ORF522 – Linear and Nonlinear Optimization

18. Acceleration schemes

Bartolomeo Stellato — Fall 2024

Today's lecture

[Section 2.2, ILCO][Chapter 1, FMO]

First-order methods acceleration

- Lower bounds
- Acceleration and convergence analysis
- Examples

Lower bounds

Sublinear convergence rates

For a convex L-smooth function f we have

Gradient descent

$$x^{k+1} = x^k - t\nabla f(x^k)$$

Proximal gradient

$$x^{k+1} = \mathbf{prox}_{tg}(x^k - t\nabla f(x^k))$$

Convergence

$$f(x^k) - f(x^\star) \leq \frac{\|x^0 - x^\star\|_2^2}{2tk}$$

distance	$O(1/k)$
iterations	$O(1/\epsilon)$

Can we do better? Is there a lower bound?

Lower bounds

First-order methods

Any algorithm that selects

$$x^{k+1} \in x_0 + \mathbf{span}\{\nabla f(x_0), \nabla f(x_1), \ldots, \nabla f(x^k)\}$$

Theorem (Nesterov ’83)

For every integer $k \leq (n-1)/2$, there exist a convex L-smooth function f such that, for any first-order method

$$f(x^k) - f(x^\star) \geq \frac{3L}{32(k+1)^2}\|x^0 - x^\star\|^2 \longrightarrow \begin{array}{ll} \textbf{distance} & O(1/k^2) \\ \textbf{iterations} & O(1/\sqrt{\epsilon}) \end{array}$$

Lower bound proof

minimize $\quad f(x) = \frac{L}{4}\left(\frac{1}{2}x^T A x - e_1^T x\right) \longrightarrow \nabla f(x) = \frac{L}{4}(Ax - e_1)$

Gil. Strang (MIT) "cupcake matrix"

$$A = \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix}, \qquad e_1 = (1, 0, \ldots, 0)$$

- f is convex and L-smooth
- $x^\star$ is the **optimizer** with $x_i^\star = 1 - \frac{i}{n+1}$ (Solves $\nabla f(x^\star) = 0 \to Ax^\star = e_1$)
- $f(x^\star) = \frac{L}{4}\left(\frac{1}{2}e_1^T x^\star - e_1^T x^\star\right) = -\frac{L}{8}x_1^\star = -\frac{L}{8}\frac{n}{n+1}, \qquad \|x^\star\|^2 \leq \frac{n+1}{3}$

Lower bound proof

Iterates

If $x^0 = 0$ then $x^k \in \mathbf{span}\{\nabla f(x^0), \ldots, \nabla f(x^{k-1})\} = \mathbf{span}\{e_1, \ldots, e_k\}$

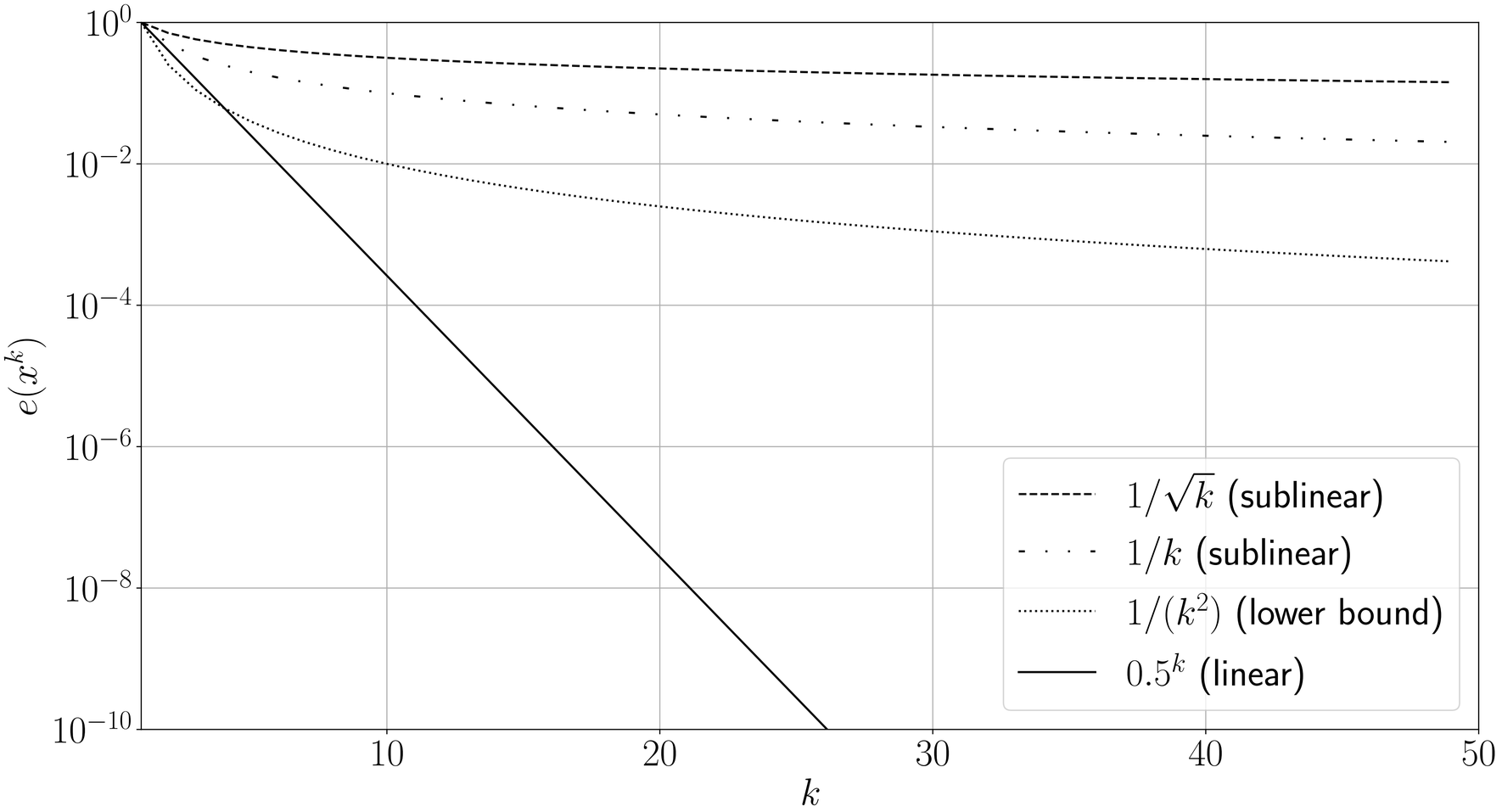
Upper bound

$$f(x^k) \geq \min_{x \in \mathbf{span}\{\nabla f(x^0), \ldots, \nabla f(x^{k-1})\}} f(x) = \min_{x_{k+1} = \cdots = x_n = 0} f(x) = -\frac{L}{8}\frac{k}{k+1}$$

For $k \approx n/2$ or $n = 2k+1$,

$$\frac{f(x^k) - f(x^\star)}{\|x^0 - x^\star\|^2} \geq \frac{L}{8}\left(-\frac{k}{k+1} + \frac{2k+1}{2k+2}\right) / \left(\frac{2k+2}{3}\right) = \frac{3L}{32(k+1)^2}$$

Convergence rates



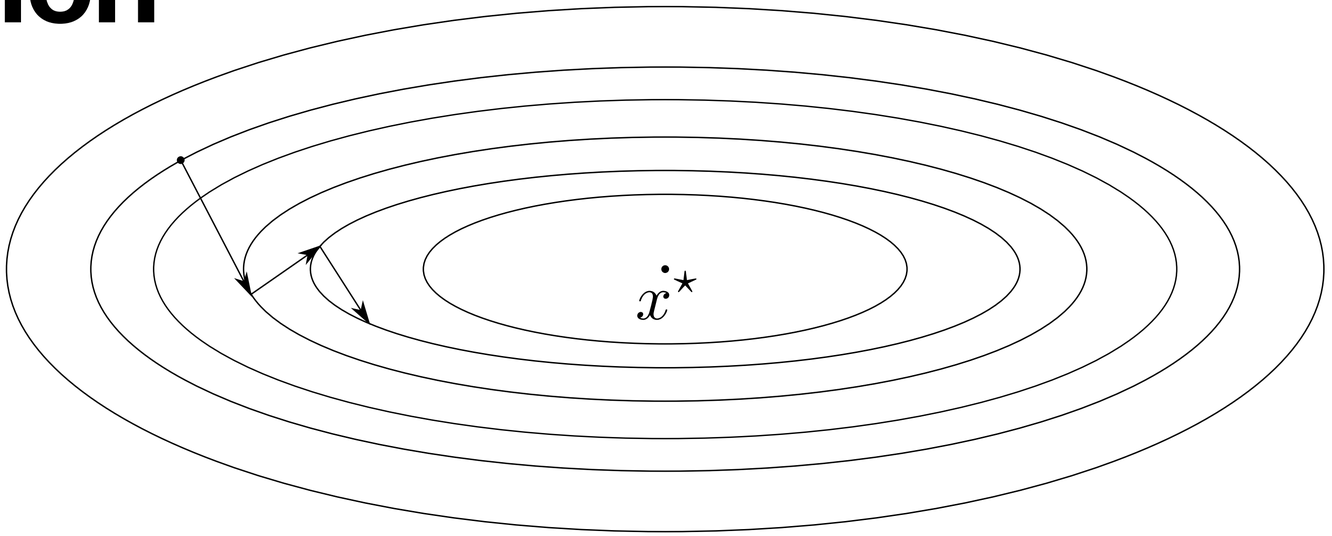
Can we achieve the lower bound?

Acceleration

Heavy ball acceleration

Gradient descent

$$x^{k+1} = x^k - t\nabla f(x^k)$$

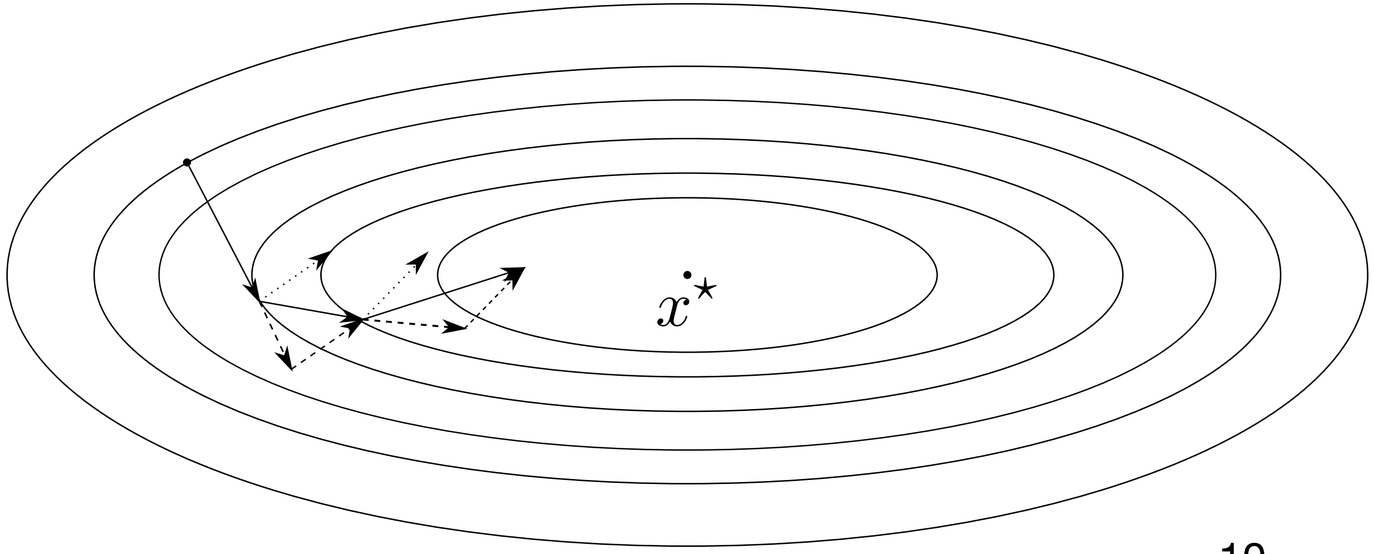


Adding momentum

$$x^{k+1} = x^k - t\nabla f(x^k) + \beta_k(x^k - x^{k-1})$$

momentum

(to mitigate zigzag)



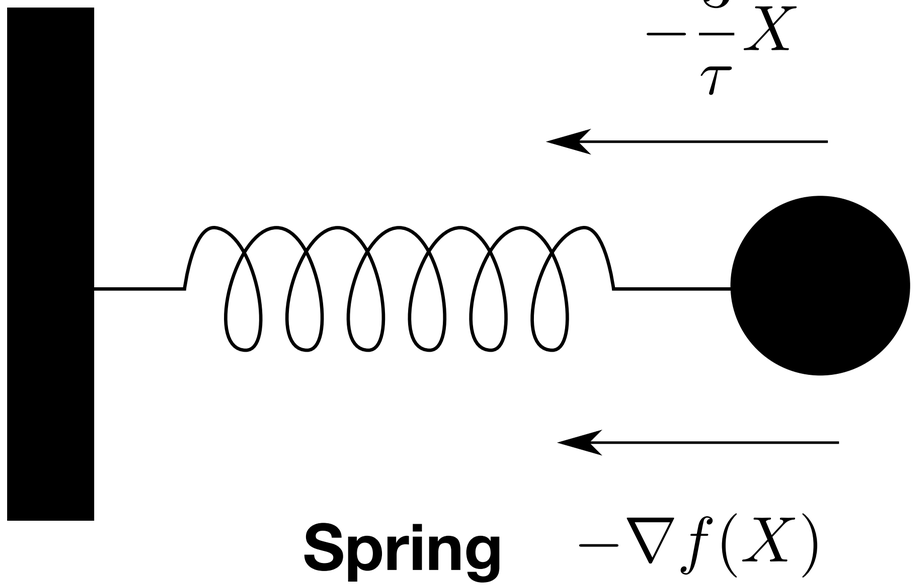
Nesterov momentum

$$x^{k+1} = y^k - t\nabla f(y^k)$$
$$y^{k+1} = x^{k+1} + \frac{k}{k+3}(x^{k+1} - x^k)$$

Properties

- Original Momentum proposed by Nesterov ('83)
- No longer a descent method (i.e., we can have $f(x^{k+1}) > f(x^k)$)
- Same complexity per iteration as gradient descent
- One of the most interesting results in optimization

ODE interpretation



Nesterov acceleration

$$x^{k+1} = y^k - t\nabla f(y^k)$$

$$y^{k+1} = x^{k+1} + \frac{k}{k+3}(x^{k+1} - x^k)$$

$$t \to 0 \quad \Big| \quad x^k \approx X(k\sqrt{t}) = X(\tau)$$

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

damping coefficient

Note: 3 is the smallest constant that guarantees $O(1/\tau^2)$ convergence

[A Differential Equation for Modeling Nesterov's Accelerated Gradient Method: Theory and Insights, Su, Boyd, Candes]

Accelerated proximal gradient method

minimize $f(x) + g(x)$

$f(x)$ convex and smooth
$g(x)$ convex (may be not differentiable)

Iterations

$$x^{k+1} = \mathbf{prox}_{tg}\left(y^k - t\nabla f(y^k)\right)$$

$$y^{k+1} = x^{k+1} + \frac{\lambda_k - 1}{\lambda_{k+1}}(x^{k+1} - x^k)$$

where $y_0 = x_0$,

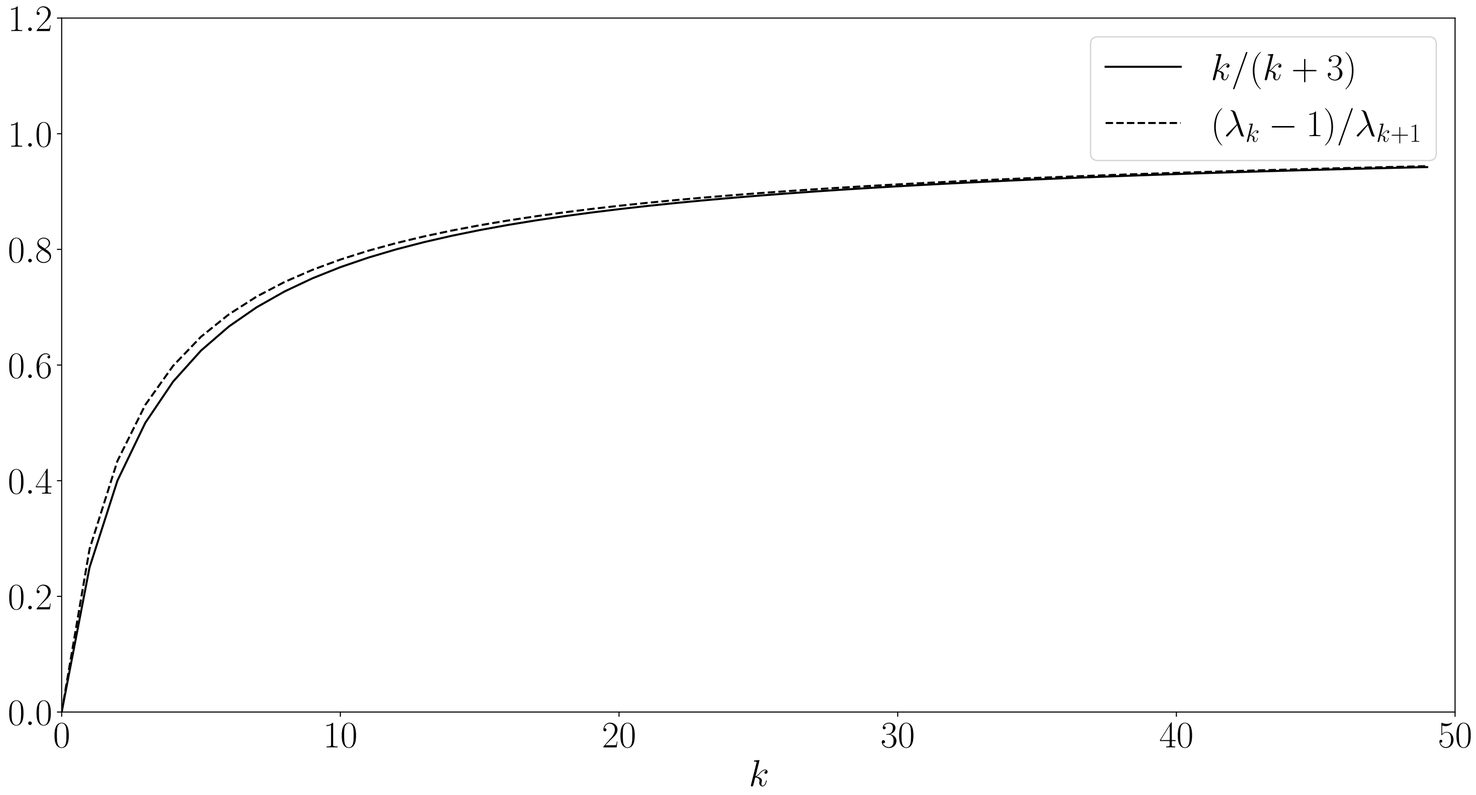
$\lambda_0 = 1$ and $\lambda_{k+1} = \dfrac{1 + \sqrt{1 + 4\lambda_k^2}}{2}$

(same as $\lambda_{k+1}^2 - \lambda_{k+1} - \lambda_k^2 = 0$)

Note: $g(x) = 0$ gives accelerated gradient descent

Proximal gradient and Nesterov weights

$$\lambda_0 = 1 \qquad \lambda_{k+1} = \frac{1 + \sqrt{1 + 4\lambda_k^2}}{2} \longrightarrow \frac{\lambda_k - 1}{\lambda_{k+1}} \approx \frac{k}{k+3} \text{ as } k \to \infty$$



Convergence rate for accelerated proximal gradient method

minimize $F(x) = f(x) + g(x)$

$f(x)$ convex and L-smooth
$g(x)$ convex (may be not differentiable)

Theorem
The accelerated proximal gradient method with step-size $t = 1/L$ satisfies

$$f(x^k) - f(x^\star) \leq \frac{2L\|x^0 - x^\star\|^2}{(k+1)^2}$$

Note (proof as exercise)
For momentum weights $\lambda_0 = 1$ and $\lambda_{k+1} = \dfrac{1 + \sqrt{1 + 4\lambda_k^2}}{2}$, we have

$$\lambda_k \geq \frac{k+2}{2} \quad \forall k \geq 1$$

Convergence rate for accelerated proximal gradient method

minimize $f(x) + g(x)$

$f(x)$ convex and L-smooth
$g(x)$ convex (may be not differentiable)

$$f(x^k) - f(x^\star) \leq \frac{2\|x^0 - x^\star\|^2}{t(k+1)^2}$$

- Better iteration complexity $O(1/k^2)$ (i.e. $O(1/\sqrt{\epsilon})$
- Fast if $\mathbf{prox}$ evaluations are cheap
- Can't do better! (from **lower bound**)

Convergence analysis

Fundamental inequality

From prox grad lecture

$$y^+ = \mathbf{prox}_{tg}\left(y - \frac{1}{L}\nabla f(y)\right)$$

$$= \underset{z}{\text{argmin}}\ \phi(z) = g(z) + f(y) + \nabla f(y^k)^T(z-y) + \frac{L}{2}\|z-y\|_2^2$$

Lemma (fundamental inequality)

Let $y^+ = \mathbf{prox}_{(1/L)g}(y - (1/L)\nabla f(y))$. then

$$F(y^+) - F(x) \leq \frac{L}{2}\|x-y\|_2^2 - \frac{L}{2}\|x-y^+\|_2^2 - h(x,y)$$

with $h(x,y) = f(x) - f(y) - \nabla f(y)^T(x-y) \geq 0$ *(by convexity)*

from previous slide

$$\phi(z) = g(z) + f(y) + \nabla f(y^k)^T(z-y) + \frac{L}{2}\|z-y\|_2^2$$

Proof of Lemma

Since ϕ is L-strongly convex and minimized at y^+, then

$$\phi(x) \geq \phi(y^+) + \frac{L}{2}\|x-y^+\|_2^2 \qquad \textit{(smoothness of } f \textit{ to strong convexity of } \phi\textit{)}$$

From smoothess of f, then

$$\begin{aligned}\phi(y^+) &= f(y) + \nabla f(y^k)^T(y^+-y) + \frac{L}{2}\|y^+-y\|_2^2 + g(y^+) \\ &\geq f(y^+) + g(y^+) = F(y^+)\end{aligned}$$

Together, we have

$$\phi(x) \geq F(y^+) + \frac{L}{2}\|x-y^+\|_2^2 \iff$$

$$f(y) + \nabla f(y)^T(x-y) + g(x) + \frac{L}{2}\|x-y\|_2^2 \geq F(y^+) + \frac{L}{2}\|x-y^+\|_2^2 \iff$$

$$F(x) - h(x,y) + \frac{L}{2}\|x-y\|_2^2 \geq F(y^+) + \frac{L}{2}\|x-y^+\|_2^2$$

■

Convergence rate accelerated proximal gradient method

Proof

1. Construct Lyapunov function

2. Lyapunov function is non-increasing when Nesterov's coefficients are used!

Constructing Lyapunov function

Lemma (Lyapunov function decrease)

Let $u^k = \lambda_{k-1} x^k - (x^\star + (\lambda_{k-1} - 1) x^{k-1}) = \lambda_{k-1}(x^k - x^\star) - (\lambda_{k-1} - 1)(x^{k-1} - x^\star)$

then

$$\|u^{k+1}\|_2^2 + \frac{2}{L}\lambda_k^2 (F(x^{k+1}) - F^\star) \leq \|u^k\|_2^2 + \frac{2}{L}\lambda_{k-1}^2 (F(x^k) - F^\star)$$

Proof of Lyapunov function decrease

Let $x = \frac{1}{\lambda_k} x^\star + \left(1 - \frac{1}{\lambda_k}\right) x^k$ and $y = y^k$. From *funamental inequality*,

$$
\begin{aligned}
F(x^{k+1}) - F(x) &\leq \frac{L}{2}\|x - y^k\|_2^2 - \frac{L}{2}\|x - x^{k+1}\|_2^2 \\
&\leq \frac{L}{2}\|\lambda_k^{-1} x^\star + (1 - \lambda_k^{-1})x^k - y^k\|_2^2 - \frac{L}{2}\|\lambda_k^{-1} x^\star + (1 - \lambda_k^{-1})x^k - x^{k+1}\|_2^2 \\
&= \frac{L}{2\lambda_k^2}\|x^\star + (\lambda_k - 1)x^k - \lambda_k y^k\|_2^2 - \frac{L}{2\lambda_k^2}\|x^\star + (\lambda_k - 1)x^k - \lambda_k x^{k+1}\|_2^2 \\
&= \frac{L}{2\lambda_k^2}(\|u^k\|_2^2 - \|u^{k+1}\|_2^2)
\end{aligned}
$$

last equality follows from

$$
\begin{aligned}
u^k &= \lambda_{k-1} x^k - (x^\star + (\lambda_{k-1} - 1)x^{k-1}) \\
y^k &= x^k + \frac{\lambda_{k-1} - 1}{\lambda_k}(x^k - x^{k-1})
\end{aligned}
$$

Proof of Lyapunov function decrease (cont.)

We can also lower bound $F(x^{k+1}) - F(x)$

By convexity of F

$$F(\lambda_k^{-1}x^\star + (1-\lambda_k^{-1})x^k) \leq \lambda_k^{-1}F(x^\star) + (1-\lambda_k^{-1})F(x^k)$$
$$= \lambda_k^{-1}F^\star + (1-\lambda_k^{-1})F(x^k)$$

Therefore

$$F(x^{k+1}) - F(x) = F(x^{k+1}) - F(\lambda_k^{-1}x^\star + (1-\lambda_k^{-1})x^k)$$
$$\geq (1-\lambda_k^{-1})(F(x^k) - F^\star) - (F(x^{k+1}) - F^\star)$$

By combining lower and upper bounds on $F(x^{k+1}) - F(x)$ and $\lambda_k^2 - \lambda_k = \lambda_{k-1}^2$,

$$\frac{L}{2}(\|u^k\|_2^2 - \|u^{k+1}\|_2^2) \geq \lambda_k^2(F(x^{k+1}) - F^\star) + (\lambda_k^2 - \lambda_k)(F(x^k) - F^\star)$$
$$= \lambda_k^2(F(x^{k+1}) - F^\star) + \lambda_{k-1}^2(F(x^k) - F^\star)$$

Convergence rate accelerated proximal gradient method

Proof

1. Construct Lyapunov function 👍

2. Lyapunov function is non-increasing when Nesterov's coefficients are used!

Convergence rate accelerated proximal gradient method

Proof

From Lyapunov function Lemma between iterations k and 0 (noting $\|u^k\| \geq 0$)

$$\frac{2}{L}\lambda_{k-1}^2(F(x^k) - F^\star) \leq \|u^1\|_2^2 + \frac{2}{L}\lambda_0^2(F(x^1) - F^\star) = \|x^1 - x^\star\|_2^2 + \frac{2}{L}(F(x^1) - F^\star)$$

Because of the fundamental inequality lemma with $y^+ = x^1, y = x^0, x = x^\star$,

$$\frac{2}{L}(F(x^1) - F^\star) \leq \|x^0 - x^\star\|_2^2 - \|x^1 - x^\star\|_2^2 \quad \iff \quad \|x^1 - x^\star\|_2^2 + \frac{2}{L}(F(x^1) - F^\star) \leq \|x^0 - x^\star\|_2^2$$

Therefore

$$\frac{2}{L}\lambda_{k-1}^2(F(x^k) - F^\star) \leq \|x^0 - x^\star\|_2^2$$

$$\Rightarrow \quad F(x^k) - F^\star \leq \frac{L\|x^0 - x^\star\|_2^2}{2\lambda_{k-1}^2} \leq \frac{L\|x^0 - x^\star\|_2^2}{2(k+1)^2}$$

(since $\lambda_k \geq \frac{k+2}{2}$)

■

Example

Example: Lasso without linear convergence

$$\text{minimize} \quad \underbrace{(1/2)\|Ax-b\|_2^2}_{f(x)} + \underbrace{\gamma\|x\|_1}_{g(x)}$$

Proximal gradient descent
(Iterative Shrinkage Thresholding Algorithm)

$$x^{k+1} = S_{\gamma t}\left(x^k - tA^T(Ax^k - b)\right)$$

ISTA

Accelerated proximal gradient descent
(Fast Iterative Shrinkage Thresholding Algorithm)

$$x^{k+1} = S_{\gamma t}\left(y^k - tA^T(Ay^k - b)\right)$$

$$y^{k+1} = x^{k+1} + \frac{\lambda_k - 1}{\lambda_{k+1}}(x^{k+1} - x^k)$$

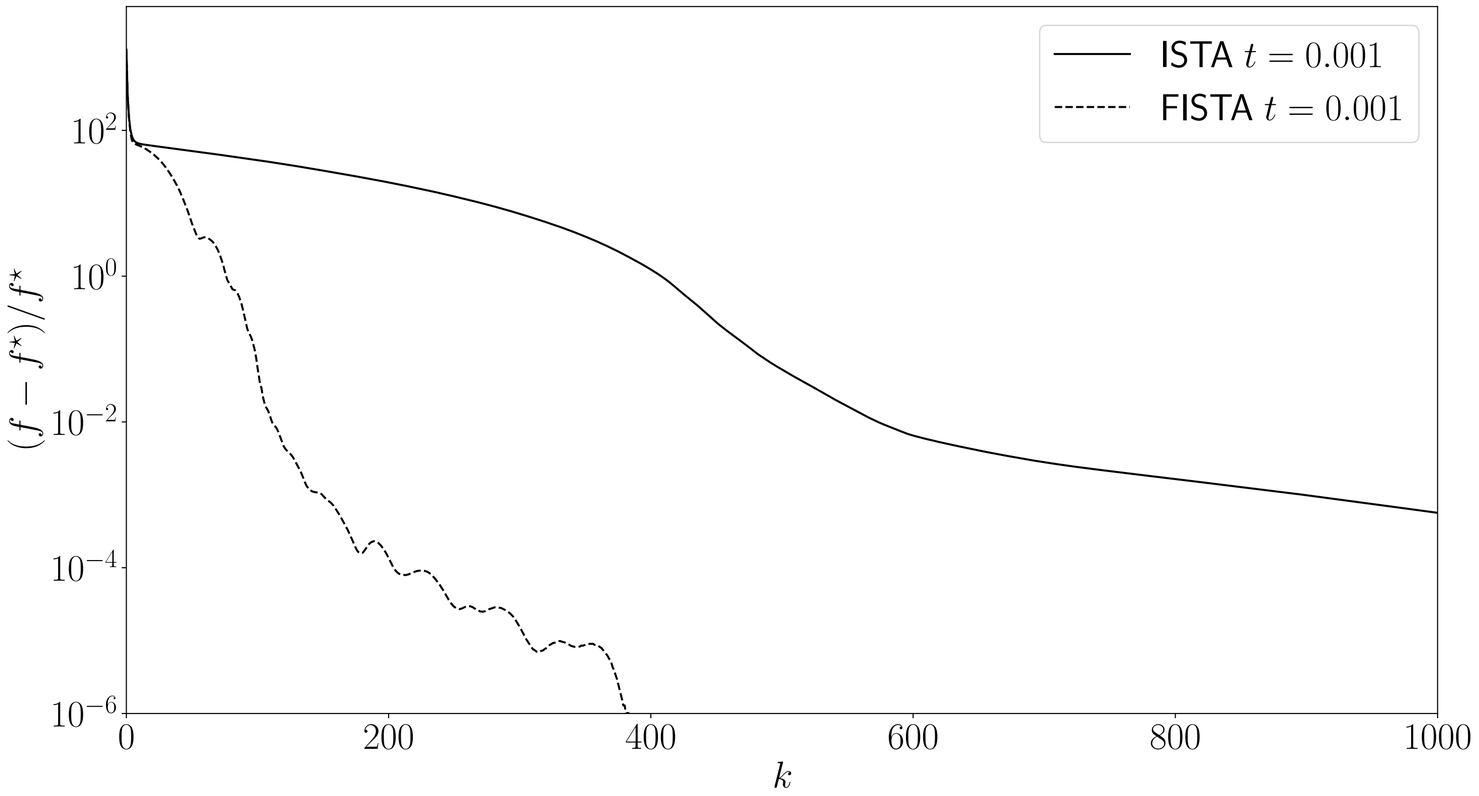
FISTA

[A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems, Beck, Teboulle]

Example: Lasso without linear convergence

Fast Iterative Soft Thresholding Algorithm (FISTA)

minimize $(1/2)\|Ax-b\|_2^2 + \gamma\|x\|_1$



Example

randomly generated $A \in \mathbf{R}^{300\times 500}$

$\Rightarrow \quad \nabla^2 f = A^T A \succeq 0$

$\Rightarrow \quad f$ not strongly convex

FISTA is much faster

Typical rippling behavior
(not a descent method)

Image deblurring

minimize $(1/2)\|Ax - b\|_2^2 + \gamma\|x\|_1$

x: reconstructed image
in wavelet basis (sparse)

original

blurred

	$k = 100$	$k = 200$
ISTA		
FISTA		

[A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems, Beck, Teboulle]

More sophisticated accelerations

Other algorithms

Acceleration can also be applied also to ADMM

[Fast Alternating Direction Optimization Methods, Goldstein, O'Donoghue, Setzer, Baraniuk]

Momentum with **restarts**
(reset momentum when it makes small progress)

Improved convergence rate

$O(1/k^2)$

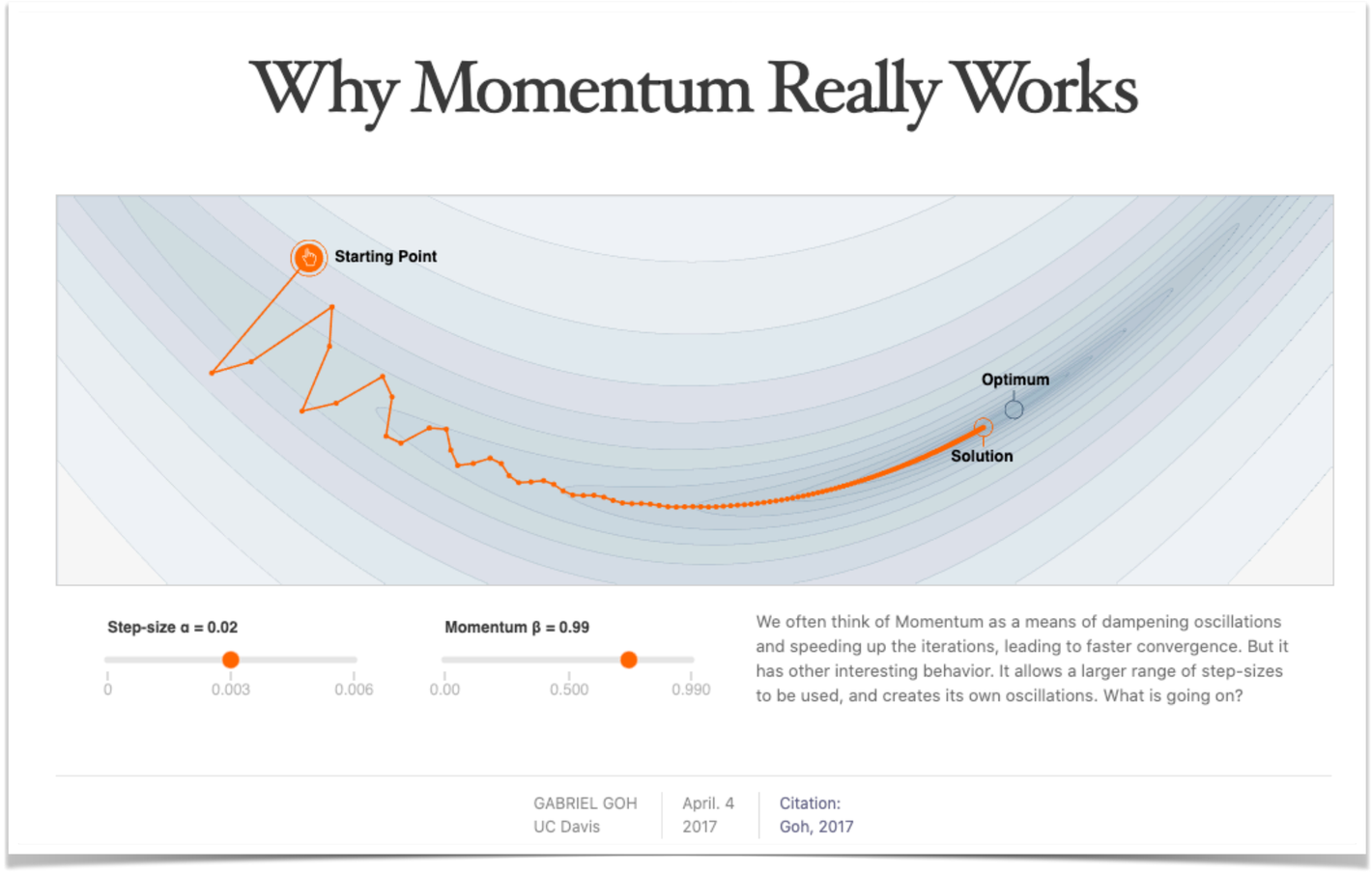
Nonlinear acceleration

(e.g., Anderson Acceleration)

Adaptively pick weights by solving a small optimization problem (usually least-squares)

[Acceleration Methods, d'Aspremont, Scieur, Taylor]

Momentum intuition and much more



All deep learning optimization algorithms
are based on
Momentum/Acceleration:

RMSprop, AdaGrad, Adam, etc.

https://distill.pub/2017/momentum/

Acceleration in nonlinear optimization

Today, we learned to:

- **Derive** lower bounds on cost optimality for first-order methods
- **Accelerate** first-order algorithms by adding momentum term
- **Apply** acceleration schemes to get the best possible convergence

Next lecture

- Computed-assisted proofs