

Taming the Devil of Gradient-based Optimization with the Angel of Differential Equations

Weijie J. Su

Department of Statistics, University of Pennsylvania

Operator Splitting Methods in Data Analysis Workshop, March 20, 2019

This talk: ordinary differential equations (ODEs) for first-order methods

- ▶ Develop ODEs as amenable surrogates for (discrete) optimization methods
- ▶ Provide intuitive and generalizable proofs
- ▶ Suggest new accelerated methods

Collaborators



Stephen Boyd (Stanford)



Emmanuel Candes
(Stanford)



Simon Du (CMU)



Michael Jordan (Berkeley)



Bin Shi (Berkeley)

Gradient descent

f is convex and L -smooth: $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$

Gradient descent for minimizing f

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

- Convergence rate

$$f(x_k) - f^* \leq O\left(\frac{1}{k}\right)$$

with step size $s = 1/L$, where $f^* = \min f(x)$

- $\nabla f(x_k)$ replaced by proximal subgradient if f is composite (lasso)

Accelerating convergence rate

Nesterov's accelerated gradient method '83

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

from $x_0 = y_0$

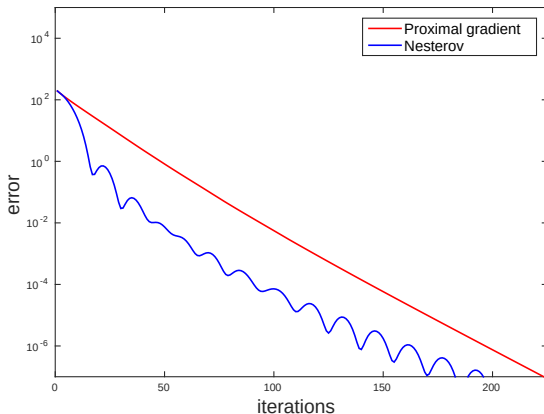
- Convergence rate with $s = 1/L$

$$f(x_k) - f^* \leq O\left(\frac{1}{k^2}\right)$$

- Optimal rate with access to first-order information
- Generalized to composite minimization (Beck and Teboulle '09)

Solving SLOPE using Nesterov's method

$$\min_x \underbrace{\frac{1}{2} \|b - Ax\|^2}_{\text{smooth}} + \underbrace{\sum_{i=1}^n \lambda_i |x|_{(i)}}_{\text{nonsmooth but convex}}$$



Error denotes $f(x_k) - f^*$; design matrix A is 1000×10000

From momentum to acceleration

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

- Common wisdom: reduce zig zags, smooth paths
- What is the underlying mechanism?
- Why $\frac{k-1}{k+2}$?

From momentum to acceleration

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

- Common wisdom: reduce zig zags, smooth paths
- What is the underlying mechanism?
- Why $\frac{k-1}{k+2}$?
- Existing approaches: generalized estimate sequence (Baes '09), Chebyshev polynomials (Hardt '13), linear coupling (Allen-Zhu and Orecchia '14), optimized first-order method (Drori and Teboulle '14), control theory (Lessard et al '16)

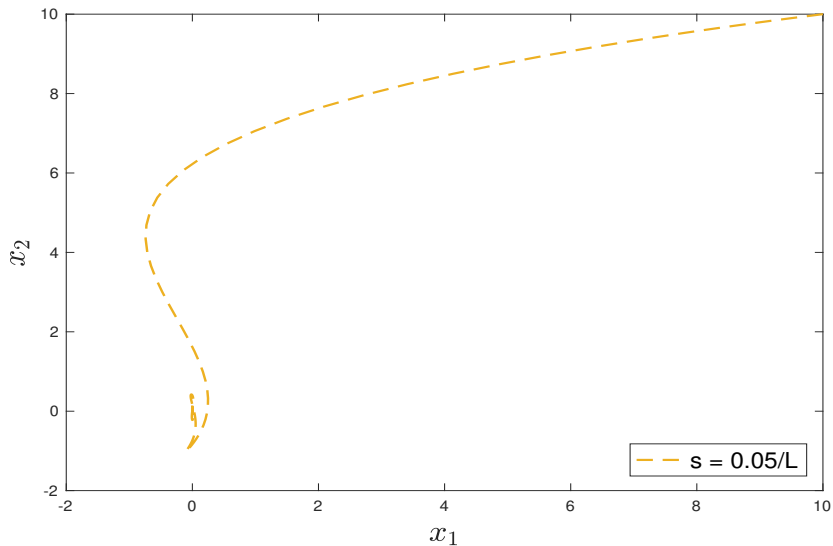
Outline

1. A second-order ODE
2. High-resolution ODEs
3. Concluding remarks

The beginning of the story...

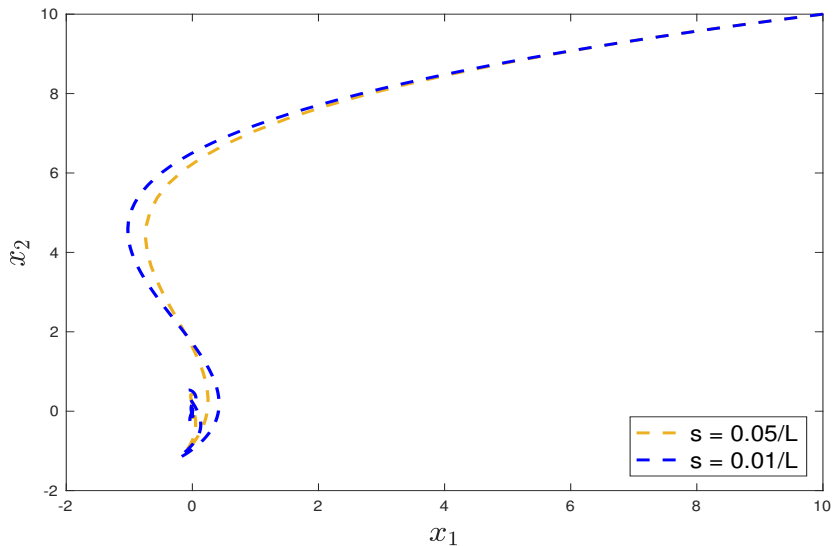
Trajectories of Nesterov's method

Iterates from minimizing $f(x) = \frac{5}{2}x_1^2 + \frac{1}{2}x_2^2$



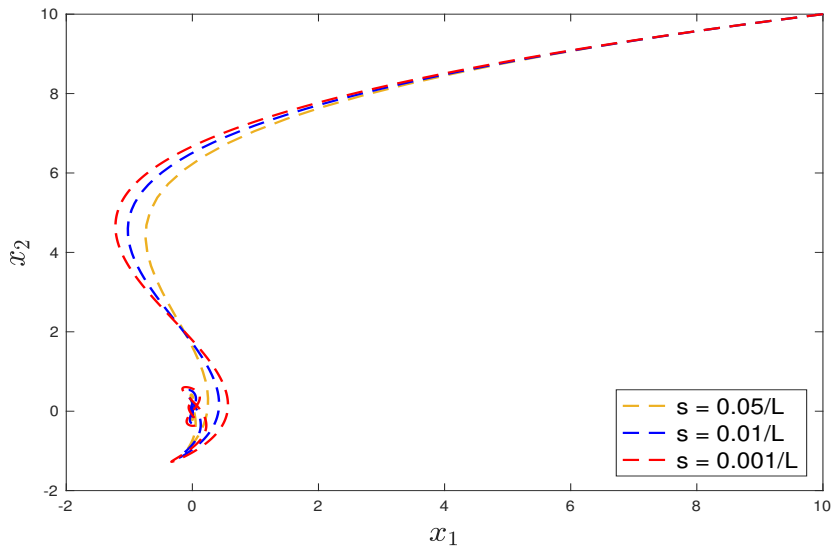
Trajectories of Nesterov's method

Iterates from minimizing $f(x) = \frac{5}{2}x_1^2 + \frac{1}{2}x_2^2$



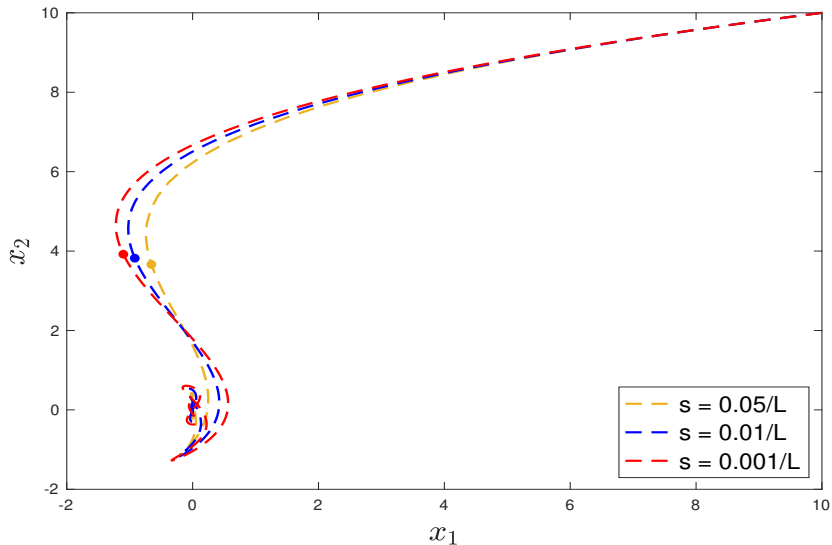
Trajectories of Nesterov's method

Iterates from minimizing $f(x) = \frac{5}{2}x_1^2 + \frac{1}{2}x_2^2$



Time scaling

Iterates at $k = 2.5/\sqrt{\text{step size}}$



The limit of Nesterov's method

Nesterov's method

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

$$y_k = x_k + \frac{k-1}{k+2}(x_k - x_{k-1})$$

Theorem (S., Boyd, and Candès)

Taking $s \rightarrow 0$, Nesterov's method converges to the ODE

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

with $X(0) = x_0, \dot{X}(0) = 0$ in the sense $\lim_{s \rightarrow 0} \max_{k \leq \frac{T}{\sqrt{s}}} \|x_k - X(k\sqrt{s})\| = 0$

The limit of Nesterov's method

Nesterov's method

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

$$y_k = x_k + \frac{k-1}{k+2}(x_k - x_{k-1})$$

Theorem (S., Boyd, and Candès)

Taking $s \rightarrow 0$, Nesterov's method converges to the ODE

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

with $X(0) = x_0, \dot{X}(0) = 0$ in the sense $\lim_{s \rightarrow 0} \max_{k \leq \frac{T}{\sqrt{s}}} \|x_k - X(k\sqrt{s})\| = 0$

- Solution exists and unique
- A second-order ODE
- Time parameter $t \approx k\sqrt{\text{step size}} \propto \sqrt{\text{step size}}$

The ODE

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

Analogous convergence rate

Theorem (S., Boyd, and Candès)

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

$\Downarrow$

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

Theorem (Nesterov)

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

$$y_k = x_k + \frac{k-1}{k+2}(x_k - x_{k-1})$$

$\Downarrow$

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

- $t^2 \approx s(k+1)^2$

A simple proof

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

- Energy functional (Lyapunov)

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

- By convexity of f

$$\begin{aligned}\frac{d\mathcal{E}}{dt} &= 2t(f(X) - f^\star) + 4\langle X - x^\star, -\frac{t}{2}\nabla f(X)\rangle \\ &= 2t(f(X) - f^\star) - 2t\langle X - x^\star, \nabla f(X)\rangle \leq 0\end{aligned}$$

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

Extending the ODE

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

Theorem (S., Boyd, and Candès)

Suppose $r > 3$. Then

$$f(X(t)) - f^* \leq \frac{(r-1)^2 \|x_0 - x^*\|^2}{2t^2}, \quad \int_0^\infty t(f(X(t)) - f^*) dt \leq \frac{(r-1)^2 \|x_0 - x^*\|^2}{2(r-3)}$$

- Acceleration remains
- If $r < 3$, *no* acceleration! (see also Attouch et al '17)
- Proof based on $\mathcal{E}(t) = \frac{2t^2}{r-1}(f(X) - f^*) + (r-1)\|X + \frac{t}{r-1}\dot{X} - x^*\|^2$

Generalized Nesterov's methods

Back to the discrete world, from $y_0 = x_0$

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

$$y_k = x_k + \frac{k-1}{k+r-1}(x_k - x_{k-1})$$

- r results from $k+r-1-(k-1)$
- Generalized to composite minimization by replacing $\nabla f(y_{k-1})$ with proximal subgradient

Generalized Nesterov's method

For $r > 3$ and $0 < s \leq 1/L$

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

- $O(1/k^2)$ convergence rate remains
- Suggests $f(x_k) - f^\star = o(1/k^2)$ asymptotically (Attouch and Peypouquet '16)
- Similar extension in Tseng '10

Outline

1. A second-order ODE
2. High-resolution ODEs
3. Concluding remarks

Methods for strongly convex functions

Let f be μ -strongly convex and L -smooth

Polyak's heavy-ball method

$$x_{k+1} = x_k + \alpha (x_k - x_{k-1}) - s \nabla f(x_k)$$

- Polyak suggests $\alpha = (1 - \sqrt{\mu/L})^2$

Nesterov's method

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

$$x_{k+1} = y_{k+1} + \frac{1 - \sqrt{\mu s}}{1 + \sqrt{\mu s}} (y_{k+1} - y_k)$$

They look similar

Let f be μ -strongly convex and L -smooth

Nesterov's method

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

Equivalent to

$$x_{k+1} = x_k + \frac{1 - \sqrt{\mu s}}{1 + \sqrt{\mu s}} (x_k - x_{k-1}) - \underbrace{s \nabla f(x_k) - \frac{1 - \sqrt{\mu s}}{1 + \sqrt{\mu s}} s (\nabla f(x_k) - \nabla f(x_{k-1}))}_{\text{gradient correction}}$$

Polyak's heavy-ball method

$$x_{k+1} = x_k + \frac{1 - \sqrt{\mu s}}{1 + \sqrt{\mu s}} (x_k - x_{k-1}) - s \nabla f(x_k)$$

- Only differ in *gradient correction*

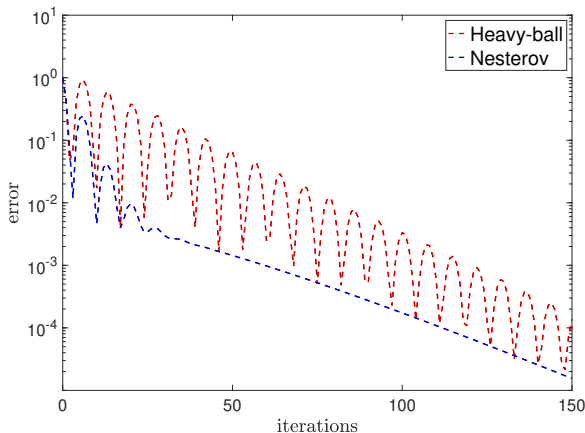
They have the same ODE

Nesterov's and Polyak's share the same ODE (Wilson et al '16)

$$\ddot{X}(t) + 2\sqrt{\mu}\dot{X}(t) + \nabla f(X(t)) = 0$$

- The gradient correction $\frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}} s (\nabla f(x_k) - \nabla f(x_{k-1}))$ is **not** reflected due to *low resolution*

But they are very different!



$f(x_1, x_2) = x_1^2 + 5 \times 10^{-3} x_2^2$, $x_0 = (1, 1)$ and step size $s = 0.09$.

- Polyak's: oscillations

Need new ODEs to capture fine-grained behaviors

High-resolution ODEs

Let s be small but non-vanishing

High-resolution ODEs (Shi, Du, Jordan, and S.)

- Polyak's

$$\ddot{X}(t) + 2\sqrt{\mu}\dot{X}(t) + (1 + \sqrt{\mu s})\nabla f(X(t)) = 0$$

- Nesterov's

$$\ddot{X}(t) + 2\sqrt{\mu}\dot{X}(t) + \sqrt{s}\nabla^2 f(X(t))\dot{X}(t) + (1 + \sqrt{\mu s})\nabla f(X(t)) = 0$$

High-resolution ODEs

Let s be small but non-vanishing

High-resolution ODEs (Shi, Du, Jordan, and S.)

- Polyak's

$$\ddot{X}(t) + 2\sqrt{\mu}\dot{X}(t) + (1 + \sqrt{\mu s})\nabla f(X(t)) = 0$$

- Nesterov's

$$\ddot{X}(t) + 2\sqrt{\mu}\dot{X}(t) + \sqrt{s}\nabla^2 f(X(t))\dot{X}(t) + (1 + \sqrt{\mu s})\nabla f(X(t)) = 0$$

- $X(0) = x_0$ and $\dot{X}(0) = -\frac{2\sqrt{s}\nabla f(x_0)}{1+\sqrt{\mu s}}$
- $\sqrt{s}\nabla^2 f(X)\dot{X}(t)$ results from $\frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}s(\nabla f(x_k) - \nabla f(x_{k-1}))$
- Derivation: carefully Taylor expand $\frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}s(\nabla f(x_k) - \nabla f(x_{k-1}))$

High-resolution ODEs

Let s be small but non-vanishing

High-resolution ODEs (Shi, Du, Jordan, and S.)

- Polyak's

$$\ddot{X}(t) + 2\sqrt{\mu}\dot{X}(t) + (1 + \sqrt{\mu s})\nabla f(X(t)) = 0$$

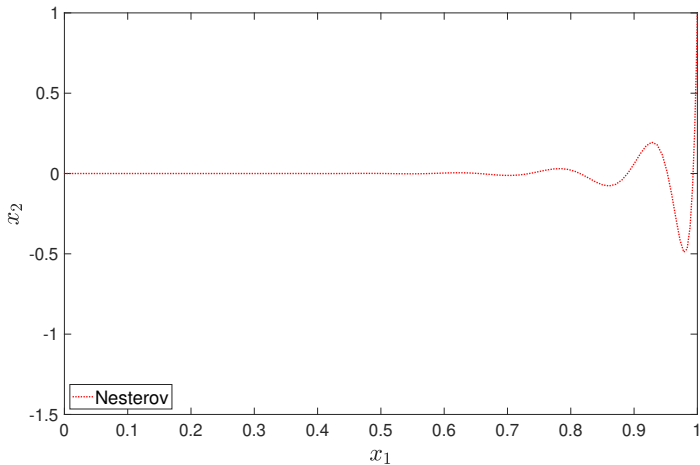
- Nesterov's

$$\ddot{X}(t) + 2\sqrt{\mu}\dot{X}(t) + \sqrt{s}\nabla^2 f(X(t))\dot{X}(t) + (1 + \sqrt{\mu s})\nabla f(X(t)) = 0$$

- If $s = 0$, high-resolution ODEs reduce to low-resolution ODE
- Modified differential equations

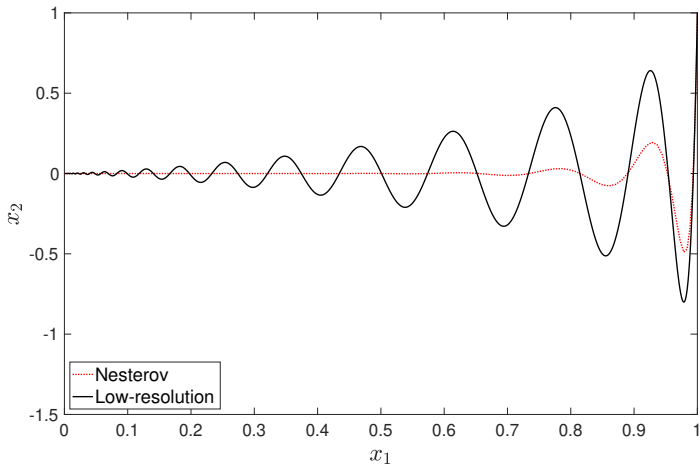
High-resolution ODEs are better surrogates

$$f(x_1, x_2) = x_1^2 + 5 \times 10^{-3} x_2^2, \quad x_0 = (1, 1)$$



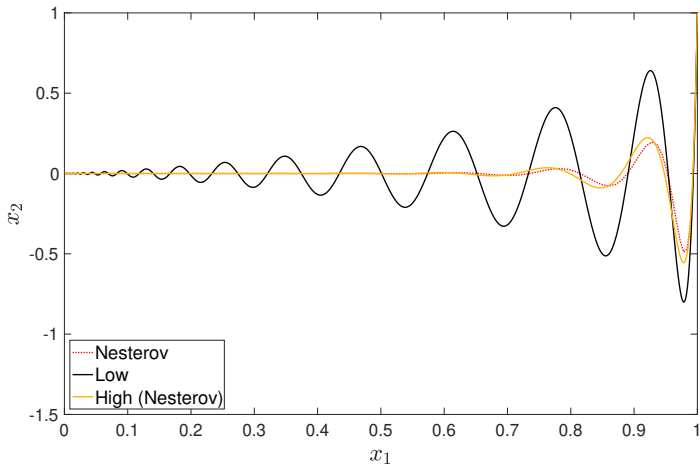
High-resolution ODEs are better surrogates

$$f(x_1, x_2) = x_1^2 + 5 \times 10^{-3} x_2^2, \quad x_0 = (1, 1)$$



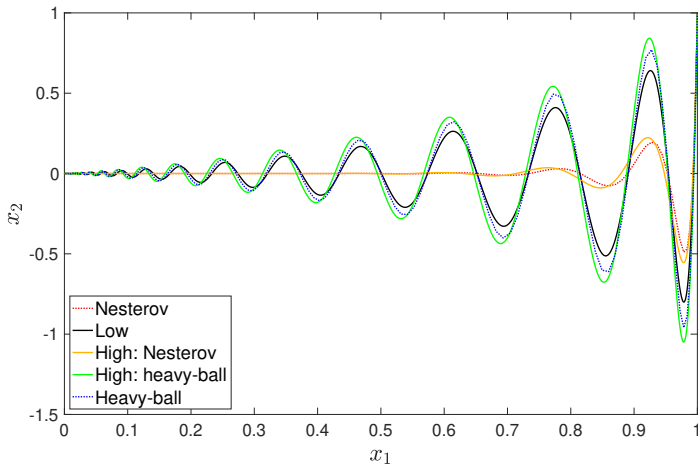
High-resolution ODEs are better surrogates

$$f(x_1, x_2) = x_1^2 + 5 \times 10^{-3} x_2^2, \quad x_0 = (1, 1)$$



High-resolution ODEs are better surrogates

$$f(x_1, x_2) = x_1^2 + 5 \times 10^{-3} x_2^2, \quad x_0 = (1, 1)$$



*Do the high-resolution ODEs distinguish acceleration
and non-acceleration?*

Gradient correction matters

The difference is in $\sqrt{s}\nabla^2 f(X(t))\dot{X}(t)$

- Polyak's

$$\ddot{X}(t) + 2\sqrt{\mu}\dot{X}(t) + (1 + \sqrt{\mu s})\nabla f(X(t)) = 0$$

- Nesterov's

$$\ddot{X}(t) + 2\sqrt{\mu}\dot{X}(t) + \sqrt{s}\nabla^2 f(X(t))\dot{X}(t) + (1 + \sqrt{\mu s})\nabla f(X(t)) = 0$$

Gradient correction matters

The difference is in $\sqrt{s}\nabla^2 f(X(t))\dot{X}(t)$

- Polyak's

$$\ddot{X}(t) + 2\sqrt{\mu}\dot{X}(t) + (1 + \sqrt{\mu s})\nabla f(X(t)) = 0$$

- Nesterov's

$$\ddot{X}(t) + 2\sqrt{\mu}\dot{X}(t) + \sqrt{s}\nabla^2 f(X(t))\dot{X}(t) + (1 + \sqrt{\mu s})\nabla f(X(t)) = 0$$

- $\sqrt{s}\nabla^2 f(X(t))\dot{X}(t)$ (gradient correction) gently adjusts the “friction”

Gradient correction matters

The difference is in $\sqrt{s}\nabla^2 f(X(t))\dot{X}(t)$

- Polyak's

$$\ddot{X}(t) + 2\sqrt{\mu}\dot{X}(t) + (1 + \sqrt{\mu s})\nabla f(X(t)) = 0$$

- Nesterov's

$$\ddot{X}(t) + 2\sqrt{\mu}\dot{X}(t) + \sqrt{s}\nabla^2 f(X(t))\dot{X}(t) + (1 + \sqrt{\mu s})\nabla f(X(t)) = 0$$

- $\sqrt{s}\nabla^2 f(X(t))\dot{X}(t)$ (gradient correction) gently adjusts the “friction”
- Fundamental to the acceleration of Nesterov's method

Energy functional for Nesterov's ODE

$$\ddot{X}(t) + 2\sqrt{\mu}\dot{X}(t) + \sqrt{s}\nabla^2 f(X(t))\dot{X}(t) + (1 + \sqrt{\mu s})\nabla f(X(t)) = 0$$

Energy functional

$$\mathcal{E}(t) = \underbrace{(1 + \sqrt{\mu s})(f(X) - f^*)}_{\text{potential}} + \underbrace{\frac{1}{4}\|\dot{X}\|^2}_{\text{kinetic}} + \frac{1}{4}\|\dot{X} + 2\sqrt{\mu}(X - x^*) + \sqrt{s}\nabla f(X)\|^2$$

Energy functional for Nesterov's ODE

$$\ddot{X}(t) + 2\sqrt{\mu}\dot{X}(t) + \sqrt{s}\nabla^2 f(X(t))\dot{X}(t) + (1 + \sqrt{\mu s}) \nabla f(X(t)) = 0$$

Energy functional

$$\mathcal{E}(t) = \underbrace{(1 + \sqrt{\mu s}) (f(X) - f^*)}_{\text{potential}} + \underbrace{\frac{1}{4} \|\dot{X}\|^2}_{\text{kinetic}} + \frac{1}{4} \|\dot{X} + 2\sqrt{\mu}(X - x^*) + \sqrt{s}\nabla f(X)\|^2$$

- $\dot{X} + 2\sqrt{\mu}(X - x^*) + \sqrt{s}\nabla f(X)$ results from integrating $\ddot{X} + 2\sqrt{\mu}\dot{X} + \sqrt{s}\nabla^2 f(X)\dot{X}$
- $\sqrt{s}\nabla f(X)$ arises from gradient correction

Convergence of Nesterov's ODE

Energy functional

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

Lemma

$$\frac{d\mathcal{E}}{dt} \leq -\frac{\sqrt{\mu}}{4}\mathcal{E} - \frac{\sqrt{s}}{2} \left[\|\nabla f(X)\|^2 + \dot{X}^\top \nabla^2 f(X) \dot{X} \right] \leq -\frac{\sqrt{\mu}}{4}\mathcal{E}$$

Convergence of Nesterov's ODE

Energy functional

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

Lemma

$$\frac{d\mathcal{E}}{dt} \leq -\frac{\sqrt{\mu}}{4}\mathcal{E} - \frac{\sqrt{s}}{2} \left[\|\nabla f(X)\|^2 + \dot{X}^\top \nabla^2 f(X) \dot{X} \right] \leq -\frac{\sqrt{\mu}}{4}\mathcal{E}$$

- $\frac{\sqrt{s}}{2}(\|\nabla f(X)\|^2 + \dot{X}^\top \nabla^2 f(X) \dot{X}) \geq 0$ arises from gradient correction
- For $s \leq 1/L$

$$f(X(t)) - f^\star \leq \frac{2\|x_0 - x^\star\|^2}{s} e^{-\frac{\sqrt{\mu}t}{4}}$$

Convergence of Polyak's ODE

Energy functional

$$\mathcal{E}(t) = (1 + \sqrt{\mu s}) (f(X) - f^*) + \frac{1}{4} \|\dot{X}\|^2 + \frac{1}{4} \|\dot{X} + 2\sqrt{\mu}(X - x^*)\|^2$$

Lemma

$$\frac{d\mathcal{E}}{dt} \leq -\frac{\sqrt{\mu}}{4} \mathcal{E}$$

- $\frac{\sqrt{s}}{2} (\|\nabla f(X)\|^2 + \dot{X}^\top \nabla^2 f(X) \dot{X})$ is not found
- For $s \leq 1/L$

$$f(X(t)) - f^* \leq \frac{7 \|x_0 - x^*\|^2}{2s} e^{-\frac{\sqrt{\mu} t}{4}}$$

Returning to the discrete world

Discrete energy functional for Nesterov's

Continuous-time

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

Discrete energy functional for Nesterov's

Continuous-time

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

Discrete-time

$$\begin{aligned} \mathcal{E}(k) = & \frac{1 + \sqrt{\mu s}}{1 - \sqrt{\mu s}} (f(x_k) - f^\star) + \frac{1}{4} \|v_k\|^2 \\ & + \frac{1}{4} \left\| v_k + \frac{2\sqrt{\mu}}{1 - \sqrt{\mu s}} (x_{k+1} - x^\star) + \sqrt{s}\nabla f(x_k) \right\|^2 - \frac{s \|\nabla f(x_k)\|^2}{2(1 - \sqrt{\mu s})} \end{aligned}$$

- Phase variable $v_k = \frac{x_{k+1} - x_k}{\sqrt{s}}$
- Seamless transform via phase space representation

Discrete energy functional for Nesterov's

$$\begin{aligned}\mathcal{E}(k) = & \frac{1 + \sqrt{\mu s}}{1 - \sqrt{\mu s}} (f(x_k) - f^\star) + \frac{1}{4} \|v_k\|^2 \\ & + \frac{1}{4} \left\| v_k + \frac{2\sqrt{\mu}}{1 - \sqrt{\mu s}} (x_{k+1} - x^\star) + \sqrt{s} \nabla f(x_k) \right\|^2 - \frac{s \|\nabla f(x_k)\|^2}{2(1 - \sqrt{\mu s})}\end{aligned}$$

Lemma

If $0 < s \leq 1/(4L)$, then $\mathcal{E}(k+1) - \mathcal{E}(k) \leq -\frac{\sqrt{\mu s}}{6} \mathcal{E}(k+1)$

Discrete energy functional for Nesterov's

$$\begin{aligned}\mathcal{E}(k) = & \frac{1 + \sqrt{\mu s}}{1 - \sqrt{\mu s}} (f(x_k) - f^\star) + \frac{1}{4} \|v_k\|^2 \\ & + \frac{1}{4} \left\| v_k + \frac{2\sqrt{\mu}}{1 - \sqrt{\mu s}} (x_{k+1} - x^\star) + \sqrt{s} \nabla f(x_k) \right\|^2 - \frac{s \|\nabla f(x_k)\|^2}{2(1 - \sqrt{\mu s})}\end{aligned}$$

Lemma

If $0 < s \leq 1/(4L)$, then $\mathcal{E}(k+1) - \mathcal{E}(k) \leq -\frac{\sqrt{\mu s}}{6} \mathcal{E}(k+1)$

- Implies

$$f(x_k) - f^\star \leq \frac{5L \|x_0 - x^\star\|^2}{\left(1 + \frac{1}{12} \sqrt{\mu/L}\right)^k}$$

- $\log(f(x_k) - f^\star) \leq -O(k\sqrt{\mu/L})$ matches the optimal bound (Nesterov '13)

Discrete energy functional for Polyak's

$$\mathcal{E}(k) = \frac{1 + \sqrt{\mu s}}{1 - \sqrt{\mu s}} (f(x_k) - f^\star) + \frac{1}{4} \|v_k\|^2 + \frac{1}{4} \left\| v_k + \frac{2\sqrt{\mu}}{1 - \sqrt{\mu s}} (x_{k+1} - x^\star) \right\|^2$$

Lemma

$$\begin{aligned} \mathcal{E}(k+1) - \mathcal{E}(k) &\leq -\sqrt{\mu s} \min \left\{ \frac{1 - \sqrt{\mu s}}{1 + \sqrt{\mu s}}, \frac{1}{4} \right\} \mathcal{E}(k+1) \\ &\quad - \left[\frac{3\sqrt{\mu s}}{4} \left(\frac{1 + \sqrt{\mu s}}{1 - \sqrt{\mu s}} \right) (f(x_{k+1}) - f^\star) - \frac{s}{2} \left(\frac{1 + \sqrt{\mu s}}{1 - \sqrt{\mu s}} \right)^2 \|\nabla f(x_{k+1})\|^2 \right] \end{aligned}$$

- Need to ensure the “annoying” term

$$\frac{3\sqrt{\mu s}}{4} \left(\frac{1 + \sqrt{\mu s}}{1 - \sqrt{\mu s}} \right) (f(x_{k+1}) - f^\star) - \frac{s}{2} \left(\frac{1 + \sqrt{\mu s}}{1 - \sqrt{\mu s}} \right)^2 \|\nabla f(x_{k+1})\|^2 \geq 0$$

Where is this “annoying” term from?

The continuous energy functional for Nesterov's

$$\frac{d\mathcal{E}}{dt} \leq -\frac{\sqrt{\mu}}{4}\mathcal{E} - \underbrace{\frac{\sqrt{s}}{2} \left[\|\nabla f(X)\|^2 + \dot{X}^\top \nabla^2 f(X) \dot{X} \right]}_D$$

- In fact, the “annoying” term appears in Nesterov's, but canceled out by D
- Recall D is due to gradient correction
- Thus, the “annoying” term is due to the lack of gradient correction in Polyak's

When is the “annoying” term nonnegative?

Lemma (Polyak's)

$$\mathcal{E}(k+1) - \mathcal{E}(k) \leq -\sqrt{\mu s} \min \left\{ \frac{1 - \sqrt{\mu s}}{1 + \sqrt{\mu s}}, \frac{1}{4} \right\} \mathcal{E}(k+1) \\ - \left[\frac{3\sqrt{\mu s}}{4} \left(\frac{1 + \sqrt{\mu s}}{1 - \sqrt{\mu s}} \right) (f(x_{k+1}) - f^*) - \frac{s}{2} \left(\frac{1 + \sqrt{\mu s}}{1 - \sqrt{\mu s}} \right)^2 \|\nabla f(x_{k+1})\|^2 \right]$$

- It is nonnegative if $s = O\left(\frac{\mu}{L^2}\right)$ in Polyak's
- $\mathcal{E}(k+1) - \mathcal{E}(k) \leq -\sqrt{\mu s} \min \left\{ \frac{1 - \sqrt{\mu s}}{1 + \sqrt{\mu s}}, \frac{1}{4} \right\} \mathcal{E}(k+1)$
- Take $s = \mu/(16L^2)$, Polyak's convergence

$$f(x_k) - f(x_0) \leq \frac{5L \|x_0 - x^*\|^2}{\left(1 + \frac{\mu}{16L}\right)^k}$$

Comparing the two methods

Nesterov's

- Contains gradient correction
- Step size $s = O\left(\frac{1}{L}\right)$
- $\log(f(x_k) - f^*) \leq -O(k\sqrt{\mu/L})$
- Achieves acceleration

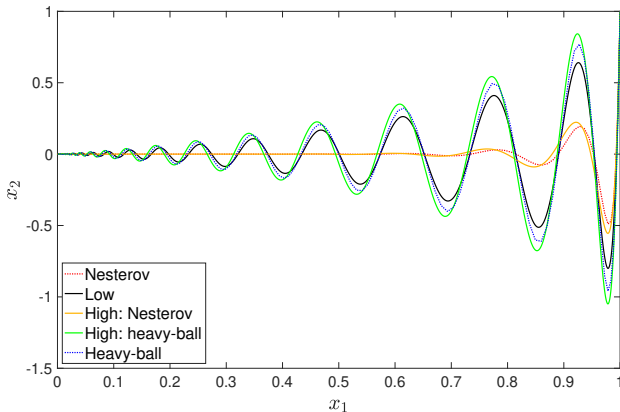
Polyak's

- No gradient correction
- Step size $s = O\left(\frac{\mu}{L^2}\right)$
- $\log(f(x_k) - f^*) \leq -O(k\mu/L)$
- No (global) acceleration

- For ill-conditioned $\mu \ll L$ cases, $O\left(\frac{1}{L}\right) \gg O\left(\frac{\mu}{L^2}\right)$

A straight or winding road?

Why Nesterov's allows a larger step size than Polyak's?



- Gradient correction in Nesterov's "smoothes out" bumps

All roads lead to Rome, but...



Yet another application of high-resolution ODEs

Make gradient small

Let f be L -smooth (non-strongly) convex

How to minimize $\|\nabla f(x)\|^2$ efficiently?

Make gradient small

Let f be L -smooth (non-strongly) convex

How to minimize $\|\nabla f(x)\|^2$ efficiently?

- A centerpiece in non-convex optimization
- Nesterov's achieves

$$\|\nabla f(x_k)\|^2 \leq O\left(\frac{1}{k^2}\right)$$

because $\|\nabla f(x_k)\|^2 \leq 2L(f(x_k) - f^*)$ and $f(x_k) - f^* \leq O(1/k^2)$. Recall

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

$$y_k = x_k + \frac{k-1}{k+2}(x_k - x_{k-1})$$

Yet another high-resolution ODE

High-resolution ODE for non-strongly convex objectives

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

for $t \geq 3\sqrt{s}/2$, with $X(3\sqrt{s}/2) = x_0$ and $\dot{X}(3\sqrt{s}/2) = -\sqrt{s}\nabla f(x_0)$

- Reduces to the low-resolution ODE if $s = 0$
- Contains gradient correction $\sqrt{s}\nabla^2 f(X)\dot{X}$

Nesterov minimizes gradient norm faster!

Theorem (Shi, Du, Jordan, and S.)

Let $s \leq 1/(3L)$, the Nesterov's method (non-strongly convex) satisfies

$$\min_{0 \leq i \leq k} \|\nabla f(x_i)\|^2 \leq \frac{c \|x_0 - x^*\|^2}{s^2 (k+1)^3}$$

Nesterov minimizes gradient norm faster!

Theorem (Shi, Du, Jordan, and S.)

Let $s \leq 1/(3L)$, the Nesterov's method (non-strongly convex) satisfies

$$\min_{0 \leq i \leq k} \|\nabla f(x_i)\|^2 \leq \frac{c \|x_0 - x^*\|^2}{s^2(k+1)^3}$$

- Improved from $O(1/k^2)$ to $O(1/k^3)$. The constant $c = 8568$
- Based on the discrete energy functional

$$\begin{aligned} \mathcal{E}(k) = & s(k+3)(k+1) (f(x_k) - f^*) \\ & + \frac{1}{2} \left\| (k+1)\sqrt{s}v_k + 2(x_{k+1} - x^*) + (k+1)s\nabla f(x_k) \right\|^2, \end{aligned}$$

which satisfies $\mathcal{E}(k+1) - \mathcal{E}(k) \leq -Cs^2k^2 \|\nabla f(x_{k+1})\|^2$

- $s^2k^2 \|\nabla f(x_{k+1})\|^2$ due to gradient correction

More comments on the improved rate

Theorem (Shi, Du, Jordan, and S.)

Let $s \leq 1/(3L)$, Nesterov's method (non-strongly convex) satisfies

$$\min_{0 \leq i \leq k} \|\nabla f(x_i)\|^2 \leq \frac{c \|x_0 - x^\star\|^2}{s^2 (k+1)^3}$$

- Previously, the best known bound of Nesterov is $o(1/k^2)$
- Sharpest known bound (without modification)

More comments on the improved rate

Theorem (Shi, Du, Jordan, and S.)

Let $s \leq 1/(3L)$, Nesterov's method (non-strongly convex) satisfies

$$\min_{0 \leq i \leq k} \|\nabla f(x_i)\|^2 \leq \frac{c \|x_0 - x^\star\|^2}{s^2 (k+1)^3}$$

- Previously, the best known bound of Nesterov is $o(1/k^2)$
- Sharpest known bound (without modification)
- Why minimization of gradient is easier?

Can the high-resolution ODEs suggest new methods?

Extensions for non-strongly convex functions

Generalized high-resolution ODE

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

for $t \geq \alpha\sqrt{s}/2$, with $X(\alpha\sqrt{s}/2) = x_0$ and $\dot{X}(\alpha\sqrt{s}/2) = -\sqrt{s}\nabla f(x_0)$

- Reduces to the original Nesterov's if $\alpha = 3, \beta = 1$

Generalized Nesterov's method

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

$$x_{k+1} = x_k - s \nabla f(x_k) + \frac{k}{k + \alpha}(y_{k+1} - y_k),$$

starting with $x_0 = y_0$

Accelerated rates

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

$$x_{k+1} = x_k - s \nabla f(x_k) + \frac{k}{k + \alpha} (y_{k+1} - y_k),$$

Theorem (Shi, Du, Jordan, and S.)

If $\alpha \geq 3$ and $\beta > \frac{1}{2}$, then

$$f(x_k) - f^* \leq O\left(\frac{1}{k^2}\right), \quad \min_{0 \leq i \leq k} \|\nabla f(x_i)\|^2 \leq O\left(\frac{1}{k^3}\right)$$

In addition, if $\alpha > 3$ then

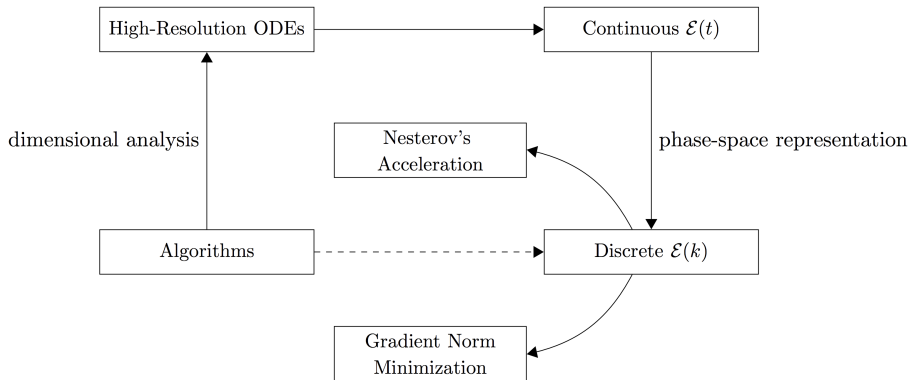
$$f(x_k) - f^* \leq o\left(\frac{1}{k^2}\right)$$

- Why $\beta > \frac{1}{2}$?
- $f(x_k) - f^* \leq o\left(\frac{1}{k^2}\right)$ for $\alpha > 3$ extends Attouch and Peypouquet '16

Outline

1. A second-order ODE
2. High-resolution ODEs
3. Concluding remarks

A framework for understanding optimization



Take home messages

- ODEs are amenable tools for analyzing gradient-based methods
- Conceptually simple, suggest new methods
- Sometimes, need to “upgrade” ODEs
- Non-convex, stochastic, sampling, constrained settings? Many research opportunities

References

- *A Differential Equation for Modeling Nesterov's Accelerated Gradient Method: Theory and Insights*
Weijie Su, Stephen Boyd, and Emmanuel Candès, Journal of Machine Learning Research, 2016
- *Understanding the Acceleration Phenomenon via High-Resolution Differential Equations*
Bin Shi, Simon Du, Michael Jordan, and Weijie Su, arXiv, 2018
- *Acceleration via Symplectic Discretization of High-Resolution Differential Equations*
Bin Shi, Simon Du, Weijie Su, and Michael Jordan, arXiv, 2019