Accelerated proximal gradient methods

Acknowledgement: slides are based on Prof. Lieven Vandenberghes.

- Nesterov's method
- analysis with fixed step size
- line search

Proximal gradient method

Results from lecture 4

- each proximal gradient iteration is a descent step (page 4.15 and 4.17):

$$f(x_{k+1}) < f(x_k), \qquad \|x_{k+1} - x^\star\|_2^2 \le c\,\|x_k - x^\star\|_2^2$$

  with $c = 1 - m/L$

- suboptimality after k iterations is $O(1/k)$ (page 4.16):

$$f(x_k) - f^\star \le \frac{L}{2k}\|x_0 - x^\star\|_2^2$$

Accelerated proximal gradient methods

- to improve convergence, we add a momentum term
- we relax the descent properties
- originated in work by Nesterov in the 1980s

Assumptions

we consider the same problem and make the same assumptions as in lecture 4:

$$\text{minimize} \quad f(x) = g(x) + h(x)$$

- h is closed and convex (so that prox_{th} is well defined)
- g is differentiable with $\mathrm{dom}\, g = \mathbf{R}^n$
- there exist constants $m \geq 0$ and $L > 0$ such that the functions

  $$g(x) - \frac{m}{2}x^T x, \qquad \frac{L}{2}x^T x - g(x)$$

  are convex
- the optimal value $f^\star$ is finite and attained at $x^\star$ (not necessarily unique)

Nesterov's method

choose $x_0 = v_0$ and $\theta_0 \in (0, 1]$, and repeat the following steps for $k = 0, 1, \ldots$

- if $k \geq 1$, define θ_k as the positive root of the quadratic equation

$$\frac{\theta_k^2}{t_k} = (1 - \theta_k)\gamma_k + m\theta_k \qquad \text{where } \gamma_k = \frac{\theta_{k-1}^2}{t_{k-1}}$$

- update x_k and v_k as follows:

$$\begin{aligned} y &= x_k + \frac{\theta_k \gamma_k}{\gamma_k + m\theta_k}(v_k - x_k) \qquad (y = x_0 \text{ if } k = 0) \\ x_{k+1} &= \mathrm{prox}_{t_k h}(y - t_k \nabla g(y)) \\ v_{k+1} &= x_k + \frac{1}{\theta_k}(x_{k+1} - x_k) \end{aligned}$$

stepsize t_k is fixed ($t_k = 1/L$) or obtained from line search

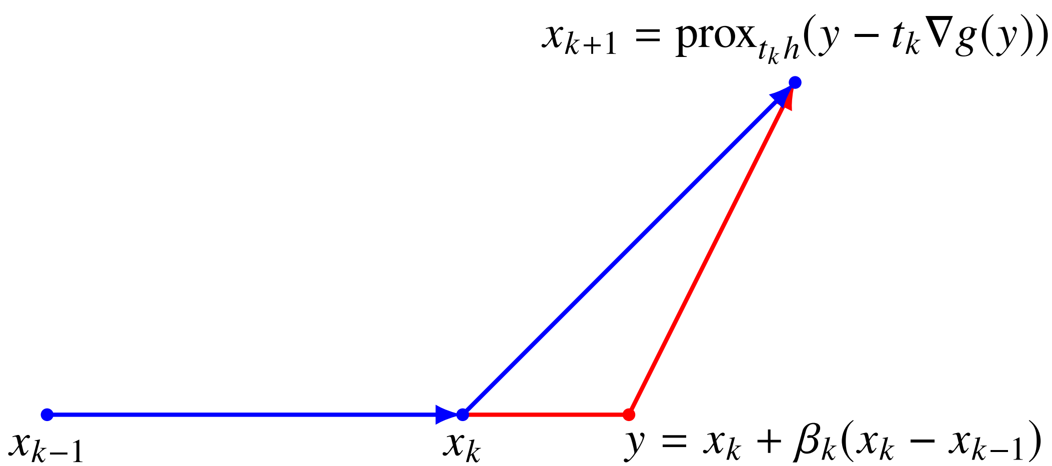
Momentum interpretation

- the first iteration ($k = 0$) is a proximal gradient step at $y = x_0$
- next iterations are proximal gradient steps at extrapolated points y:

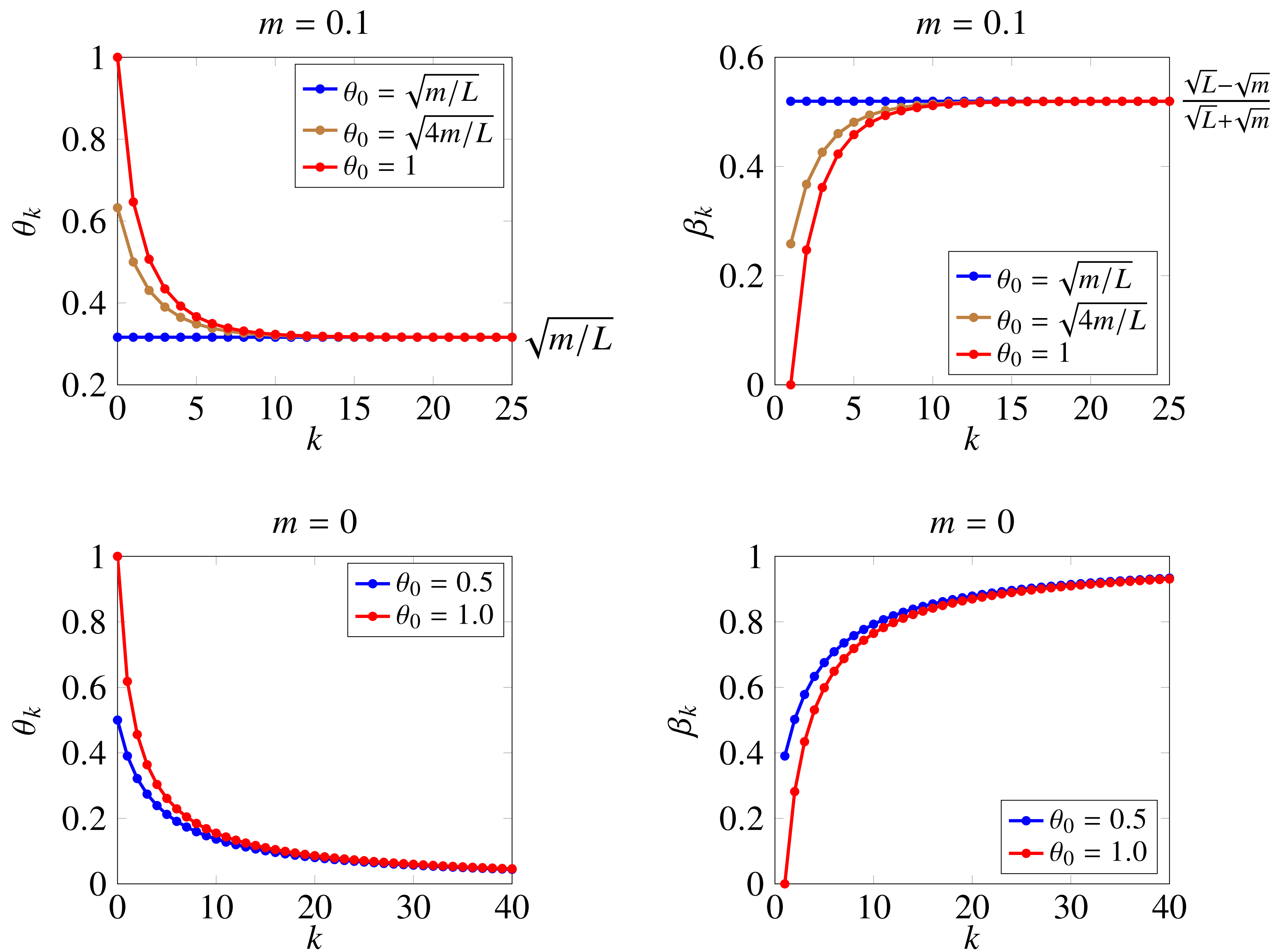
$$y = x_k + \frac{\theta_k \gamma_k}{\gamma_k + m\theta_k}(v_k - x_k) = x_k + \beta_k (x_k - x_{k-1})$$

where

$$\beta_k = \frac{\theta_k \gamma_k}{\gamma_k + m\theta_k} \left(\frac{1}{\theta_{k-1}} - 1 \right) = \frac{t_k \theta_{k-1}(1 - \theta_{k-1})}{t_{k-1}\theta_k + t_k \theta_{k-1}^2}$$



Parameters θ_k and β_k (for fixed stepsize $t_k = 1/L = 1$)



Parameter θ_k

- for $k \geq 1$, θ_k is the positive root of the quadratic equation

$$\frac{\theta_k^2}{t_k} = (1 - \theta_k)\frac{\theta_{k-1}^2}{t_{k-1}} + m\theta_k$$

- if $m > 0$ and $\theta_0 = \sqrt{mt_0}$, then $\theta_k = \sqrt{mt_k}$ for all k
- $\theta_k < 1$ if $mt_k < 1$
- for constant t_k, sequence θ_k is completely determined by θ_0

FISTA

if we take $m = 0$ on page 7.4, the expression for y simplifies:

$$
\begin{aligned}
y &= x_k + \theta_k(v_k - x_k) \\
x_{k+1} &= \mathrm{prox}_{t_k h}(y - t_k \nabla g(y)) \\
v_{k+1} &= x_k + \frac{1}{\theta_k}(x_{k+1} - x_k)
\end{aligned}
$$

eliminating the variables $v^{(k)}$ gives the equivalent iteration

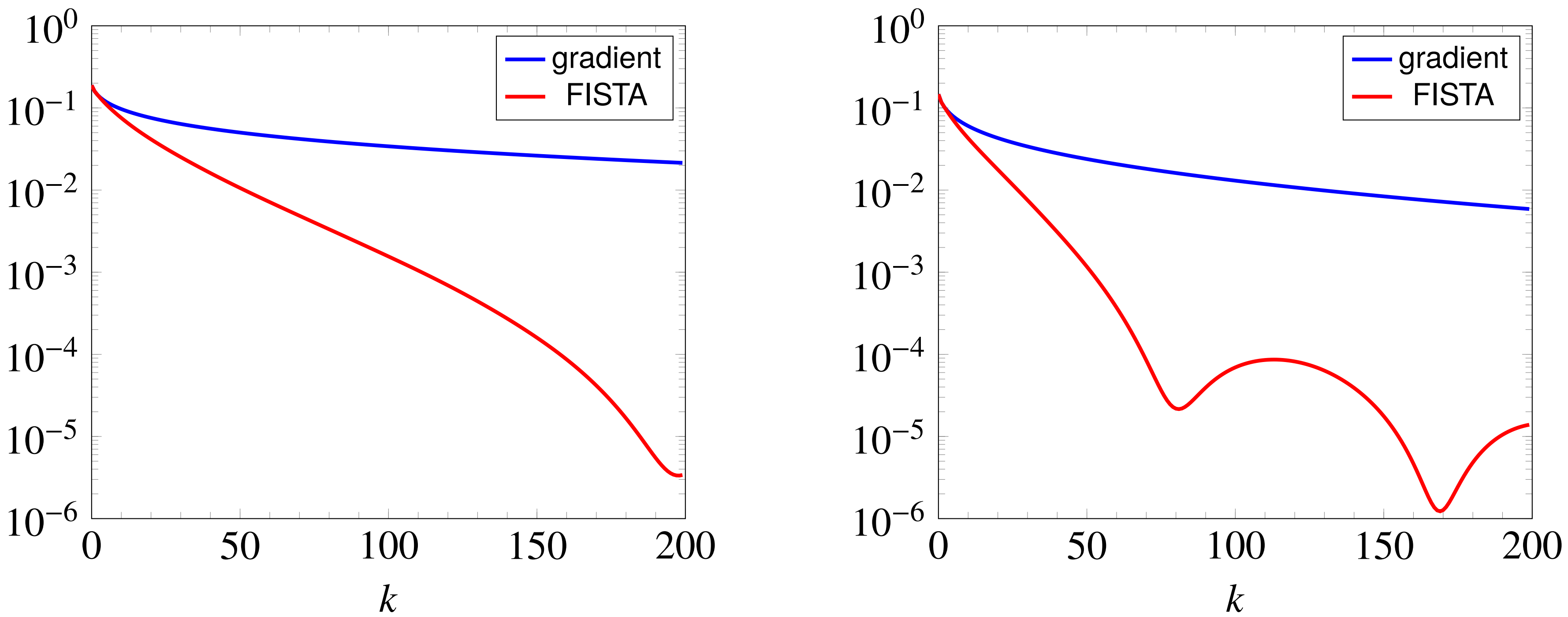
$$
\begin{aligned}
y &= x_k + \theta_k(\frac{1}{\theta_{k-1}} - 1)(x_k - x_{k-1}) \qquad (y = x_0 \text{ if } k = 0) \\
x_{k+1} &= \mathrm{prox}_{t_k h}(y - t_k \nabla g(y))
\end{aligned}
$$

this is known as **FISTA** (Fast Iterative Shrinkage-Thresholding Algorithm)

Example

$$\text{minimize} \quad \log \sum_{i=1}^{p} \exp(a_i^T x + b_i)$$

- two randomly generated problems with $p = 2000$, $n = 1000$
- same fixed step size used for gradient method and FISTA
- figures show $(f(x^{(k)}) - f^\star)/f^\star$



A simplification for strongly convex problems

- if $m > 0$ and we choose $\theta_0 = \sqrt{mt_0}$, then

$$\gamma_k = m, \qquad \theta_k = \sqrt{mt_k} \qquad \text{for all } k \geq 1$$

- the algorithm on page 7.4 and page 7.5 simplifies:

$$\begin{aligned} y &= x_k + \frac{\sqrt{t_k}}{\sqrt{t_{k-1}}} \frac{1 - \sqrt{mt_{k-1}}}{1 + \sqrt{mt_k}} (x_k - x_{k-1}) \qquad (y = x_0 \text{ if } k = 0) \\ x_{k+1} &= \operatorname{prox}_{t_k h}(y - t_k \nabla g(y)) \end{aligned}$$

- with constant stepsize $t_k = 1/L$, the expression for y reduces to

$$y = x_k + \frac{1 - \sqrt{m/L}}{1 + \sqrt{m/L}} (x_k - x_{k-1}) \qquad (y = x_0 \text{ if } k = 0)$$

Outline

- Nesterov's method
- **analysis with fixed step size**
- line search

Overview

- we show that if $t_i = 1/L$, the following inequality holds at iteration i:

$$
\begin{aligned}
& f(x_{i+1}) - f^\star + \frac{\gamma_{i+1}}{2}\|v_{i+1} - x^\star\|_2^2 \\
& \quad \leq \quad (1-\theta_i)(f(x_i) - f^\star) + \frac{\gamma_{i+1} - m\theta_i}{2}\|v_i - x^\star\|_2^2 \\
& \quad = \quad (1-\theta_i)\left(f(x_i) - f^\star) + \frac{\gamma_i}{2}\|v_i - x^\star\|_2^2\right) \qquad \text{if } i \geq 1
\end{aligned}
$$

- combining the inequalities from $i = 0$ to $i = k-1$ shows that

$$
\begin{aligned}
f(x_k) - f^\star \quad & \leq \quad \lambda_k \left((1-\theta_0)(f(x_0) - f^\star) + \frac{\gamma_1 - m\theta_0}{2}\|x_0 - x^\star\|_2^2\right) \\
& \leq \quad \lambda_k \left((1-\theta_0)(f(x_0) - f^\star) + \frac{\theta_0^2}{2t_0}\|x_0 - x^\star\|_2^2\right)
\end{aligned}
$$

where $\lambda_1 = 1$ and $\lambda_k = \prod_{i=1}^{k-1}(1-\theta_i)$ for $k > 1$

(here we assume $x_0 \in \operatorname{dom} f$)

Notation for one iteration

quantities in iteration i of the algorithm on page 7.4

- define $t = t_i$, $\theta = \theta_i$, $\gamma^+ = \gamma_{i+1} = \theta^2/t$
- if $i \geq 1$, define $\gamma = \gamma_i$ and note that $\gamma^+ - m\theta = (1-\theta)\gamma$
- define $x = x_i$, $x^+ = x_{i+1}$, $v = v_i$, and $v^+ = v_{i+1}$:

$$\begin{aligned} y &= \frac{1}{\gamma + m\theta}\left(\gamma^+ x + \theta\gamma v\right) \qquad (y = x = v \text{ if } i = 0) \\ x^+ &= y - tG_t(y) \\ v^+ &= x + \frac{1}{\theta}(x^+ - x) \end{aligned}$$

- v^+, v, and y are related as

$$\gamma^+ v^+ = \gamma^+ v + m\theta(y - v) - \theta G_t(y) \tag{1}$$

Proof (last identity):

- combine v and x updates and use $\gamma^+ = \theta^2/t$:

$$\begin{aligned} v^+ &= x + \frac{1}{\theta}(y - tG_t(y) - x) \\ &= \frac{1}{\theta}(y - (1-\theta)x) - \frac{\theta}{\gamma^+}G_t(y) \end{aligned}$$

- for $i = 0$, the equation (1) follows because $y = x = v$
- for $i \geq 1$, multiply with $\gamma^+ = \gamma + m\theta - \theta\gamma$:

$$\begin{aligned} \gamma^+ v^+ &= \frac{\gamma^+}{\theta}(y - (1-\theta)x) - \theta G_t(y) \\ &= \frac{(1-\theta)}{\theta}((\gamma + m\theta)y - \gamma^+ x) + \theta m y - \theta G_t(y) \\ &= (1-\theta)\gamma v + \theta m y - \theta G_t(y) \\ &= (\gamma^+ - m\theta)\gamma v + \theta m y - \theta G_t(y) \end{aligned}$$

Bound on objective function

recall the results on the proximal gradient update (page 4.13):

- if $0 < t \leq 1/L$ then $g(x^+) = g(y - tG_t(y))$ is bounded by

$$g(x^+) \leq g(y) - t\nabla g(y)^T G_t(y) + \frac{t}{2}\|G_t(y)\|_2^2 \tag{2}$$

- if the inequality (2) holds, then $mt \leq 1$ and, for all z,

$$f(z) \geq f(x^+) + \frac{t}{2}\|G_t(y)\|_2^2 + G_t(y)^T(z - y) + \frac{m}{2}\|z - y\|_2^2$$

- add $(1 - \theta)$ times the inequality for $z = x$ and θ times the inequality for $z = x^\star$:

$$\begin{aligned} f(x^+) - f^\star \quad &\leq \quad (1 - \theta)(f(x) - f^\star) - G_t(y)^T \left((1 - \theta)x + \theta x^\star - y\right) \\ &\qquad - \frac{t}{2}\|G_t(y)\|_2^2 - \frac{m\theta}{2}\|x^\star - y\|_2^2 \end{aligned}$$

Bound on distance to optimum

- it follows from (1) that

$$
\begin{aligned}
\frac{\gamma^+}{2}\|v^+ - x^\star\|_2^2 &= \frac{\gamma^+ - m\theta}{2}\|v - x^\star\|_2^2 + \theta G_t(y)^T(x^\star - v - \frac{m\theta}{\gamma^+}(y - v)) \\
&\quad - \frac{m\theta(\gamma^+ - m\theta)}{2\gamma^+}\|y - v\|_2^2 + \frac{t}{2}\|G_t(y)\|_2^2 + \frac{m\theta}{2}\|x^\star - y\|_2^2 \\
&\le \frac{\gamma^+ - m\theta}{2}\|v - x^\star\|_2^2 + \theta G_t(y)^T(x^\star - v - \frac{m\theta}{\gamma^+}(y - v)) \\
&\quad + \frac{t}{2}\|G_t(y)\|_2^2 + \frac{m\theta}{2}\|x^\star - y\|_2^2
\end{aligned}
$$

- γ^+ and y are chosen so that $\theta(\gamma^+ - m\theta)(y - v) = \gamma^+(1 - \theta)(x - y)$; hence

$$
\begin{aligned}
\frac{\gamma^+}{2}\|v^+ - x^\star\|_2^2 &\le \frac{\gamma^+ - m\theta}{2}\|v - x^\star\|_2^2 + G_t(y)^T(\theta x^\star + (1 - \theta)x - y) \\
&\quad + \frac{t}{2}\|G_t(y)\|_2^2 + \frac{m\theta}{2}\|x^\star - y\|_2^2
\end{aligned}
$$

Progress in one iteration

- combining the bounds on page 7.15 and 7.16 gives

$$
\begin{aligned}
& f(x^+) - f^\star + \frac{\gamma^+}{2}\|v^+ - x^\star\|_2^2 \\
\leq \quad & (1-\theta)(f(x) - f^\star) + \frac{\gamma^+ - m\theta}{2}\|v - x^\star\|_2^2
\end{aligned}
$$

  this is the first inequality on page 7.12

- if $i \geq 1$, we use $\gamma^+ - m\theta = (1-\theta)\gamma$ to write this as

$$
\begin{aligned}
& f(x^+) - f^\star + \frac{\gamma^+}{2}\|v^+ - x^\star\|_2^2 \\
\leq \quad & (1-\theta)\left(f(x) - f^\star + \frac{\gamma}{2}\|v - x^\star\|_2^2\right)
\end{aligned}
$$

Analysis for fixed step size

the product $\lambda_k = \prod_{i=1}^{k-1}(1-\theta_i)$ determines the rate of convergence (page 7.12)

- the sequence λ_k satisfies the following bound (proof on next page)

$$\lambda_k \leq \frac{4}{(2 + \sqrt{\gamma_1}\sum\limits_{i=1}^{k-1}\sqrt{t_i})^2} = \frac{4t_0}{(2\sqrt{t_0} + \theta_0\sum\limits_{i=1}^{k-1}\sqrt{t_i})^2} \tag{3}$$

- for constant step size and $\theta_0 = 1$, we obtain

$$\lambda_k \leq \frac{4}{(k+1)^2}$$

- with $t_0 = 1/L$, the inequality on page 7.12 shows a $1/k^2$ convergence rate

$$f(x_k) - f^\star \leq \frac{2L}{(k+1)^2}\|x_0 - x^\star\|_2^2$$

Proof.

- recall that for $k \geq 1$,

$$\gamma_{k+1} = (1 - \theta_k)\gamma_k + \theta_k m, \qquad \gamma_k = \theta_{k-1}^2 / t_{k-1}$$

- we first note that $\lambda_k \leq \gamma_k / \gamma_1$; this follows from

$$\lambda_{i+1} = (1 - \theta_i)\lambda_i = \frac{\gamma_{i+1} - \theta_i m}{\gamma_i}\lambda_i \leq \frac{\gamma_{i+1}}{\gamma_i}\lambda_i$$

- the inequality (3) follows by combining from $i = 1$ to $i = k - 1$ the inequalities

$$\begin{aligned} \frac{1}{\sqrt{\lambda_{i+1}}} - \frac{1}{\sqrt{\lambda_i}} &\geq \frac{\lambda_i - \lambda_{i+1}}{2\lambda_i\sqrt{\lambda_{i+1}}} \\ &= \frac{\theta_i}{2\sqrt{\lambda_{i+1}}} \\ &\geq \frac{\theta_i}{2\sqrt{\gamma_{i+1}/\gamma_1}} \\ &= \frac{1}{2}\sqrt{\gamma_1 t_i} \end{aligned}$$

Strongly convex functions

the following bound on λ_k is useful for strongly convex functions ($m > 0$)

- if $\theta_0 \geq \sqrt{m t_0}$, then $\theta_k \geq \sqrt{m t_k}$ for all k and

$$\lambda_k \leq \prod_{i=1}^{k-1} (1 - \sqrt{m t_i})$$

  (proof on next page)

- for constant step size $t_k = 1/L$, we obtain

$$\lambda_k \leq \left(1 - \sqrt{m/L}\right)^{k-1}$$

- combined with the inequality on page 7.12, this shows linear convergence

$$f(x_k) - f^\star \leq \left(1 - \sqrt{\frac{m}{L}}\right)^{k-1} \left((1 - \theta_0)(f(x_0) - f^\star) + \frac{\theta_0^2}{2t_0} \|x_0 - x^\star\|_2^2\right)$$

Proof.

- if $\theta_{k-1} \geq \sqrt{mt_{k-1}}$, then $\theta_k \geq \sqrt{mt_k}$:

$$\begin{aligned}\frac{\theta_k^2}{t_k} &= (1-\theta_k)\frac{\theta_{k-1}^2}{t_{k-1}} + m\theta_k \\ &\geq (1-\theta_k)m + m\theta_k \\ &= m\end{aligned}$$

- if $\theta_0 \geq \sqrt{mt_0}$, then $\theta_k \geq \sqrt{mt_k}$ for all k and

$$\lambda_k = \prod_{i=1}^{k-1}(1-\theta_i) \leq \prod_{i=1}^{k-1}(1-\sqrt{mt_i})$$

Outline

- Nesterov’s method
- analysis with fixed step size
- **line search**

Line search

- the analysis for fixed step size starts with the inequality (2):

$$g(x - tG_t(y)) \leq g(y) - t\nabla g(y)^T G_t(y) + \frac{t}{2}\|G_t(y)\|_2^2$$

  this inequality is known to hold for $0 \leq t \leq 1/L$

- if L is not known, we can satisfy (2) by a backtracking line search:

  start at some $t := \hat{t} > 0$ and backtrack ($t := \beta t$) until (2) holds

- step size selected by the line search satisfies $t \geq t_{\min} = \min\{\hat{t}, \beta/L\}$

- for each tentative t_k we need to recompute θ_k, y, x_{k+1} in the algorithm on p. 7.4

- requires evaluations of ∇g, prox_{th}, and g (twice) per line search iteration

Analysis with line search

- from page 7.18, if $\theta_0 = 1$:

$$\lambda_k \le \frac{4t_0}{(2\sqrt{t_0} + \sum\limits_{i=1}^{k-1} \sqrt{t_i})^2} \le \frac{4\hat{t}/t_{\min}}{(k+1)^2}$$

- from page 7.20, if $\theta_0 \ge \sqrt{mt_0}$:

$$\lambda_k \le \prod_{i=1}^{k-1}(1 - \sqrt{mt_i}) \le \left(1 - \sqrt{mt_{\min}}\right)^{k-1}$$

- therefore the results for fixed step size hold with $1/t_{\min}$ substituted for L

References

Most of the material in the lecture is from §2.2 in Nesterov's *Lectures on Convex Optimization*.

FISTA

- A. Beck, *First-Order Methods in Optimization* (2017), §10.7.
- A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. on Imaging Sciences (2009).
- A. Beck and M. Teboulle, *Gradient-based algorithms with applications to signal recovery*, in: Y. Eldar and D. Palomar (Eds.), *Convex Optimization in Signal Processing and Communications* (2009).

Accelerated proximal gradient methods

- S. Bubeck, *Convex Optimization: Algorithms and Complexity*, Foundations and Trends in Machine Learning (2015), §3.7.
- P. Tseng, *On accelerated proximal gradient methods for convex-concave optimization* (2008).

Line search strategies

- FISTA papers by Beck and Teboulle.
- D. Goldfarb and K. Scheinberg, *Fast first-order methods for composite convex optimization with line search* (2011).
- O. Güler, *New proximal point algorithms for convex minimization*, SIOPT (1992).
- Yu. Nesterov, *Gradient methods for minimizing composite functions* (2013).

Implementation

- S. Becker, E.J. Candès, M. Grant, *Templates for convex cone problems with applications to sparse signal recovery*, Mathematical Programming Computation (2011).
- B. O'Donoghue, E. Candès, *Adaptive restart for accelerated gradient schemes*, Foundations of Computational Mathematics (2015).
- T. Goldstein, C. Studer, R. Baraniuk, *A field guide to forward-backward splitting with a FASTA implementation,* arXiv:1411.3406 (2016).