes

In this section we show how to exploit the power of the ODE for deriving variants of Nesterov's scheme. One would be interested in studying the ODE (2) with the number 3 appearing in the coefficient of $\dot{X}/t$ replaced by a general constant r as in

$$\ddot{X} + \frac{r}{t}\dot{X} + \nabla f(X) = 0, \ X(0) = x_0, \dot{X}(0) = 0. \tag{9}$$

Using arguments similar to those in the proof of Theorem 2.1, this new ODE is guaranteed to assume a unique global solution for any $f \in \mathcal{F}_\infty$.

4.1 Continuous optimization

To begin with, we consider a modified energy functional defined as

$$\mathcal{E}(t) = \frac{2t^2}{r-1}(f(X(t)) - f^\star) + (r-1)\left\|X(t) + \frac{t}{r-1}\dot{X}(t) - x^\star\right\|^2.$$

Since $r\dot{X} + t\ddot{X} = -t\nabla f(X)$, the time derivative $\dot{\mathcal{E}}$ is equal to

$$\frac{4t}{r-1}(f(X) - f^\star) + \frac{2t^2}{r-1}\langle \nabla f, \dot{X}\rangle + 2\langle X + \frac{t}{r-1}\dot{X} - x^\star, r\dot{X} + t\ddot{X}\rangle$$
$$= \frac{4t}{r-1}(f(X) - f^\star) - 2t\langle X - x^\star, \nabla f(X)\rangle. \tag{10}$$

A consequence of (10) is this:

Theorem 4.1. *Suppose $r > 3$ and let X be the unique solution to (9) for some $f \in \mathcal{F}_\infty$. Then X obeys*

$$f(X(t)) - f^\star \leq \frac{(r-1)^2\|x_0 - x^\star\|^2}{2t^2}$$

and

$$\int_0^\infty t(f(X(t)) - f^\star)\mathrm{d}t \leq \frac{(r-1)^2\|x_0 - x^\star\|^2}{2(r-3)}.$$

Proof of Theorem 4.1. By (10), the derivative $\mathrm{d}\mathcal{E}/\mathrm{d}t$ equals

$$2t(f(X) - f^\star) - 2t\langle X - x^\star, \nabla f(X)\rangle - \frac{2(r-3)t}{r-1}(f(X) - f^\star) \leq -\frac{2(r-3)t}{r-1}(f(X) - f^\star), \tag{11}$$

where the inequality follows from the convexity of f. Since $f(X) \geq f^\star$, (11) implies that $\mathcal{E}$ is non-increasing. Hence

$$\frac{2t^2}{r-1}(f(X(t)) - f^\star) \leq \mathcal{E}(t) \leq \mathcal{E}(0) = (r-1)\|x_0 - x^\star\|^2,$$

yielding the first inequality of the theorem as desired. To complete the proof, by (10) it follows that

$$\int_0^\infty \frac{2(r-3)t}{r-1}(f(X) - f^\star)\mathrm{d}t \leq -\int_0^\infty \frac{\mathrm{d}\mathcal{E}}{\mathrm{d}t}\mathrm{d}t = \mathcal{E}(0) - \mathcal{E}(\infty) \leq (r-1)\|x_0 - x^\star\|^2,$$

as desired for establishing the second inequality. $\qquad\square$

We now demonstrate faster convergence rates under the assumption of strong convexity. Given a strongly convex function f, consider a new energy functional defined as

$$\tilde{\mathcal{E}}(t) = t^3(f(X(t)) - f^\star) + \frac{(2r-3)^2 t}{8}\left\|X(t) + \frac{2t}{2r-3}\dot{X}(t) - x^\star\right\|^2.$$

As in Theorem 4.1, a more refined study of the derivative of $\tilde{\mathcal{E}}(t)$ gives

Theorem 4.2. *For any $f \in \mathcal{S}_{\mu,L}(\mathbb{R}^n)$, the unique solution X to (9) with $r \geq 9/2$ obeys*

$$f(X(t)) - f^\star \leq \frac{Cr^{\frac{5}{2}}\|x_0 - x^\star\|^2}{t^3\sqrt{\mu}}$$

for any $t > 0$ and a universal constant $C > 1/2$.

The restriction $r \geq 9/2$ is an artifact required in the proof. We believe that this theorem should be valid as long as $r \geq 3$. For example, the solution to (9) with $f(x) = \|x\|^2/2$ is

$$X(t) = \frac{2^{\frac{r-1}{2}}\Gamma((r+1)/2)J_{(r-1)/2}(t)}{t^{\frac{r-1}{2}}}x_0, \tag{12}$$

where $J_{(r-1)/2}(\cdot)$ is the first kind Bessel function of order $(r-1)/2$. For large t, this Bessel function obeys $J_{(r-1)/2}(t) = \sqrt{2/(\pi t)}(\cos(t - (r-1)\pi/4 - \pi/4) + O(1/t))$. Hence,

$$f(X(t)) - f^\star \lesssim \|x_0 - x^\star\|^2/t^r,$$

in which the inequality fails if $1/t^r$ is replaced by any higher order rate. For general strongly convex functions, such refinement, if possible, might require a construction of a more sophisticated energy functional and careful analysis. We leave this problem for future research.

4.2 Composite optimization

Inspired by Theorem 4.2, it is tempting to obtain such analogies for the discrete Nesterov's scheme as well. Following the formulation of [1], we consider the composite minimization:

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad f(x) = g(x) + h(x),$$

where $g \in \mathcal{F}_L$ for some $L > 0$ and h is convex on $\mathbb{R}^n$ with possible extended value ∞. Define the proximal subgradient

$$G_s(x) \triangleq \frac{x - \operatorname{argmin}_z\left[\|z - (x - s\nabla g(x))\|^2/(2s) + h(z)\right]}{s}.$$

Parametrizing by a constant r, we propose a generalized Nesterov's scheme,

$$\begin{aligned} x_k &= y_{k-1} - sG_s(y_{k-1}) \\ y_k &= x_k + \frac{k-1}{k+r-1}(x_k - x_{k-1}), \end{aligned} \tag{13}$$

starting from $y_0 = x_0$. The discrete analog of Theorem 4.1 is below, whose proof is also deferred to the supplementary materials as well.

Theorem 4.3. *The sequence $\{x_k\}$ given by (13) with $r > 3$ and $0 < s \leq 1/L$ obeys*

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

and

$$\sum_{k=1}^\infty (k+r-1)(f(x_k) - f^\star) \leq \frac{(r-1)^2\|x_0 - x^\star\|^2}{2s(r-3)}.$$

The idea behind the proof is the same as that employed for Theorem 4.1; here, however, the energy functional is defined as

$$\mathcal{E}(k) = 2s(k + r - 2)^2(f(x_k) - f^\star)/(r - 1) + \|(k + r - 1)y_k - kx_k - (r - 1)x^\star\|^2/(r - 1).$$

The first inequality in Theorem 4.3 suggests that the generalized Nesterov's scheme still achieves $O(1/k^2)$ convergence rate. However, if the error bound satisfies

$$f(x_{k'}) - f^\star \geq \frac{c}{k'^2}$$

for some $c > 0$ and a dense subsequence $\{k'\}$, i.e., $|\{k'\} \cap \{1, \ldots, m\}| \geq \alpha m$ for any positive integer m and some $\alpha > 0$, then the second inequality of the theorem is violated. Hence, the second inequality is not trivial because it implies the error bound is in some sense $O(1/k^2)$ suboptimal.

In closing, we would like to point out this new scheme is equivalent to setting $\theta_k = (r-1)/(k+r-1)$ and letting $\theta_k(\theta_{k-1}^{-1} - 1)$ replace the momentum coefficient $(k - 1)/(k + r - 1)$. Then, the equal sign " $=$ " in (5) has to be replaced by " $\geq$ ". In examining the proof of Theorem 1(b) in [21], we can get an alternative proof of Theorem 4.3 by allowing (5), which appears in Eq. (36) in [21], to be an inequality.

5   Accelerating to linear convergence by restarting

Although an $O(1/k^3)$ convergence rate is guaranteed for generalized Nesterov's schemes (13), the example (12) provides evidence that $O(1/\texttt{poly}(k))$ is the best rate achievable under strong convexity. In contrast, the vanilla gradient method achieves linear convergence $O((1 - \mu/L)^k)$ and [12] proposed a first-order method with a convergence rate of $O((1 - \sqrt{\mu/L})^k)$, which, however, requires knowledge of the condition number μ/L. While it is relatively easy to bound the Lipschitz constant L by the use of backtracking [3, 20], estimating the strong convexity parameter μ, if not impossible, is very challenging. Among many approaches to gain acceleration via adaptively estimating μ/L, [15] proposes a restarting procedure for Nesterov's scheme in which (1) is restarted with $x_0 = y_0 := x_k$ whenever $\nabla f(y_k)^T(x_{k+1} - x_k) > 0$. In the language of ODEs, this gradient based restarting essentially keeps $\langle \nabla f, \dot{X} \rangle$ negative along the trajectory. Although it has been empirically observed that this method significantly boosts convergence, there is no general theory characterizing the convergence rate.

In this section, we propose a new restarting scheme we call the *speed restarting scheme*. The underlying motivation is to maintain a relatively high velocity $\dot{X}$ along the trajectory. Throughout this section we assume $f \in \mathcal{S}_{\mu,L}$ for some $0 < \mu \leq L$.

Definition 5.1. *For ODE* (2) *with* $X(0) = x_0, \dot{X}(0) = 0$, *let*

$$T = T(f, x_0) = \sup\{t > 0 : \forall u \in (0, t), \frac{\mathrm{d}\|\dot{X}(u)\|^2}{\mathrm{d}u} > 0\}$$

be the speed restarting time.

In words, T is the first time the velocity $\|\dot{X}\|$ decreases. The definition itself does not imply that $0 < T < \infty$, which is proven in the supplementary materials. Indeed, $f(X(t))$ is a decreasing function before time T; for $t \leq T$,

$$\frac{\mathrm{d}f(X(t))}{\mathrm{d}t} = \langle \nabla f(X), \dot{X} \rangle = -\frac{3}{t}\|\dot{X}\|^2 - \frac{1}{2}\frac{\mathrm{d}\|\dot{X}\|^2}{\mathrm{d}t} \leq 0.$$

The speed restarted ODE is thus

$$\ddot{X}(t) + \frac{3}{t_{\mathrm{sr}}}\dot{X}(t) + \nabla f(X(t)) = 0, \tag{14}$$

where t_{sr} is set to zero whenever $\langle \dot{X}, \ddot{X} \rangle = 0$ and between two consecutive restarts, t_{sr} grows just as t. That is, $t_{\mathrm{sr}} = t - \tau$, where τ is the latest restart time. In particular, $t_{\mathrm{sr}} = 0$ at $t = 0$. The theorem below guarantees linear convergence of the solution to (14). This is a new result in the literature [15, 10].

Theorem 5.2. *There exists positive constants c_1 and c_2, which only depend on the condition number L/μ, such that for any $f \in \mathcal{S}_{\mu,L}$, we have*

$$f(X^{\mathrm{sr}}(t)) - f(x^\star) \leq \frac{c_1 L\|x_0 - x^\star\|^2}{2}\mathrm{e}^{-c_2 t\sqrt{L}}.$$

5.1 Numerical examples

Below we present a discrete analog to the restarted scheme. There, $k_{\min}$ is introduced to avoid having consecutive restarts that are too close. To compare the performance of the restarted scheme with the original (1), we conduct four simulation studies, including both smooth and non-smooth objective functions. Note that the computational costs of the restarted and non-restarted schemes are the same.

Algorithm 1 Speed Restarting Nesterov's Scheme

---

input: $x_0 \in \mathbb{R}^n, y_0 = x_0, x_{-1} = x_0, 0 < s \le 1/L, k_{\max} \in \mathbb{N}^+$ and $k_{\min} \in \mathbb{N}^+$
$j \leftarrow 1$
for $k = 1$ to $k_{\max}$ **do**
   $x_k \leftarrow \mathrm{argmin}_x(\frac{1}{2s}\|x - y_{k-1} + s\nabla g(y_{k-1})\|^2 + h(x))$
   $y_k \leftarrow x_k + \frac{j-1}{j+2}(x_k - x_{k-1})$
   if $\|x_k - x_{k-1}\| < \|x_{k-1} - x_{k-2}\|$ **and** $j \ge k_{\min}$ **then**
      $j \leftarrow 1$
   else
      $j \leftarrow j + 1$
   end if
end for

Quadratic. $f(x) = \frac{1}{2}x^T A x + b^T x$ is a strongly convex function, in which A is a 500×500 random positive definite matrix and b a random vector. The eigenvalues of A are between 0.001 and 1. The vector b is generated as i. i. d. Gaussian random variables with mean 0 and variance 25.

Log-sum-exp.

$$f(x) = \rho \log \Big[\sum_{i=1}^m \exp((a_i^T x - b_i)/\rho) \Big],$$

where $n = 50, m = 200, \rho = 20$. The matrix $A = \{a_{ij}\}$ is a random matrix with i. i. d. standard Gaussian entries, and $b = \{b_i\}$ has i. i. d. Gaussian entries with mean 0 and variance 2. This function is not strongly convex.

Matrix completion. $f(X) = \frac{1}{2}\|X_{\mathrm{obs}} - M_{\mathrm{obs}}\|_F^2 + \lambda\|X\|_*$, in which the ground truth M is a rank-5 random matrix of size 300×300. The regularization parameter is set to $\lambda = 0.05$. The 5 singular values of M are $1, \ldots, 5$. The observed set is independently sampled among the 300×300 entries so that 10% of the entries are actually observed.

Lasso in ℓ_1–constrained form with large sparse design. $f = \frac{1}{2}\|Ax - b\|^2$ s.t. $\|x\|_1 \le \delta$, where A is a 5000×50000 random sparse matrix with nonzero probability 0.5% for each entry and b is generated as $b = Ax^0 + z$. The nonzero entries of A independently follow the Gaussian distribution with mean 0 and variance $1/25$. The signal x^0 is a vector with 250 nonzeros and z is i. i. d. standard Gaussian noise. The parameter δ is set to $\|x^0\|_1$.

In these examples, $k_{\min}$ is set to be 10 and the step sizes are fixed to be $1/L$. If the objective is in composite form, the Lipschitz bound applies to the smooth part. Figures 1(a), 1(b), 1(c) and 1(d) present the performance of the speed restarting scheme, the gradient restarting scheme proposed in [15], the original Nesterov's scheme and the proximal gradient method. The objective functions include strongly convex, non-strongly convex and non-smooth functions, violating the assumptions in Theorem 5.2. Among all the examples, it is interesting to note that both speed restarting scheme empirically exhibit linear convergence by significantly reducing bumps in the objective values. This leaves us an open problem of whether there exists provable linear convergence rate for the gradient restarting scheme as in Theorem 5.2. It is also worth pointing that compared with gradient restarting, the speed restarting scheme empirically exhibits more stable linear convergence rate.

6 Discussion

This paper introduces a second-order ODE and accompanying tools for characterizing Nesterov's accelerated gradient method. This ODE is applied to study variants of Nesterov's scheme. Our
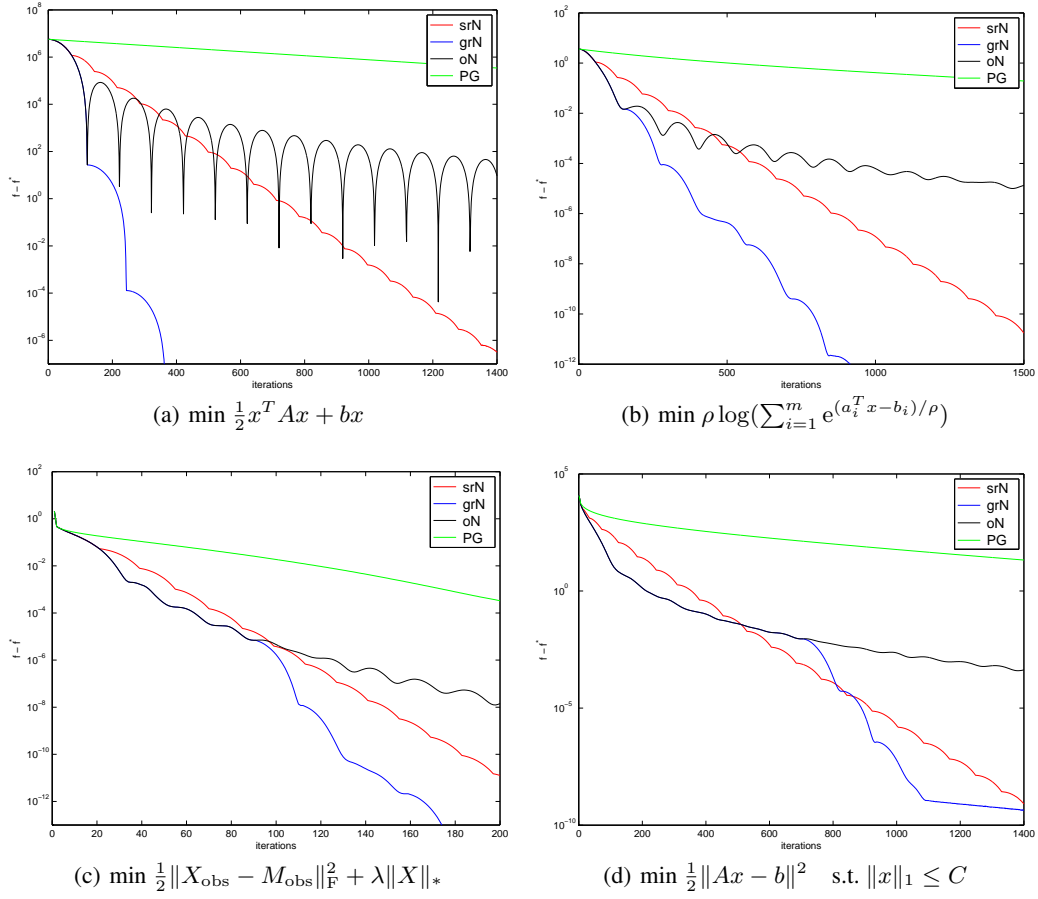
Figure 1: Numerical performance of speed restarting (srN), gradient restarting (grN) proposed in [15], the original Nesterov's scheme (oN) and the proximal gradient (PG)

approach suggests (1) a large family of generalized Nesterov's schemes that are all guaranteed to converge at the rate $1/k^2$, and (2) a restarted scheme provably achieving a linear convergence rate whenever f is strongly convex.

In this paper, we often utilize ideas from continuous-time ODEs, and then apply these ideas to discrete schemes. The translation, however, involves parameter tuning and tedious calculations. This is the reason why a general theory mapping properties of ODEs into corresponding properties for discrete updates would be a welcome advance. Indeed, this would allow researchers to only study the simpler and more user-friendly ODEs.

7 Acknowledgements

References

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

[2] 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.

[3] S. Becker, E. J. Candès, and M. Grant. Templates for convex cone problems with applications to sparse signal recovery. *Mathematical Programming Computation*, 3(3):165–218, 2011.

[4] A. Bloch (Editor). *Hamiltonian and gradient flows, algorithms, and control*, volume 3. American Mathematical Soc., 1994.

[5] F. H. Branin. Widely convergent method for finding multiple solutions of simultaneous nonlinear equations. *IBM Journal of Research and Development*, 16(5):504–522, 1972.

[6] A. A. Brown and M. C. Bartholomew-Biggs. Some effective methods for unconstrained optimization based on the solution of systems of ordinary differential equations. *Journal of Optimization Theory and Applications*, 62(2):211–224, 1989.

[7] R. Hauser and J. Nedic. The continuous Newton–Raphson method can look ahead. *SIAM Journal on Optimization*, 15(3):915–925, 2005.

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

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

[10] R. Monteiro, C. Ortiz, and B. Svaiter. An adaptive accelerated first-order method for convex optimization, 2012.

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

[12] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, Boston, MA, 2004.

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

[14] Y. Nesterov. Gradient methods for minimizing composite objective function. CORE Discussion Papers, 2007.

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

[16] Y.-G. Ou. A nonmonotone ODE-based method for unconstrained optimization. *International Journal of Computer Mathematics*, (ahead-of-print):1–21, 2014.

[17] N. Parikh and S. Boyd. Proximal algorithms. In *Foundations and Trends in Optimization*, volume 1, pages 123–231. 2013.

[18] R. T. Rockafellar. *Convex analysis*. Princeton Landmarks in Mathematics. Princeton University Press, Princeton, NJ, 1997. Reprint of the 1970 original, Princeton Paperbacks.

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

[20] P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. submitted to SIAM J. 2008.

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

Supplementary Materials

A   Proof of Theorem 2.1

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

A.1   Existence

In this section we aim to prove

Lemma A.1. *For any $f \in \mathcal{F}_\infty(\mathbb{R}^n)$ and any $x_0 \in \mathbb{R}^n$, ODE (2) with initial conditions $X(0) = x_0, \dot{X}(0) = 0$ has at least one solution X in $C^2(0, \infty) \cap C^1[0, \infty)$. Recall $C^2(0, \infty)$ is the set of functions, taking values in $\mathbb{R}^n$, defined on $[0, \infty)$ and twice continuously differentiable on $(0, \infty)$. Similarly $C^1[0, \infty)$ is the set of continuously differentiable functions from $[0, \infty)$ to $\mathbb{R}^n$.*

To begin with, for any $\delta > 0$ consider the smoothed ODE

$$\ddot{X} + \frac{3}{\max(\delta, t)}\dot{X} + \nabla f(X) = 0 \tag{15}$$

with $X(0) = x_0, \dot{X}(0) = 0$. Denoting by $Z = \dot{X}$, then (15) is equivalent to

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

As functions of (X, Z), both (Z and $-3Z/\max(\delta, t) - \nabla f(X)$) are Lipschitz continuous with constant at most $\max(1, L) + 3/\delta$. Hence by standard ODE theory (15) has a unique global solution in $C^2[0, \infty)$, which is 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 $u \in (0, t]$. We remark 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 $u = o(1)$. We given an upper bound for $M_\delta(t)$ in the following lemma.

Lemma A.2. *For $\delta < \sqrt{6/L}$ one has*

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

The proof of Lemma A.2 relies on a simple lemma.

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

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

Proof of Lemma A.3. By Lipschitz continuity,

$$\|\nabla f(X_\delta(u)) - \nabla f(x_0)\| \le L\|X_\delta(u) - x_0\| = \left\|\int_0^u \dot{X}_\delta(v)dv\right\| \le \int_0^u v\frac{\|\dot{X}_\delta(v)\|}{v}dv \le \frac{1}{2}LM_\delta(u)u^2.$$

$\square$

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

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

which yields

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

Hence, by Lemma A.3

$$\frac{\|\dot{X}_\delta(t)\|}{t} \leq \frac{1}{t}\mathrm{e}^{-3t/\delta}\|\nabla f(x_0)\| \int_0^t \mathrm{e}^{3u/\delta}du + \frac{1}{t}\mathrm{e}^{-3t/\delta} \int_0^t \frac{1}{2}LM_\delta(u)u^2\mathrm{e}^{3u/\delta}du$$

$$\leq \|\nabla f(x_0)\| + \frac{LM_\delta(\delta)\delta^2}{6}.$$

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

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

Lemma A.4. *For $\delta < \sqrt{6/L}$ and $\delta < t < \sqrt{12/L}$, one has*

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

Proof of Lemma A.4. When $t > \delta$ the smoothed ODE reads

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

which is equivalent to

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

By integration,

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

Therefore by Lemmas A.3 and A.2 we have

$$\frac{\|\dot{X}_\delta(t)\|}{t} \leq \frac{t^4 - \delta^4}{4t^4}\|\nabla f(x_0)\| + \frac{1}{t^4}\int_\delta^t \frac{1}{2}LM_\delta(u)u^5du + \frac{\delta^4}{t^4}\frac{\|\dot{X}_\delta(\delta)\|}{\delta}$$

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

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

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

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

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

The desired result follows from rearranging the inequality. □

Lemma A.5. *Consider the set of continuous functions $\mathcal{F} = \{X_\delta : [0, \sqrt{6/L}] \to \mathbb{R}^n | \delta = \sqrt{3/L}/2^m, m = 0, 1, \ldots\}$ is uniformly bounded and equicontinuous.*

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

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

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

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

□

11

Now it is ready to give

Proof of Lemma A.1. By the Arzelá–Ascoli theorem and Lemma A.5, $\mathcal{F}$ contains a sequence converge uniformly on $[0, \sqrt{6/L}]$. Denote by $\{X_{\delta_{m_i}}\}_{i \in \mathbb{N}}$ the convergent sequence 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 (2) 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 $X_{t_0}^D$. Without loss of generality, assume the subsequence is the original sequence. Denote by $\tilde{X}$ the local solution to (2) with $X(t_0) = \breve{X}(t_0)$ and $\dot{X}(t_0) = X_{t_0}^D$. On the other hand, recall $X_{\delta_{m_i}}$ is the solution to (2) 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 $\dot{X}_{\delta_{m_i}}(t_0)$ go to $\breve{X}(t_0)$ and $X_{t_0}^D$, respectively, there exits $\epsilon_0 > 0$ such that

$$\sup_{t \in (t_0 - \epsilon_0, t_0 + \epsilon_0)} \|X_{\delta_{m_i}}(t) - \tilde{X}(t)\| \to 0$$

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

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

Therefore $\breve{X}$ and $\tilde{X}$ have to be identical on $(t_0 - \epsilon_0, t_0 + \epsilon_0)$. So $\breve{X}$ satisfies (2) at t_0. Since t_0 is arbitrary, we conclude that $\breve{X}$ is a solution to (2) on $(0, \sqrt{6/L})$. By extension, $\breve{X}$ can be a global solution to (2) 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 one, pick a small $t > 0$ and note that

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

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

A.2 Uniqueness

In this section we prove the uniqueness of the solution to (2).

Lemma A.6. *For any initial point $x_0 \in \mathbb{R}^n$, ODE (2) with initial conditions $X(0) = x_0, \dot{X}(0) = 0$ has at most one local solution near $t = 0$.*

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

Lemma A.7. *For any $t \in (0, \alpha)$ one has*

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

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

$$\|\nabla f(X(t)) - \nabla f(Y(t))\| \le L\|X(t) - Y(t)\| = L\left\| \int_0^t \dot{X}(u) - \dot{Y}(u)du + X(0) - Y(0)\right\|$$

$$\le L \int_0^t \|\dot{X}(u) - \dot{Y}(u)\|du \le Lt\tilde{M}(t).$$

$\square$

12

Proof of Lemma A.6. Similar to the proof of Lemma A.4, one has

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

Applying Lemma A.7 gives

$$t^3\|\dot{X}(t) - \dot{Y}(t)\| \leq \int_0^t Lu^4\tilde{M}(u)du \leq \frac{1}{5}Lt^5\tilde{M}(t),$$

which reads $\|\dot{X}(t) - \dot{Y}(t)\| \leq Lt^2\tilde{M}(t)/5$. Thus for any $t' \leq t$ it is true that $\|\dot{X}(t') - \dot{Y}(t')\| \leq Lt^2\tilde{M}(t)/5$. Taking the supremum of $\|\dot{X}(t') - \dot{Y}(t')\|$ over $t' \in (0, t)$ gives $\tilde{M}(t) \leq Lt^2\tilde{M}(t)/5$. Therefore $\tilde{M}(t) = 0$ 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. $\qquad\square$

Proof of Theorem 2.1. Lemma A.1 together with Lemma A.6 completes the proof of Theorem 2.1.
$\qquad\square$

B Proof of Theorem 4.2

Proof of Theorem 4.2. The derivative of $\tilde{\mathcal{E}}$ reads

$$\frac{d\tilde{\mathcal{E}}(t)}{dt} = 3t^2(f(X) - f^\star) + t^3\langle \dot{X}, \nabla f(X)\rangle + \frac{(2r-3)^2}{8}\left\langle X + \frac{2t}{2r-3}\dot{X} - x^\star, \frac{4t^2}{2r-3}\ddot{X} + \frac{4rt}{2r-3}\dot{X} + X - x^\star\right\rangle$$

$$= 3t^2(f(X) - f^\star) - \frac{(2r-3)t^2}{2}\langle X - x^\star, \nabla f(X)\rangle + \frac{(2r-3)^2}{8}\|X - x^\star\|^2 + \frac{(2r-3)t}{4}\langle \dot{X}, X - x^\star\rangle. \tag{16}$$

By convexity and strong convexity of f, the second term of the RHS of (16) meets

$$\frac{(2r-3)t^2}{2}\langle X - x^\star, \nabla f(X)\rangle \geq \frac{(2r-3)t^2}{2}(f(X) - f^\star) + \frac{\mu(2r-3)t^2}{4}\|X - x^\star\|^2.$$

Since $r \geq 4$, substituting the above into (16) yields

$$\frac{d\tilde{\mathcal{E}}(t)}{dt} \leq \left[3t^2 - \frac{(2r-3)t^2}{2}\right](f(X) - f^\star) - \frac{2(2r-3)\mu t^2 - (2r-3)^2}{8}\|X - x^\star\|^2 + \frac{(2r-3)t}{8}\frac{d\|X - x^\star\|^2}{dt}$$

$$\leq -\frac{2(2r-3)\mu t^2 - (2r-3)^2}{8}\|X - x^\star\|^2 + \frac{(2r-3)t}{8}\frac{d\|X - x^\star\|^2}{dt}.$$

Hence if $t \geq t' \triangleq \sqrt{(2r-3)/(2\mu)}$, we obtain

$$\frac{d\tilde{\mathcal{E}}(t)}{dt} \leq \frac{(2r-3)t}{8}\frac{d\|X - x^\star\|^2}{dt}. \tag{17}$$

For $t > t'$, integrating (17) over (t', t) gives

$$\tilde{\mathcal{E}}(t) \leq \tilde{\mathcal{E}}(t') + \frac{2r-3}{8}t\|X(t) - x^\star\|^2 - \frac{2r-3}{8}t'\|X(t') - x^\star\|^2 - \frac{2r-3}{8}\int_{t'}^t \|X(u) - x^\star\|^2du$$

$$\leq \tilde{\mathcal{E}}(t') + \frac{2r-3}{8}t\|X(t) - x^\star\|^2 \leq \tilde{\mathcal{E}}(t') + \frac{2r-3}{4\mu}t(f(X(t)) - f^\star)$$

$$\leq \tilde{\mathcal{E}}(t') + \frac{(2r-3)(r-1)^2\|x_0 - x^\star\|^2}{8\mu t} \leq \tilde{\mathcal{E}}(t') + \frac{(2r-3)(r-1)^2\|x_0 - x^\star\|^2}{8\mu t'}, \tag{18}$$

13

where the second last inequality follows from Theorem 4.1. We can make use of $\mathcal{E}(t')$ to bound $\tilde{\mathcal{E}}(t')$ in (18). Indeed we have

$$
\begin{aligned}
\tilde{\mathcal{E}}(t') &= t'^3(f(X(t')) - f^\star) + \frac{(2r-3)^2 t'}{8}\|X(t') + \frac{2t'}{2r-3}\dot{X}(t') - x^\star\|^2 \\
&\leq t'^3(f(X(t')) - f^\star) + \frac{(2r-3)^2 t'}{4}\left\|\frac{2r-2}{2r-3}X(t') + \frac{2t'}{2r-3}\dot{X}(t') - \frac{2r-2}{2r-3}x^\star\right\|^2 \\
&\quad + \frac{(2r-3)^2 t'}{4}\left\|\frac{1}{2r-3}X(t') - \frac{1}{2r-3}x^\star\right\|^2 \\
&\leq (r-1)t'\mathcal{E}(t') + \frac{t'}{4}\|X(t') - x^\star\|^2 \leq (r-1)^2 t'\|x_0 - x^\star\|^2 + \frac{(r-1)^2\|x_0 - x^\star\|^2}{4\mu t'},
\end{aligned}
$$

which combined with (18) yields

$$
\tilde{\mathcal{E}}(t) \leq (r-1)^2 t'\|x_0 - x^\star\|^2 + \frac{(2r-1)(r-1)^2\|x_0 - x^\star\|^2}{8\mu t'} = O(\frac{r^{\frac{5}{2}}\|x_0 - x^\star\|^2}{\sqrt{\mu}}).
$$

It completes the proof for $t \geq \sqrt{(2r-3)/(2\mu)}$ by noting $f(X(t)) - f^\star \leq \tilde{\mathcal{E}}(t)/t^3$, whereas for $t < \sqrt{(2r-3)/(2\mu)}$ by Theorem 4.1 we have

$$
f(X(t)) - f^\star \leq \frac{(r-1)^2\|x_0 - x^\star\|^2}{2t^2} \leq \frac{(r-1)^2\sqrt{\mu}\sqrt{(2r-3)/(2\mu)}}{2Cr^{\frac{5}{2}}}\frac{Cr^{\frac{5}{2}}\|x_0 - x^\star\|^2}{t^3\sqrt{\mu}} \leq \frac{Cr^{\frac{5}{2}}\|x_0 - x^\star\|^2}{t^3\sqrt{\mu}}.
$$

$\square$

C Proof of Theorem 4.3

Proof of Theorem 4.3. In parallel to the proof of Theorem 4.1, we propose an energy function defined as

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

where $z_k = (k+r-1)y_k/(r-1) - kx_k/(r-1)$. Suppose we have

$$
\mathcal{E}(k) + \frac{2s[(r-3)(k+r-2)+1]}{r-1}(f(x_{k-1}) - f^\star) \leq \mathcal{E}(k-1). \tag{19}
$$

Then it immediately yields the desired results by summing over (19). To be specific, by recursively applying (19) we see

$$
\mathcal{E}(k) + \sum_{i=1}^{k} \frac{2s[(r-3)(i+r-2)+1]}{r-1}(f(x_{i-1}) - f^\star) \leq \mathcal{E}(0) = \frac{2(r-2)^2 s}{r-1}(f(x_0) - f^\star) + (r-1)\|x_0 - x^\star\|^2,
$$

which is equivalent to

$$
\mathcal{E}(k) + \sum_{i=1}^{k-1} \frac{2s[(r-3)(i+r-1)+1]}{r-1}(f(x_i) - f^\star) \leq (r-1)\|x_0 - x^\star\|^2. \tag{20}
$$

Noting that the LHS of (20) is lower bounded by $2s(k+r-2)^2(f(x_k) - f^\star)/(r-1)$ gives the first desired inequality. With $\mathcal{E}(k) \geq 0$, the second one is obtained via taking the limit $k \to \infty$ in (20) and replacing $(r-3)(i+r-1)+1$ by $(r-3)(i+r-1)$.

To complete, we aims to establish (19) in the rest of the proof. For $s \leq 1/L$ it is well-known in proximal gradient literature, for example [17], that

$$
f(y - sG_s(y)) \leq f(x) + G_s(y)^T(y - x) - \frac{s}{2}\|G_s(y)\|^2 \tag{21}
$$

14

for any x and y. Note that $y_{k-1} - sG_s(y_{k-1})$ actually coincides with x_k. Summing of $(k-1)/(k+r-2) \times (21)$ with $x = x_{k-1}, y = y_{k-1}$ and $(r-1)/(k+r-2) \times (21)$ with $x = x^\star, y = y_{k-1}$ gives

$$
\begin{aligned}
f(x_k) &\leq \frac{k-1}{k+r-2}f(x_{k-1}) + \frac{r-1}{k+r-2}f^\star \\
&\quad + \frac{r-1}{k+r-2}G_s(y_{k-1})^T\left(\frac{k+r-2}{r-1}y_{k-1} - \frac{k-1}{r-1}x_{k-1} - x^\star\right) - \frac{s}{2}\|G_s(y_{k-1})\|^2 \\
&= \frac{k-1}{k+r-2}f(x_{k-1}) + \frac{r-1}{k+r-2}f^\star + \frac{(r-1)^2}{2s(k+r-2)^2}\left(\|z_{k-1} - x^\star\|^2 - \|z_k - x^\star\|^2\right),
\end{aligned}
$$

where we use $z_{k-1} - s(k+r-2)G_s(y_{k-1})/(r-1) = z_k$. Rearranging the above inequality with multiplying by $2s(k+r-2)^2/(r-1)$ gives the desired (19).

$\square$

D   Proof of Theorem 5.2

Remark D.1. *Indeed the linear convergence of X^{sr} remains for generalized ODE (9) with $r > 3$. Only minor modifications in proof such as replacing u^3 with u^r in the definition of $I(t)$ in Lemma D.1 are required to get analogous convergence rate for the speed restarting version of (9).*

Lemma D.1. *The speed restarting time T obeys*

$$
T(x_0, f) \geq \frac{4}{5\sqrt{L}}.
$$

Proof. Denote by $M(t)$ the supremum of $\|\dot{X}(u)\|/u$ over $u \in (0, t]$ and

$$
I(t) \triangleq \int_0^t u^3(\nabla f(X(u)) - \nabla f(x_0))du.
$$

By the proof of Lemma A.5 it is guaranteed that M defined above is finite. M is useful in that it gives a bound on the gradient of f:

$$
\|\nabla f(X(t)) - \nabla f(x_0)\| \leq L\|X(t) - x_0\| = L\left\|\int_0^t \dot{X}(u)du\right\| \leq L\int_0^t u\frac{\|\dot{X}(u)\|}{u}du \leq \frac{LM(t)t^2}{2}. \tag{22}
$$

By (22), it is easy to see that I can also be bounded via M:

$$
\|I(t)\| \leq \int_0^t u^3\|\nabla f(X(u)) - \nabla f(x_0)\|du \leq \int_0^t \frac{LM(u)u^5}{2}du \leq \frac{LM(t)t^6}{12}. \tag{23}
$$

To fully facilitate these bounds, we need to bound M as

$$
M(t) \leq \frac{\|\nabla f(x_0)\|}{4(1 - Lt^2/12)} \tag{24}
$$

for any $t < \sqrt{12/L}$.

To this end, note that indeed ODE (2) is equivalent to $d(t^3\dot{X}(t))/dt = -t^3\nabla f(X(t))$, which by integration leads to

$$
t^3\dot{X}(t) = -\frac{t^4}{4}\nabla f(x_0) - \int_0^t u^3(\nabla f(X(u)) - \nabla f(x_0))du = -\frac{t^4}{4}\nabla f(x_0) - I(t). \tag{25}
$$

Dividing (25) by t^4 and applying (23), we obtain

$$
\frac{\|\dot{X}(t)\|}{t} \leq \frac{\|\nabla f(x_0)\|}{4} + \frac{\|I(t)\|}{t^4} \leq \frac{\|\nabla f(x_0)\|}{4} + \frac{LM(t)t^2}{12}.
$$

Note that the RHS of the above is monotonically increasing in t. Hence by taking the supremum of the LHS over $(0, t]$ we obtain

$$
M(t) \leq \frac{\|\nabla f(x_0)\|}{4} + \frac{LM(t)t^2}{12},
$$

which gives the desired (24) by rearranging the inequality for $t < \sqrt{12/L}$.

Having established (24), we proceed to lower bound T via studying $\langle \dot{X}(t), \ddot{X}(t) \rangle$. Dividing (25) by t^3, one has an expression for $\dot{X}$, which reads

$$\dot{X}(t) = -\frac{t}{4}\nabla f(x_0) - \frac{1}{t^3}\int_0^t u^3(\nabla f(X(u)) - \nabla f(x_0))du. \tag{26}$$

Differentiating the above, we also obtain an expression for $\ddot{X}$:

$$\ddot{X}(t) = -\nabla f(X(t)) + \frac{3}{4}\nabla f(x_0) + \frac{3}{t^4}\int_0^t u^3(\nabla f(X(u)) - \nabla f(x_0))du. \tag{27}$$

Using the two expressions for $\dot{X}$ and $\ddot{X}$ we will show that $d\|\dot{X}\|^2/dt = 2\langle \dot{X}(t), \ddot{X}(t) \rangle > 0$ for $0 < t < 4/(5\sqrt{L})$. To this end, noting that (26) and (27) yield

$$
\begin{aligned}
\langle \dot{X}(t), \ddot{X}(t) \rangle =& \Big\langle -\frac{t}{4}\nabla f(x_0) - \frac{1}{t^3}I(t), \ -\nabla f(X(t)) + \frac{3}{4}\nabla f(x_0) + \frac{3}{t^4}I(t) \Big\rangle \\
\geq & \frac{t}{4}\langle \nabla f(x_0), \nabla f(X(t)) \rangle - \frac{3t}{16}\|\nabla f(x_0)\|^2 - \frac{1}{t^3}\|I(t)\|\Big(\|\nabla f(X(t))\| + \frac{3}{2}\|\nabla f(x_0)\| \Big) - \frac{3}{t^7}\|I(t)\|^2 \\
\geq & \frac{t}{4}\|\nabla f(x_0)\|^2 - \frac{t}{4}\|\nabla f(x_0)\|\|\nabla f(X(t)) - \nabla f(x_0)\| - \frac{3t}{16}\|\nabla f(x_0)\|^2 \\
& - \frac{LM(t)t^3}{12}\Big(\|\nabla f(X(t)) - \nabla f(x_0)\| + \frac{5}{2}\|\nabla f(x_0)\| \Big) - \frac{L^2M(t)^2t^5}{48} \\
\geq & \frac{t}{16}\|\nabla f(x_0)\|^2 - \frac{LM(t)t^3\|\nabla f(x_0)\|}{8} - \frac{LM(t)t^3}{12}\Big(\frac{LM(t)t^2}{2} + \frac{5}{2}\|\nabla f(x_0)\| \Big) - \frac{L^2M(t)^2t^5}{48} \\
= & \frac{t}{16}\|\nabla f(x_0)\|^2 - \frac{LM(t)t^3}{3}\|\nabla f(x_0)\| - \frac{L^2M(t)^2t^5}{16},
\end{aligned}
$$

where we use (23) and (22). To complete the proof, applying (24) in the above inequality yields

$$\langle \dot{X}(t), \ddot{X}(t) \rangle \geq \Big(\frac{1}{16} - \frac{Lt^2}{12(1-Lt^2/12)} - \frac{L^2t^4}{256(1-Lt^2/12)^2} \Big)\|\nabla f(x_0)\|^2 t \geq 0$$

for $t < \min\{\sqrt{12/L}, 4/(5\sqrt{L})\} = 4/(5\sqrt{L})$, where the positiveness follows from the fact that

$$\frac{1}{16} - \frac{Lt^2}{12(1-Lt^2/12)} - \frac{L^2t^4}{256(1-Lt^2/12)^2} > 0$$

for $0 < t \leq 4/(5\sqrt{L})$. $\qquad\square$

Next we give a lemma which claims that the objective function decays by a constant through each speed restarting.

Lemma D.2. *There is a universal constant $C > 0$ such that*

$$f(X(T)) - f(x^\star) \leq \Big(1 - \frac{C\mu}{L} \Big)(f(x_0) - f(x^\star)).$$

Proof. By (25), (23) and (24) in Lemma D.1, for $t < \sqrt{12/L}$ one has

$$\Big\| \dot{X}(t) + \frac{t}{4}\nabla f(x_0) \Big\| = \frac{1}{t^3}\|I(t)\| \leq \frac{LM(t)t^3}{12} \leq \frac{L\|\nabla f(x_0)\|t^3}{48(1-Lt^2/12)},$$

which gives

$$0 \leq \frac{t}{4}\|\nabla f(x_0)\| - \frac{L\|\nabla f(x_0)\|t^3}{48(1-Lt^2/12)} \leq \|\dot{X}(t)\| \leq \frac{t}{4}\|\nabla f(x_0)\| + \frac{L\|\nabla f(x_0)\|t^3}{48(1-Lt^2/12)} \tag{28}$$

for $t < \sqrt{12/L}$. By Lemma D.1 $d\|\dot{X}\|^2/dt \geq 0$ for $0 < t < 4/(5\sqrt{L})$ because $T \geq 4/(5\sqrt{L})$. Hence for $0 < t < 4/(5\sqrt{L})$ it yields that

$$\frac{df(X(t))}{dt} = -\frac{3}{t}\|\dot{X}\|^2 - \frac{1}{2}\frac{d}{dt}\|\dot{X}\|^2 \leq -\frac{3}{t}\|\dot{X}\|^2 \leq -\frac{3}{t}\Big(\frac{t}{4}\|\nabla f(x_0)\| - \frac{L\|\nabla f(x_0)\|t^3}{48(1-Lt^2/12)} \Big)^2 \leq -ct\|\nabla f(x_0)\|^2,$$

where $c > 0$ is an absolute constant and the second last inequality follows from (28). Therefore we have

$$f(X\Big(\frac{4}{5\sqrt{L}}\Big)) - f(x_0) \le \int_0^{\frac{4}{5\sqrt{L}}} -cu\|\nabla f(x_0)\|^2 du$$

$$= -\frac{c'}{L}\|\nabla f(x_0)\|^2 \le -\frac{2c'\mu}{L}(f(x_0) - f^\star),$$

where the last step follows from the μ–strong convexity of f. Above $c' > 0$ is an absolute constant. Thus we have

$$f(X\Big(\frac{4}{5\sqrt{L}}\Big)) - f^\star \le \Big(1 - \frac{2c'\mu}{L}\Big)(f(x_0) - f^\star).$$

Last, recall that $f(X(t))$ decreases on $(4/(5\sqrt{L}), T)$, which finishes the proof by noting

$$f(X(T)) - f^\star \le f(X\Big(\frac{4}{5\sqrt{L}}\Big)) - f^\star \le \Big(1 - \frac{2c'\mu}{L}\Big)(f(x_0) - f^\star).$$

$\square$

To establish the linear convergence, we also need to ensure that T can not be too large. To this end, we give the following lemma.

Lemma D.3. *The speed restarting time T satisfies*

$$T \le \frac{4}{5\sqrt{L}}\exp\frac{C'L}{\mu}.$$

Proof. For $4/(5\sqrt{L}) \le t \le T$, we have

$$\frac{df(X(t))}{dt} \le -\frac{3}{t}\|\dot{X}(t)\|^2 \le -\frac{3}{t}\|\dot{X}(4/(5\sqrt{L}))\|^2,$$

which implies

$$f(X(T)) - f(x_0) \le f(X(T)) - f(X(4/(5\sqrt{L}))) \le -\int_{\frac{4}{5\sqrt{L}}}^T \frac{3}{t}\|\dot{X}(4/(5\sqrt{L}))\|^2 dt$$

$$= -3\|\dot{X}(4/(5\sqrt{L}))\|^2 \log\frac{5T\sqrt{L}}{4}.$$

Hence we get an upper bound for T which reads

$$T \le \frac{4}{5\sqrt{L}}\exp\Big(\frac{f(x_0) - f(X(T))}{3\|\dot{X}(4/(5\sqrt{L}))\|^2}\Big) \le \frac{4}{5\sqrt{L}}\exp\Big(\frac{f(x_0) - f^\star}{3\|\dot{X}(4/(5\sqrt{L}))\|^2}\Big). \tag{29}$$

Plugging $t = 4/(5\sqrt{L})$ in (28) gives

$$\|\dot{X}(4/(5\sqrt{L}))\| \ge \frac{c}{\sqrt{L}}\|\nabla f(x_0)\| \tag{30}$$

for some universal constant $c > 0$. Substituting (30) in (29) yields

$$T \le \frac{4}{5\sqrt{L}}\exp\Big(\frac{L(f(x_0) - f^\star)}{3c^2\|\nabla f(x_0)\|^2}\Big) \le \frac{4}{5\sqrt{L}}\exp\frac{L}{6c^2\mu}.$$

$\square$

It readily gives the proof of Theorem 5.2 by combining Lemmas D.2 and D.3.

17

Proof of Theorem 5.2. According to Lemma D.3, by time t there are at least $n^\star \triangleq \lfloor 5t\sqrt{L}e^{-C'L/\mu}/4 \rfloor$ restartings for X^{sr}. By Lemma D.2 and monotonically decreasing of f before restarting, we have

$$f(X^{\mathrm{sr}}(t)) - f(x^\star) \le f(X^{\mathrm{sr}}(\sum_{i=1}^{n^\star} T_i)) - f(x^\star)$$

$$\le (1 - \frac{C\mu}{L})(f(X^{\mathrm{sr}}(\sum_{i=1}^{n^\star-1} T_i)) - f(x^\star))$$

$$\le \dots$$

$$\le (1 - \frac{C\mu}{L})^{n^\star}(f(x_0) - f(x^\star))$$

$$\le \exp(-\frac{C\mu n^\star}{L})(f(x_0) - f(x^\star))$$

$$\le c_1(f(x_0) - f(x^\star))e^{-c_2 t\sqrt{L}},$$

where $c_1 = \exp(C\mu/L)$ and $c_2 = 5C\mu e^{-C'\mu/L}/(4L)$. $\qquad\square$