Accelerated gradient methods

Yuxin Chen

Princeton University, Fall 2019

Outline

- Heavy-ball methods
- Nesterov's accelerated gradient methods
- Accelerated proximal gradient methods (FISTA)
- Convergence analysis
- Lower bounds

(Proximal) gradient methods

Iteration complexities of (proximal) gradient methods

- strongly convex and smooth problems

$$O\left(\kappa \log \frac{1}{\varepsilon}\right)$$

- convex and smooth problems

$$O\left(\frac{1}{\varepsilon}\right)$$

Can one still hope to further accelerate convergence?

Issues and possible solutions

Issues:

- GD focuses on improving the cost per iteration, which might sometimes be too "short-sighted"
- GD might sometimes zigzag or experience abrupt changes

Solutions:

- exploit information from the history (i.e. past iterates)
- add buffers (like momentum) to yield smoother trajectory

Heavy-ball methods

— *Polyak '64*

Heavy-ball method

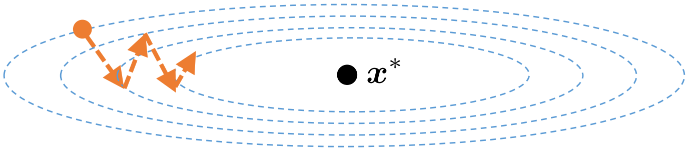
B. Polyak

$$\text{minimize}_{\boldsymbol{x} \in \mathbb{R}^n} \quad f(\boldsymbol{x})$$

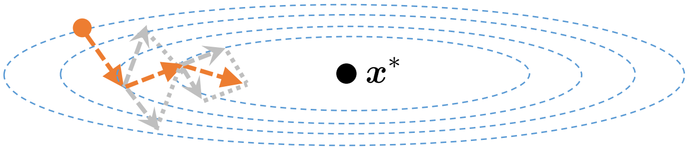
$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta_t \nabla f(\boldsymbol{x}^t) + \underbrace{\theta_t(\boldsymbol{x}^t - \boldsymbol{x}^{t-1})}_{\text{momentum term}}$$

- add inertia to the "ball" (i.e. include a momentum term) to mitigate zigzagging

Heavy-ball method



gradient descent



heavy-ball method

State-space models

$$\text{minimize}_{\boldsymbol{x}} \quad \frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}^*)^\top \boldsymbol{Q}(\boldsymbol{x}-\boldsymbol{x}^*)$$

where $\boldsymbol{Q} \succ \mathbf{0}$ has a condition number κ

One can understand heavy-ball methods through dynamical systems

State-space models

Consider the following dynamical system

$$\begin{bmatrix} \boldsymbol{x}^{t+1} \\ \boldsymbol{x}^{t} \end{bmatrix} = \begin{bmatrix} (1+\theta_t)\boldsymbol{I} & -\theta_t \boldsymbol{I} \\ \boldsymbol{I} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{x}^{t} \\ \boldsymbol{x}^{t-1} \end{bmatrix} - \begin{bmatrix} \eta_t \nabla f(\boldsymbol{x}^t) \\ \boldsymbol{0} \end{bmatrix}$$

or equivalently,

$$\begin{aligned} \underbrace{\begin{bmatrix} \boldsymbol{x}^{t+1} - \boldsymbol{x}^* \\ \boldsymbol{x}^{t} - \boldsymbol{x}^* \end{bmatrix}}_{\text{state}} &= \begin{bmatrix} (1+\theta_t)\boldsymbol{I} & -\theta_t \boldsymbol{I} \\ \boldsymbol{I} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{x}^{t} - \boldsymbol{x}^* \\ \boldsymbol{x}^{t-1} - \boldsymbol{x}^* \end{bmatrix} - \begin{bmatrix} \eta_t \nabla f(\boldsymbol{x}^t) \\ \boldsymbol{0} \end{bmatrix} \\ &= \underbrace{\begin{bmatrix} (1+\theta_t)\boldsymbol{I} - \eta_t \boldsymbol{Q} & -\theta_t \boldsymbol{I} \\ \boldsymbol{I} & \boldsymbol{0} \end{bmatrix}}_{\text{system matrix}} \begin{bmatrix} \boldsymbol{x}^{t} - \boldsymbol{x}^* \\ \boldsymbol{x}^{t-1} - \boldsymbol{x}^* \end{bmatrix} \end{aligned}$$

System matrix

$$\begin{bmatrix} \boldsymbol{x}^{t+1} - \boldsymbol{x}^* \\ \boldsymbol{x}^t - \boldsymbol{x}^* \end{bmatrix} = \underbrace{\begin{bmatrix} \big(1+\theta_t\big)\boldsymbol{I} - \eta_t \boldsymbol{Q} & -\theta_t \boldsymbol{I} \\ \boldsymbol{I} & \boldsymbol{0} \end{bmatrix}}_{=:\boldsymbol{H}_t \text{ (system matrix)}} \begin{bmatrix} \boldsymbol{x}^t - \boldsymbol{x}^* \\ \boldsymbol{x}^{t-1} - \boldsymbol{x}^* \end{bmatrix} \tag{7.1}$$

implication: convergence of heavy-ball methods depends on the spectrum of the system matrix $\boldsymbol{H}_t$

key idea: find appropriate stepsizes η_t and momentum coefficients θ_t to control the spectrum of $\boldsymbol{H}_t$

Convergence for quadratic problems

Theorem 7.1 (Convergence of heavy-ball methods for quadratic functions)

Suppose f is a L-smooth and μ-strongly convex quadratic function. Set $\eta_t \equiv 4/(\sqrt{L}+\sqrt{\mu})^2$, $\theta_t \equiv \max\left\{|1-\sqrt{\eta_t L}|, |1-\sqrt{\eta_t \mu}|\right\}^2$, and $\kappa = L/\mu$. Then

$$\left\| \begin{bmatrix} \boldsymbol{x}^{t+1} - \boldsymbol{x}^* \\ \boldsymbol{x}^t - \boldsymbol{x}^* \end{bmatrix} \right\|_2 \lesssim \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} \right)^t \left\| \begin{bmatrix} \boldsymbol{x}^{1} - \boldsymbol{x}^* \\ \boldsymbol{x}^0 - \boldsymbol{x}^* \end{bmatrix} \right\|_2$$

- iteration complexity: $O(\sqrt{\kappa}\log\frac{1}{\varepsilon})$
- significant improvement over GD: $O(\sqrt{\kappa}\log\frac{1}{\varepsilon})$ vs. $O(\kappa\log\frac{1}{\varepsilon})$
- relies on knowledge of both L and μ

Proof of Theorem 7.1

In view of (7.1), it suffices to control the spectrum of $\boldsymbol{H}_t$ (which is time-invariant). Let λ_i be the ith eigenvalue of $\boldsymbol{Q}$ and set $\boldsymbol{\Lambda} := \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}$, then the spectral radius (denoted by $\rho(\cdot)$) of $\boldsymbol{H}_t$ obeys

$$\begin{aligned} \rho(\boldsymbol{H}_t) &= \rho\left(\begin{bmatrix} (1+\theta_t)\boldsymbol{I} - \eta_t \boldsymbol{\Lambda} & -\theta_t \boldsymbol{I} \\ \boldsymbol{I} & \boldsymbol{0} \end{bmatrix}\right) \\ &\leq \max_{1 \leq i \leq n} \rho\left(\begin{bmatrix} 1+\theta_t - \eta_t \lambda_i & -\theta_t \\ 1 & 0 \end{bmatrix}\right) \end{aligned}$$

To finish the proof, it suffices to show

$$\max_i \rho\left(\begin{bmatrix} 1+\theta_t - \eta_t \lambda_i & -\theta_t \\ 1 & 0 \end{bmatrix}\right) \leq \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} \tag{7.2}$$

Proof of Theorem 7.1

To show (7.2), note that the two eigenvalues of $\begin{bmatrix} 1+\theta_t-\eta_t\lambda_i & -\theta_t \\ 1 & 0 \end{bmatrix}$ are the roots of

$$z^2-(1+\theta_t-\eta_t\lambda_i)z+\theta_t=0 \tag{7.3}$$

If $(1+\theta_t-\eta_t\lambda_i)^2\leq 4\theta_t$, then the roots of this equation have the same magnitudes $\sqrt{\theta_t}$ (as they are either both imaginary or there is only one root).

In addition, one can easily check that $(1+\theta_t-\eta_t\lambda_i)^2\leq 4\theta_t$ is satisfied if

$$\theta_t\in\left[\left(1-\sqrt{\eta_t\lambda_i}\right)^2,\left(1+\sqrt{\eta_t\lambda_i}\right)^2\right], \tag{7.4}$$

which would hold if one picks $\theta_t=\max\left\{\left(1-\sqrt{\eta_t L}\right)^2,\left(1-\sqrt{\eta_t\mu}\right)^2\right\}$

Proof of Theorem 7.1

With this choice of θ_t, we have (from (7.3) and the fact that two eigenvalues have identical magnitudes)

$$\rho(\boldsymbol{H}_t) \leq \sqrt{\theta_t}$$

Finally, setting $\eta_t = \frac{4}{(\sqrt{L}+\sqrt{\mu})^2}$ ensures $1 - \sqrt{\eta_t L} = -(1 - \sqrt{\eta_t \mu})$, which yields

$$\theta_t = \max\left\{ \left(1 - \frac{2\sqrt{L}}{\sqrt{L}+\sqrt{\mu}}\right)^2, \left(1 - \frac{2\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^2 \right\} = \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^2$$

This in turn establishes

$$\rho(\boldsymbol{H}_t) \leq \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$$

Nesterov's accelerated gradient methods

Convex case

$$\text{minimize}_{\boldsymbol{x}\in\mathbb{R}^n} \quad f(\boldsymbol{x})$$

For a positive definite quadratic function f, including momentum terms allows to improve the iteration complexity from $O\big(\kappa \log \frac{1}{\varepsilon}\big)$ to $O\big(\sqrt{\kappa} \log \frac{1}{\varepsilon}\big)$

Can we obtain improvement for more general convex cases as well?

Nesterov's idea

Y. Nesterov

— *Nesterov '83*

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

- alternates between gradient updates and *proper* extrapolation
- each iteration takes nearly the same cost as GD
- not a descent method (i.e. we may not have $f(\boldsymbol{x}^{t+1}) \le f(\boldsymbol{x}^t)$)
- one of the most *beautiful* and *mysterious* results in optimization

  ...

Convergence of Nesterov's accelerated gradient method

Suppose f is convex and L-smooth. If $\eta_t \equiv \eta = 1/L$, then

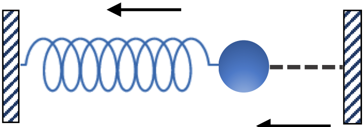
$$f(\boldsymbol{x}^t) - f^{\mathsf{opt}} \leq \frac{2L\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2}{(t+1)^2}$$

- iteration complexity: $O(\frac{1}{\sqrt{\varepsilon}})$
- much faster than gradient methods
- we'll provide proof for the (more general) proximal version later

Interpretation using differential equations

Nesterov's momentum coefficient $\frac{t}{t+3} = 1 - \frac{3}{t}$ is particularly mysterious

Interpretation using differential equations



To develop insight into why Nesterov's method works so well, it's helpful to look at its continuous limits ($\eta_t \to 0$), which is given by second-order ordinary differential equations (ODE)

$$\ddot{\boldsymbol{X}}(\tau) + \underbrace{3/\tau}_{\text{dampling coefficient}} \dot{\boldsymbol{X}}(\tau) + \nabla \underbrace{f(\boldsymbol{X}(\tau))}_{\text{potential}} = \boldsymbol{0}$$

— *Su, Boyd, Candes '14*

Heuristic derivation of ODE

To begin with, Nesterov's update rule is equivalent to

$$\frac{\boldsymbol{x}^{t+1}-\boldsymbol{x}^{t}}{\sqrt{\eta}}=\frac{t-1}{t+2}\frac{\boldsymbol{x}^{t}-\boldsymbol{x}^{t-1}}{\sqrt{\eta}}-\sqrt{\eta}\nabla f(\boldsymbol{y}^{t}) \tag{7.5}$$

Let $t=\frac{\tau}{\sqrt{\eta}}$. Set $\boldsymbol{X}(\tau)\approx\boldsymbol{x}^{\tau/\sqrt{\eta}}=\boldsymbol{x}^{t}$ and $\boldsymbol{X}(\tau+\sqrt{\eta})\approx\boldsymbol{x}^{t+1}$. Then the Taylor expansion gives

$$\frac{\boldsymbol{x}^{t+1}-\boldsymbol{x}^{t}}{\sqrt{\eta}}\approx\dot{\boldsymbol{X}}(\tau)+\frac{1}{2}\ddot{\boldsymbol{X}}(\tau)\sqrt{\eta}$$
$$\frac{\boldsymbol{x}^{t}-\boldsymbol{x}^{t-1}}{\sqrt{\eta}}\approx\dot{\boldsymbol{X}}(\tau)-\frac{1}{2}\ddot{\boldsymbol{X}}(\tau)\sqrt{\eta}$$

which combined with (7.5) yields

$$\dot{\boldsymbol{X}}(\tau)+\frac{1}{2}\ddot{\boldsymbol{X}}(\tau)\sqrt{\eta}\approx\left(1-\frac{3\sqrt{\eta}}{\tau}\right)\left(\dot{\boldsymbol{X}}(\tau)-\frac{1}{2}\ddot{\boldsymbol{X}}(\tau)\sqrt{\eta}\right)-\sqrt{\eta}\nabla f\big(\boldsymbol{X}(\tau)\big)$$

$$\Longrightarrow\quad\ddot{\boldsymbol{X}}(\tau)+\frac{3}{\tau}\dot{\boldsymbol{X}}(\tau)+\nabla f\big(\boldsymbol{X}(\tau)\big)\approx\boldsymbol{0}$$

Convergence rate of ODE

$$\ddot{\boldsymbol{X}} + \frac{3}{\tau}\dot{\boldsymbol{X}} + \nabla f(\boldsymbol{X}) = \mathbf{0} \tag{7.6}$$

Standard ODE theory reveals that

$$f(\boldsymbol{X}(\tau)) - f^{\mathsf{opt}} \leq O\left(\frac{1}{\tau^2}\right) \tag{7.7}$$

which somehow explains Nesterov's $O(1/t^2)$ convergence

Proof of (7.7)

Define $\underbrace{\mathcal{E}(\tau) := \tau^2\big(f(\boldsymbol{X}) - f^{\mathsf{opt}}\big) + 2\big\|\boldsymbol{X} + \frac{\tau}{2}\dot{\boldsymbol{X}} - \boldsymbol{X}^*\big\|_2^2}_{\text{Lyapunov function / energy function}}$. This obeys

$$\dot{\mathcal{E}} = 2\tau\big(f(\boldsymbol{X}) - f^{\mathsf{opt}}\big) + \tau^2\big\langle \nabla f(\boldsymbol{X}), \dot{\boldsymbol{X}}\big\rangle + 4\Big\langle \boldsymbol{X} + \frac{\tau}{2}\dot{\boldsymbol{X}} - \boldsymbol{X}^*, \frac{3}{2}\dot{\boldsymbol{X}} + \frac{\tau}{2}\ddot{\boldsymbol{X}}\Big\rangle$$

$$\overset{\text{(i)}}{=} 2\tau\big(f(\boldsymbol{X}) - f^{\mathsf{opt}}\big) - 2\tau\big\langle \boldsymbol{X} - \boldsymbol{X}^*, \nabla f\big(\boldsymbol{X}\big)\big\rangle \overset{\text{(by convexity)}}{\leq} 0$$

where (i) follows by replacing $\tau\ddot{\boldsymbol{X}} + 3\dot{\boldsymbol{X}}$ with $-\tau\nabla f\big(\boldsymbol{X}\big)$

This means $\mathcal{E}$ is non-decreasing in τ, and hence

$$f(\boldsymbol{X}(\tau)) - f^{\mathsf{opt}} \overset{\text{(defn of } \mathcal{E})}{\leq} \frac{\mathcal{E}(\tau)}{\tau^2} \leq \frac{\mathcal{E}(0)}{\tau^2} = O\left(\frac{1}{\tau^2}\right)$$

Magic number 3

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

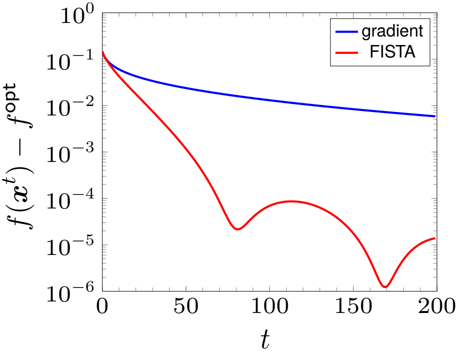
- 3 is the smallest constant that guarantees $O(1/\tau^2)$ convergence, and can be replaced by any other $\alpha \geq 3$
- in some sense, 3 minimizes the pre-constant in the convergence bound $O(1/\tau^2)$ (see Su, Boyd, Candes '14)

Numerical example

taken from UCLA EE236C

$$\text{minimize}_{\boldsymbol{x}} \quad \log\left(\sum_{i=1}^{m} \exp(\boldsymbol{a}_i^\top \boldsymbol{x} + b_i)\right)$$

with randomly generated problems and $m = 2000$, $n = 1000$



Extension to composite models

$$\begin{array}{rl} \text{minimize}_{\boldsymbol{x}} & F(\boldsymbol{x}) := f(\boldsymbol{x}) + h(\boldsymbol{x}) \\ \text{subject to} & \boldsymbol{x} \in \mathbb{R}^n \end{array}$$

- f: convex and smooth
- h: convex (may not be differentiable)

let $F^{\mathsf{opt}} := \min_{\boldsymbol{x}} F(\boldsymbol{x})$ be the optimal cost

FISTA (Beck & Teboulle '09)

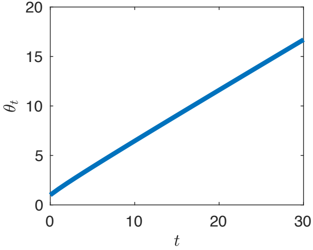
Fast iterative shrinkage-thresholding algorithm

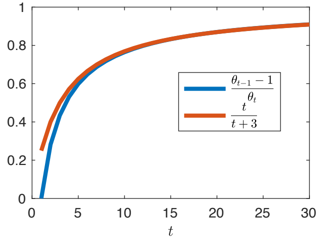
$$\boldsymbol{x}^{t+1} = \mathsf{prox}_{\eta_t h}(\boldsymbol{y}^t - \eta_t \nabla f(\boldsymbol{y}^t))$$
$$\boldsymbol{y}^{t+1} = \boldsymbol{x}^{t+1} + \frac{\theta_t - 1}{\theta_{t+1}}(\boldsymbol{x}^{t+1} - \boldsymbol{x}^t)$$

where $\boldsymbol{y}^0 = \boldsymbol{x}^0$, $\theta_0 = 1$ and $\theta_{t+1} = \frac{1+\sqrt{1+4\theta_t^2}}{2}$

- adopt the momentum coefficients originally proposed by Nesterov '83

Momentum coefficient



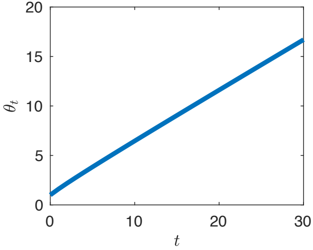


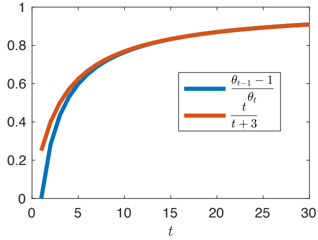
$$\theta_{t+1} = \frac{1+\sqrt{1+4\theta_t^2}}{2} \qquad \text{with } \theta_0 = 1$$

coefficient $\frac{\theta_t - 1}{\theta_{t+1}} = 1 - \frac{3}{t} + o\left(\frac{1}{t}\right)$ (homework)

- asymptotically equivalent to $\frac{t}{t+3}$

Momentum coefficient





$$\theta_{t+1} = \frac{1+\sqrt{1+4\theta_t^2}}{2} \qquad \text{with } \theta_0 = 1$$

Fact 7.2

For all $t \geq 1$, one has $\theta_t \geq \frac{t+2}{2}$ (homework)

Convergence analysis

Convergence for convex problems

Theorem 7.3 (Convergence of accelerated proximal gradient methods for convex problems)

Suppose f is convex and L-smooth. If $\eta_t \equiv 1/L$, then

$$F(\boldsymbol{x}^t) - F^{\mathsf{opt}} \le \frac{2L\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2}{(t+1)^2}$$

- improved iteration complexity (i.e. $O(1/\sqrt{\varepsilon})$) than proximal gradient method (i.e. $O(1/\varepsilon)$)
- fast if prox can be efficiently implemented

Recap: the fundamental inequality for proximal method

Recall the following fundamental inequality shown in the last lecture:

Lemma 7.4

Let $\boldsymbol{y}^+ = \mathrm{prox}_{\frac{1}{L}h}\big(\boldsymbol{y} - \frac{1}{L}\nabla f(\boldsymbol{y})\big)$, *then*

$$F(\boldsymbol{y}^+) - F(\boldsymbol{x}) \le \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 - \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}^+\|_2^2$$

Proof of Theorem 7.6

1. build a discrete-time version of "Lyapunov function"

2. ***magic happens!***
    - "Lyapunov function" is non-increasing when Nesterov's momentum coefficients are adopted

Proof of Theorem 7.6

Key lemma: monotonicity of a certain "Lyapunov function"

Lemma 7.5

Let $\boldsymbol{u}^t = \underbrace{\theta_{t-1}\boldsymbol{x}^t - \left(\boldsymbol{x}^* + (\theta_{t-1} - 1)\boldsymbol{x}^{t-1}\right)}_{\text{or } \theta_{t-1}(\boldsymbol{x}^t - \boldsymbol{x}^*) - (\theta_{t-1}-1)(\boldsymbol{x}^{t-1} - \boldsymbol{x}^*)}$. *Then*

$$\|\boldsymbol{u}^{t+1}\|_2^2 + \frac{2}{L}\theta_t^2\left(F(\boldsymbol{x}^{t+1}) - F^{\mathsf{opt}}\right) \leq \|\boldsymbol{u}^t\|_2^2 + \frac{2}{L}\theta_{t-1}^2\left(F(\boldsymbol{x}^t) - F^{\mathsf{opt}}\right)$$

- quite similar to $2\left\|\boldsymbol{X} + \frac{\tau}{2}\dot{\boldsymbol{X}} - \boldsymbol{X}^*\right\|_2^2 + \tau^2\left(f(\boldsymbol{X}) - f^{\mathsf{opt}}\right)$ (Lyapunov function) as discussed before (think about $\theta_t \approx t/2$)

Proof of Theorem 7.6

With Lemma 7.5 in place, one has

$$\begin{aligned}\frac{2}{L}\theta_{t-1}^2\big(F(\boldsymbol{x}^t)-F^{\mathsf{opt}}\big) &\le \|\boldsymbol{u}^1\|_2^2+\frac{2}{L}\theta_0^2\big(F(\boldsymbol{x}^1)-F^{\mathsf{opt}}\big)\\ &= \|\boldsymbol{x}^1-\boldsymbol{x}^*\|_2^2+\frac{2}{L}\big(F(\boldsymbol{x}^1)-F^{\mathsf{opt}}\big)\end{aligned}$$

To bound the RHS of this inequality, we use Lemma 7.4 and $\boldsymbol{y}^0=\boldsymbol{x}^0$ to get

$$\frac{2}{L}\big(F(\boldsymbol{x}^1)-F^{\mathsf{opt}}\big)\le \|\boldsymbol{y}^0-\boldsymbol{x}^*\|_2^2-\|\boldsymbol{x}^1-\boldsymbol{x}^*\|_2^2=\|\boldsymbol{x}^0-\boldsymbol{x}^*\|_2^2-\|\boldsymbol{x}^1-\boldsymbol{x}^*\|_2^2$$

$$\Longleftrightarrow\qquad \|\boldsymbol{x}^1-\boldsymbol{x}^*\|_2^2+\frac{2}{L}\big(F(\boldsymbol{x}^1)-F^{\mathsf{opt}}\big)\le\|\boldsymbol{x}^0-\boldsymbol{x}^*\|_2^2$$

As a result,

$$\frac{2}{L}\theta_{t-1}^2\big(F(\boldsymbol{x}^t)-F^{\mathsf{opt}}\big)\le\|\boldsymbol{x}^1-\boldsymbol{x}^*\|_2^2+\frac{2}{L}\big(F(\boldsymbol{x}^1)-F^{\mathsf{opt}}\big)\le\|\boldsymbol{x}^0-\boldsymbol{x}^*\|_2^2,$$

$$\Longrightarrow\qquad F(\boldsymbol{x}^t)-F^{\mathsf{opt}}\le\frac{L\|\boldsymbol{x}^0-\boldsymbol{x}^*\|_2^2}{2\theta_{t-1}^2}\overset{\text{(Fact 7.2)}}{\le}\frac{2L\|\boldsymbol{x}^0-\boldsymbol{x}^*\|_2^2}{(t+1)^2}$$

Proof of Lemma 7.5

Take $\boldsymbol{x} = \frac{1}{\theta_t}\boldsymbol{x}^* + \big(1 - \frac{1}{\theta_t}\big)\boldsymbol{x}^t$ and $\boldsymbol{y} = \boldsymbol{y}^t$ in Lemma 7.4 to get

$$
\begin{aligned}
& F(\boldsymbol{x}^{t+1}) - F\Big(\theta_t^{-1}\boldsymbol{x}^* + \big(1 - \theta_t^{-1}\big)\boldsymbol{x}^t\Big) && (7.8) \\
&\leq \frac{L}{2}\big\|\theta_t^{-1}\boldsymbol{x}^* + \big(1 - \theta_t^{-1}\big)\boldsymbol{x}^t - \boldsymbol{y}^t\big\|_2^2 - \frac{L}{2}\big\|\theta_t^{-1}\boldsymbol{x}^* + \big(1 - \theta_t^{-1}\big)\boldsymbol{x}^t - \boldsymbol{x}^{t+1}\big\|_2^2 \\
&= \frac{L}{2\theta_t^2}\big\|\boldsymbol{x}^* + \big(\theta_t - 1\big)\boldsymbol{x}^t - \theta_t\boldsymbol{y}^t\big\|_2^2 - \frac{L}{2\theta_t^2}\big\|\underbrace{\boldsymbol{x}^* + \big(\theta_t - 1\big)\boldsymbol{x}^t - \theta_t\boldsymbol{x}^{t+1}}_{=-\boldsymbol{u}^{t+1}}\big\|_2^2 \\
&\overset{\text{(i)}}{=} \frac{L}{2\theta_t^2}\big(\|\boldsymbol{u}^t\|_2^2 - \|\boldsymbol{u}^{t+1}\|_2^2\big), && (7.9)
\end{aligned}
$$

where (i) follows from the definition of $\boldsymbol{u}^t$ and $\boldsymbol{y}^t = \boldsymbol{x}^t + \frac{\theta_{t-1}-1}{\theta_t}(\boldsymbol{x}^t - \boldsymbol{x}^{t-1})$

Proof of Lemma 7.5 (cont.)

We will also lower bound (7.8). By convexity of F,

$$
\begin{aligned}
F\Big(\theta_t^{-1}\boldsymbol{x}^* + \big(1-\theta_t^{-1}\big)\boldsymbol{x}^t\Big) &\le \theta_t^{-1}F(\boldsymbol{x}^*) + \big(1-\theta_t^{-1}\big)F(\boldsymbol{x}^t) \\
&= \theta_t^{-1}F^{\mathsf{opt}} + \big(1-\theta_t^{-1}\big)F(\boldsymbol{x}^t)
\end{aligned}
$$

$$
\begin{aligned}
\iff \quad & F\Big(\theta_t^{-1}\boldsymbol{x}^* + \big(1-\theta_t^{-1}\big)\boldsymbol{x}^t\Big) - F(\boldsymbol{x}^{t+1}) \\
& \le \big(1-\theta_t^{-1}\big)\big(F(\boldsymbol{x}^t) - F^{\mathsf{opt}}\big) - \big(F(\boldsymbol{x}^{t+1}) - F^{\mathsf{opt}}\big)
\end{aligned}
$$

Combining this with (7.9) and $\theta_t^2 - \theta_t = \theta_{t-1}^2$ yields

$$
\begin{aligned}
\frac{L}{2}\big(\|\boldsymbol{u}^t\|_2^2 - \|\boldsymbol{u}^{t+1}\|_2^2\big) &\ge \theta_t^2\big(F(\boldsymbol{x}^{t+1}) - F^{\mathsf{opt}}\big) - \big(\theta_t^2 - \theta_t\big)\big(F(\boldsymbol{x}^t) - F^{\mathsf{opt}}\big) \\
&= \theta_t^2\big(F(\boldsymbol{x}^{t+1}) - F^{\mathsf{opt}}\big) - \theta_{t-1}^2\big(F(\boldsymbol{x}^t) - F^{\mathsf{opt}}\big),
\end{aligned}
$$

thus finishing the proof

Convergence for strongly convex problems

$$\boldsymbol{x}^{t+1} = \mathsf{prox}_{\eta_t h}\big(\boldsymbol{y}^t - \eta_t \nabla f(\boldsymbol{y}^t)\big)$$
$$\boldsymbol{y}^{t+1} = \boldsymbol{x}^{t+1} + \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}(\boldsymbol{x}^{t+1} - \boldsymbol{x}^t)$$

Theorem 7.6 (Convergence of accelerated proximal gradient methods for strongly convex case)

Suppose f is μ-strongly convex and L-smooth. If $\eta_t \equiv 1/L$, then

$$F(\boldsymbol{x}^t) - F^{\mathsf{opt}} \le \left(1 - \frac{1}{\sqrt{\kappa}}\right)^t \left(F(\boldsymbol{x}^0) - F^{\mathsf{opt}} + \frac{\mu\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2}{2}\right)$$

A practical issue

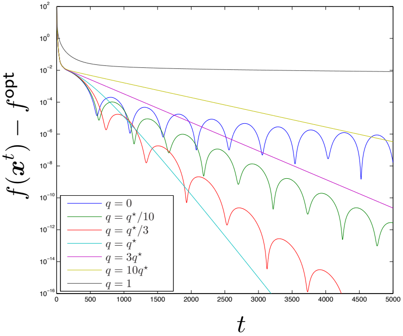
Fast convergence requires knowledge of $\kappa = L/\mu$

- in practice, estimating μ is typically very challenging

A common observation: ripples / bumps in the traces of cost values

Rippling behavior

Numerical example: take $\boldsymbol{y}^{t+1} = \boldsymbol{x}^{t+1} + \frac{1-\sqrt{q}}{1+\sqrt{q}}(\boldsymbol{x}^{t+1} - \boldsymbol{x}^t)$; $q^* = 1/\kappa$



period of ripples is often proportional to $\sqrt{L/\mu}$

O'Donoghue, Candes '12

- when $q > q^*$: we underestimate momentum $\longrightarrow$ slower convergence
- when $q < q^*$: we overestimate momentum ($\frac{1-\sqrt{q}}{1+\sqrt{q}}$ is large) $\longrightarrow$ overshooting / rippling behavior

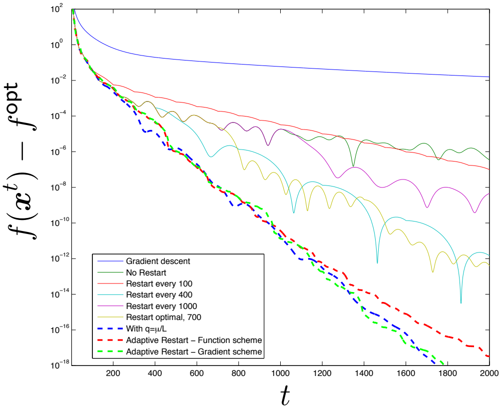
Adaptive restart (O'Donoghue, Candes '12)

When a certain criterion is met, restart running FISTA with

$$
\begin{aligned}
\boldsymbol{x}^0 &\leftarrow \boldsymbol{x}^t \\
\boldsymbol{y}^0 &\leftarrow \boldsymbol{x}^t \\
\theta_0 &= 1
\end{aligned}
$$

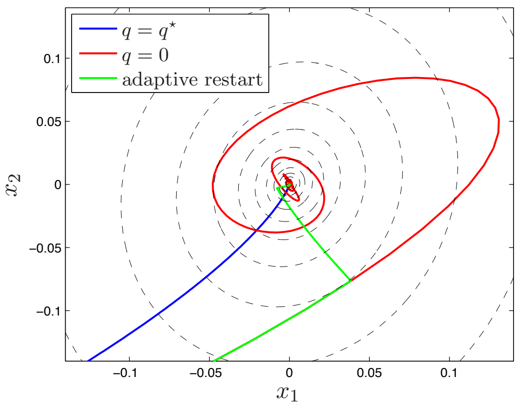
- take the current iterate as a new starting point
- erase all memory of previous iterates and reset the momentum back to zero

Numerical comparisons of adaptive restart schemes



- function scheme: restart when $f(\boldsymbol{x}^t) > f(\boldsymbol{x}^{t-1})$
- gradient scheme: restart when $\underbrace{\langle \nabla f(\boldsymbol{y}^{t-1}), \boldsymbol{x}^t - \boldsymbol{x}^{t-1} \rangle > 0}_{\text{restart when momentum lead us towards a bad direction}}$

Illustration



- with overestimated momentum (e.g. $q = 0$), one sees spiralling trajectory
- adaptive restart helps mitigate this issue

Lower bounds

Optimality of Nesterov's method

Interestingly, no first-order methods can improve upon Nesterov's results in general

More precisely, $\exists$ convex and L-smooth function f s.t.

$$f(\boldsymbol{x}^t) - f^{\mathsf{opt}} \geq \frac{3L\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2}{32(t+1)^2}$$

as long as $\underbrace{\boldsymbol{x}^k \in \boldsymbol{x}^0 + \mathsf{span}\{\nabla f(\boldsymbol{x}^0), \cdots, \nabla f(\boldsymbol{x}^{k-1})\}}_{\text{definition of first-order methods}}$ for all $1 \leq k \leq t$

— *Nemirovski, Yudin '83*

Example

$$\text{minimize}_{\boldsymbol{x}\in\mathbb{R}^{(2n+1)}} \quad f(\boldsymbol{x}) = \frac{L}{4}\left(\frac{1}{2}\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} - \boldsymbol{e}_1^\top \boldsymbol{x}\right)$$

where $$\boldsymbol{A} = \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix} \in \mathbb{R}^{(2n+1)\times(2n+1)}$$

- f is convex and L-smooth
- the optimizer $\boldsymbol{x}^*$ is given by $x_i^* = 1 - \frac{i}{2n+2}$ $(1 \le i \le n)$ obeying

$$f^{\mathsf{opt}} = \frac{L}{8}\left(\frac{1}{2n+2} - 1\right) \quad \text{and} \quad \|\boldsymbol{x}^*\|_2^2 \le \frac{2n+2}{3}$$

Example

$$\text{minimize}_{\boldsymbol{x}\in\mathbb{R}^{(2n+1)}} \quad f(\boldsymbol{x}) = \frac{L}{4}\left(\frac{1}{2}\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} - \boldsymbol{e}_1^\top \boldsymbol{x}\right)$$

$$\text{where } \boldsymbol{A} = \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix} \in \mathbb{R}^{(2n+1)\times(2n+1)}$$

- $\nabla f(\boldsymbol{x}) = \frac{L}{4}\boldsymbol{A}\boldsymbol{x} - \frac{L}{4}\boldsymbol{e}_1$
- $\underbrace{\text{span}\{\nabla f(\boldsymbol{x}^0), \cdots, \nabla f(\boldsymbol{x}^{k-1})\}}_{=:\mathcal{K}_k} = \text{span}\{\boldsymbol{e}_1, \cdots, \boldsymbol{e}_k\}$ if $\boldsymbol{x}^0 = \boldsymbol{0}$
  - every iteration of first-order methods expands the search space by *at most* one dimension

Example (cont.)

If we start with $\boldsymbol{x}^0 = \mathbf{0}$, then

$$f(\boldsymbol{x}^n) \geq \inf_{\boldsymbol{x} \in \mathcal{K}_n} f(\boldsymbol{x}) = \frac{L}{8}\left(\frac{1}{n+1} - 1\right)$$

$$\Longrightarrow \qquad \frac{f(\boldsymbol{x}^n) - f^{\mathsf{opt}}}{\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2} \geq \frac{\frac{L}{8}\left(\frac{1}{n+1} - \frac{1}{2n+2}\right)}{\frac{1}{3}(2n+2)} = \frac{3L}{32(n+1)^2}$$

Summary: accelerated proximal gradient

	stepsize rule	convergence rate	iteration complexity
convex & smooth problems	$\eta_t = \frac{1}{L}$	$O\left(\frac{1}{t^2}\right)$	$O\left(\frac{1}{\sqrt{\varepsilon}}\right)$
strongly convex & smooth problems	$\eta_t = \frac{1}{L}$	$O\left(\left(1 - \frac{1}{\sqrt{\kappa}}\right)^t\right)$	$O\left(\sqrt{\kappa}\log\frac{1}{\varepsilon}\right)$

Reference

[1] "*Some methods of speeding up the convergence of iteration methods*," B. Polyak, *USSR Computational Mathematics and Mathematical Physics*, 1964

[2] "*A method of solving a convex programming problem with convergence rate* $O(1/k^2)$," Y. Nestrove, *Soviet Mathematics Doklady*, 1983.

[3] "*A fast iterative shrinkage-thresholding algorithm for linear inverse problems*," A. Beck and M. Teboulle, *SIAM journal on imaging sciences*, 2009.

[4] "*First-order methods in optimization*," A. Beck, Vol. 25, SIAM, 2017.

[5] "*Mathematical optimization, MATH301 lecture notes*," E. Candes, Stanford.

[6] "*Gradient methods for minimizing composite functions*,", Y. Nesterov, *Technical Report*, 2007.

Reference

[7] "*Large-scale numerical optimization, MS&E318 lecture notes*," M. Saunders, Stanford.

[8] "*Analysis and design of optimization algorithms via integral quadratic constraints*," L. Lessard, B. Recht, A. Packard, *SIAM Journal on Optimization*, 2016.

[9] "*Proximal algorithms*," N. Parikh and S. Boyd, *Foundations and Trends in Optimization*, 2013.

[10] "*Convex optimization: algorithms and complexity*," S. Bubeck, Foundations and trends in machine learning, 2015.

[11] "*Optimization methods for large-scale systems, EE236C lecture notes*," L. Vandenberghe, UCLA.

[12] "*Problem complexity and method efficiency in optimization*," A. Nemirovski, D. Yudin, Wiley, 1983.

Reference

[13] "*Introductory lectures on convex optimization: a basic course*," Y. Nesterov, 2004

[14] "*A differential equation for modeling Nesterov's accelerated gradient method*," W. Su, S. Boyd, E. Candes, NIPS, 2014.

[15] "*On accelerated proximal gradient methods for convex-concave optimization*," P. Tseng, 2008

[16] "*Adaptive restart for accelerated gradient schemes*," B. O'donoghue, and E. Candes, *Foundations of computational mathematics*, 2012