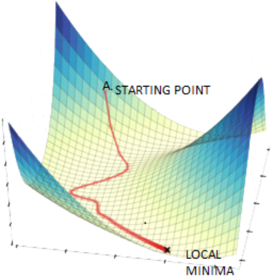
# Continuous-Time Perspectives on First-Order Optimization Methods

Weijie Su

University of Pennsylvania

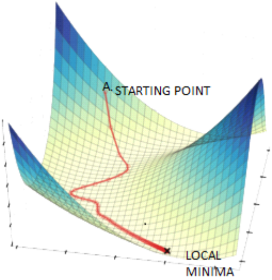
# Gradient-based optimization

minimize $f(x)$ using $\nabla f(x)$



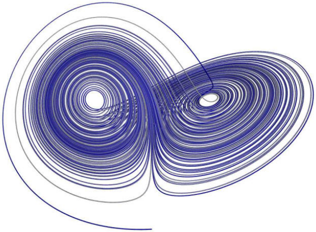
# Gradient-based optimization

minimize $f(x)$ using $\nabla f(x)$




- Simplest example: gradient descent
- Almost entirely focused on differentiation
- Toolkit is (relatively) small

# Dynamical systems



- Simplest example: ordinary differential equation (ODE)
- Interplay between differentiation and integration
- A much larger toolkit

# Connecting dynamical systems with optimization?

*Leverage the power of ODEs to analyze optimization methods*

# Connecting dynamical systems with optimization?

*Leverage the power of ODEs to analyze optimization methods*

- Long history (see monograph of Helmke and Moore ’96)

# This talk: connecting ODEs with gradient-based methods

*A framework for modeling, analyzing, interpreting,*
*and designing accelerated optimization methods*

# This talk: connecting ODEs with gradient-based methods

*A framework for modeling, analyzing, interpreting, and designing accelerated optimization methods*

- Develop ODEs as amenable surrogates for accelerated optimization methods

# This talk: connecting ODEs with gradient-based methods

*A framework for modeling, analyzing, interpreting, and designing accelerated optimization methods*

- Develop ODEs as amenable surrogates for accelerated optimization methods
- Provide intuitive and generalizable proofs

# This talk: connecting ODEs with gradient-based methods

*A framework for modeling, analyzing, interpreting, and designing accelerated optimization methods*

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

# Collaborators

- Stephen Boyd (Stanford)
- Emmanuel Candès (Stanford)
- Shuxiao Chen (UPenn)
- Simon Du (CMU)
- Yicong Jiang (Harvard)
- Michael Jordan (Berkeley)
- Bin Shi (Berkeley)
- Da Wu (UPenn)

# Gradient descent

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

- $\|b - Ax\|^2$: least squares
- $-b^T(Ax + x_0) + 1^T \log(1 + \exp(Ax + x_0))$: logistic regression
- $\frac{1}{2}\|b - Ax\|^2 + \lambda\|x\|_1$: lasso

## Gradient descent for minimizing $f$

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

# Gradient descent

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

- $\|b - Ax\|^2$: least squares
- $-b^T(Ax + x_0) + 1^T \log(1 + \exp(Ax + x_0))$: logistic regression
- $\frac{1}{2}\|b - Ax\|^2 + \lambda\|x\|_1$: lasso

**Gradient descent for minimizing $f$**

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

- Convergence rate

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

  if $s = 1/L$, where $f^\star = \min f(x)$
- $\nabla f(x_k)$ replaced by proximal subgradient if $f$ is composite (lasso)

# Accelerating the convergence

## Nesterov's accelerated gradient method '83

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

from $x_0 = y_0$

# Accelerating the convergence

## Nesterov's accelerated gradient method '83

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

from $x_0 = y_0$

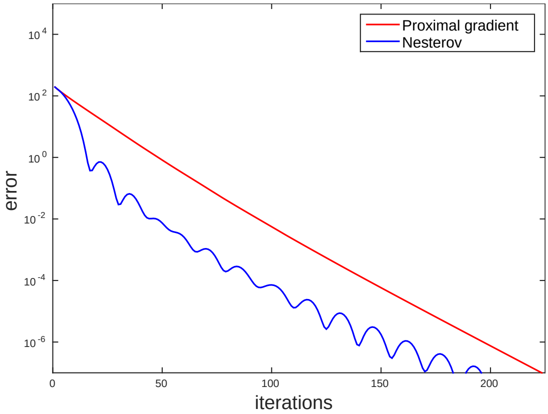
- For $L$-smooth convex $f$, Nesterov proved that for $0 < s \leq 1/L$

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

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

# Solving SLOPE using Nesterov's method

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




Error denotes $f(x_k) - f^\star$; design matrix A is $1000 \times 10000$

# Mysteries of acceleration

Common wisdom: momentum reduces zig zags and smooths paths

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

# Mysteries of acceleration

Common wisdom: momentum reduces zig zags and smooths paths

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

- What is the underlying mechanism?
- Why $\frac{k-1}{k+2}$?

# Mysteries of acceleration

Common wisdom: momentum reduces zig zags and smooths paths

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

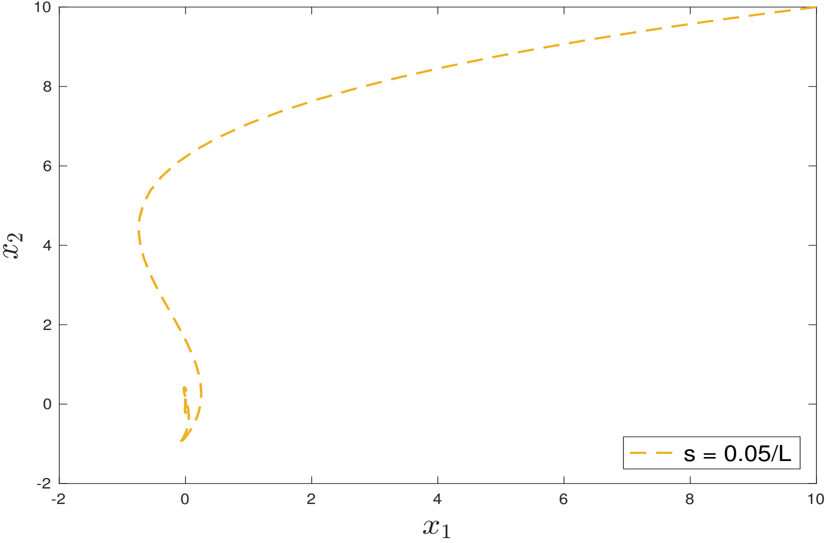
- What is the underlying mechanism?
- Why $\frac{k-1}{k+2}$?
- Existing approaches: generalized estimate sequence (Baes '09), Chebyshev polynomials (Hardt '13), linear coupling (Allen-Zhu and Orecchia '14), optimized first-order method (Drori and Teboulle '14), control theory (Lessard et al '16)

# Outline

*The beginning of the story...*

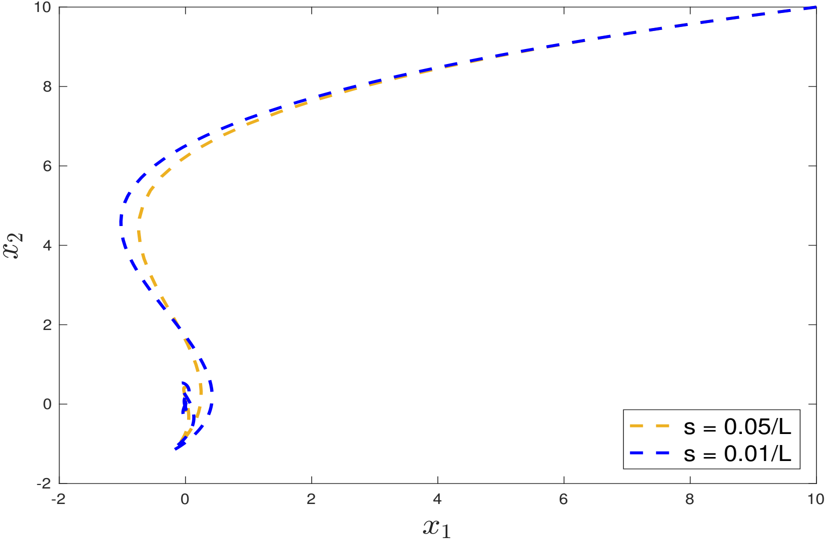
# Trajectories of Nesterov's method

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



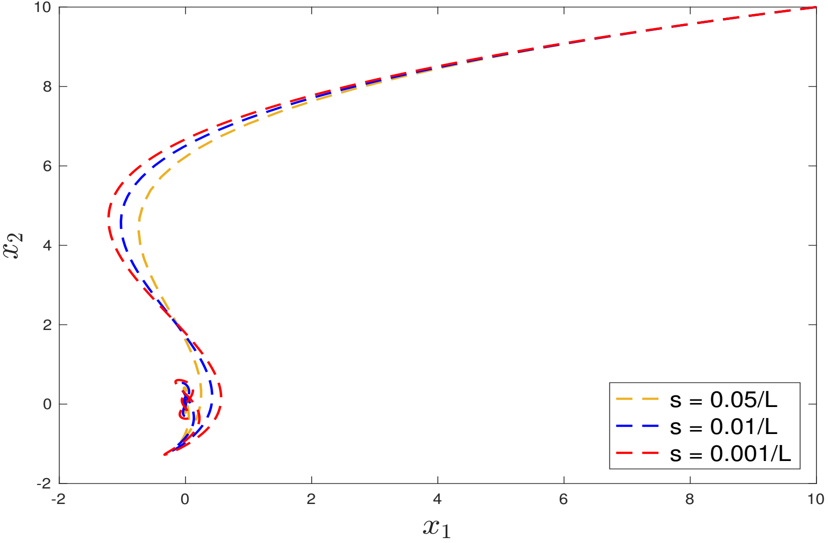
# Trajectories of Nesterov's method

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

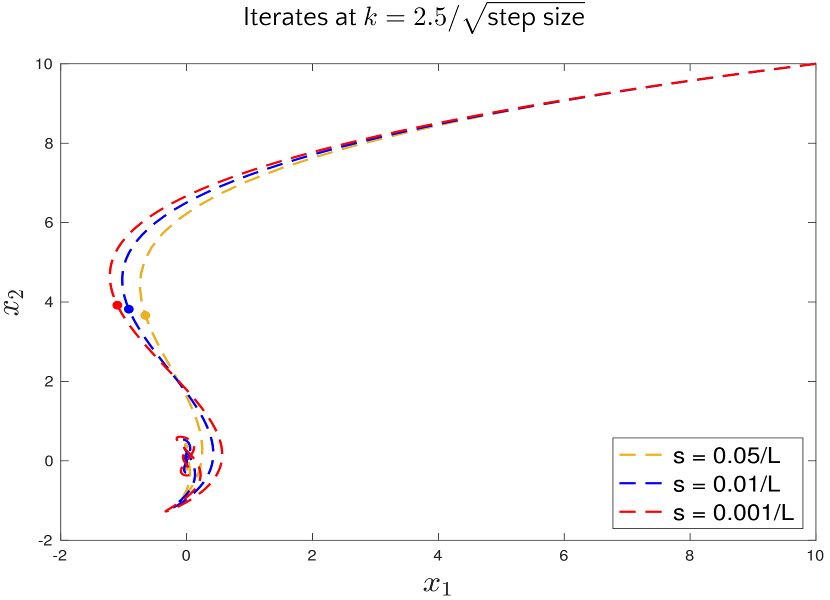


# Trajectories of Nesterov's method

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



# Time scaling



# 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**

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

# 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**

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

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

# Derivation I

Nesterov's method in one-line

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

$$\Downarrow$$

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

# Derivation II

Let $t_k = k\sqrt{s}$. Assume $x_k = X(t_k)$ for some smooth curve $X$

$$\frac{x_{k+1} - x_k}{\sqrt{s}} = \dot{X}(t_k) + \frac{1}{2}\ddot{X}(t_k)\sqrt{s} + o(\sqrt{s})$$
$$\frac{x_k - x_{k-1}}{\sqrt{s}} = \dot{X}(t_k) - \frac{1}{2}\ddot{X}(t_k)\sqrt{s} + o(\sqrt{s})$$
$$\sqrt{s}\nabla f(y_k) = \sqrt{s}\nabla f(X(t_k)) + o(\sqrt{s})$$

Comparing coefficients of $\sqrt{s}$ in Nesterov's method gives

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

# Ask me anything

## The Nesterov ODE

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

A useful surrogate for Nesterov's method?

# Ask me anything

## The Nesterov ODE

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

A useful surrogate for Nesterov's method? I think so

# Ask me anything

## The Nesterov ODE

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

A useful surrogate for Nesterov's method? I think so

Simplify some proofs for Nesterov's method?

# Ask me anything

## The Nesterov ODE

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

| | |
|---|---|
| A useful surrogate for Nesterov's method? | I think so |
| Simplify some proofs for Nesterov's method? | Yes |

# Ask me anything

## The Nesterov ODE

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

| | |
|---|---|
| A useful surrogate for Nesterov's method? | I think so |
| Simplify some proofs for Nesterov's method? | Yes |
| Suggest new accelerated methods? | |

# Ask me anything

## The Nesterov ODE

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

| | |
|---|---|
| A useful surrogate for Nesterov's method? | I think so |
| Simplify some proofs for Nesterov's method? | Yes |
| Suggest new accelerated methods? | Yes |

# Ask me anything

## The Nesterov ODE

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

| | |
|---|---|
| A useful surrogate for Nesterov's method? | I think so |
| Simplify some proofs for Nesterov's method? | Yes |
| Suggest new accelerated methods? | Yes |

Can the ODE do everything for the method?

# Ask me anything

**The Nesterov ODE**

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

| | |
|---|---|
| A useful surrogate for Nesterov's method? | I think so |
| Simplify some proofs for Nesterov's method? | Yes |
| Suggest new accelerated methods? | Yes |
| Can the ODE do everything for the method? | Of course not, but we've *upgraded* the ODE |

# Ask me anything

## The Nesterov ODE

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

| | |
|---|---|
| A useful surrogate for Nesterov's method? | I think so |
| Simplify some proofs for Nesterov's method? | Yes |
| Suggest new accelerated methods? | Yes |
| Can the ODE do everything for the method? | Of course not, but we've *upgraded* the ODE |
| Can the upgraded ODEs do something new? | |

# Ask me anything

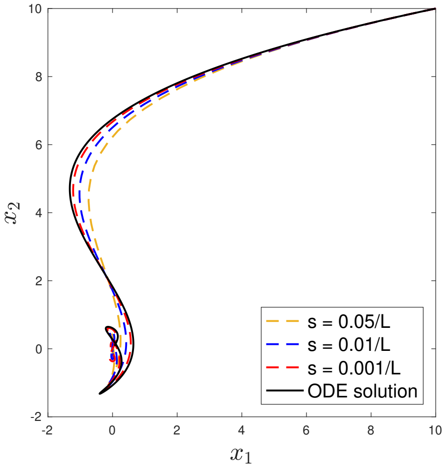
**The Nesterov ODE**

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

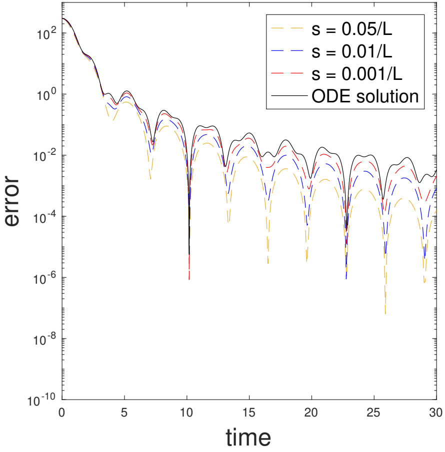
| | |
|---|---|
| A useful surrogate for Nesterov's method? | I think so |
| Simplify some proofs for Nesterov's method? | Yes |
| Suggest new accelerated methods? | Yes |
| Can the ODE do everything for the method? | Of course not, but we've *upgraded* the ODE |
| Can the upgraded ODEs do something new? | Yes |

# A faithful surrogate

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




Trajectories




Convergence

# Analogous convergence rate

**Theorem (Our)**

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

$$\Downarrow$$

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

# Analogous convergence rate

**Theorem (Our)**

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

$$\Downarrow$$

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

**Theorem (Nesterov)**

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

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

$$\Downarrow$$

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

# Analogous convergence rate

**Theorem (Our)**

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

$$\Downarrow$$

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

**Theorem (Nesterov)**

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

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

$$\Downarrow$$

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

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

# A simple proof

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

# A simple proof

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

- Energy functional (Lyapunov)

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

# A simple proof

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 $f$

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

# A simple proof

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

- 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 $f$

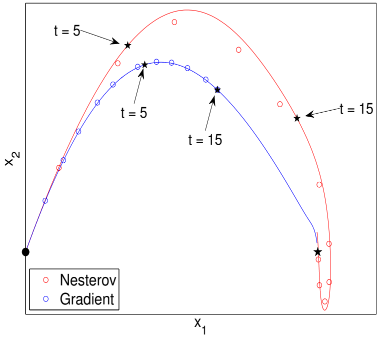
$$\begin{aligned}\frac{\mathrm{d}\mathcal{E}}{\mathrm{d}t} &= 2t(f(X) - f^\star) + 4\langle X - x^\star, -\frac{t}{2}\nabla f(X)\rangle \\ &= 2t(f(X) - f^\star) - 2t\langle X - x^\star, \nabla f(X)\rangle \le 0\end{aligned}$$

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

# Comparing gradient descent with Nesterov's method

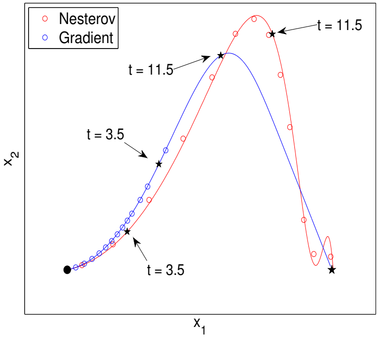
Gradient descent ODE

- $\dot{X} + \nabla f(X) = 0$
- Euler stable step size $O(1/L)$
- Each iteration moves $\propto s$




Nesterov ODE

- $\ddot{X} + \frac{3}{t}\dot{X} + \nabla f(X) = 0$
- Euler stable step size $O(1/\sqrt{L})$
- Each iteration moves $\propto \sqrt{s}$



# Suggesting new methods

New ODE

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

# Suggesting new methods

New ODE

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

**Theorem**

*Suppose* $r > 3$. *Then*

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

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

# 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 $0 < s \leq 1/L$**

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

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

# Generalized Nesterov's method

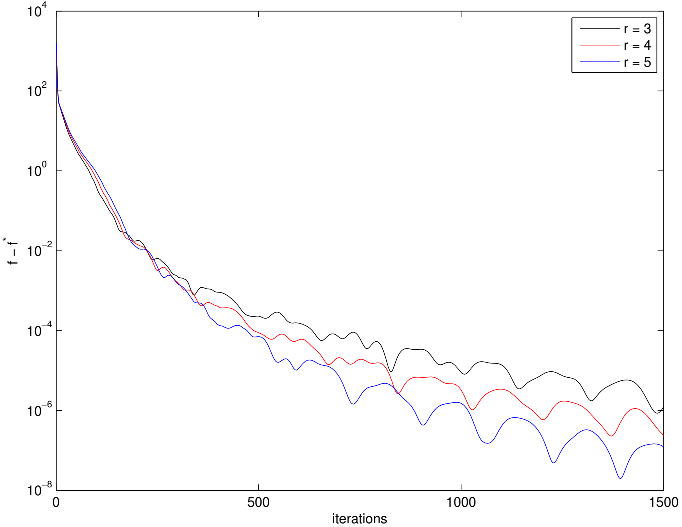
**For $r > 3$ and $0 < s \le 1/L$**

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

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

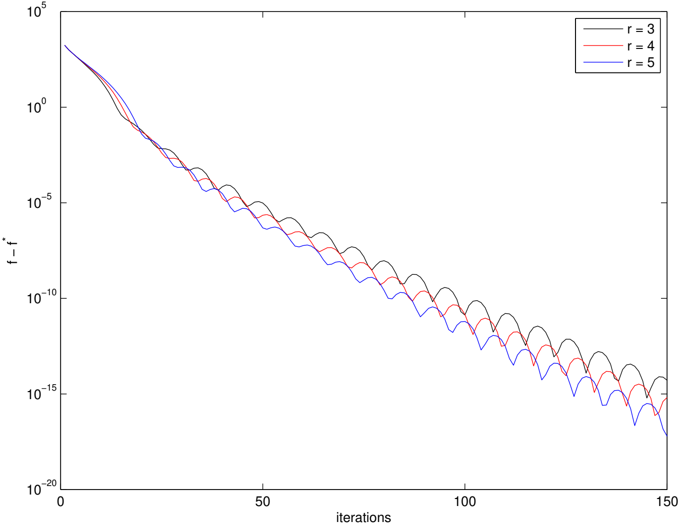
- $O(1/k^2)$ convergence rate remains
- Suggests $f(x_k) - f^\star = o(1/k^2)$ asymptotically (Attouch and Peypouquet '16)

# Numerical Examples I



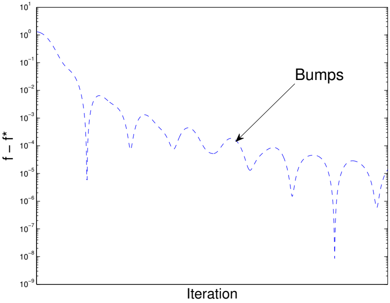

$$\min \frac{1}{2}\|Ax-b\|^2+\lambda\|x\|_1$$

# Numerical Examples II




$$\min \frac{1}{2}\|Ax - b\|^2, \quad \text{s.t. } x \succeq 0$$

# Restarting Nesterov's method I



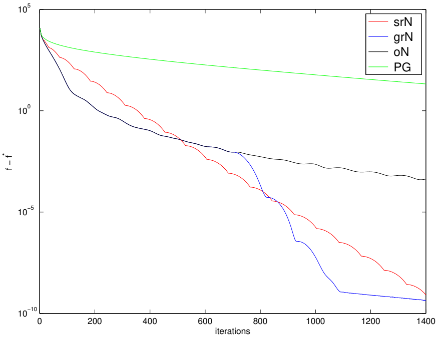

## Cause

If $\frac{3}{t}$ is small, friction is too low

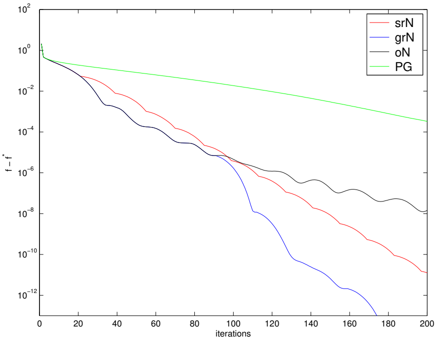
- Time is set to zero whenever velocity starts to decreases
- Early restarting ideas (O'Donoghue and Candès '12)

# Restarting Nesterov's method II

Our restarting (srN), gradient restarting (grN) ( O'Donoghue and Candès '12), Nesterov's method (oN), and proximal gradient (PG)




min $\frac{1}{2}\|Ax - b\|^2$ s.t. $\|x\|_1 \le C$




min $\frac{1}{2}\|X_{\text{obs}} - M_{\text{obs}}\|_{\text{F}}^2 + \lambda\|X\|_*$

# Acceleration and monotonicity simultaneously?

- Nesterov's method achieves acceleration, but is not monotone
- Gradient descent is monotone, but not accelerated

# Acceleration and monotonicity simultaneously?

- Nesterov's method achieves acceleration, but is not monotone
- Gradient descent is monotone, but not accelerated

**Theorem**

*If a first-order method can be represented as a linear combination of several iterates and the gradient, then it **cannot** be both accelerated and monotone*

# Outline

# Methods for strongly convex functions

Let $f$ be $\mu$-strongly convex and $L$-smooth

**Polyak's heavy-ball method**

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

**Nesterov's method**

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

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

# They look similar

Let $f$ be $\mu$-strongly convex and $L$-smooth

## Nesterov's method

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

Equivalent to

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

## 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*

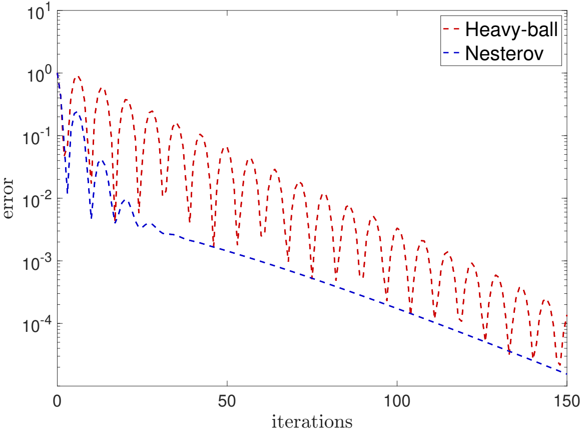
# They have the same ODE

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

- 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

- 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 $\frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}} s\left(\nabla f(x_k) - \nabla f(x_{k-1})\right)$
- Derivation: carefully Taylor expand $\frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}} s\left(\nabla f(x_k) - \nabla f(x_{k-1})\right)$

# High-resolution ODEs

Let $s$ be small but non-vanishing

## High-resolution ODEs

- Polyak's

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

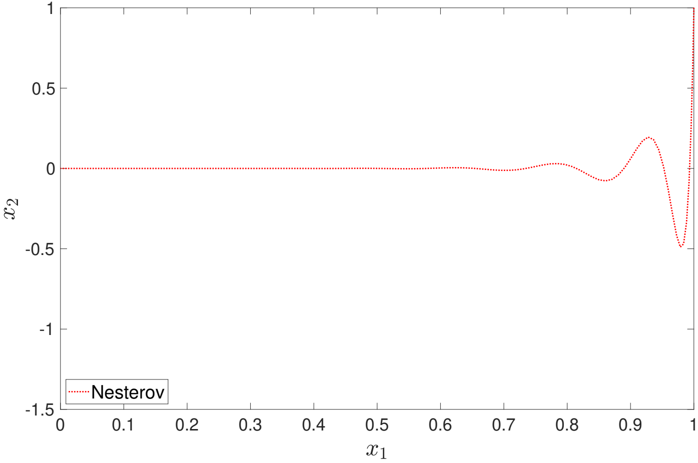
- Nesterov's

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

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

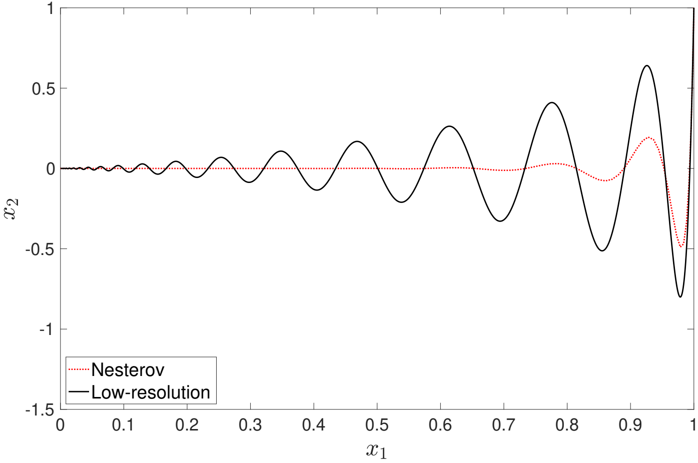
# High-resolution ODEs are better surrogates

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



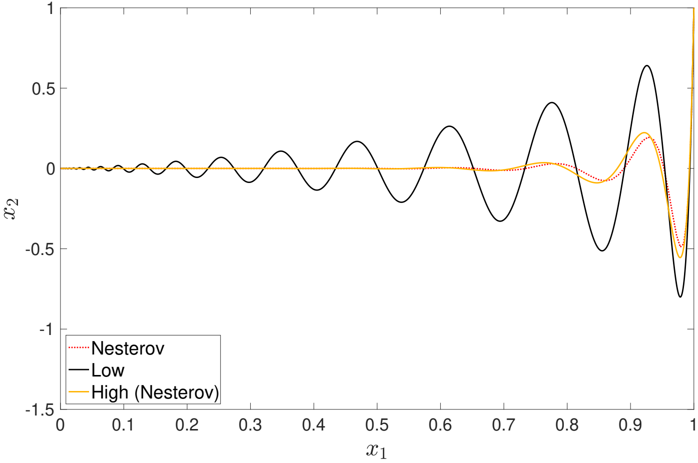
# High-resolution ODEs are better surrogates

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



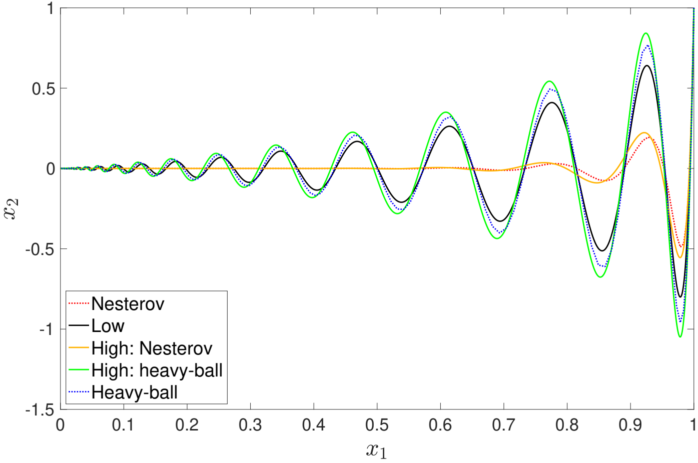
# High-resolution ODEs are better surrogates

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



# 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?*

# The answer is in the *gradient correction*

## 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

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

# The answer is in the *gradient correction*

**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"

# The answer is in the *gradient correction*

**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

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

# 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

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

- $\dot{X} + 2\sqrt{\mu}(X - x^\star) + \sqrt{s}\nabla f(X)$ 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**

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

# Convergence of Nesterov's ODE

Energy functional

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

**Lemma**

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

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

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

# Convergence of Polyak's ODE

Energy functional

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

**Lemma**

$$\frac{\mathrm{d}\mathcal{E}}{\mathrm{d}t} \le -\frac{\sqrt{\mu}}{4}\mathcal{E}$$

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

$$f(X(t)) - f^\star \le \frac{7\,\|x_0 - x^\star\|^2}{2s}\mathrm{e}^{-\frac{\sqrt{\mu}t}{4}}$$

*Returning to the discrete world*

# Discrete energy functional for Nesterov's

## Continuous-time

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

# Discrete energy functional for Nesterov's

## Continuous-time

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

## Discrete-time

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

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

# Discrete energy functional for Nesterov's

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

**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

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

## 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

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

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

# Discrete energy functional for Polyak's

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

## Lemma

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

- Need to ensure the "annoying" term
$\frac{3\sqrt{\mu s}}{4}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\left(f(x_{k+1}) - f^\star\right) - \frac{s}{2}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)^2 \left\|\nabla f(x_{k+1})\right\|^2 \geq 0$

# Where is this "annoying" term from?

The continuous energy functional for Nesterov's

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

# Where is this “annoying” term from?

The continuous energy functional for Nesterov’s

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

- 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)**

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

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

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

# It is the gradient correction that matters

**Nesterov's**

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

**Polyak's**

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

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

# Numerical stability

Forward Euler scheme on Nesterov's

$$\frac{X(t+\sqrt{s})-2X(t)+X(t-\sqrt{s})}{s} + (2\sqrt{\mu} + \sqrt{s}\nabla^2 f(X(t-\sqrt{s}))) \cdot \frac{X(t)-X(t-\sqrt{s})}{\sqrt{s}} + (1+\sqrt{\mu s})\nabla f(X(t-\sqrt{s})) = 0$$

**Stable step sizes for solving Nesterov's**

$$s \leq O\left(\frac{1}{L}\right)$$

# Numerical stability

Forward Euler scheme on Nesterov's

$$\frac{X(t+\sqrt{s})-2X(t)+X(t-\sqrt{s})}{s} + (2\sqrt{\mu} + \sqrt{s}\nabla^2 f(X(t-\sqrt{s}))) \cdot \frac{X(t)-X(t-\sqrt{s})}{\sqrt{s}} + (1+\sqrt{\mu s})\nabla f(X(t-\sqrt{s})) = 0$$

**Stable step sizes for solving Nesterov's**

$$s \leq O\left(\frac{1}{L}\right)$$

Forward Euler scheme on Polyak's

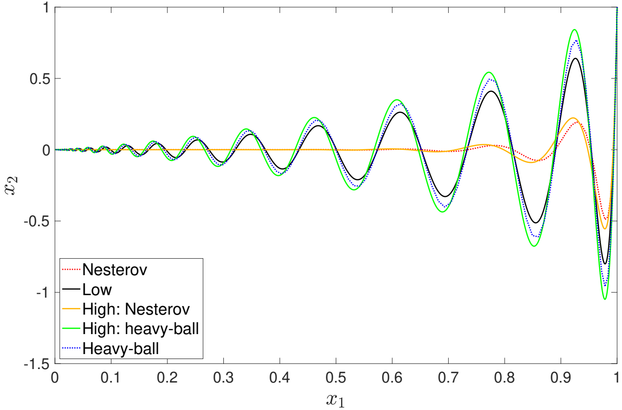
$$\frac{X(t+\sqrt{s})-2X(t)+X(t-\sqrt{s})}{s} + 2\sqrt{\mu}\frac{X(t)-X(t-\sqrt{s})}{\sqrt{s}} + (1+\sqrt{\mu s})\nabla f(X(t-\sqrt{s})) = 0$$

**Stable step sizes for solving Polyak's**

$$s \leq O\left(\frac{\mu}{L^2}\right)$$

# 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

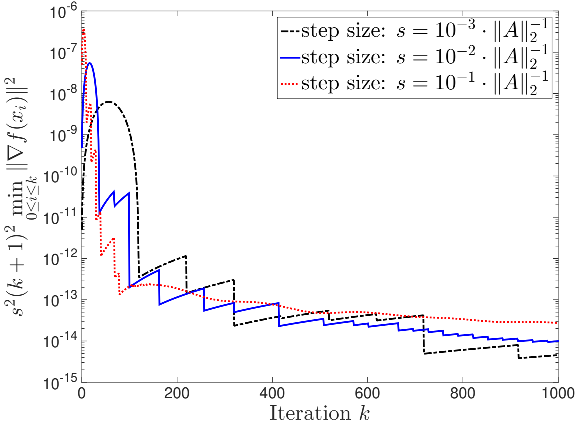
$$\|\nabla f(x_k)\|^2 \le O\left(\frac{1}{k^2}\right)$$

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

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

# Is $O(1/k^2)$ the right rate?

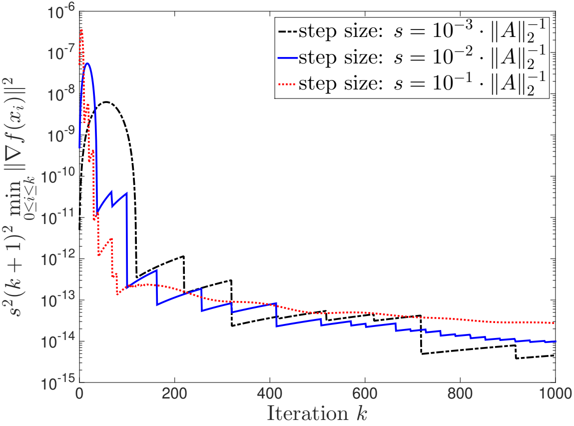
Scaled squared gradient norm $s^2(k+1)^2 \min_{0\le i\le k} \|\nabla f(x_i)\|^2$




$f(x) = \frac{1}{2}\langle Ax, x\rangle + \langle b, x\rangle$, where $A$ is $500 \times 500$

# Is $O(1/k^2)$ the right rate?

Scaled squared gradient norm $s^2(k+1)^2 \min_{0\le i\le k} \|\nabla f(x_i)\|^2$




$f(x) = \frac{1}{2}\langle Ax, x\rangle + \langle b, x\rangle$, where $A$ is $500 \times 500$

- Unfortunately, the low-resolution ODE *cannot* explain

# 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 $\sqrt{s}\nabla^2 f(X)\dot{X}$

# High-resolution energy functional of Nesterov's method

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

**Lemma**

$$\frac{\mathrm{d}\mathcal{E}(t)}{\mathrm{d}t} \le -\left[\sqrt{s}t^2 + \left(\frac{1}{L} + \frac{s}{2}\right)t + \frac{\sqrt{s}}{2L}\right]\|\nabla f(X)\|^2$$

# High-resolution energy functional of Nesterov's method

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

**Lemma**

$$\frac{\mathrm{d}\mathcal{E}(t)}{\mathrm{d}t} \leq -\left[\sqrt{s}t^2 + \left(\frac{1}{L} + \frac{s}{2}\right)t + \frac{\sqrt{s}}{2L}\right]\|\nabla f(X)\|^2$$

- $\|\nabla f(X)\|^2$ arises from gradient correction. Thus does *not* apply to low-resolution ODE

# High-resolution energy functional of Nesterov's method

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

**Lemma**

$$\frac{\mathrm{d}\mathcal{E}(t)}{\mathrm{d}t} \le -\left[\sqrt{s}t^2 + \left(\frac{1}{L} + \frac{s}{2}\right)t + \frac{\sqrt{s}}{2L}\right]\|\nabla f(X)\|^2$$

- $\|\nabla f(X)\|^2$ arises from gradient correction. Thus does *not* apply to low-resolution ODE
- Observe

$$\inf_{t_0 \le u \le t}\|\nabla f(X(u))\|^2 \int_{t_0}^{t}\left[\sqrt{s}u^2 + \left(\frac{1}{L} + \frac{s}{2}\right)u + \frac{\sqrt{s}}{2L}\right]\mathrm{d}u$$
$$\le \int_{t_0}^{t}\left[\sqrt{s}u^2 + \left(\frac{1}{L} + \frac{s}{2}\right)u + \frac{\sqrt{s}}{2L}\right]\|\nabla f(X(u))\|^2\,\mathrm{d}u$$

# An improved rate in continuous case

**Theorem**

*Let $s = 1/L$. The squared gradient norm in the high-resolution ODE satisfies*

$$\inf_{t_0 \le u \le t} \|\nabla f(X(u))\|^2 = O\left(\frac{\sqrt{L}}{t^3}\right)$$

# An improved rate in continuous case

**Theorem**

*Let $s = 1/L$. The squared gradient norm in the high-resolution ODE satisfies*

$$\inf_{t_0 \le u \le t} \|\nabla f(X(u))\|^2 = O\left(\frac{\sqrt{L}}{t^3}\right)$$

- Improved from $O(1/t^2)$ to $O(1/t^3)$

# An improved rate in continuous case

**Theorem**

*Let $s = 1/L$. The squared gradient norm in the high-resolution ODE satisfies*

$$\inf_{t_0 \le u \le t} \|\nabla f(X(u))\|^2 = O\left(\frac{\sqrt{L}}{t^3}\right)$$

- Improved from $O(1/t^2)$ to $O(1/t^3)$
- Possible to extend the result to discrete cases?

# Returning to the discrete world (which we care about)

**Theorem**

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

$$\min_{0 \leq i \leq k} \left\| \nabla f(x_i) \right\|^2 \leq \frac{8568 \left\| x_0 - x^\star \right\|^2}{s^2 (k+1)^3}$$

# Returning to the discrete world (which we care about)

**Theorem**

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

$$\min_{0 \leq i \leq k} \|\nabla f(x_i)\|^2 \leq \frac{8568 \|x_0 - x^\star\|^2}{s^2(k+1)^3}$$

- Improved from $O(1/k^2)$ to $O(1/k^3)$
- Based on the discrete energy functional

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

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**

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

$$\min_{0 \leq i \leq k} \|\nabla f(x_i)\|^2 \leq \frac{8568 \|x_0 - x^\star\|^2}{s^2 (k+1)^3}$$

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

# More comments on the improved rate

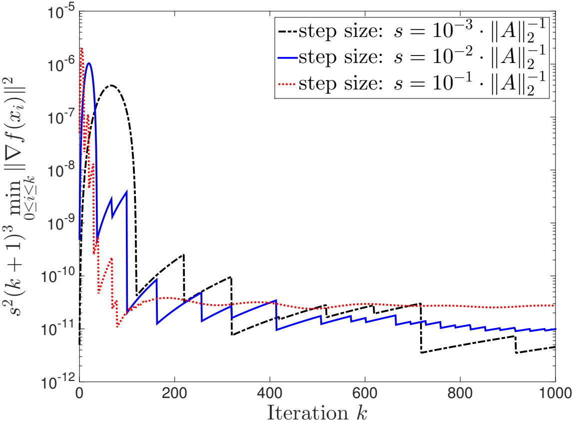
**Theorem**

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

$$\min_{0 \leq i \leq k} \|\nabla f(x_i)\|^2 \leq \frac{8568 \|x_0 - x^\star\|^2}{s^2(k+1)^3}$$

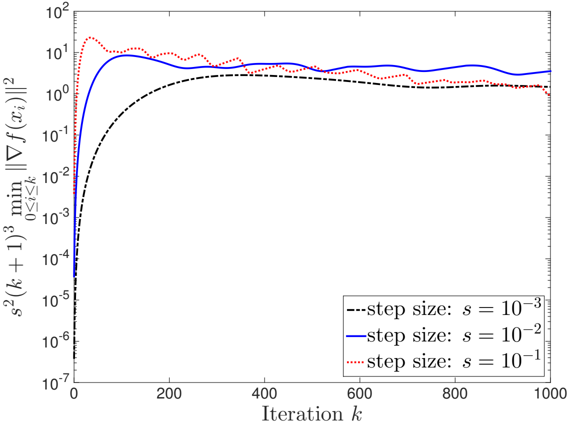
- Previously, the best known bound of Nesterov is $o(1/k^2)$
- Sharpest known bound (without modification)
- Why minimization of gradient is easier?

# Simulations I




Scaled squared gradient norm $s^2(k+1)^3 \min_{0\le i\le k} \|\nabla f(x_i)\|^2$.
$f(x) = \frac{1}{2}\langle Ax, x\rangle + \langle b, x\rangle$, where $A$ is $500 \times 500$

# Simulations II




Scaled squared gradient norm $s^2(k+1)^3 \min_{0\le i\le k} \|\nabla f(x_i)\|^2$.

$f(x) = \rho \log\left\{\sum_{i=1}^{200} \exp\left[\left(\langle a_i, x\rangle - b_i\right)/\rho\right]\right\}$, where $A = [a_1, \ldots, a_{200}]'$ is $200 \times 50$ and $\rho = 20$

*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 $\alpha = 3, \beta = 1$

# 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**

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

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

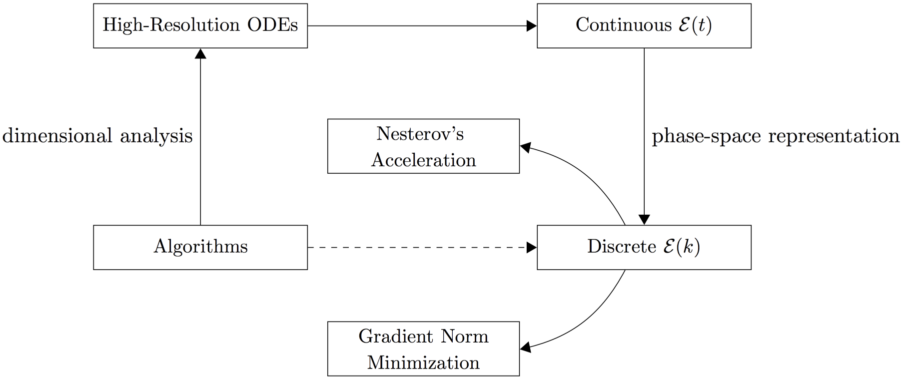
*In addition, if $\alpha > 3$ then*

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

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

# Outline

# A new framework for understanding optimization



# Rambling thoughts

Hermann Weyl

*In these days the angel of topology and the devil of abstract algebra fight for the soul of every individual discipline of mathematics*

# Rambling thoughts

Hermann Weyl

*In these days the angel of topology and the devil of abstract algebra fight for the soul of every individual discipline of mathematics*

- *Algebraic topology*. What is dynamical systems + optimization?

# Rambling thoughts

Hermann Weyl

> *In these days the angel of topology and the devil of abstract algebra fight for the soul of every individual discipline of mathematics*

- *Algebraic topology*. What is dynamical systems + optimization?
- Gaps between discrete and continuous worlds exist

# Rambling thoughts

Hermann Weyl

*In these days the angel of topology and the devil of abstract algebra fight for the soul of every individual discipline of mathematics*

- *Algebraic topology*. What is dynamical systems + optimization?
- Gaps between discrete and continuous worlds exist
- Need more research efforts

# Take home messages

- ODEs are amenable tools for analyzing gradient-based methods

# Take home messages

- ODEs are amenable tools for analyzing gradient-based methods
- Conceptually simple, suggest new methods

# Take home messages

- ODEs are amenable tools for analyzing gradient-based methods
- Conceptually simple, suggest new methods
- Sometimes, need to “upgrade” ODEs

# 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, constrained settings? Many research opportunities

# Thank you!

- *A Differential Equation for Modeling Nesterov's Accelerated Gradient Method: Theory and Insights*
  with Boyd and Candès, Journal of Machine Learning Research, 2016
- *Understanding the Acceleration Phenomenon via High-Resolution Differential Equations*
  with Shi, Du, and Jordan, arXiv, 2018
- *Acceleration via Symplectic Discretization of High-Resolution Differential Equations*
  with Shi, Du, and Jordan, NeurIPS, 2019
- Acknowledgment: Sloan Research Fellowship, NSF CAREER Award, and Wharton Dean's Research Fund